University of Maribor

Faculty of Electrical Engineering
and Computer Science

# EJC 2015

## Proceedings of the 25th International Conference on Information Modelling and Knowledge Bases

8-12 June, 2015, Maribor, Slovenia

# Proceedings of the 25th International Conference on Information Modelling and Knowledge Bases

# EJC 2015

T. Welzer
M. Hölbl
Y. Kiyoki
B. Thalheim
H. Jaakkola

8-12 June, 2015, Maribor, Slovenia

CIP - Katadložni zapis o publikaciji
Univerzitetna knjižnica Maribor

004(082)

INTERNATIONAL Conference on Information Modelling and Knowledge Bases (25 ; 2015 ; Maribor)
Proceedings of the 25th International Conference on Information Modelling and Knowledge Bases - EJC 2015 [Elektronski vir] : Maribor, 8 - 12 June 2015 / [editors T. Welzer ... [et al.]. - 1st ed. - Maribor : Faculty of Electrical Engineering and Computer science 2015, 2015

50 USB ključev

ISBN 978-961-248-486-6

1. Welzer-Družovec, Tatjana
COBISS.SI-ID 82674433

## Impressum





| | |
|---|---|
| **Title:** | Proceedings of the 25th International Conference on Information Modelling and Knowledge Bases (EJC 2015) |
| **Editors:** | Tatjana Welzer, University of Maribor, Slovenia<br>Marko Hölbl, University of Maribor, Slovenia<br>Yasushi Kiyoki, Keio University, Japan<br>Bernhard Thalheim, Christian-Albrechts University at Kiel, Germany<br>Hannu Jaakkola, Tampere University of Technology (Pori), Finland |
| **Technical Editors:** | Lili Nemec Zlatolas, University of Maribor, Slovenia<br>Naofumi Yoshida, Komazawa University, Japan |
| **Page count:** | 324 |
| **Edition:** | 1st |
| **First published:** | 2015 |
| **Publisher:** | University of Maribor, Faculty of Electrical Engineering and Computer Science |
| **Impression:** | 50 USB flash drive units |

## PREFACE

In the last three decades information modelling and knowledge bases have become essentially important subjects not only in academic communities related to information systems and computer science but also in the business area where information technology is applied.

The series of European – Japanese Conference on Information Modelling and Knowledge Bases (EJC) originally started as a co-operation initiative between Japan and Finland in 1982. The practical operations were then organised by professor Ohsuga in Japan and Professors Hannu Kangassalo and Hannu Jaakkola in Finland (Nordic countries). Geographical scope was first expanded to cover Nordic countries, then Europe and finally also other countries. Workshop characteristic - discussion, enough time for presentations and limited number of participants (50) / papers (30) - is typical for the conference.

The 25$^{th}$ International Conference on Information Modelling and Knowledge Bases (EJC 2015) constitute a world-wide research forum for the exchange of scientific results and experiences achieved in computer science and other related disciplines using innovative methods and progressive approaches. In this way a platform has been established drawing together researches as well as practitioners dealing with information modelling and knowledge bases. The main topics of EJC conferences target the variety of themes in the topics include, but are not limited to:

1. **Conceptual modelling:** Modelling and specification languages; Domain-specific conceptual modelling; Concepts, concept theories and ontologies; Conceptual modelling of large and heterogeneous systems; Conceptual modelling of spatial, temporal and biological data; Methods for developing, validating and communicating conceptual models.
2. **Knowledge and information modelling and discovery:** Knowledge discovery, knowledge representation and knowledge management; Advanced data mining and analysis methods; Conceptions of knowledge and information; Modelling information requirements; Intelligent information systems; Information recognition and information modelling.
3. **Linguistic modelling:** Models of HCI; Information delivery to users; intelligent informal querying; Linguistic foundation of information and knowledge; Fuzzy linguistic models; Philosophical and linguistic foundations of conceptual models.
4. **Cross-cultural communication and social computing:** Cross-cultural support systems; Integration, evolution and migration of systems; Collaborative societies; Multicultural web-based software systems; Intercultural collaboration and support systems; Social computing, behavioural modelling and prediction.
5. **Environmental modelling and engineering:** Environmental information systems (architecture); Spatial, temporal and observational information systems; Large-scale environmental systems; Collaborative knowledge base systems; Agent concepts and conceptualisation; Hazard prediction, prevention and steering systems.
6. **Multimedia data modelling and systems:** Modelling multimedia information and knowledge; Content-based multimedia data management; Content-based multimedia retrieval; Privacy and context enhancing technologies; Semantics and pragmatics of multimedia data; Metadata for multimedia information systems.

Careful evaluation gave us the following program: 11 papers have been selected as long paper, 10 papers as short papers, 2 papers as position papers. We have one invited talk in the program as well as panel session and presentation of Slovenian research projects. We thank all colleagues for their support of this issue of the EJC conference, especially the program committee, the organising committee, and the programme coordination team. The conference proceedings that will include after the conference revised presentations, will be published in the Series of "Frontiers in Artificial Intelligence" by IOS Press (Amsterdam). The books "Information Modelling and Knowledge Bases" are edited by the Editing Committee of the conference.

We believe that the conference will be productive and fruitful in the advance of research and application of information modelling and knowledge bases.

Tatjana Welzer,
Marko Hölbl,
Yasushi Kiyoki,
Bernhard Thalheim,
Hannu Jaakkola.

## General Programme Chair

Hannu Kangassalo, University of Tampere, Finland

## Program Committee Co-chairs

Yasushi Kiyoki, Keio University, Japan
Bernhard Thalheim, Christian-Albrechts University at Kiel, Germany

## Members of the Programme Committee

Boštjan Brumen
Pierre-Jean Charrel
Xing Chen
Marie Duží
Anneli Heimbürger
Jaak Henno
Yoshihide Hosokawa
Hannu Jaakkola
Yasushi Kiyoki
Sebastian Link
Heinrich C. Mayr
Tommi Mikkonen
Jørgen Fischer Nilsson
Tomoya Noro
Jari Palomäki
Bernhard Rumpe
Tetsuya Suzuki
Bernhard Thalheim
Naofumi Yoshida

## External Reviewers

Robert Eikermann
Robert Heim
Christoph Schulze

## General Organizing Chair

Hannu Jaakkola, Tampere University of Technology (Pori), Finland

## Organizing Committee

Tatjana Welzer, University of Maribor, Slovenia
Lili Nemec Zolatolas, University of Maribor, Slovenia
Marko Hölbl, University of Maribor, Slovenia
Borut Zlatolas, University of Maribor, Slovenia
Xing Chen, Kanagawa Institute of Technology, Japan
Ulla Nevanranta (Publication), Tampere University of Technology (Pori), Finland

## Programme Coordination Team

Naofumi Yoshida, Komazawa University, Japan (chair)
Anneli Heimbürger, University of Jyväskylä, Finland

# TABLE OF CONTENTS

# A Dynamic Dengue Fever Risk Area Model for Real-Time Control in Local-Global Spreading

Wahjoe T Sesulihatien[1,2], Yasushi Kiyoki[2]
[1] *Politeknik Elektronika Negri Surabaya*
[2] *Keio University, Shonan Fujisawa Campus, Japan*

**Abstract.** *Dengue fever is the fastest spreading communicable disease in the world. The virus has been increasing its geographic reach, partly due to increasing human moving. In many dengue-endemic countries, problem in dengue spreading is laid on determining spot infected area and strategy to prevent the disease. Determining spot infected area relates with pattern of human moving while strategy to prevent is depend on vulnerability of area due to dengue contraction. In this paper we proposed a real time model of local-global dengue spreading based on human movement. This method combines a spatially-explicit, area-based mathematical model with discrete life-cycle of virus. The proposed method includes (1) state-space model of routine movement cycle (2) algorithm of local-global scenario of spreading (3) prediction of next infection area by relation in graph ,and(4) determine vulnerability value of suspected area. There are two important features in this method: real-time prediction of infected area and rank of suspected area vulnerability. To perform the simulation we utilize real data of infected people in Surabaya in January 2011.The result shows that this method is suitable to represents real time both of local and global spreading. However, we have to consider about random movement, especially in week end.*

**Keywords.** *Real time, state space, local-global spreading, routine human moving*

## 1. Introduction

Dengue fever is a painful, debilitating mosquito-borne disease caused by one of four closely related dengue viruses. It is transmitted by the bite of an infected Aedes mosquito. The mosquito becomes infected when it bites a person with dengue virus present in their blood. Until now, more than 100 million cases of dengue fever occur worldwide in The Indian subcontinent, Southeast Asia, Southern China, Taiwan, The Pacific Islands, The Caribbean Mexico, Africa, Central and South America southern United States, and southern Australia [1]. In Indonesia, dengue cases increase yearly in almost all regions [2] The virus has been spreading in geographic reach, partly due to increasing urbanization and partly due to climate change [3].

Due to high chance of serious damage impact in dengue spreading all over the world, the effective strategy for controlling dengue is vital. The important issue in strategy to combat dengue spreading is realizing quick and effective action [1]. Quick action particularly intended for preventing another area from next spreading. Effective relate with minimize effect to environmental. Based on this issue some research was conducted

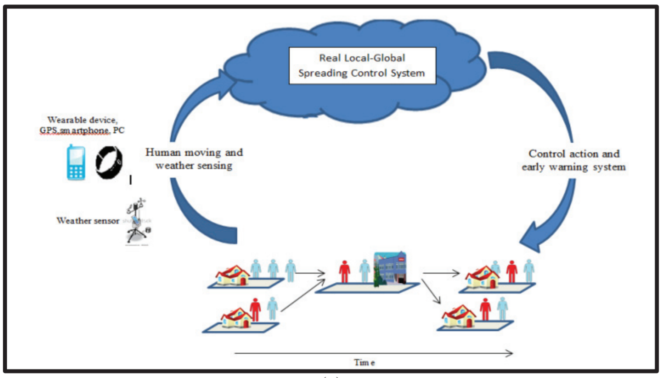
Previous study about dengue focused on strategy based on "static situation". Static means connection between host (human) and vector (mosquito) occurs under assumption that people never moving across the border areas. By this point of view some strategy are established. Some published strategies are prediction based on history of dengue occurrences [4]-[8], association among spatial attributes [9]-[11], statistic approach of significant variable [12]-[15], modelling in vector (mosquito) behavior [16]-[18], relation between geographical distribution [19][20] and analysis in population density to dengue case [21]-[23]. Modelling by static condition resulting variety in conclusion and sometimes bring contrast result on. The result is not suitable to apply due to increasing number of traveler. If number of people who travel out of area increase, we cannot keep "static situation" anymore.

The afterward research, then, focused on "dynamic situation'. Dynamic means people, with their attributes, moving across their border area. During traveling time, they possible to interact with infected people. Disease is transmitted through mosquito's bite. Modelling of spreading is focused on the interaction process, for instance by social structure [24][25], scaling range of contact [26][27]. All models were conducted by agent base model [28]. Then the agent base model expanded in network to analyze interplay between network and epidemiology [29]. This model is suitable to understand phenomena of transmission, but since it developed by hypothesis information, level of readiness is low relatively [30]. The difficulties caused by different pattern of personal contact among agent every day.

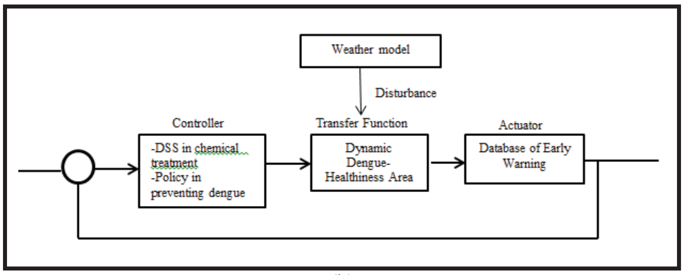
To build realistic model, contact among agent should be transformed in dynamic relation among area in the system. There are several models that take into risk diseases approach in network but not in spatial heterogeneity, such as in based on knowledge [31], social network information [32] and social contact [33]. In the other hand, models [34], [35] take into account spatial heterogeneity but refer to compartment model of individual in spreading. This research deals with spatial heterogeneity and area spreading model in network. This model consist of (1) state-space model of routine movement cycle (2) algorithm of local-global scenario of spreading (3) prediction of next infection area by relation in graph of human moving, and (4) determine vulnerability value of suspected area. Through this transfer function we perform controllability and observability of near future spreading area. Because it is based on data-driven model, this model is flexible in changing of pattern. Therefore, the control action can be performed quickly and effectively according to the situation.

## 2. Proposed Method

The new approach we propose in dengue spreading is Adaptive Real Local-Global Spreading Control System. Basic concept of this system is figures out physical phenomena of disease transmission in network and transform to Linier Time Invariant state. This state is transfer function system that represents behavior of system. Figure 1(a) describes overall system. Spreading disease in this study involved 3 components: human moving, weather and disease contracting location. Attribute of human moving are origin location, destination location and number of people. Weather is variable that reflect mosquito existence. Mosquito life cycle is strong influenced by weather, while number of mosquito is linier to probability of infection. Physical phenomena of spreading are captured by human sensor (such as GPS, smart phone, PC or wearable device) and weather sensor. Then all data are processed and controlled. The output is control action and early warning system for public.




(a)




(b)

Figure 1. Real Local-Global Spreading Control System Framework

(a) Overall system
(b) Control system

Figure 1(b) shows block diagram of control system. This system consists of transfer function plant, controller and actuator. Transfer function is mathematic model that represent phenomena of disease transmission by local and global people interaction, In this paper transfer function is Dynamic Dengue-Healthiness Area. Controller plays important role in managing or regulating the behavior of system. In this case, it is chemical treatment or policy to prevent dengue. Actuator translates mathematic model to real action. In this case, it has a shape of information services for public in term of early warning system.

In order to build dynamic control system, the first step is composing transfer function of disease transmission. This transfer function associates input in time (t) in to output in time (t+1). Inputs are population of infected and heath people, weather and risk contracting-area level .Output are number of infected people and infected area. This transfer function consists of differential time of movement from state (n) to state (n+1), mathematic model of input when it going in compartment and going out from compartment ,and algorithm of sketching next infected area

### 2.1. Variable of System

Variable of system in this paper is defined as a set of named values that has relation with number of infected people. Then the candidate variable is verified by some method with data from 2007 to 2012 to confirm the significances. As the system is dynamic, the variable should interpret time-vary of changes in dengue infected people. In this case, we chose contracting-area and weather as most significant variable.

#### *2.1.1. Dengue Location-Contraction Risk*

Dengue Location-Contraction Risk, L (*n, t*), [36] is chosen as variable to measure risk area *n* causing by human involvement. This model is built by modelling contagious places. Value of every cell is shown as different color in map as shown in figure 2


**Figure 2.** Dengue Location-Contracting Risk Map

Every color represents contagious-place risk-ranking that explicitly represents regular and routine movement. This method is perform by (1) statistical analysis about correlations between contagious places and Dengue fever cases, (2) a multi-layer weighting method for determining the weight of each cell (3) a ranking and classification method for places' human-mingling.

#### *2.1.2 Dynamic weather*

Dynamic weather symbolizes number of mosquito[37][38]. Then, in this research, existence of mosquito is represented by weather variable: humidity, temperature and rainfall. We use dynamic projecting model to identify a meaningful and significant indicator in weather component that relate with occurrence of dengue fever. All data were normalized, transformed in Hilbert space and analyzed by Hilbert transform. Then, we utilized projection in Mathematical Method of Meaning in some spaces to find the relation between pattern of weather data and infected people data. The result is shown in figure 3. shows that accumulation of temperature, humidity and rainfall conveniently feasible to represent weather variability.

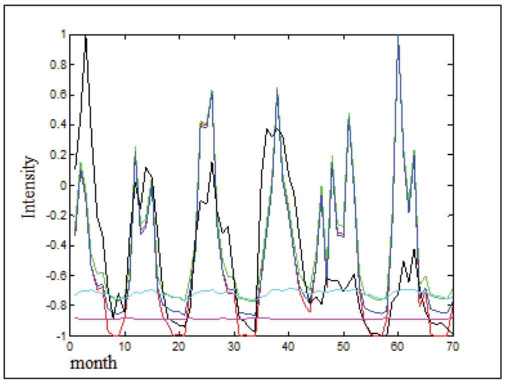



**Figure 3** Correlation between weather and occurrence of dengue

Figure 3 shows correlation between weather and dengue occurrence in Surabaya. The black line is dengue occurrence, blue one is subspace of temperature, humidity and rainfall , cyan line is humidity and temperature line in magenta color. Insight, dengue occurrence represents existence of mosquito. It is clear that accumulation of temperature, humidity and rainfall conveniently feasible to represent weather variability.

### 2.2 Individual Risk

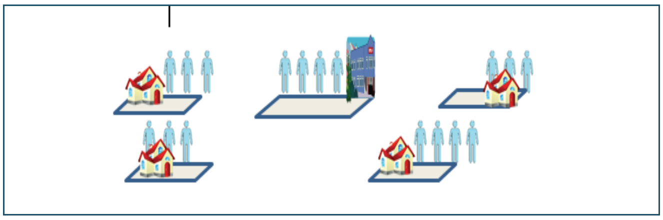
Individual risk, $P_d$ (t), in this paper is defined as probability the risk of exposure individual (i.e., being bitten by a vector) for individual P, in time *t* at area *n* and written as [39]

$$P(n,t) = \sum_j^n \left(\frac{L_j}{h_x}\right)^2 \left(\frac{1}{h_x}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2)$$

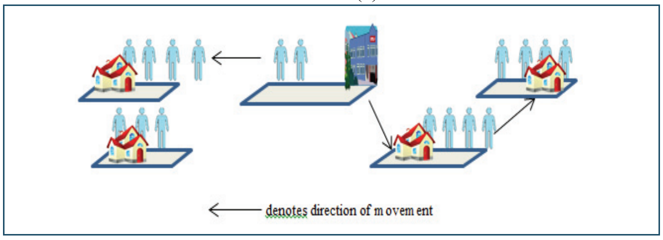
Where : $L_n$ = Dengue Location-Contraction Risk value
$P$ = individual risk
$h_x$ = proportional use of that site within some interval by people
x = number of individual
n = area identity (example n=4 means area no 4)
(for example if a site is used by 2 individuals for 6 hours each over a week,
$h_x$=(2 humans * 6 hours)/(24 hours/day * 7 days) = 0.07 humans.)

### 2.3 Healthiness of Area

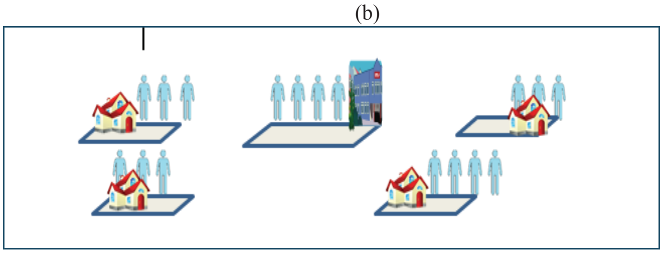
We define Healthiness of Area as an accumulation of individual disease risk when they stay in the area. That value reflecting resistance of area in dengue transmission Accumulation happened as a consequence of human movement. The movement in this study is assumed over from origin (home) to destination (school, office etc) in the morning and from destination to origin in the evening return to home as shown in figure 3



(a)




(b)



(c )

**Figure 4** Dynamic routine-movement population
(a) Population in day 1, evening
(b) Population in day 2, morning
(c ) Population in day 2, evening

From figure 3, it is clear that population of every area is different in the morning and in the evening. Every individual who is entering the destination area brings Individual risk, T, as attribute. Therefore, healthiness of area is vary depends on pattern of movement. Healthiness of area will be described more in section 2.3.2

### *2.3.1 Moving-flow of Area*

This model is built from dynamic population changes. We assume that number of people visited area *n* at time (*t*) and spend their time in destination at interval $\Delta t$ , before returning home at time (*t+1*) . Then number of people in area *n* during time $\Delta t$ is $X_n(\Delta t)$, written as :

$$X_n(t+1) - X_n(t) = X_n(\Delta t) \quad \text{......... ......................(3)}$$

Number of people entering area *n* or leaving area *n* is illustrated in figure 3. Figure 3(a) shows flow of moving people. *Xn*(t) represent number of individual from area *n*, *Pn* represent individual risk when they enter an area

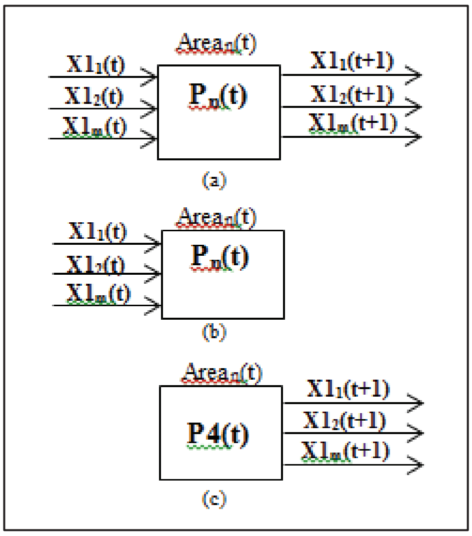


Figure 4. Process in every area

Figure 4 shows flow of people entering area $n$. From this picture, we can perform relation between input and output as follows:

$$X1_m(t+1) = P_n X1_m(\text{t}) \qquad \text{.........................(4)}$$

Where : n = identify number of area where individual leaving from
m = identify number of area where individual coming from
m=n means individual stay at this area

### *2.3.2 Flow-Cycle Human Moving*

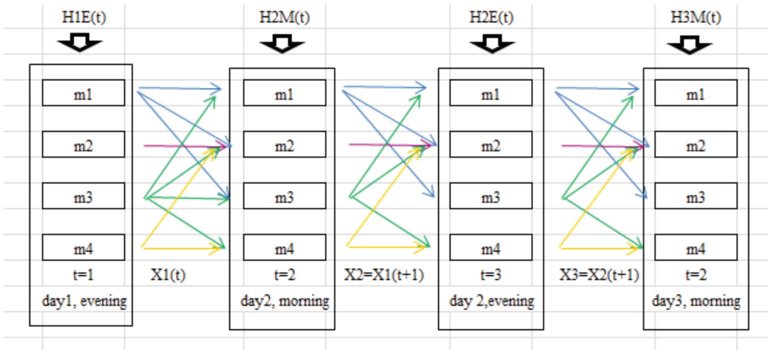
Flow-cycle of human moving is illustrated in figure 5. One cycle is defined as origin – destination-origin moving. We assume that individuals leave their origin *m* (m=1,2,3…number of area) to visit area n(n=1,2,3…number of area) in the morning. The visit time is declared as *t*. After that, in the evening they return home in the same day. The return time is *t+1*. Then tomorrow morning they will visit the same area in different day. The time visit is *t+2*. This cycle figure cycle-flow of moving people as illustrated in figure 4.

Destination area (D)

Origin area (O)

| | total pop O | T | m1 | m2 | m3 | m4 |
|---|---|---|---|---|---|---|
| m1 | 10 | 0.2 | 5 | 3 | 2 | 0 |
| m2 | 4 | 0.1 | 0 | 4 | 0 | 0 |
| m3 | 17 | 0.5 | 7 | 1 | 8 | 1 |
| m4 | 5 | 0.3 | 0 | 4 | 0 | 1 |
| | Total pop D | | 12 | 12 | 10 | 2 |

(a) Initial Matrix (M ): origin-destination matrix




(b) Flow-cycle graph

**Figure 5**. Cycle-flow of moving people

Figure 5 (a) shows example of moving activity people. The green one indicates origin area, and the yellow one is destination area. The blue color indicates population of Origin (O) in the evening (population that is registered by government). The brown color indicates population of Destination (D) in the morning. T is Individual Transmission risk. This matrix is called initial matrix M. The arrow designates direction of moving people. Then, moving activity can be translated in graph as shown in figure 5(b).

Then, we composed state space to describe inner system behavior of dynamic-population flow model. It is important because through this transfer function, controllability or observability can be defined and determined. Then, we can monitor and control dengue spreading.

The equation in state-space is written as:

$$X2_m = X1_m(t+1) \quad = P2M_n X1_m(t) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(6)$$

$$X3_m = X2_m(t+1) \quad = P2E_n X2_m(t) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

We can represent this equation in matrix

$$\begin{bmatrix} \frac{dX1_m}{dt}(t) \\ \frac{dX2_m}{dt}(t) \end{bmatrix} = \begin{bmatrix} \frac{1}{P2M_m} & 0 \\ 0 & \frac{1}{P2E_m} \end{bmatrix} \begin{bmatrix} X1_m(t) \\ X2_m(t) \end{bmatrix} \qquad \ldots\ldots(8)$$

with : $X1_m(t) = H1E_m X0_m(t)$ as initial condition
where $X0_m(t)$ = population in origin O

As an output we choose H2M and H2E. H2M represent resistance of area in the morning, it is written as:

$$H2M(t) = \sum_{1}^{m} P2M_n X1_m(t) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

While H2E represent resistance of area in the evening, it is written as:

$$H2E(t) = \sum_{1}^{m} P2E_n X1_m(t) \qquad \text{.........................(10)}$$

Healthiness matrix will be applied as parameter to determine which area is infected by compare Healthiness standard (not infected) with Healthiness infected. Comparison process will be given detail in next section.

### *2.3.3 Scenario of Local-Global Spreading*

The life cycle of dengue virus is strictly dependent on only mosquitoes and humans as shown in figure 6.

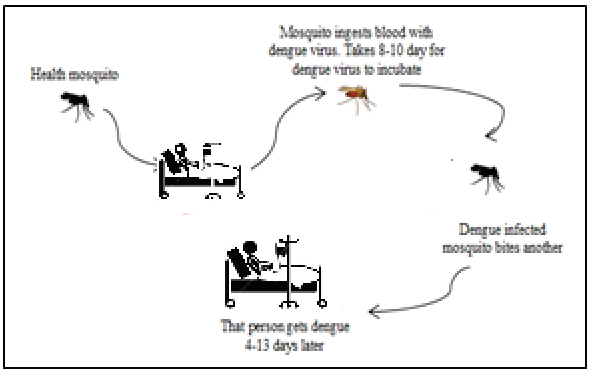



**Figure 6.** Dengue spreading cycle

Transmission of dengue fever Transmission of dengue virus occurs when health mosquito bites infected people and the health mosquito becomes infected mosquito. Time duration from health mosquito to be in infected mosquito so-called mosquito incubation time, it is 7 days. If infected mosquito bites health people, the health people become infected people. Time duration from health people to be infected people namely human incubation time, it takes 3-14 day. After the incubation time, the person can be detected as a dengue patient. Then people struggle against dengue virus and recovered, called recovery time is 7-10 days [404].

From this information above, it is important to handling dengue based on cycle of dengue disease. By combining cycle of human moving and cycle of dengue disease we write scenario of spreading. This scenario is illustrated in figure 7

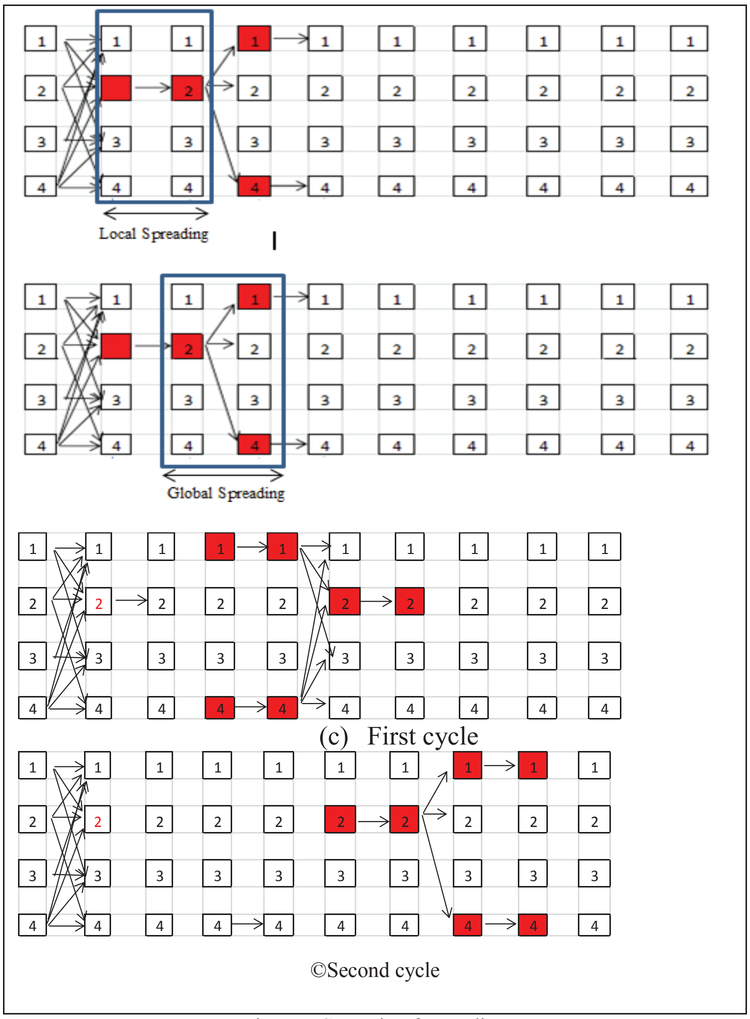


Figure 7 Scenario of spreading

In first cycle, area number 2 is infected. Then spreading happened. Local spreading is denoted by arrow from area with same number. In the same time, global spreading probably happened. Global spreading means disease transmit to another area. It takes a time from somebody catching an infection and symptoms appearing, it called incubation time (Ti).

To analyze the next spreading we investigate number of people in area2, area1 and area4. If somebody sick we assume that they stay at home. When they absent, number of people in one area is increase while others is decrease. We can see the differences in matrix H2M, H2E and H3M. Differences between current matrixes to standard matrix happened only in infected area.

The next step is determine query Then we can focus on area where area number 2 is connected with. The area connection is area1 and area 4. So, next spreading will be happened in area 1 or 4. After 4-13 days, we will get information of infected people in suspected area. Then, second cycle is started. During second cycle people in first cycle was recovered or die. This process is repeated until no one is infected.

### *2.3.4. Algorithm of Local-Global Spreading*

Algorithm of spreading is run under assumption bellows:

- The system is isolated, during spreading there is no migration in system
- pattern of human moving is same everyday

Step in spreading

1. Initialization: - weather factor: W(t-1)
   - Dengue-contracting risk value $\vec{L} = (L_1(t), L_2(t), L_3(t), \dots L_n(t))$
   - Initial origin-destination vector $\vec{M}(t) = (M_1(t), M_2(t), \dots M_n(t))$

2. Calculate $P_n(t)$ by eq (2)
3. Calculate Risk Area H1M(t), H2M(t), H2E(t) standard with equation (6)-(8).
4. Input : Infected people $\overrightarrow{MI_n}(t+1)$
   if Infected people = $\overrightarrow{MI_n}(t+1) = \big(M_1(t+1), \dots M_n(t+1)\big)$
   then $T_n(t+1) = 1$
   else $T_n(t+1) = T_n$
5. Repeat step 3 for t= t+1
6. Calculate $DH1M_m = H1M_m(t+1) - H1M_m(t)$
   $DH2E_m = H2E_m(t+1) - H2E_m(t)$
   $DH2M_m = H2M_m(t+1) - H2M_m(t)$
7. Check : if $DH1M_m = 0,\ DH2E_m = 0\ ,\ DH2M_m = 0$
   then stop
   else select from vector $\vec{M}(t)$ to get $AI_n(t+1)$
   $AI_n(t+1)$are Area Infected with specification
   $DH1M_m \neq 0,\ \ DH2E_m \neq 0,\ DH2M_m \neq 0$
8. Select next suspected area from vector $\overrightarrow{MI_n}(t+1)$by inputting $AI_n(t+1)$
9. Update standard H1M(t), H2M(t), H2E(t) with H1M(t+1), H2M(t+1), H2E(t+1)
10. Repeat process for t = t+n with condition :
    H1M(t-2)= H1M(t), H2M(t-2)= H2M(t), H2E(t-2)= H2E(t)

## 3.Simulation and Analysis

Area of study is Surabaya as shown in figure 8. We divided the population of Surabaya into 2251 patches with size 40m x 40m and performed simulation by real data of infected people in Surabaya in 2011. The simulation was conducted for 4 weeks in January 2011. We applied proposed algorithm with assumption that during incubation time there are no overlapping infection with other disease. Also, pattern of human moving is always same during week day. In week end, we assumed that people do not making their daily commute and stay at their area, therefore the evening force of infection remains the same as during the week

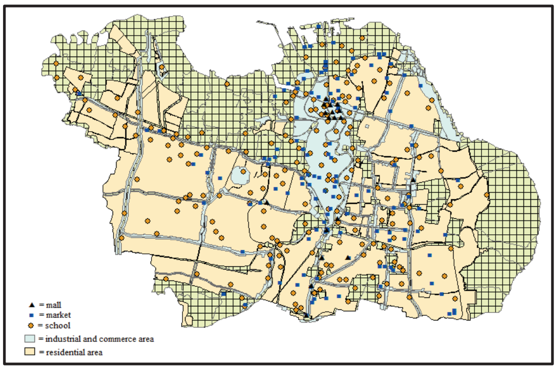


Figure 8 Area of study

### 3.1. Human Moving Pattern

Data infected people is collected from Ministry of Health in Surabaya. Data population is gathered from Surabaya Municipal government .We also conducted interview to get information about the place where people usually visit in between 09.00 to 18.00. In this case, we choose places that people spend 4-8 hours in the morning, in week day and week end. Table 1shows data gathered from 161 house,857people in Manukan Area.

**Table 1**. Questionnaire

| | | | | | Address of activity | |
|---|---|---|---|---|---|---|
| No | Name | Gender | Age | Address | Monday-Friday | Saturday-Sunday |
| 1 | Puji | M | 43 | Manukan Yoso 7G/1 | Perak | Home |
| | Retno | F | 42 | | Home | Home |
| | Karin | F | 18 | | SMA 11 | Home |
| | Zulfah | F | 15 | | SMP | Home |
| | Balqis | F | 9 | | SD Manukan | Home,scholl |
| | Bintang | M | 5 | | TK Manukan | Home |
| 2 | Zainul A | M | 38 | Manukan Yoso 7G/2 | Sikatan | Home |
| | Ferliana | F | 34 | | Tsanawiyah Rungkut | Home |
| | Sadian | M | 5 | | TK Melati | Home |

From the questionnaire result we build table of origin-destination that as shown in figure 9. Figure 9 shows fraction of population in the morning. This fraction becomes parameter in human side.

| Origin | 1292 | 1294 | 1296 | 1298 | 1301 | 1354 | 1355 | 1356 | 1357 | 1358 | 1359 | 1360 | 1361 | 1417 | 1476 | tot_O | tot_D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1292 | 12 | 0 | 23 | 12 | 0 | 0 | 8 | 0 | 0 | 12 | 15 | 0 | 0 | 0 | 0 | 82 | 52 |
| 1294 | 3 | 8 | 13 | 0 | 1 | 1 | 10 | 5 | 8 | 2 | 0 | 0 | 0 | 0 | 0 | 51 | 49 |
| 1296 | 6 | 2 | 15 | 0 | 2 | 0 | 3 | 0 | 8 | 4 | 6 | 2 | 8 | 12 | 2 | 70 | 102 |
| 1298 | 0 | 3 | 12 | 21 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 38 | 77 |
| 1301 | 2 | 0 | 0 | 2 | 13 | 1 | 1 | 9 | 0 | 12 | 0 | 10 | 1 | 1 | 1 | 53 | 40 |
| 1354 | 4 | 16 | 0 | 3 | 0 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 53 | 36 |
| 1355 | 7 | 3 | 0 | 10 | 12 | 0 | 12 | 0 | 0 | 0 | 2 | 0 | 4 | 0 | 2 | 52 | 43 |
| 1356 | 0 | 6 | 4 | 1 | 0 | 0 | 0 | 18 | 2 | 4 | 8 | 2 | 0 | 0 | 0 | 45 | 42 |
| 1357 | 2 | 10 | 2 | 0 | 2 | 0 | 1 | 0 | 16 | 0 | 6 | 0 | 2 | 0 | 0 | 41 | 51 |
| 1358 | 5 | 0 | 8 | 0 | 2 | 0 | 0 | 0 | 6 | 21 | 11 | 6 | 0 | 1 | 4 | 64 | 74 |
| 1359 | 2 | 0 | 6 | 10 | 5 | 4 | 4 | 2 | 6 | 10 | 16 | 12 | 2 | 0 | 0 | 70 | 87 |
| 1360 | 4 | 0 | 6 | 2 | 1 | 0 | 1 | 0 | 3 | 4 | 0 | 14 | 0 | 1 | 2 | 38 | 73 |
| 1361 | 4 | 0 | 2 | 2 | 0 | 0 | 0 | 2 | 2 | 0 | 8 | 12 | 26 | 0 | 6 | 64 | 55 |
| 1417 | 0 | 0 | 10 | 13 | 0 | 0 | 3 | 6 | 0 | 0 | 12 | 12 | 12 | 30 | 2 | 100 | 45 |
| 1476 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 6 | 2 | 3 | 0 | 0 | 12 | 27 | 31 |

**Figure 9**. Table of fraction population

### 3.2 Simulation Result

Purpose of simulation is to analyze reliability of algorithm. First step, we perform standard for normal condition: there is no infected people. Then, were inputting infected people (number of infected people and the grid they live in) in first week to get suspect of next week infected area. The result is shown in figure 10

| Number of grid | | 1292 | 1294 | 1296 | 1298 | 1301 | 1354 | 1355 | 1356 | 1357 | 1358 | 1359 | 1360 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | I1E | 0.01274038 | 0.211823 | 0.25088 | 2.048299 | 0.047184 | 0.047184 | 0.449631 | 0.944327 | 0.91731 | 0.42875 | 0.32768 | 2.0483 |
| | I2M | 0.01537494 | 0.224418 | 0.25516 | 2.061281 | 0.052972 | 0.055124 | 0.491564 | 1.005577 | 0.94244 | 0.443377 | 0.33668 | 2.0635 |
| | I2E | 0.02811533 | 0.436241 | 0.50604 | 4.10958 | 0.100156 | 0.102308 | 0.941196 | 1.949905 | 1.85975 | 0.872127 | 0.66436 | 4.1118 |
| standard | H2M | 22.0946793 | 27.51468 | 58.9969 | 60.95343 | 17.19796 | 2.985253 | 13.11863 | 20.64705 | 31.6072 | 39.35189 | 34.8362 | 50.381 |
| | H2E | 45.091711 | 23.63641 | 31.5649 | 47.12798 | 43.28308 | 11.48975 | 34.64018 | 33.00771 | 21.193 | 42.62527 | 67.5861 | 44.591 |
| | H3M | 45.010442 | 56.127 | 119.951 | 124.05 | 35.16124 | 6.263086 | 27.22942 | 43.07813 | 64.116 | 80.00193 | 72.1432 | 102.91 |
| Infected | H2M | 22.0946793 | 27.51468 | 58.9969 | 60.95343 | 17.19796 | 2.985253 | 13.11863 | 20.64705 | 31.6072 | 39.78064 | 34.4074 | 50.381 |
| | H2E | 45.091711 | 23.63641 | 31.5649 | 47.12798 | 43.28308 | 11.48975 | 34.64018 | 33.00771 | 21.193 | 42.73197 | 67.5861 | 44.591 |
| | H3M | 45.010442 | 56.127 | 119.951 | 124.05 | 35.16124 | 6.263086 | 27.22942 | 43.07813 | 64.116 | 80.87406 | 71.2711 | 102.91 |
| | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.87213 | 0.87213 | 0 |

**Figure 10** Result of Healthiness Area

Standard result is healthiness area without infected people. Infected represents value of healthiness area after we inputting infected people. Figure 10 shows that risk of some area that connected with infected people will increase next week. Risk of area every week is illustrated in figure 11.

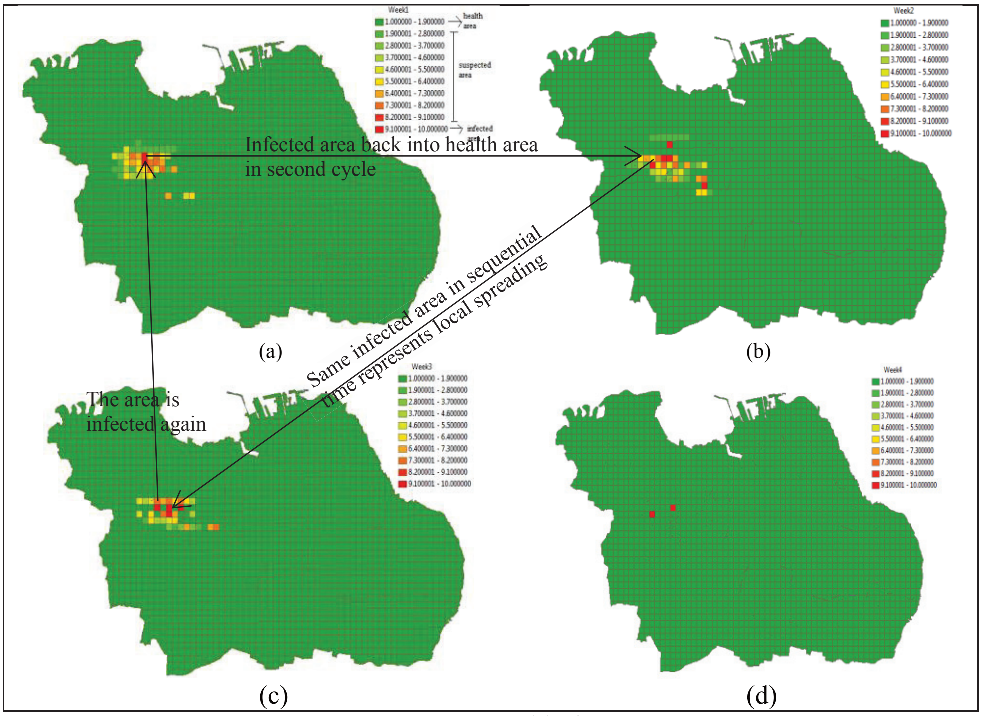



**Figure 11**. .Risk of area
(a) 1 st week January 2011
(b) 2nd week January 2011
(c) 3rd week January 2011
(d) 4th week January 2011

Our simulation shows that the next suspected area has lesser probability compares with standard. We found that the next infected area mostly coming from the lowest healthiness area. Variation of healthiness value is important to decide mathematic model of controller. In our study, controller are policy in chemical treatment, self-protection and promote awareness in hygiene. In this simulation, people's destination is close to their house. It is possible due to easiness to get facility for living there. Therefore, simulation result showed that disease possible to return as cycle. Pattern of disease transmission reflects house to house movement [41]. Therefore, transformation of disease is influenced by social structure [42]. Also, probability of local spreading is bigger than global spreading. For this type, it easy to localizes the disease.

To validate our simulation we compare result of simulation with real infected people. There are some correct patterns, however there are several errors. This error is particularly in consequence of absence of weekend pattern. The error is shown in red circle.

### 4. Conclusion and Future Work

We already conducted simulation based on proposed algorithm of real time local-global spreading in Surabaya. The result shows that routine movement play important role in spreading. Our algorithm suitable to represents real time both of local and global spreading. However, we have to consider about random movement, especially in week end. This finding simplifies the control of dengue spreading by adaptive prediction for next near future because the model is data-driven base.

Furthermore, this study provides a new realistic approach to control dengue. Through this model, quick action will be executed. This action is set up by controller. Our next research will be focused on build control of this transfer function.

**Acknowledgments**
This research is mainly supported by Ministry of Education Indonesia scheme INTERNATIONAL RESEARCH COLLABORATION between EEPIS and KEIO University and partly by MEXT Grant-in-Aid for the Program for Leading Graduate School. The authors thanks the Ministry of Health district Surabaya for providing dengue case data.

**References**

[1] www.who.int/csr/disease/dengue

[2] Departemen Kesehatan, 2010, Buletin Jendela Epidemologi, Volume 2 tahun 2010, http://www.depkes.go.id/downloads/publikasi/buletin/BULETIN%20DBD.pdf

[3] www.cdc.gov/Dengue

[4] A Anna L Buczak, 2012**,** A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data, *BMC Medical Informatics and Decision Making*

[5] Daniel Parker ,Darryl Holman ,2012, Event history analysis of dengue fever epidemic and inter-epidemic spells in Barbados, Brazil, and Thailand, International Journal of Infectious Diseases 16 (2012) e793–e798

[6] E.L. Estallo, A.E. Carbajo, M.G. Grech, M. Frías-Céspedes, L. López, M.A. Lanfri, F.F. Ludueña-Almeida, W.R. Almirón, Spatio-temporal dynamics of dengue 2009 outbreak in Córdoba City, Argentina, Acta Tropica, Available online 29 April 2014

[7] M. Sriprom, K Chalvet-Monfray, T Chaimane, K Vongsawat, D.J. Bicout, 2010, Monthly district level risk of dengue occurrences in Sakon Nakhon Province, Thailand, Science of The Total EnvironmentVolume 408, Issue 22, 15 October 2010, Pages 5521–5528

[8] S. Ly1,∗, V. Duong1, A. Rachmat2, C. Yasuda2, C. Ngan3, R. Huy3,S. Ong1, W. Rogers2, P. Buchy1, A. Tarantola1,2012, Dengue attack rate and proportion of asymptomatic infections in the Kampong Cham prospective community study 2011: Preliminary Data, International Journal of Infectious Diseases Volume 16, Supplement 1, June 2012, Pages e14

[9] Aziz Shafie, 2011, Evaluation of the Spatial Risk Factors for High Incidence of Dengue Fever and Dengue Hemorrhagic Fever Using GIS Application, Sains Malaysiana 40(8)(2011): 937–943

[10] Gonzalo M Vazkuez-Prokopec ,2011, et all Quantifying the Spatial Dimension of Dengue Virus Epidemic Spread within a Tropical Urban Environment, http://www.plosntds.org /rticle/info%3Adoi%2F10.1371%2Fjournal.pntd.0000920, 2011

[11] AAbdulatif Alharty, 2009, Role of GIS in Dengue Control Management Strategy at Jeddah Municipality,www.saudis.org/ FCLFiles/File/ 33_E_ AbdullatifAlharty_ KSA.pdf

[12] Liliam Cesar et all, 2012, Modelling the Dynamic Transmission of Dengue Fever : Investigating Dissease Persistence, www.plosntds.org/ article /info%3Adoi% 2F10.1371 % 2 Fjournal.pntd.0000942

[13] Myriam Gharbi1, Philippe Quenel, Joël Gustave3 Sylvie Cassadou, Guy La Ruche4, Laurent Girdary and Laurence Marrama, Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors, BMC Infectious Diseases 2011, 11:166

[14] Md. Shahin Mia, Rawshan Ara Begum, A.C. Er, Raja Datuk Zaharaton Raja Zainal Abidin, Joy Jacqueline Pereira, 2013, Trends of dengue infections in Malaysia,Neli R.S. Ortega, et all, 2008, A Fuzzy Reed–Frost Model for Epidemic Spreading, Bulletin of Mathematical Biology (2008) 70: 1925–1936

[15] N W. Y. Wan Fairos, W. H. Wan Azaki, L. Mohamad Alias, and Y. Bee Wah,2010, Modelling Dengue Fever (DF) and Dengue Haemorrhagic Fever (DHF) Outbreak Using Poisson and Negative Binomial Model, World Academy of Science, Engineering and Technology 38 2010

[16] AAzad Mohammed, Dave D. Chadee∗, 2011, Effects of different temperature regimens on the development of Aedes aegypti (L.) (Diptera: Culicidae) mosquitoes, Acta Tropica 119 (2011) 38–43

[17] Marcello Otero, Nicolás Schweigmann, Hernán G. Solari, A Stochastic Spatial Dynamical Model for *Aedes Aegypti,* Bulletin of Mathematical Biology, (2008), DOI 10.1007/ s11538-008-9300-y

[18] Roberto Barrera*, Manuel Amador, Andrew J. MacKay, Population Dynamics of Aedes aegypti and Dengue as Influenced by Weather and Human Behavior in San Juan, Puerto Rico, 2011, http://www.plosntds.org

[19] Sair Arboleda 1, Nicolas Jaramillo-O. 1 and A. Townsend Peterson, Mapping Environmental Dimensions of Dengue Fever Transmission Risk in the Aburrá Valley, Colombia, Int. J. Environ. Res. Public Health 2009, 6, 3040-3055

[20] Misael Enrique Oviedo Pastrana, Rachel Lage Brito, Rafael Romero Nicolino, Camila Stefanie Fonseca de Oliveira, João Paulo Amaral Haddad, 2014, Spatial and statistical methodologies to determine the distribution of dengue in Brazilian municipalities and relate incidence with the Health Vulnerability Index, Spatial and Spatio-temporal Epidemiology, Available online 13 April 2011

[21] Wolf-Peter Schmidt1, et all, 2011, Population Density, Water Supply, and the Risk of Dengue Fever in Vietnam: Cohort Study and Spatial Analysis, PLoS Medicine | August 2011 | Volume 8 | Issue 8 | e1001082

[22] Abhishek Pandey, Anuj Mubayi, Jan Medlock,2013, Comparing vector–host and SIR models for dengue transmission , Mathematical Biosciences, Volume 246, Issue 2, December 2013, Pages 252-259

[23] Mathuros Tipayamongkholgul1, 2011, Socio-geographical factors in vulnerability to dengue in Thai villages: a spatial regression analysis, Geospatial Health 5(2), 2011, pp. 191-198

[24] Robert C. Reiner Jr, Steven T. Stoddard, Thomas W. Scott, 2014, Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal, Epidemic, Volume 6, March 2014, Pages 30–36

[25] M Magali Teurlei et all,2012, Can Human Movement Explain Heterogenous Propagation of Dengue Fever in Cambodia?, www.plosntds.org/ article / info% 3Adoi% 2F10.1371% 2Fjournal .pntd.0001957

[26] M Andraud, Niehl Hens, Christiaan Marais, Phillipe Beutel, 2012, Dynamic Epidemilogical Models for Dengue Transmission: Systematic Review of Structural Approaches PLoS ONE7(11): e49085. doi:10.1371 /journal.pone.0049085
[27] Vazquez-Prokopec G.M., et all,2013, Linking fine scale human mobility and social contacts to understand infectious disease dynamics within a resource-poor urban center.PLoS ONE.8(4):e58802
[28] Mniszewski, A. ManoreTulane , S. Y. Del Valle,D. Roberts, Towards a hybrid agent-based model for mosquito borne disease SummerSim '14 Proceedings of the 2014 Summer Simulation Multiconference,2014
[29] Courtney D. Corley, Laura L Pullum, David M Hartley, Corey Benedum, Christine Noonan, Peter M Rabinowitz, Mary J Lancaster, Disease Prediction Models and Operational Readiness, http:// journals.plos.org/plosone/ article?id=10.1371/ journal.pone.0091989#pone.0091989-Pitman1
[30] Timothy A. McLennan-Smith, Geoffry N. Mercer, Complex behaviour in a dengue model with a seasonally varying vector population, Mathematical Biosciences 248 (2014) 22–30]
[31] N.K.Kameswara Rao* et al, Network Architecture to Identify Spatial Knowledge for Dengue (IJITR) INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND RESEARCH,Volume No. 1, Issue No. 2, February - March 2013, 155 – 16010
[32] Enrique Fr´ıas-Mart´ınez , Graham Williamson , Vanessa Fr´ıas-Mart´ınez ,An Agent-Based Model of Epidemic Spread using Human Mobility and Social Network Information, 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing,page 58-64
[33] Yuanzheng Ge, Zhichao Song, Xiaogang Qiu , Hongbin Song, Yong Wang, Modular and hierarchical structure of social contact networks, Physica A 392 (2013) 4619–4628
[34] Mariagrazia Mecoli a,n, VandaDeAngelis a, SallyC.Brailsford , Using system dynamics to evaluate control strategies for mosquito-borne diseases spread by human travel, Computers & Operations Research 40 (2013) 2219–2228
[35] Vitaly Belik, Theo Geisel, and Dirk Brockmann, Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases, PHYSICAL REVIEW X 1, 011001 (2011)
[36] Wahjoe T Sesulihatien, Yasushi Kiyoki , A Dengue Location-Contraction Risk Calculation Method for Analyzing Disease-Spread, Information, Information Modelling and Knowledge Base XXVI, 2014
[37] Wiwiek Setya Winahju,, Adatul Mukarromah, Modeling Dengue Cases Using Poisson INAR, Procedia Engineering 50 ( 2012 ) 837
[38] P Ezanno , M.Aubry-Kientz , SArnoux , P Cailly , G LAmbert , C. Toty , T. Balenghien , A. Tran, A generic weather-driven model to predict mosquito population dynamics applied to species of *Anopheles*, *Culex* and *Aedes* genera of southern France, Preventive Veterinary Medicine, Volume 120, Issue 1, 1 June 2015, Pages 39–50
[39] Steven T. Stoddard1*, Amy C. Morrison1, Gonzalo M. Vazquez-Prokopec2, Valerie Paz Soldan3, Tadeusz J.Kochel4, Uriel Kitron2, John P. Elder, The Role of Human Movement in the Transmission of Vector-Borne Pathogens, www.plosntds.org, July 2009
[40] http ://www.cdc. gov/dengue/entomologyecology

[41] Steven T Stoddart et all, 2012, House to House Human Movement Drive Dengue VirusTransmission, http://www.pnas.org
[42] Robert C. Reiner Jr, Steven T. Stoddard, Thomas W. Scott, 2014, Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal, Epidemic, Volume 6, March 2014, Pages 30–36

# Multi-Dimensional Semantic Computing with Spatial-Temporal and Semantic Axes for Multi-spectrum Images in Environment Analysis

Yasushi KIYOKI,* Xing CHEN,** Shiori SASAKI* and Chawan KOOPIPAT***

* *Department of Environment and Information Studies,*
*Keio University*
*Fujisawa, Kanagawa 252-8520, Japan*
*kiyoki@sfc.keio.ac.jp*
** *Department of Information & Computer Sciences*
*Kanagawa Institute of Technology*
*1030 Simo-Ogino, Atsugi-shi, Kanagawa 243-0292, Japan*
*chen@ic.kanagawa-it.ac.jp*
*** *Department of Imaging and Printing Tech.*
*Faculty of Science, Chulalongkorn University*
*Bangkok, Thailand 10330*
*chawan.k@chula.ac.th*

**Abstract.** Semantic computing is an important and promising approach to semantic analysis for various environmental phenomena and changes in real world. This paper presents a new semantic computing method with multi-spectrum images for analyzing and interpreting environmental phenomena and changes occurring in the physical world.

We have already presented a concept of "Semantic Computing System" for realizing global environmental analysis. This paper presents a new semantic computing method to realize semantic associative search for the multiple-colours-spectrum images in the multi-dimensional semantic space, that is *"multi-spectrum semantic-image space"* with semantic projection functions. This space is created for dynamically computing semantic equivalence, similarity and difference between multi-spectrum images and environmental situations.

We apply this system to global environmental analysis as a new platform of environmental computing. We have already presented the 5D World Map System, as an international research environment with spatio-temporal and semantic analysers. We also present several new approaches to global environmental-analysis for multi-spectrum images in *"multi-spectrum semantic-image space."*

## 1. Introduction

Multi-spectrum images are important sensing data in the natural environment research field. A lot of algorithms, methods and tools have been developed for analyzing multi-spectrum images and interpreting various images related to natural phenomena. Prior research has shown that those tools are efficient to find natural phenomena and environmental situations from the multi-spectrum images.

We have already presented a new concept of "Semantic Computing System [1,2,7]" for realizing global environmental analysis. Semantic computing is an important and promising approach for analyzing various environmental phenomena and changes in the real world.

This paper presents a new semantic computing method with multi-spectrum images for interpreting the meanings of environmental phenomena and changes occurring in the physical world. This method focuses on a new concept connecting semantics with the "*waving features*" of multi-spectrum images. In the method, the semantic associative search method is applied for the multiple-colors-spectrum images. The "*multi-spectrum semantic-image space*" is created with semantic projection functions for dynamically computing semantic equivalence, similarity and difference between multi-spectrum images and linguistic words. Furthermore, we apply this method to a "global environmental system" as a new platform of environmental computing. Several new approaches are also presented for analyzing multi-spectrum images in the "*n*-spectrum color-dimensional semantic space" using analyzers in a 5D World Map System[4,8], which is an international research environment with spatio-temporal and semantic analyzers developed in our prior work.

The computation tools of multi-spectrum image are widely utilized to assist researchers to analyze observed nature environment data and to explain environment changes, for example, the global woodland change. The essential computation in environmental study is context-dependent computation to analyze the changes of various situations (temperature, color, $CO_2$, places of livings, sea level, coral area, etc. [11]). It is important to realize global environmental computing methodology for analyzing difference and diversity of nature and livings in a context dependent way with a large amount of information resources in terms of global environments. That is because humankind, the dominant species on Earth, faces the most essential and indispensable mission; we must endeavor on a global scale to perpetually restore and improve our natural and social environments.

There are many existing image processing [3] and analyzing tools dealing with multi-spectrum images. Prior research has shown that those tools are efficient to help human being finding semantics from the multi-spectrum images. It is very important to find semantics from sensing data in various aspects and contexts, in order to discover what are happening in the nature of our planet. From this background, it is essentially important to realize semantic computing methodology for analyzing sensing data of nature and livings in a context dependent way.

We have various (almost infinite) aspects and contexts in environmental changes in our planet, and it is essential to realize a new analyzer for computing semantics in those situations for discovering actual aspects and contexts existing in the nature. We have proposed a method for semantic computing in our Multi-dimensional World map [4,8]. We utilize a multi-dimensional computing model, the Mathematical Model of Meaning (MMM) [1,2,7], and a multi-dimensional space filtering method with adaptive axis adjustment mechanisms to implement semantic computing. This system realizes semantic associative computing and search for media data and it is applied to compute semantic correlations between keywords, images, music and documents dynamically in a context-dependent way. The main feature of this system is to realize semantic associative search in the 2000 dimensional orthogonal semantic space with semantic projection functions. This space is created for dynamically computing semantic equivalence or similarity between keywords and media data. The semantic correlated factors will be highlighted after semantic filtering, and the highlighted factors are visualized using our Multi-dimensional World Map.

In the design of environment-analytical systems for interpreting the meanings from multi-spectrum images, one of the most important issues is how to search and analyze the images according to user's contexts. After a brief review of the outline of MMM, we will present a new semantic computing method with multi-spectrum images in the following sections. A new concept applied in the new method will be presented which connects semantics with the "*waving features*" of multi-spectrum images.

## 2. Semantic Computing and Multi-spectrum Image Space

In this section, the outline of our semantic associative search method based on the Mathematical Model of Meaning is briefly reviewed. This model has been presented in [1,2,7] in detail. We will also introduce the basic idea of the Multi-dimensional World map [4, 8].

### 2.1 Semantic Computing

In the Mathematical Model of Meaning (MMM) [1,2,7], an orthogonal semantic space is created for semantic associative search. Retrieval candidates and queries are mapped onto the semantic space. The semantic associative search is performed by calculating the correlation of the retrieval semantic space.

In MMM, the acquisition of information or knowledge is performed by semantic computations. Context-dependent interpretation means that information is dynamically extracted by a semantic computation with context-recognition. The method realizes the computational machinery for recognizing the meaning of contexts and obtaining the semantically related information to the given context. MMM is essentially different from those methods. The essential difference is that this method provides dynamic recognition of the context. That is, the "context-dependent interpretation" is realized by dynamically selecting a certain subspace from the entire semantic space. The other methods do not provide the context dependent interpretation, that is, their space is fixed and static. The outline of MMM [1,2,7] is summarized as the following:

(1) A set of $m$ words is given, and each word is characterized by $n$ features. That is, an $m$ by $n$ matrix $\boldsymbol{M}$ is given as the data matrix.
(2) The correlation matrix $\boldsymbol{M}^T\boldsymbol{M}$ with respect to the $n$ features is constructed from the matrix $\boldsymbol{M}$. Then, the eigenvalue decomposition of the correlation

matrix is computed and the eigenvectors are normalized. The orthogonal semantic space ***MDS*** is created as the span of the eigenvectors which correspond to nonzero eigenvalues.

(3) **“Context words” and “image”** are characterized as **“context”** by using the *n* features and representing them as *n*-dimensional vectors.

(4) The context words and **“image”** are mapped into the orthogonal semantic space by computing the Fourier expansion for the *n*-dimensional vectors.

(5) A set of all the projections from the orthogonal semantic space to the invariant subspaces (eigen spaces) is defined. Each subspace represents a phase of meaning, and it corresponds to “**context.”**

**(6) A subspace of the orthogonal semantic space is selected according to the given “context” expressed in *n*-dimensional vectors,** which are given as **“context”** represented by **“a sequence of words”** and **“image.”**

(7) The most correlated information resources to the given **“context”** are extracted as the selected subspace by applying the metric defined in the semantic space.

The advantages and original points of this method are summarized as follows:

(1) The semantic associative media search based on semantic computation is realized by a mathematical approach. This media search method surpasses the search methods which use pattern matching for associative search. Users can use their own words or images for representing impression and data contents for media retrieval, and do not need to know how the metadata of media data of retrieval candidates are characterized in databases.

(2) Dynamic context recognition is realized using a mathematical foundation. The context recognition can be used for obtaining multimedia information by giving the user's impression and the contents of the information as **“context.”** A semantic space is created as a space for representing various contexts which correspond to its subspaces. A context is recognized by the computation for selecting a subspace.

The essential advantage is that this method provides the important function for semantic projections which realizes the dynamic recognition of **“context.”** That is, the “context-dependent interpretation” is dynamically performed by computing the distance between different media data, information resources and words in a context-dependent way. The context-dependency is realized by dynamically selecting a subspace from the entire orthogonal semantic space, according to **"context."** In this method, the number of phases of contexts is almost infinite (currently $2^{2000}$ in the general English word space [5] and $2^{180}$ in the color-image space, approximately). For semantic associative computations of "Kansei" information[9], this method constructed several actual semantic spaces, such as the general English-word space in 2115 dimensions, the color-image space in 183 dimensions, and music space in 8 dimensions in the current implementations.

### 2.2 *"multi-spectrum semantic-image space"*

In this section, we propose a *"multi-spectrum semantic-image space"* in 6 dimensions ((a) IR cut-off filtered axis, (b) Blue filtered axis, (c) Blue-Green filtered axis, (d) Green filtered axis, (e) Red filtered axis, and (f) Infra-Red pass filtered axis). The brightness in the corresponding pixel in each filtered image is mapped as the value of the corresponding axis. For example, when "Spot-S" in some place is mapped into 6-dimensional multi-spectrum image space, the brightness of the pixel-A in the axis is defined as a value of Spot-A in this axis. That is, six values are defines for Spot-S in six axes in this semantic space, and we can apply semantic computing in MMM to the multi-spectrum images.

The 6-dimentional *"multi-spectrum semantic-image spac*e" consists of 6 axes, each of which corresponds to a single spectrum image filtered by a spectrum (color) filter, as shown in Figure 1.

(Axis-A) IR cut-off filtered axis: This axis is used as similar to human eye, because we use a common Digital single lens reflex (DSLR) camera in which IR cut filter is removed to detect IR spectra using IR pass filter. That is, the camera's IR cutting function (the sensor to cut off the IR spectra) is customized by expanding its wavelength responses to the IR region to acquire multispectral images.

(Axis-B) Blue filtered axis: With Blue filter: wave length 0.45-0.515μm (Visible): this axis of the spectrum is used to analyse the characteristics of the water, the land-use, the soil and the plants. This part is easy to be effected by the air diffusion.

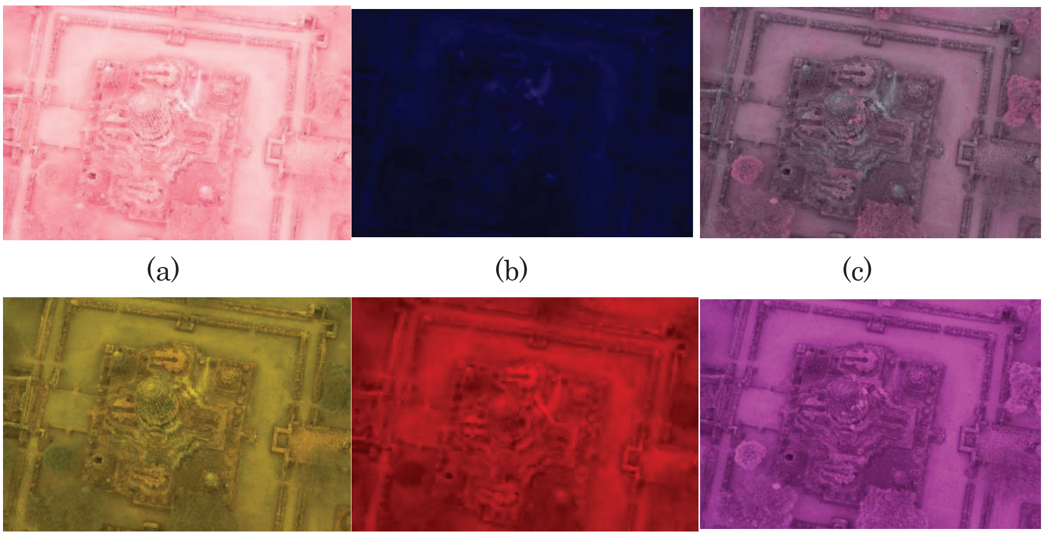
(Axis-C) Blue-Green filtered axis: With Blue-Green filter: wave length

0.515-0.525μm (Visible): this axis of the spectrum is also used to analyse the characteristics of the water, the land-use, the soil and the plants. The leaves turn orange and look bright, while the water turns cyan and looks darker by this filter.

(Axis-D) Green filtered axis: With Green filter: wave length 0.52-0.60μm (Visible): this axis of the spectrum reflects the chlorophyll of the plants with leaves in the color between blue to red. This filter is useful to analysis the plant's activity.

(Axis-E) Red filtered axis: With Red filter: wave length 0.63-0.69μm (Visible): this axis reflects the chlorophyll of the plants with green leaves. This axis is the most important for vegetation analysis, and extraction of the waterline and the differentiation of the borderline of the different soil's nature – dry or wet.

(Axis-F) Infra-Red pass filtered axis: With IR pass filter: wave length 0.76-0.90μm (Invisible): this axis of the spectrum is especially important for ecology because healthy plants reflect it – the water in their leaves scatters the wavelengths back into the sky. By comparing it with other bands, we get indexes, which let us measure plant health more precisely than if we only looked at visible greenness. The leaves are brighter than the normal image, while the water is darker by this filter. This axis corresponds to biomass. This axis is useful to distinguish the land and the water area.



(a) (b) (c)

(d) (e) (f)

**Figure 1**. Examples of multispectral images using: (a) IR cut-off filter, (b) Blue filter, (c) Blue-green filter (BPB50), (d) Green filter, (e) Red filter, (f) Infra-Red pass filter (IR76), (Taken at Wat Phra Ram, Ayutthaya: latitude 14.354152, longitude 100.561763, February, 1st, 2014, 08:50)

We also define environmental-objects in the 6-dimensions, and map them onto the multi-spectrum image space. Those environmental-objects (air-dust, smoke, water, ocean, dry-soil, wet-soil, plant-surface-wax, living-plant, green-vegetation, red-green-vegetation, biomass, water-in-leaves,…) are mapped onto 6-dimensional multi-spectrum image space, as shown in Figure 2.

In this space, an orthogonal semantic space is defined and each pixel is mapped onto the space for computing semantic correlations between pixel and environmental-objects semantic objects, according to various contexts.

To compute semantic correlations, a context image or context words that represent environmental situations and phenomenon are given as a context. According to this context, a "*semantic subspace*" is selected dynamically. Then, the related environmental-objects to each pixel in this context are extracted by computing semantic correlations in the selected semantic subspace.

| Pixel ID | filter color (sensor) | wave length (μ m) | air_dust | smoke | shallow_water | water | ocean | dry_soil | wet_soil | plant_surface_wax | living_plant |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | deep_blue & violet | 0.433–0.453 | 1 | 1 | 1 | | | | | 1 | 1 |
| 2 | blue | 0.450–0.515 | | | | 1 | 1 | −1 | | | |
| 3 | green | 0.525–0.600 | | | | | | | | | 1 |
| 4 | red | 0.630–0.680 | | | | | | 1 | | | |
| 5 | Near InfraRed (NIR) | 0.845–0.885 | | | | | | | | | 1 |
| 6 | Shortwave InfraRed (SWIR) | 1.560–1.660 | | | | | | −1 | 1 | | |
| 7 | Shortwave InfraRed (SWIR) | 2.100–2.300 | | | | | | 1 | −1 | | −1 |
| 8 | panchromatic, b/w | 0.500–0.680 | | | | | | | | | |
| 9 | (very thin slice of wavelengths) | 1.360–1.390 | | | | 1 | | | | | |
| 10 | Thermal InfraRed (TIR) | 10.6-11.2 | | | | | | | | | |
| 11 | Thermal InfraRed (TIR) | 11.5-12.5 | | | | | | | | | |

| Pixel ID | green_vegetation | red_green_vegetation | biomass | water_in_leaves | healthy_plants | scar_of_fire | cloud | cirrus | snow |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | 1 | −1 | | | | | | | |
| 4 | −1 | 1 | | | | | | | |
| 5 | | | 1 | 1 | 1 | | | | |
| 6 | | | | | | −1 | 1 | | 1 |
| 7 | | | | −1 | −1 | 1 | | | |
| 8 | | | | | | | | | |
| 9 | | | | | | | 1 | 1 | 1 |
| 10 | | | | | | 1 | | | |
| 11 | | | | | | 1 | | | |

| Pixel ID | ice | light | temparature | geothermal_heat | stress_of_vegetation | stone | rock | clay | mineral |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | | | | | | | | | |
| 4 | | | | | | | | | |
| 5 | | | | | | | | | |
| 6 | 1 | | | | | | | | |
| 7 | | | | | | 1 | 1 | 1 | 1 |
| 8 | | 1 | | | | | | | |
| 9 | 1 | | | | | | | | |
| 10 | | | 1 | 1 | 1 | 1 | 1 | | |
| 11 | | | 1 | 1 | 1 | 1 | 1 | | |

**Figure 2**. Environmental-objects (air-dust, smoke, water, ocean, dry-soil, wet-soil, plant-surface-wax, living-plant, green-vegetation, red-green-vegetation, biomass, water-in-leaves,…) are mapped onto 6-dimensional multi-spectrum image space.

## 2.3 Multi-dimensional World Map

We have introduced the architecture of a multi-visualized and dynamic knowledge representation system "*5D World Map System*" [4,8] applied to cross-cultural multimedia computing. The basic space of this system consists of a temporal (1st dimension), spatial (2nd, 3rd and 4th dimensions) and semantic dimensions (5th dimension, representing a large-scale and multiple-dimensional semantic space that is based on our semantic associative computing system (MMM). This space memorizes and recalls various multimedia information resources with temporal, spatial and semantic correlation computing functions, and realizes a 5D World Map for dynamically creating temporal-spatial and semantic multiple views applied for various "*environmental information resources*."

We apply the dynamic evaluation and mapping functions of multiple views of temporal-spatial metrics, and integrate the results of semantic evaluation to analyze environmental multimedia information resources. MMM is applied as a semantic associative search method for realizing the concept that "*semantics*" of environmental multimedia information resources, according to the "*context*". The main feature of this system is to create world-wide and global maps and views of environmental features expressed in multimedia information resources (image, music, text and video) dynamically, according to analyst's viewpoints. Spatially, temporally, semantically and evaluated and analyzed environmental multimedia information resources are mapped onto a 5D time-series multi-geographical space. The basic concept of the 5D World Map System is shown in Figure 3.

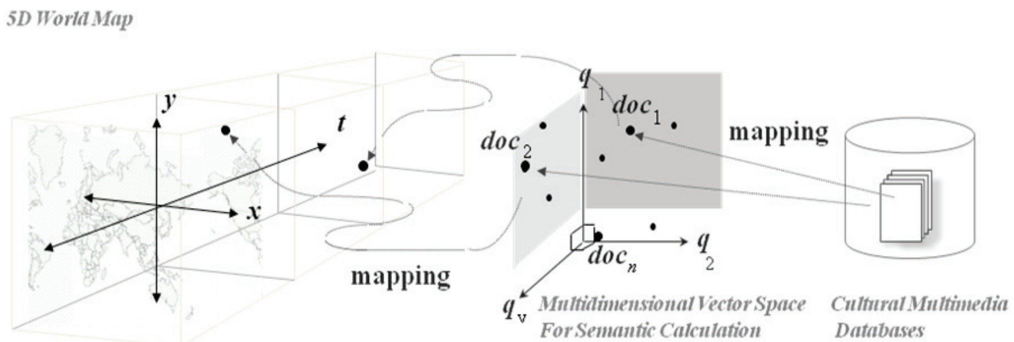



**Figure 3**: 5D World Map System with Spatio, Temporal and Semantic Axes

The 5D World Map system applied to environmental multimedia computing visualizes world-wide and global relations among different areas and times in environmental aspects, by using dynamic mapping functions with temporal, spatial, semantic and color-based computations. We have developed a 5D World Map System for global viewing of environmental multimedia information resources as shown in Figure 4.

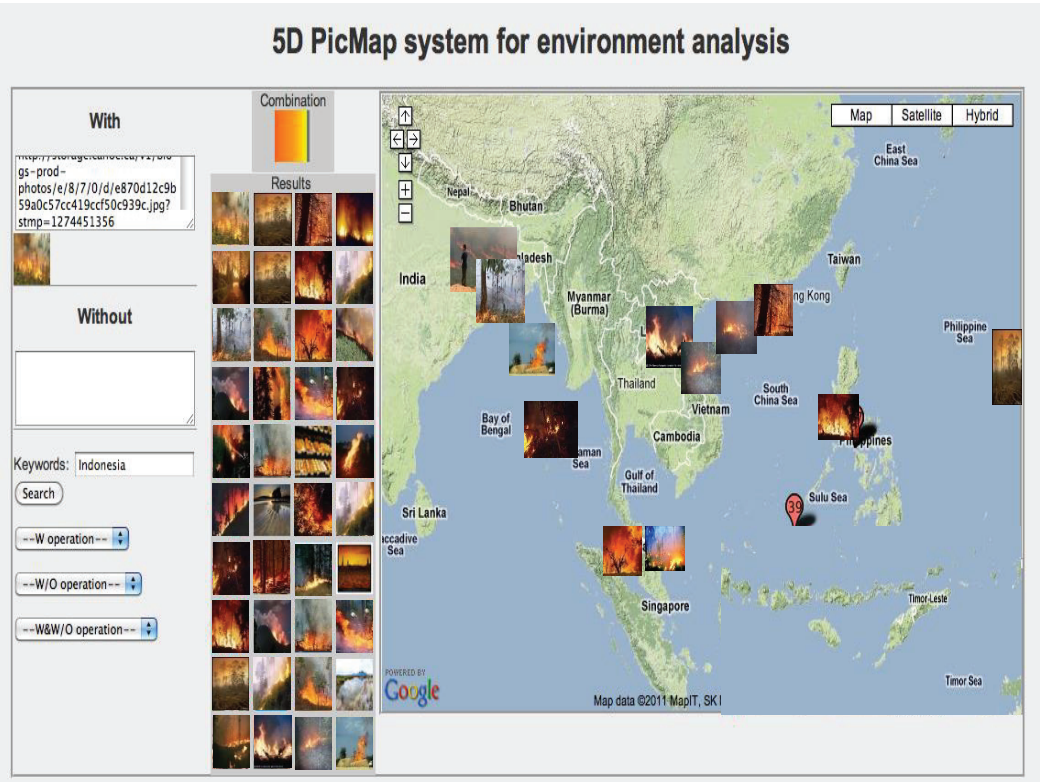



**Figure 4**. The 5D World Map System for global viewing of environmental multimedia information resources

We have realized the “multi-spectrum semantic-image space” in 6 dimensions in the 5D World Map system as shown in Figures 5 and 6. As shown in the figures, the brightness in the corresponding pixel in each filtered image is mapped as the value of the corresponding axis ((a) IR cut-off filtered axis, (b) Blue filtered axis, (c) Green filtered axis, (d) Red filtered axis, and (e) Infra-Red pass filtered axis (IR76) ).

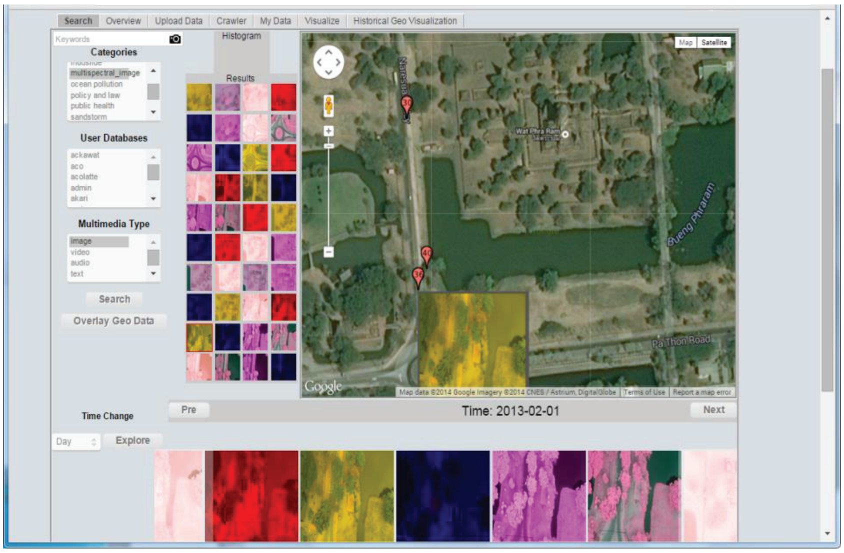


**Figure 5**. Mapping and visualization results of multispectral images using spatiotemporal information (Exif) in 5D World Map. (The photos are taken at Wat Phra Ram, Ayutthaya: latitude 14.354152, longitude 100.561763, February, 1st, 2014, 08:34. The markers on the map indicate the photo-shooting points, where multiple (six) images were taken almost at once.)

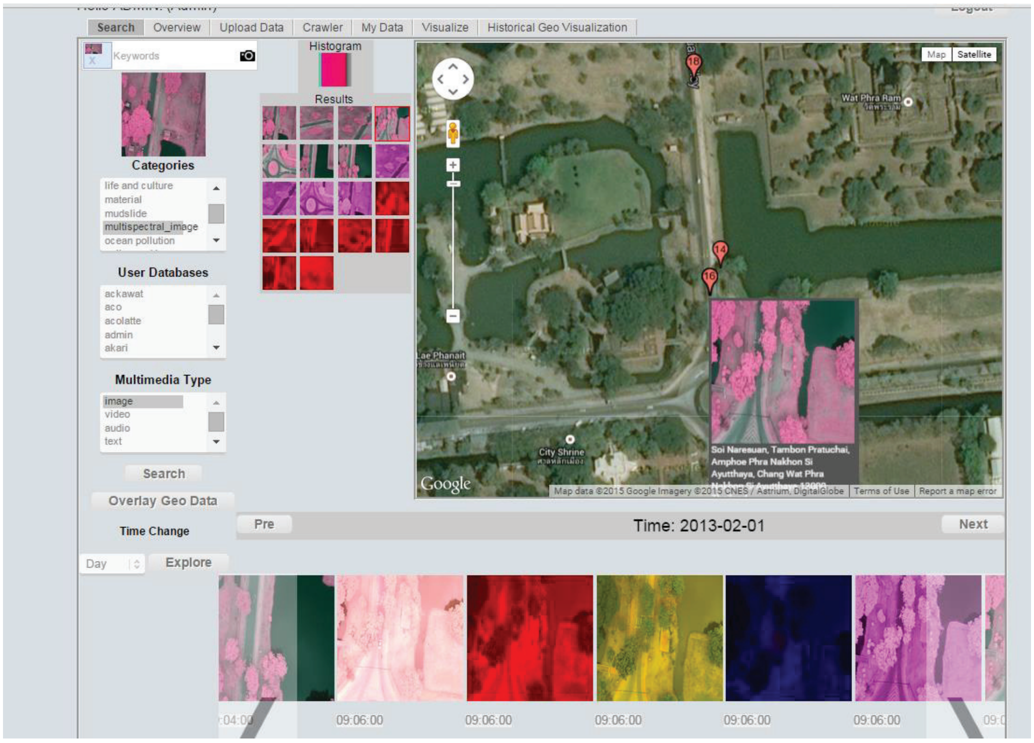



**Figure 6**. Image retrieval for multispectral images using color information in 5D World Map system (The analysts acquire the similar images with the same filter by color information, and also the similar images with the same spatial information by temporal information.)

## 3. Creating Semantic Vectors of Multi-Spectrum Images

In this section, we present the MMM application to Multi-spectrum images. Multi-spectrum images are taken by the multi-spectrum cameras in different ranges of frequencies, referred to as bands, along the electromagnetic spectrum. A lot of multi-spectrum images are opened to publics and accessible through internet. For example, NASA opened the multi-spectrum images taken by their satellites, the Landsat serials [10]. The new Landsat serials is Landsat 8 which has 11 bands. Landsat numbers its red, green, and blue sensors as 4, 3, and 2. The 11 bands are summarized in NASA's Landsat site [10] as the following:

- Band 1 senses deep blues and violets. It has two main uses: imaging shallow water, and tracking fine particles like dust and smoke.
- Band 1 minus Band 2. The ocean and living plants reflect more deep blue-violet hues. Most plants produce surface wax (for example, the frosty coating on fresh plums) as they grow, to reflect harmful ultraviolet light away.
- Bands 2, 3, and 4 are visible blue, green, and red. But while we're revisiting them, let's take a reference section of Los Angeles, with a range of different land uses, to compare against other bands:
- Band 5 measures the near infrared, or NIR. This part of the spectrum is especially important for ecology because healthy plants reflect it – the water in their leaves scatters the wavelengths back into the sky. By comparing it with other bands, we get indexes like NDVI, which let us measure plant health more precisely than if we only looked at visible greenness.
- Bands 6 and 7 cover different slices of the shortwave infrared, or SWIR. They are particularly useful for telling wet earth from dry earth, and for geology: rocks and soils that look similar in other bands often have strong contrasts in SWIR.
- Band 7 is for fire. The fire scar reflects strongly in Band 7 and hardly at all in the others, making it red.
- Band 8 is the panchromatic – or just pan – band. It works just like black and white film: instead of collecting visible colors separately, it combines them into one channel.
- Band 9 is for clouds.
- Bands 10 and 11 are in the thermal infrared, or TIR – they see heat. Instead of measuring the temperature of the air, like weather stations do, they report on the ground itself, which is often much hotter.

Each band measures special wavelength of light. For example, Band 1 measures the light with the wavelength between 0.433–0.453 micrometers. Special wavelength light can be used to recognized special objects on the earth. For example, Band 5 measures the near infrared, or NIR. Healthy plants reflect this part of the spectrum because the water in their leaves scatters the wavelength back into the sky. By comparing it with other bands, we can measure plant health. This leads us an idea of connecting the spectrum with semantics.

Our idea is to create matrix with multiple axes based on spectrum of multi-spectrum images and performing differential computing to connect spectrum of images to semantics. Based on the matrix, an "*n*-spectrum color-dimensional semantic space" is created with semantic projection functions for dynamically computing semantic equivalence, similarity and difference between multi-spectrum images and linguistic words.

We use **P** to represent a vector space, noted as space-P and $\mathbf{P}^{(e)}$ to represent pixels of multi-spectrum images. If the multiplication of $\mathbf{P}^{(e)}$ and **R** can be applied, we define the result of the multiplication as $\mathbf{P}^{(r)}$:

$$\mathbf{P}^{(e)}\mathbf{R} = \mathbf{P}^{(r)}.$$

The multiplication result $\mathbf{P}^{(r)}$ represents the pixels on the semantic space **R**. It is clear that when the semantic space **R** is created from spectrum correlated data sets, pixels will be represented by the spectrum correlated vectors.

The scalar product between pixels on the semantic space **R** shows the correlation between them. Next, we will use an example to illustrate it. If the scalar product of pixel vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ is zero,

$$\mathbf{p}_1{'}\times\mathbf{p}_2 = 0 \quad ,$$

they are orthogonal to each other. That is, the correlation between them cannot be found. When they are transformed to new vectors, their scalar product will be,

$$\mathbf{p}_1^{(r)} = (\mathbf{p}_1{'}\times\mathbf{R})',$$

$$\mathbf{p}_2^{(r)} = (\mathbf{p}_2{'}\times\mathbf{R})',$$

$$\mathbf{p}_1^{(r)'}\times\mathbf{p}_2^{(r)} = (\mathbf{p}_1{'}\times\mathbf{R})''\times(\mathbf{p}_2{'}\times\mathbf{R})'$$

$$= \mathbf{p}_1^{(r)'}\times\mathbf{R}\times\mathbf{R}'\times\mathbf{p}_2^{(r)''}.$$

Non-zero scores will be given to the pixels on the semantic space. That is, the correlation between pixels cannot be found.

When two pixel vectors $\mathbf{p}_i$ and $\mathbf{p}_j$ represent a same spectrum, semantic projection is performed to extract context-specific semantic sub-space. On the selected sub-space, the distribution of the pixels is check to see if there are '*noise*' data exist among them or not. If there are '*noise*' data, the basis vectors which represent the '*noise*' factors are searched and removed from the space, making the sub-space optimal. The event selection, sub-space extraction and sub-space optimization are the basic operations for differential computation.

In our method, the sub-space selection operations are divided into two processing steps. The first step is to extract sub-space based on pixel vectors. The second step is to optimize the sub-space by removing the axes representing the 'noise' factors on the sub-space. These steps are important for highlighting the important spectrum factors. In the following, we describe technical details.

We use a data set $e_i$ to represent a spectrum. We use another data set $E$ to represent total spectrum vector sets. We suppose that $e_i$ is a subset of $E$. That is, data set $E$ includes all spectrum vectors. We construct a matrix, $\mathbf{M}_i$, by using two data sets, $e_i$ and $e_j$. If $E$ is an $n$ rows matrix, $\mathbf{M}_i$ is an $n$ rows and two columns matrix. Each of the rows corresponds to a pixel data set $d_i$, and the first column of the matrix corresponds to the set $e_i$ and the second column corresponds to the set $e_j$. The value of the $(i,j)^{\text{th}}$ entry of the matrix is the spectrum value of a pixel.

If there are $m$ data sets, $e_1$, $e_2$, …, $e_m$, we construct the semantic vector transform matrix $\mathbf{M}$ by uniting the matrixes $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_m$ in the following method:

The $j^{\text{th}}$ column of the matrix $\mathbf{M}$ is defined as the first column of the matrix $\mathbf{S}_j$.

The semantic matrix $\mathbf{M}$ has multiple axes for expressing multi-spectrum images. If a subspace with the basis of vectors $\mathbf{f}_i$, $\mathbf{f}_j$, …, $\mathbf{f}_k$ represents a same feature, a subspace-R, is created by rotating and combining those basis vectors on the semantic space into a new basis vector $\mathbf{r}_i$. Let $B$ represents the set of the vectors $\mathbf{f}_i$, $\mathbf{f}_j$, …, $\mathbf{f}_k$, the basis vectors of the new created semantic vector space can be represented as

$$\mathbf{r}_l = \mathbf{f}_l,\ \ if\ \mathbf{f}_l \notin B,$$

$$\mathbf{r}_i = \frac{1}{\sqrt{\sum_{\mathbf{f}_i \in B} |\mathbf{f}_i|}} \left( \sum_{\mathbf{f}_i \in B} \mathbf{f}_i \right),\ \ \mathbf{f}_i \in B\,.$$

On the subspace, subspace-R, vectors $\mathbf{r}_l$ and $\mathbf{r}_i$ are pair-wise orthogonal because $\mathbf{r}_i$ is on the subspace [$\mathbf{f}_i$, $\mathbf{f}_j$, …, $\mathbf{f}_k$] and the vector $\mathbf{r}_l$ is orthogonal to the subspace.

In the case that each of the pixel value is represented by the $n$-feature vectors,

$$\mathbf{d}_i = d_{1,i}\mathbf{f}_1 + d_{2,i}\mathbf{f}_2 + \cdots + d_{n,i}\mathbf{f}_n,$$

an $n \times n$ semantic matrix $\mathbf{R}$ is created. For subspace selection, an index vector $\mathbf{C}_r$

is generated from a learning data set $L$, which contains the index of all the features in the learning data set $L$. The elements in the learning data set $L$ are the vectorized metadata of the environmental data. The indexing set $\mathbf{C}_r$ is created through the following steps:

**Step-1**: Set $\mathbf{C}_r$ as an $n$-dimensional vector. Each element of the vector $\mathbf{C}_r$ is set to '0'.

**Step-2**: For each element $d_j$ in the learning data set $L$, if the feature value $d_{k,j}$ is greater than a threshold $e$, the $k$-th element of the vector $\mathbf{C}_r$ is set to '1'.

The non-zero elements of the index vector $\mathbf{C}_r$ indicate the feature vectors that should be rotated and combined. After the index vector $\mathbf{C}_r$ is created, the rotating and the combining operations are implemented through the following steps:

**Step-3:** If the $k$-th element is the first non-zero element of the index vector $\mathbf{C}_r$, the $k$-th column of the matrix $\mathbf{R}$ is replaced by the vector $\mathbf{C}_r$. This column is identified as $\mathbf{R}_{Cr}$.

**Step-4**: Remove all the columns of the matrix $\mathbf{R}$ which are indicated by the non-zero elements of the index vector $\mathbf{C}_r$ except the column $\mathbf{R}_{Cr}$ which is replaced by the vector $\mathbf{C}_r$.

After the above four processing steps, the rotating and the combination operations are finished. The removing operation is implemented through the following steps:

**Step-1**: Set $\mathbf{q}$ as an $n$-dimensional zero vector.

**Step-2**: If the $k$-th element is the first non-zero element of the index vector $\mathbf{C}_r$, set the $k$-th element of the vector $\mathbf{q}$ to '1'.

**Step-3**: For each metadata $\mathbf{d}_i$, calculate the inner product of $\mathbf{d}_i$ with the vector $\mathbf{q}$ on the semantic space $\mathbf{R}$:

$$p_i = (\mathbf{d}_i \times \mathbf{R}) \times (\mathbf{q}' \times \mathbf{R})'.$$

This step will be repeated until the inner products of all the metadata are calculated.

**Step-4**: Reversely sort the metadata based on the inner products calculated by **Step-3**: The ranked position of $\mathbf{d}_i$ is stored into a variable $Rank_i$, where, $i$ is the index used to indicate the metadata $\mathbf{d}_i$, its inner product and its ranking position value $Rank_i$. If the ranking value of $\mathbf{d}_i$ is one, $Rank_i = 1$, $\mathbf{d}_i$ is ranked at the top position.

**Step-5**: For each combined feature vectors in the semantic space **R**, remove one of the combined feature vector $\mathbf{f}_k$ from the space. Store all the ranking values of the metadata in the learning data set *L*. If $\mathbf{d}_i$ is the element of the set *L*, $Rank_i$ is stored into the variable $Prev\_Rank_i$. Excuse Step-3 and Step-4 again. For each element $\mathbf{d}_i$ in the set *L*, if the new ranking position of $\mathbf{d}_i$ lower than its previous ranking position after the feature vector $\mathbf{f}_k$ is removed from the space **R**, that is, $Rank_i$ is greater than $Prev\_Rank_i$, it means that the feature vector $\mathbf{f}_k$ cannot be removed from the space. Otherwise, the feature vector $\mathbf{f}_k$ is removed from the space **R**. This step will be continued until all the combined feature vectors are tested whether they can be removed from the space **R** or not.

After the above removing processing, the optimal semantic subspace **R** is created.

Multi-spectrum pixels are projected onto the semantic space according to the projection method of MMM. On the selected subspace, the differential computing on two multi-spectrum pixels $\mathbf{p}_i$ and $\mathbf{p}_j$ are performed. During the following steps, a common vector and feature vectors of $\mathbf{p}_i$ and $\mathbf{p}_j$ are calculated.

**Step-1**: The differential vector Diff is created by calculating the difference of two multi-spectrum pixels $\mathbf{p}_i$ and $\mathbf{p}_j$ which are selected from the selected subspace:

$$\mathbf{Diff} = |\mathbf{p}_\mathrm{i} - \mathbf{p}_\mathrm{j}|;$$

Where, "||"represents the absolute value.

**Step-2**: The common vector Rc is created with the given threshold ε**:**

$$\mathbf{R}_\mathrm{c} = \mathbf{Diff} \quad <= \varepsilon,$$

$$\mathbf{R}_c = \{r_k \mid if\ k_{Diff} \le \varepsilon\ then\ r_k = 1\ else\ r_k = 0\ end\,if,\ where\ k_{Diff} \in \mathbf{Diff}\};$$

Where, $k_{\mathrm{Diff}}$ is an element of the vector **Diff**.

**Step-3**: The feature vectors $\mathbf{R}i$ and $\mathbf{R}j$ are created with the given threshold $\varepsilon_\mathrm{i}$ and $\varepsilon_\mathrm{j}$:

$$\mathbf{R}_i = \mathbf{E}_i \quad > \varepsilon_\mathrm{i}\ ,$$

$$\mathbf{R}_i = \{r_k \mid if\ k_{E_i} > \varepsilon_i\ then\ r_k = 1\ else\ r_k = 0\ end\,if,\ where\ k_{E_i} \in E_i\},$$

$$\mathrm{R}_j = \mathrm{E}_j \quad > \varepsilon_\mathrm{j}\ ,$$

$$\mathbf{R}_j = \{r_k \mid if\ k_{Ej} > \varepsilon_j\ then\ r_k = 1\ else\ r_k = 0\ end\,if,\ where\ k_{E_j} \in E_j\},$$

Where, $k_{\mathrm{E}i}$ is an element of the vector $\mathbf{E}i$ and $k_{\mathrm{E}j}$ is an element of the vector $\mathbf{E}j$.

**Figure 7** shows an example of spectrum by using "USGS Spectral Viewer" [10]. In the example, the bands of Landsat 8 OLI is selected. Four spectra, "Alunite", "Aspen Leaf 1", "H2O Ice", and "Desert Varnish" are selected. The figure shows that the optical characteristics of objects are different. A pixel with the same optical characteristics of "Alunite" will be represented as a 9 dimensional vector **p**:

$$\mathbf{p} = [0.65, 0.75, 0.80, 0.83, 0.88, 0.81, 0.55, 0.83, 0.87],$$

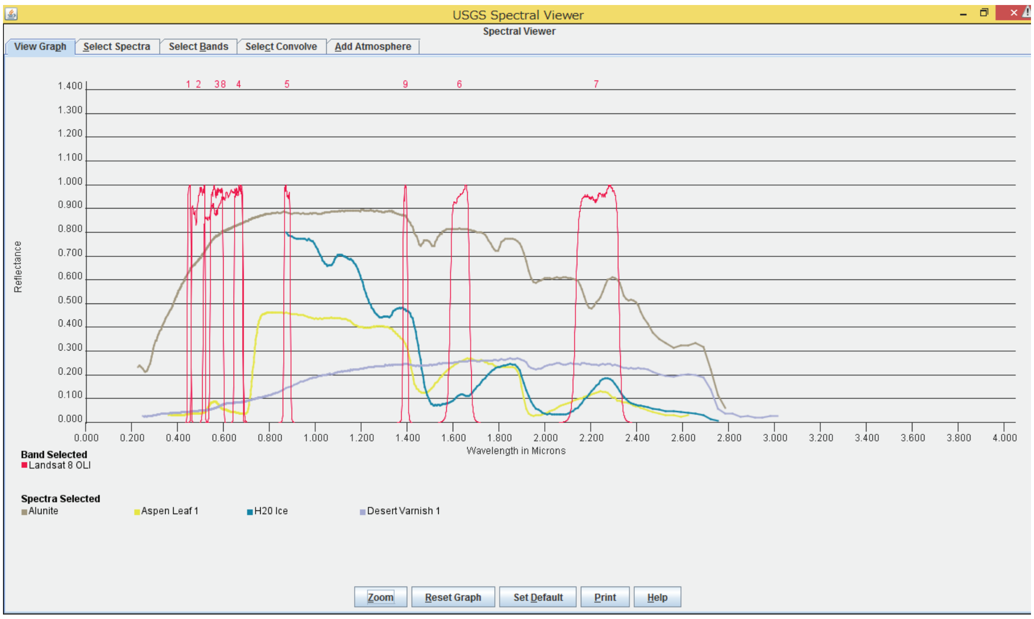



**Figure 7.** An example of spectrum by using "USGS Spectral Viewer" [10]

## 4. Application to Environmental Analysis

The important application of the semantic associative computing system is "*Global Environment-Analysis,*" which aims to evaluate various influences caused by natural disasters in global environments. We have several experiments for a global environment-analysis system based on MMM for natural disasters, especially for mud-flow disasters, as shown in Figure 8. Those results show the feasibility and effectiveness of our "*Differential Computing System*" with MMM for realizing deep analysis of global environments.

In the design of environmental analysis systems, one of the most important issues is how to search and analyze environmental data (temperature, color, CO2, places of livings, sea level, coral area, etc.), according to the changes of environmental data with time. We have created a multi-dimensional world map system, and we use this model to realize environmental data visualization and semantic associative processing for environmental changes, according to differential computing. Not only for multi-spectrum image analysis, we have applied this model to compute semantic correlations between temperature, color, CO2, places of livings, sea level, coral area, etc.

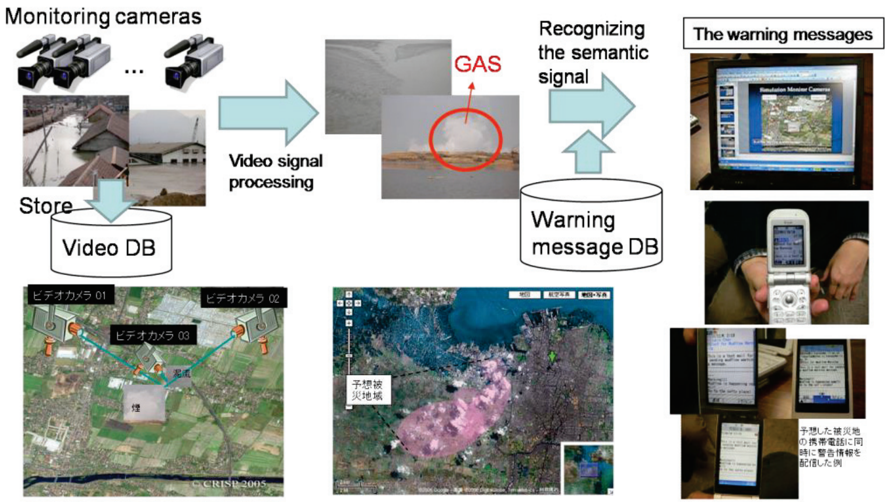



**Figure 8**. Environmental Analysis with the semantic associative computing system

We have constructed "*Environmental Change Computing System*" for analyzing environment changes with MMM functions applied to "*environmental observed data,*" as a new platform of nature environmental collaborative environments. This environment enables to create a remote, interactive and real-time nature environmental research among different countries.

We have designed this system to promote nature environmental change understanding and visualization. The system consists of differential analysis, search and visualization functions, characterized by three main features: (1) an environmental-dependent semantic metadata extraction method, which extracts both environmental elements (e.g., temperature, color, CO2, places of livings, sea level, coral area, etc.) and semantic subspaces corresponding to properties of learning data sets, (2) a differential computing mechanism to represent differences and similarities among various natural environments, and (3) easy-to-use interfaces designed for helping users to analyse the observed nature data and join the data database creation process. This system extracts features of natural events and expresses environmental changes by multi-dimension world map, and makes it possible to compare event difference and similarity with times.

The important objective of this differential computing system is to visualize important environmental events and changes by representing various difference-based events to multi-dimensional world map from observed environmental data. The main characteristics designed to visualize important events are (1) interpreting important events among different environment factors and (2) removing noise factors.

The system realizes event extraction, search, visualization, and differential computing functions which have been designed in a subspace-oriented way. Learning data sets based on environmental research works are utilized to create semantic space and also utilized for subspace selection. In this system, learning data set is important to deal with semantic heterogeneity when observed nature data are variously expressed among different environments. This system provides a subspace-dependent event extraction method to tackle this challenge.

## 5. Conclusion

As the progress of technologies, large amount of multi-spectrum images are taken and shared in the natural environment research field. Finding semantics from the multi-spectrum images is becoming an important research issues in the computer science research field. We have presented a new semantic computing method with multi-spectrum images for interpreting the meanings of environmental phenomena and changes occurring in the physical world. A new concept that connects semantics with the "*waving features*" of multi-spectrum images is proposed and applied in the method. This new concept will lead to new environmental-analysis tools developments. We have also presented a new "Semantic Computing System" with multi-spectrum images for realizing global environmental analysis. The main feature of this system is to realize semantic associative search in the multiple-colors-spectrum based dimensional orthogonal semantic space with semantic projection functions. This space is created for dynamically computing semantic equivalence or similarity between keywords and "*waving features*" of multi-spectrum images. We have applied this system to "global environmental system research and education" as a new platform of environmental computing. This system enables to create a remote, interactive and real-time cultural and academic research exchange from a global viewpoint. We have also presented the 5D World Map System, as an international and cross-cultural research environment with spatio-temporal and semantic analyzers. We have also presented several new approaches to global environmental-analysis with multimedia system architectures and "spatio-temporal & semantic computing."

As our future work, we will extend our "Semantic Computing System" as a multiple-colors-spectrum analyzer platform for realizing sensing data understanding on environmental issues in the environmental research field.

## References


[1] Kiyoki, Y., Kitagawa, T. and Hayama, T., "*A metadatabase system for semantic image search by a mathematical model of meaning*," ACM SIGMOD Record, Vol.23, No. 4, pp.34-41, Dec. 1994.

[2] Kiyoki, Y. and Kitagawa, T., "*A semantic associative search method for knowledge acquisition*," Information Modelling and Knowledge Bases (IOS Press), Vol. VI, pp.121-130, 1995.

[3] Chen, X. and Kiyoki, Y., "*A Visual and Semantic Image Retrieval Method Based on Similarity Computing with Query-Context Recognition*," Information Modelling and Knowledge Bases, Vol. XVIII, pp.245-252, May 2007.

[4] Sasaki, S., Takahashi, Y., Kiyoki,Y.: "*The 4D World Map System with Semantic and Spatio-temporal Analyzers*," Information Modelling and Knowledge Bases (IOS Press), Vol. XXI, pp.1-18, 2010

[5] "*Longman Dictionary of Contemporary English*", Prentice Hall College Div, 2006.

[6] Kiyoki, Y., S. Sasaki, N.N. Trang, Nguyen Thi Ngoc Diep, "Cross-cultural Multimedia Computing with Impression-based Semantic Spaces", Conceptual Modelling and Its Theoretical Foundations, Lecture Notes in Computer Science, Springer, pp.316-328, March 2012.

[7] Kiyoki, Y. and Ishihara, S., "*A Semantic Search Space Integration Method for Meta-level Knowledge Acquisition from Heterogeneous Databases*," Information Modelling and Knowledge Bases (IOS Press), Vol. 14, pp.86-103, May 2002.

[8] Kiyoki, Y. and Chen, X., "Contextual and Differential Computing for the Multi-dimensional World Map with Spatial-Temporal and Semantic Axes", Information Modelling and Knowledge Bases (IOS Press), Vol. XXV, 20 pages, March 2014.

[9] Harada, A. (eds.), "*Report of modeling the evaluation structure of KANSEI*", Univ. of Tsukuba, 1997.

[10] U.S. Geological Survey (USGS), " USGS Spectral Viewer", http://landsat.usgs.gov/tools_viewer.php, (accessed Jan. 17, 2015).

[11] Barbara Rose Johnston, Lisa Hiwasaki, Irene J. Klaver, Ameyali Ramos Castillo, Veronica Strang, "*Water, Cultural Diversity, and Global Environmental Change: Emerging Trends, Sustainable Futures?*" Feb., 2012 ( ISBN-10: 9400717733 | ISBN-13: 978-9400717732)

# Wide-Area River-Water Quality Analysis and Visualization with 5D World Map System

Chalisa VEESOMMAI[a,1], Yasushi KIYOKI[b], Shiori SASAKI[c] and Petchporn CHAWAKITCHAREON[d]
[a] *Graduate School of Media and Governance, Keio University 5332 Endo, Fujisawa, Kanagawa 252-0011 Japan*
[b] *Professor, Graduate School of Media and Governance, Keio University 5332 Endo, Fujisawa, Kanagawa 252-0011 Japan*
[c] *Assistant Professor, Graduate School of Media and Governance, Keio University 5332 Endo, Fujisawa, Kanagawa 252-0011 Japan*
[d] *Associate Professor, Department of Environmental Engineering, Faculty of Engineering, Chulalongkorn University 245 Phayathai Rd., Bangkok 10330, Thailand*
*Email: {chalisav, kiyoki, sashiori@sfc.keio.ac.jp}, petchporn.c@chula.ac.th*

**Abstract.** This paper presents the analysis and visualization of river-water quality in 25 rivers in Thailand by using 5D World Map system. Water pollution is analyzed by using Water Quality Index (WQI) and Metal Index (MI), which focus on Ping, Nan and Chao Phraya River (the important rivers of Thai). The WQI indicator was used to evaluate water quality by conductivity, $NO_3$-N, $NO_2$-N, $NH_3$-N, Cd, Cr, Mn, Ni, Pb, Zn and As. The MI indicator was used to estimate concentration of metal in the river. The results on 5D World Map System show that several actual values assigned to water-quality parameters are shown in snapshots. The results of Water Quality Index (WQI) show the WQI levels 32.697 at Chaophraya river (Bangkok, 2004) and 38.534 at Ping river (Nakhonsawan, 2014) for Irrigation and Aquatic life respectively, and can be classified into categories of quality-levels for Irrigation and Aquatic life. The results of Metal Index (MI) show that the MI level reaches 92.902 at Ping River (Nakhonsawan, 2014) and 1803.303 at Ping River (Nakhonsawan, 2014) for Irrigation and Aquatic life respectively.

**Keywords.** Water Quality, 5D World Map System, WQI and MI

## 1. Introduction

Nowadays, the environmental problems are more complex and affected globally. The causes of these problems relate to an unbalance use of resources and human activities. The rapid growth of population, economic expansion, industrialization, urbanization and energy over-consumption are main driving forces. These processes cause environmental pollutions in the phase of water quality. Moreover, some types of pollutants like heavy metal and chemical are recalcitrant and hardly decomposed in nature, and they are accumulated or contaminated into environments and human beings.

The water contamination with heavy metals, chemical and microbial has been the factor in determining human care. Actually, the water system containing different anions form and heavy metals (including cadmium, chromium, manganese, iron and lead) is useful and gives effects on human care. Some of heavy metals are essentially required for body growth, such as cobalt, copper and zinc, while the high concentration of other metals are toxic, such as cadmium, chromium, manganese, iron and lead.

The characteristics of physical, chemical and microbiological are used to determine water quality. In the river, water quality is characterized with wide various parameters. The physical characteristics are caused by some particles that we can be aware of 5 senses of human, such Conductivity, Taste-Odor, Temperature, Turbidity, Total Dissolved Solids (TDS) and Total Solid (TS). The chemical characteristics, caused by mineral or chemical compound that are dissolved in the water, rock or/and soils, are reflection in the water, such Alkalinity, Biological Oxygen Demand (BOD), Dissolve Oxygen (DO), Hardness, Nitrate ($NO_3$) and Potential of Hydrogen ion (pH). Moreover, the treated waste-water by agricultural, municipal, industrial and urban runoff leads to some impacts to water quality. The microbiological characteristics are mainly important in water because it causes disease in human, like "Water Born Disease", such as cholera disease and typhoid disease. The water quality filed usually analyzes biological characteristics in term of total coliform bacterial and fecal coliform bacterial. The important microbiological parameters for water quality considerations are Coliform bacterial (*Citrobacter, Enterobacter, Hafnia, Klebsiella* and *Serratia*) and Fecal Coliform bacterial (*Escherichia coli*).

The different purposes of water-use should be estimated in term of the specific water-quality parameters for the effect of water-use. The definitions of characteristics for physical, chemical and microbiological are shown as Table 1. This paper presents water quality characteristics in terms of physical, chemical and biological properties. Most map of using for water quality analysis is satellite images. The satellite images can be determined water quality in overview of water-body. A bit of difference, in this paper we study water quality in several points in river and we found the 5D World Map System is a powerful tool for study in several points in river all of the map. So we focus on water quality by using 5D World Map System for visualization and data analysis, and evaluate water quality by using Water Quality Index (WQI) and Metal Index (MI) in Ping, Nan and Chao Phraya River which important rivers of Thai.

### *1.1. The definition of Water Quality and Water Pollution and 5D World Map System*

Water Quality and Water Pollution in Thailand. The Kingdom of Thailand covers a land of area around 513,000 square kilometers. The country extends from north to south around 1,500 kilometers and from east to west around 800 kilometers. The geography region of Thailand can be divided to 5 main regions that consist of the north, the east, the west, the central and the south. Geographical characteristics in Thailand can be divided into 25 river basins, shown as Table 1. The north of Thailand is mainly mountainous where is the origin of 4 major rivers (Ping, Wang, Yom and Nan rivers) and converges to Chao Phraya River in the center region of Thailand. The Chao Phraya River is the main of water source for Thai. The development of city, domestic, industry and land use is the primary factor causing poor water quality in the river [1].

The causes of water pollution can be divided into 3 sources:

- The amount from Industry was approximately 2.8 M $m^3$/d from Factories (>120,000 factories).
- The amount from Agriculture was approximately 150.1 M $m^3$/d from the paddy field (150 M $m^3$/d) and the pig farm and aquaculture (0.1 M $m^3$/d), respectively.
- The amount from Domestic was approximately 14 M $m^3$/d from the municipalities (>1687 municipalities) and Pattaya city [2].

5D World Map System. The 5D world Map System has been proposed for sharing, mapping and visualization data [3] onto 5D time-series which temporal (1 Dimension) spatial (2-4 Dimensions) and semantic (5 Dimension) [4]

**Table 1.** The definitions of characteristic for physical, chemical and microbiological contamination.

| Parameter | Definition |
|---|---|
| Alkalinity | The alkalinity is an ability of water that can be provided neutral water, which can refer to 3 types of alkalinity in aqueous, is Hydroxide ($OH^-$), Carbonate ($CO_3^{2-}$) and Bicarbonate ($HCO^{3-}$). The concentration of this parameter is reported in mg/l as $CaCO_3$. |
| Biological Oxygen Demand (BOD) | The biological oxygen demand is a measure of total oxygen used by indigenous microbial population (aqueous life) in water. The concentration of this parameter is reported in mg/l. |
| Conductivity | The conductivity is related between total dissolved solids and electrical conductivity. The concentration of this parameter is reported in (μS). |
| Coliform bacterial | The kind of bacterial that live in intestines of warm-blooded animal. This parameter used to present the pathogenic organisms of human, and the concentration of this parameter is reported in MPN/100 ml or CFU/100 ml. |
| Fecal Coliform bacterial | The king of bacterial that growth and live with animal or/and human waste. This parameter used to present the pathogenic organisms of human, and the concentration of this parameter is reported in MPN/100 ml or CFU/100 ml |
| Dissolve Oxygen (DO) | The concentration of oxygen that require by microorganisms, fish and another aqueous life in aquatic system. The concentration of this parameter is reported in mg/l. |
| Hardness | The values of hardness are represented to total concentration of metal anion ($X^{-2}$ form). Hardness is the main parameter of potential precipitation of calcium carbonate to dregs in tubes, boilers, cooling towers and poor taste in water. The concentration of this parameter is reported in mg/l as $CaCO_3$. |
| Nitrate-nitrogen ($NO_3$-N)/ Nitrite-nitrogen ($NO_2$-N) | The nitrate anions are resulted of the bacteriological oxidation nitrogenous in soil. The nitrate anions are one of the indicators for the degree of the pollution in water with nitrate-content substances (the highly values of nitrate anion can be caused "Algae Bloom Crisis" and "Acid Precipitation"). The concentration of this parameter is reported in mg/l. |
| Potential of Hydrogen ion (pH) | The measurement of acidity and basicity in aqueous solution. From the theory, pH in water should be between 0-14 and pure water should be in pH = 7.0. |
| Total Dissolved Solids (TDS) | Total Dissolved Solids which refers to solid compound or article in the solid phase is dissolved in aqueous, such inorganic acid and organic compound. The concentration of this parameter is reported in mg/l. |
| Total Solid (TS) | Total Solid can refer to total solid compound or article in the solid phase in aqueous, after evaporation of the water and dry the solid compound or article in the solid phase at 103 $^0$C -105 $^0$C. The concentration of this parameter is reported in mg/l. |

**Table 1.** The definitions of characteristic for physical, chemical and microbiological contamination (Continue).

| Parameter | Definition |
|---|---|
| Suspended Solids (SS) | The Suspended Solids which refer to solid compound or article in the solid phase isn't dissolved in aqueous and suspended in aqueous. The concentration of this parameter is reported in mg/l. |
| Turbidity | The turbidity can by caused by infection of soil, sand, algae, plankton, diatom and colloidal, and is an efficiency indicator for water analysis in Environmental field, which is measured by the light-transmitting properties in the water. The concentration of this parameter is reported in NTU. |

## 2. Materials and Methods

In this section, we analyze water quality with Water Quality Index (WQI), Metal Index (MI) and 5D World Map System (5DWM),

- Analysis by using 5D World Map System (5DWM). 5D World Map System is useful as a data base system and evaluates water quality each parameter in term of Physical, Chemical and Biological characteristic 25 rivers of Thailand by visualization function.
- Analysis by using Water Quality Index (WQI) and Metal Index (MI). In this part we focus on water quality in each spot so we use Water Quality Index (WQI) and Metal Index (MI) for evaluation. The water-body target is Ping, Nan and Chao Phraya River because this water-body is long from North of Thailand to Gulf of Thailand. This waterbody are directly related with people because it passes community, industry and water supply plan.

### *2.1. Data source*

The water quality is monitored in data resources from the Pollution Control Department, Ministry of Natural Resource and Environment [5]. The water quality data are collected in 2004-2014. The collection of historical data and their sources are shown as Table 2. And the standard rang of water parameter compared to guidelines used in WQI and MI computations show as Table 3.

Pollution Control Department, Ministry of Natural Resource and Environment. The Department of Pollution Control was established on June 4, 1992 under the Royal Decree on the Organization Division of Pollution Control Department, Ministry of Science, Technology and Environment B.E. 2535 (1992), due to the enhancement and conversation of the National Environment Quality Act B.E. 2535 (1992).

The Department of Pollution Control is an organization for trusts and confidence in the management of pollution for a better environment and better quality of life. The Pollution Control Organization works on 2 missions:  to control, prevent, reduce and eliminate pollution, and to conserve and rehabilitate the environment conducive for human life.

**Table 2.** Collection of historical data and their sources [5], [6]

| Type of Data | Review Document | Water Quality Data | GIS Data: Latitude, Longitude |
|---|---|---|---|
| Sources of Data | Review on Journals and textbook | Pollution Control Department, Thailand | http://www.pcd.go.th http://maps.google.com |

**Table 3.** The standard rang of water parameter compared to guidelines used in WQI and MI computations.

| Parameter | Standard | |
|---|---|---|
| | Irrigation [7] | Aquatic live [8] |
| Temperature ($^0$C) | | 8-28 |
| pH range | 8.5 | 6.5-9 |
| Turbidity (NTU) | | |
| Conductivity (*m*S) | 3000 | |
| Dissolved Oxygen, DO (mg/l) | | 5.5 |
| Biological Oxygen Demand, BOD (mg/l) | | |
| Total Phosphate (mg/l) | 2 | |
| Nitrate nitrogen, $NO_3$-N (mg/l) | 10 | 2.93 |
| Nitrite-nitrogen, $NO_2$-N (mg/l) | | 0.06 |
| Ammonia-nitrogen, $NH_3$-N (mg/l) | | 1.37 |
| Suspended Solid (mg/l) | | 25 |
| Total Dissolved Solid (mg/l) | 2000 | 500 |
| Iron, Fe (mg/L) | 5 | 0.3 |
| Cadmium, Cd (mg/L) | 0.01 | 0.001 |
| Chromium, Cr (mg/L) | | |
| Manganese, Mn (mg/L) | 0.2 | 0.05 |
| Nickel, Ni (mg/L) | 0.2 | 0.025 |
| Lead, Pb (mg/L) | 5 | 0.007 |
| Zinc, Zn (mg/L) | 2 | 0.05 |
| Copper, Cu (mg/L) | 0.2 | 0.994 |

*2.2. Description of study area*

The study areas of water quality are in the rivers of Thailand as Ping, Wang, Yom, Nan, Kuang, Kok, Kwan Phrayao, Mae Jang, Ing, Chee, Mun, Lamtakhon, Loei, Rayong, Welu, Chao Phraya, Kheaw Yai, Kheaw Noi, Kui Bure, Petchabun, Thajen, Pha Sak, Chumporn, Pattane and Trang. The description of the sampling points for Pollution Control Department is shown as table 4.

**Table 4.** Description of the sampling points for Rives [5], [6]

| River | Watershed area (sq.km) | Latitude | Longitude |
|---|---|---|---|
| **North** | | | |
| Ping | 12,426.20 | 5.720039-19.453152 | 100.144566-98.99523 |
| Wang | 6,860.50 | 17.136486- 18.695487 | 99.104920-99.569321 |
| Yom | 15,283.65 | 15.912479-16.799990 | 100.25069-100.28512 |
| Nan | 17,809.81 | 15.652099-19.125264 | 100.14493-100.81062 |
| Kuang | 2,876.70 | 18.541809-18.964389 | 98.940019-99.237288 |
| Kok | 2,055.09 | 19.919988-20.227243 | 99.846032-100.12879 |
| Kwan Phrayao | ND | 19.157037-19.166234 | 99.917641-99.897902 |
| Mae Jang | 1,196.62 | 18.114083-18.163942 | 99.413816-99.659983 |
| Ing | 3,105.72 | 19.832829-20.136734 | 100.20013-100.42028 |
| **East** | | | |
| Chee | 17,433.03 | 15.180476-16.250521 | 98.719076-97.072446 |
| Mun | 69,700.44 | 14.561359-15.432204 | 96.175253-97.015786 |
| Lamtakhon | 3,310.64 | 14.636499- 15.019828 | 101.42214-101.72879 |
| Loei | ND | 17.492023-17.858344 | 101.73752-101.61446 |
| **West** | | | |
| Rayong | ND | 12.656131-12.847320 | 101.28099-101.30407 |
| Welu | ND | 12.335528-12.458591 | 102.26598-102.31289 |
| **Center** | | | |
| Chao Phraya | | | |
| Kheaw Yai | 15,875.14 | 5.510661-6.008650 | 110.07648-108.032592 |
| Kheaw Noi | 9091.68 | 14.022690-14.399515 | 99.526127-99.138532 |
| Kui Bure | 7,498.92 | 14.014146-14.891130 | 99.525857-98.520744 |
| Petchabun | ND | 12.039423-12.059113 | 99.909704-99.859663 |
| Thajen | 5,132.15 | 12.813409-13.223570 | 99.794241-99.990462 |
| Pha Sak | 11,561.78 | 13.510005-15.209973 | 100.27485-100.07426 |
| | 12,432.86 | 14.349358-16.648211 | 100.58469-101.21680 |
| **South** | | | |
| Chumporn | ND | 10.439674-10.576708 | 100.14109-99.250837 |
| Pattane | 3,654.87 | 6.134718-6.895482 | 101.27478-101.25358 |
| Thrang | 3,449.27 | 7.313309-7.9420199 | 99.507822-99.581391 |

## 2.3. Data Structure

We collected 700 files in CSV form (28 parameter from 25 rivers) and added semantic and spatiotemporal metadata, such Category, Location, Date, Description for each data in which the data structures, shown as Figure 1, are based on 5D World Map System.

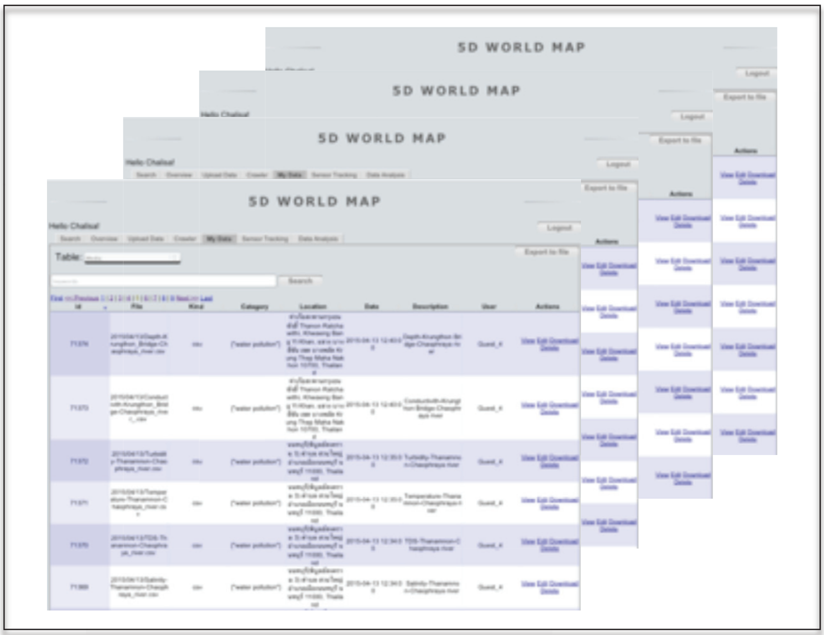



**Figure 1.** The data structures on 5D World Map System

## 2.4. Water Quality analysis

1. Water Quality Index (WQI) is a mathematical instrument used to transform large qualities of water quality data into single number for using determines water quality [9] [10] [11]. The formula for calculation is Eq. 2

$$WQI = \sum_{i=1}^{n} Q_i W_i \Big/ \sum_{i=1}^{n} W_i \qquad (1)$$

when $Q_i$ is a sub quality index of I parameter
$W_i$ is a weight unit of each parameter, and
n is a number of parameter

Calculation of
$Q_i$ = $[(V_i\text{-}V_0)/(S_i\text{-}V_0)]$ ; $V_i$ is the value of measure of i parameter, $S_i$ is standards of i-th parameter, and $V_0$ is an ideal value of i-th parameter in distilled waters and $V_0$ = 0 except pH = 7.0 and DO = 14.6 mg/l [12].

$W_i$ = K/Si, K is a constant of weights for various water quality

$$K = \frac{1}{1/S_i}$$

The WQI scores are classified into 5 classes of the water quality in excellent, good, poor, very poor and unfit. [12], The Water Quality rating as per weight arithmetic water quality index method show as Table 5.

**Table 5.** The Water Quality rating as per weight arithmetic water quality index method. [13]

| Grading | WQI Value | Rating of Water Quality |
|---|---|---|
| A | 0-25 | Excellent |
| B | 26-50 | Good |
| C | 51-75 | Poor |
| D | 76-100 | Very Poor |
| E | Above 100 | Unfit |

2. Metal Index (MI) is an indicator for determining the level of metal concentration in the river. The threshold of warming is MI ≥ 1. The formula for calculation is Eq. 1

$$MI = \sum_{i=1}^{n} \frac{C_i}{MAC_i} \tag{2}$$

when $C_i$ is concentration of each metal,
$MAC_i$ is maximum allowable concentration of each metal according the standard of water quality of waterbody in Thailand. 1994 [14], FAO (1994) [7] and CCME (2007) [8].

## 3. Results and discussion

### *3.1. 5D World Map System*

The Physical, Chemical and Biological parameters of water quality between 2004-2014 is analyzed and visualized on 5DWM, and those results are shown in Figures 2. We compare the values in each station and mount by using the different levels of colors. The parameters are turbidity, pH, conductivity, DO, BOD, TS, TDS, SS, total coliform bacterial, fecal coliform bacterial and heavy metal etc. The concentration of water pollutions is at down-side more than up-side in each river because the pollution is accumulated in rivers and utilization of water is around rivers in agriculture, domestic and industry [2]. The water pollution source is related to the population rate (0.91 in 2004, 0.62 in 2009 and 0.54 in 2012) and the types of agriculture and industry.

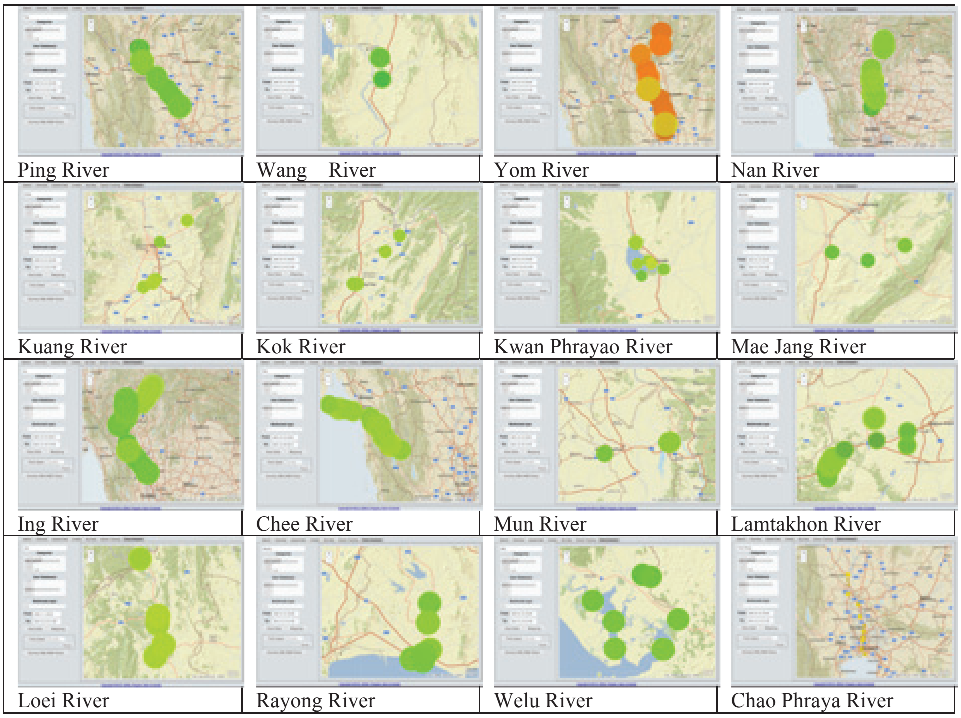
### *3.2. Water Quality Index (WQI)*

This paper focuses on 2004, 2006, 2008, 2010, 2012 and 2014 in Ping, Nan and Chao Phraya rivers which important rivers of Thai by using WQI indicator. WQI for Irrigation and Aquatic life was computed, using guidelines of the standard of water quality of water-body in Thailand [14], FAO (1994) [7] and CCME (1999) [8]. The table 6. Shown as WQI of Ping, Nan and Chao Phraya River range between 0.315-32.697 and 0.038-38.534 for Irrigation and Aquatic life respectively. Ping, Nan and Chao Phraya rivers could be classified into the “excellence” to “good” for the Irrigation and aquatic life.

*3.3. Metal Index (MI)*

This paper focuses on 2004, 2006, 2008, 2010, 2012 and 2014 in Ping, Nan and Chao Phraya rivers which important rivers of Thai by using MI indicator. MI for Irrigation and Aquatic life was computed, using guidelines of the standard of water quality of water-body in Thailand [14], FAO (1994) [7] and CCME (1999) [8]. In the Metal Index (MI) for Irrigation and Aquatic life, the MI reaches 92.902 and 1803.032 for Irrigation and Aquatic life at Ping river in 2014 because of the high of concentration of Cadmium, Manganese and Zinc. From the result, we found that the value of MI $\geq$ 1 shows the threshold of warning by using PostgreSQL. The threshold of warning of Metal Index (MI) for Irrigation and Aquatic life is shown in Figure 3 (a) and (b).

In addition, the results of WQI and MI are shown that the water quality is at down-side more than up-side of river, because the pollutions are accumulated in water-body that supports the results on 5D World map System



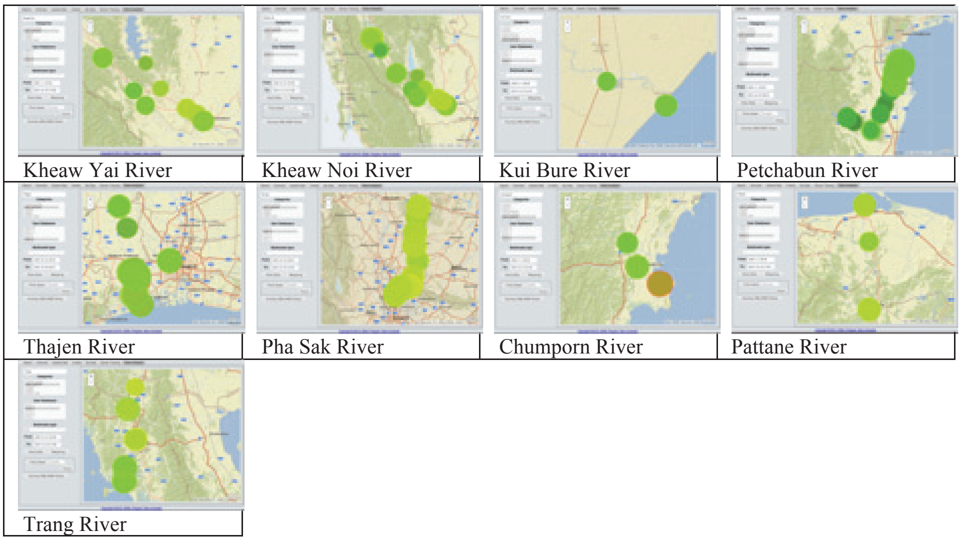
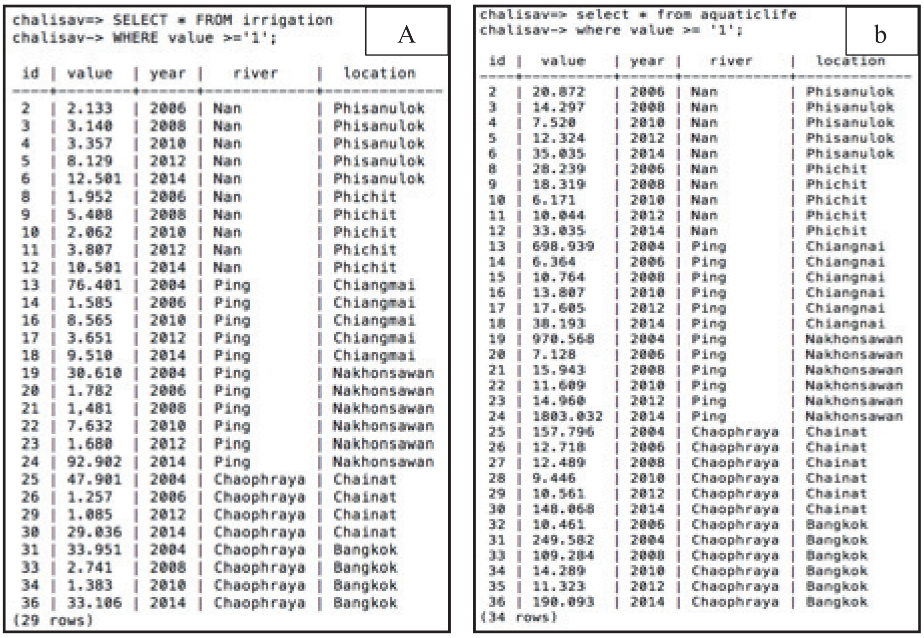



**Table 6.** Water Quality Index (WQI) for Irrigation and Aquatic life.

| ID | WQI | | Year | River | Location | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| | Irrigation | Aquatic live | | | | | |
| 1 | 0.623 | 0.038 | 2004 | Nan | Phitsanulok | 17.0481337108011 | 100.1815660869856 |
| 2 | 0.664 | 0.849 | 2006 | Nan | Phitsanulok | 17.0481337108012 | 100.181566086987 |
| 3 | 0.315 | 4.559 | 2008 | Nan | Phitsanulok | 17.0481337108013 | 100.181566086988 |
| 4 | 0.372 | 4.928 | 2010 | Nan | Phitsanulok | 17.0481337108014 | 100.181566086989 |
| 5 | 0.948 | 6.768 | 2012 | Nan | Phitsanulok | 17.0481337108015 | 100.181566086990 |
| 6 | 0.358 | 4.764 | 2014 | Nan | Phitsanulok | 17.0481337108016 | 100.181566086991 |
| 7 | 1.997 | 0.044 | 2004 | Nan | Phichit | 16.3597027941911 | 100.1119741613277 |
| 8 | 2.141 | 1.118 | 2006 | Nan | Phichit | 16.3597027941912 | 100.111974161329 |
| 9 | 2.998 | 6.466 | 2008 | Nan | Phichit | 16.3597027941913 | 100.111974161330 |
| 10 | 0.674 | 3.833 | 2010 | Nan | Phichit | 16.3597027941914 | 100.111974161331 |
| 11 | 0.736 | 2.932 | 2012 | Nan | Phichit | 16.3597027941915 | 100.111974161332 |
| 12 | 1.533 | 2.968 | 2014 | Nan | Phichit | 16.3597027941916 | 100.111974161333 |
| 13 | 1.996 | 10.214 | 2004 | Ping | Chiang Mai | 19.1497126669801 | 98.98877851325373 |
| 14 | 1.792 | 0.343 | 2006 | Ping | Chiang Mai | 19.1497126669802 | 98.9887785132538 |
| 15 | 14.927 | 0.914 | 2008 | Ping | Chiang Mai | 19.1497126669803 | 98.9887785132539 |
| 16 | 1.239 | 0.737 | 2010 | Ping | Chiang Mai | 19.1497126669804 | 98.9887785132540 |
| 17 | 9.624 | 0.947 | 2012 | Ping | Chiang Mai | 19.1497126669805 | 98.9887785132541 |
| 18 | 0.678 | 0.539 | 2014 | Ping | Chiang Mai | 19.1497126669806 | 98.9887785132542 |
| 19 | 8.986 | 17.738 | 2004 | Ping | Nakhon Sawan | 15.7200393987260 | 100.1445661702587 |
| 20 | 9.649 | 3.122 | 2006 | Ping | Nakhon Sawan | 15.720039398727 | 100.144566170260 |
| 21 | 0.577 | 2.534 | 2008 | Ping | Nakhon Sawan | 15.720039398728 | 100.144566170261 |

**Table 6.** Water Quality Index (WQI) for Irrigation and Aquatic life. (Continue)

| ID | WQI | | Year | River | Location | Latitude | Longitude |
|---|---|---|---|---|---|---|---|
| | Irrigation | Aquatic live | | | | | |
| 22 | 0.457 | 8.334 | 2010 | Ping | Nakhon Sawan | 15.720039398729 | 100.144566170262 |
| 23 | 0.715 | 4.101 | 2012 | Ping | Nakhon Sawan | 15.720039398730 | 100.144566170263 |
| 24 | 0.603 | 38.534 | 2014 | Ping | Nakhon Sawan | 15.720039398731 | 100.144566170264 |
| 25 | 0.726 | 9.506 | 2004 | Chao Phraya | Chai nat | 15.1559182073315 | 100.1839720070586 |
| 26 | 0.389 | 0.552 | 2006 | Chao Phraya | Chai nat | 15.1559182073316 | 100.183972007060 |
| 27 | 1.525 | 0.466 | 2008 | Chao Phraya | Chai nat | 15.1559182073317 | 100.183972007061 |
| 28 | 0.418 | 0.288 | 2010 | Chao Phraya | Chai nat | 15.1559182073318 | 100.183972007062 |
| 29 | 3.114 | 0.317 | 2012 | Chao Phraya | Chai nat | 15.1559182073319 | 100.183972007063 |
| 30 | 0.274 | 3.991 | 2014 | Chao Phraya | Chai nat | 15.1559182073320 | 100.183972007064 |
| 31 | 32.697 | 7.748 | 2004 | Chao Phraya | Bangkok | 13.7028182431342 | 100.5699896563129 |
| 32 | 21.414 | 0.478 | 2006 | Chao Phraya | Bangkok | 13.7028182431344 | 100.569989656314 |
| 33 | 24.020 | 0.554 | 2008 | Chao Phraya | Bangkok | 13.7028182431345 | 100.569989656315 |
| 34 | 1.418 | 0.659 | 2010 | Chao Phraya | Bangkok | 13.7028182431346 | 100.569989656316 |
| 35 | 8.723 | 0.567 | 2012 | Chao Phraya | Bangkok | 13.7028182431347 | 100.569989656317 |
| 36 | 11.965 | 5.369 | 2014 | Chao Phraya | Bangkok | 13.7028182431348 | 100.569989656318 |



A

```
chalisav=> SELECT * FROM irrigation
chalisav-> WHERE value >='1';

 id | value  | year |   river    |  location
----+--------+------+------------+-------------
 2  | 2.133  | 2006 | Nan        | Phisanulok
 3  | 3.140  | 2008 | Nan        | Phisanulok
 4  | 3.357  | 2010 | Nan        | Phisanulok
 5  | 8.129  | 2012 | Nan        | Phisanulok
 6  | 12.501 | 2014 | Nan        | Phisanulok
 8  | 1.952  | 2006 | Nan        | Phichit
 9  | 5.408  | 2008 | Nan        | Phichit
 10 | 2.062  | 2010 | Nan        | Phichit
 11 | 3.807  | 2012 | Nan        | Phichit
 12 | 10.501 | 2014 | Nan        | Phichit
 13 | 76.401 | 2004 | Ping       | Chiangmai
 14 | 1.585  | 2006 | Ping       | Chiangmai
 16 | 8.565  | 2010 | Ping       | Chiangmai
 17 | 3.651  | 2012 | Ping       | Chiangmai
 18 | 9.510  | 2014 | Ping       | Chiangmai
 19 | 30.610 | 2004 | Ping       | Nakhonsawan
 20 | 1.782  | 2006 | Ping       | Nakhonsawan
 21 | 1,481  | 2008 | Ping       | Nakhonsawan
 22 | 7.632  | 2010 | Ping       | Nakhonsawan
 23 | 1.680  | 2012 | Ping       | Nakhonsawan
 24 | 92.902 | 2014 | Ping       | Nakhonsawan
 25 | 47.901 | 2004 | Chaophraya | Chainat
 26 | 1.257  | 2006 | Chaophraya | Chainat
 29 | 1.085  | 2012 | Chaophraya | Chainat
 30 | 29.036 | 2014 | Chaophraya | Chainat
 31 | 33.951 | 2004 | Chaophraya | Bangkok
 33 | 2.741  | 2008 | Chaophraya | Bangkok
 34 | 1.383  | 2010 | Chaophraya | Bangkok
 36 | 33.106 | 2014 | Chaophraya | Bangkok
(29 rows)
```

b

```
chalisav=> select * from aquaticlife
chalisav-> where value >= '1';

 id |  value   | year |   river    |  location
----+----------+------+------------+-------------
 2  | 20.872   | 2006 | Nan        | Phisanulok
 3  | 14.297   | 2008 | Nan        | Phisanulok
 4  | 7.520    | 2010 | Nan        | Phisanulok
 5  | 12.324   | 2012 | Nan        | Phisanulok
 6  | 35.035   | 2014 | Nan        | Phisanulok
 8  | 28.239   | 2006 | Nan        | Phichit
 9  | 18.319   | 2008 | Nan        | Phichit
 10 | 6.171    | 2010 | Nan        | Phichit
 11 | 10.044   | 2012 | Nan        | Phichit
 12 | 33.035   | 2014 | Nan        | Phichit
 13 | 698.939  | 2004 | Ping       | Chiangnai
 14 | 6.364    | 2006 | Ping       | Chiangnai
 15 | 10.764   | 2008 | Ping       | Chiangnai
 16 | 13.807   | 2010 | Ping       | Chiangnai
 17 | 17.605   | 2012 | Ping       | Chiangnai
 18 | 38.193   | 2014 | Ping       | Chiangnai
 19 | 970.568  | 2004 | Ping       | Nakhonsawan
 20 | 7.128    | 2006 | Ping       | Nakhonsawan
 21 | 15.943   | 2008 | Ping       | Nakhonsawan
 22 | 11.609   | 2010 | Ping       | Nakhonsawan
 23 | 14.960   | 2012 | Ping       | Nakhonsawan
 24 | 1803.032 | 2014 | Ping       | Nakhonsawan
 25 | 157.796  | 2004 | Chaophraya | Chainat
 26 | 12.718   | 2006 | Chaophraya | Chainat
 27 | 12.489   | 2008 | Chaophraya | Chainat
 28 | 9.446    | 2010 | Chaophraya | Chainat
 29 | 10.561   | 2012 | Chaophraya | Chainat
 30 | 148.068  | 2014 | Chaophraya | Chainat
 32 | 10.461   | 2006 | Chaophraya | Bangkok
 31 | 249.582  | 2004 | Chaophraya | Bangkok
 33 | 109.284  | 2008 | Chaophraya | Bangkok
 34 | 14.289   | 2010 | Chaophraya | Bangkok
 35 | 11.323   | 2012 | Chaophraya | Bangkok
 36 | 190.093  | 2014 | Chaophraya | Bangkok
(34 rows)
```

**Figure 3**. The threshold of warning of Metal Index (MI) for (a) Irrigation and (b) Aquatic life by PostgreSQL.

## 4. Conclusion

The Chao Phraya River is the main of water sources in Thai and it is converged by Ping and Nan rivers. Regional and industrial activities are the primary factors causing poor water quality in the river, and the different purposes of water use has been evaluated in terms of the specific water-quality parameters for the effect of water-use.

Water quality pollution in rivers of Thailand can be analyzed by using 5D World Map System in several parameters and points in Thailand. 5D World Map System analyzes and visualizes useful data of Temperature, pH, Turbidity, Conductivity, Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), Total Coliform Bacterial, Fecal Coliform Bacterial, Nitrate-nitrogen, Nitrite-nitrogen, Ammonia-nitrogen, Suspended Solid, Total Dissolved Solid, Iron, Cadmium, Chromium, Manganese, Nickel, Lead, Zinc, Copper, WQI and MI for the bigger study area in the future.

The WQI results have shown that the water quality in the water body is good and excellent for Irrigation and Aquatic life.

The MI results have shown that the water body is obviously polluted by various metals as cadmium, chromium, manganese, nickel and zinc. This paper has presented the water quality evaluation with thresholds for warning of Metal Index (MI) of Irrigation and Aquatic life, and has evaluated the effects of water-use for estimating the appropriate and effective protection of the water body in the future.

## 5. Acknowledgement

This work is supported by Pollution Control Department, Ministry of Natural Resource and Environment, Thailand and Multimedia Database Laboratory, Graduate School of Media and Governance, Keio University. We thank the anonymous reviewers for their valuable comments and suggestions.

## 6. References

[1] J. Kibena., I Nhapi., W. Qumindoga. 2014. Assessing the relationship between water quality parameter And changes in landuse patterns in the Upper Manyame River, Zimbabwe. Physics and Chemistry of the Earth (67-69), 153-163.

[2] K. Intarukdach., S. Songprakhop. Wastewater Management Division, Pollution Control Department. 2010. Water Quality Management Measures for Urban Areas.

[3] Sasaki, S., Takahashi, Y., Kiyoki, Y.: "*The 4D World Map System with Semantic and Spatiotemporal Analyzers*," Information Modelling and Knowledge Bases (IOS Press), Vol. XXI, pp. 1-18, 2010.

[4] Kiyoki, Y. and Chen, X., "Contextual and Differential Computing for the Multi-dimensional World Map with Spatial-Temporal and Semantic Axes", Information Modelling and Knowledge Bases (IOS Press), Vol. XXV, 20 pages, March 2014.

[5] Annual reports 2004-2014 of Pollution Control Department, Ministry of Natural Resource and Environment. Thailand.

[6] Bureau of Research, Development and Hydrology, Department of Water Resources, Thailand. 2012. Map 25 Basin of Thailand Branch L7018 1:50,000 WGS84 UTM Zone 47N. Nonthabure . Sahamit Priting and Publicing.

[7] R.S. Ayers., D.W. Westco. 1994. Food, Agriculture Organization of the United Nations (FAO), Water Quality for Agriculture, Irrigation and Drainage.

[8] Canadian Council of Ministers of the Environment. 2007. Canadian environmental quality guidelines.

[9] I.N. Balan., M. Shivakumar., P.D.M. Kumar. 2012. An assessment of ground water quality using water quality index in Chennai Tamil Nadu, India. Chronicles Young Science. 3(2), 146-150.

[10] N. Bharti., D. Katyal. 2011. Water quality indices used for surface water vulnerability assessment. Int. J. Environ. Sci. 2(1), 154-173.

[11] Fabiano dos Santos Simoes., Altair B. Moreira., Marcia Cristina Bisinoti., Sonia M. Nobre Gimenez, Maria Josefa Santos Yabe. 2008. Water quality index as a simple indicator of aquaculture effect on aquatic bodies. Ecological indicators (8), 476-484.
[12] S. Nazeer., M. Zaffar Hashmi., R. Naseem Malik. 2014. Heavy metal distribution, risk assess-ment and Water quality characterization by water quality index of the River Soan, Pakistan. Ecological indicators (43)262-270.
[13] G. E. Mohamed., M. H. Ali., A. A. Ibrahim., H. F. Ayman., M.E. Seliem. 2014. Evaluation of surface water quality and heavy netal indices of Ismailia Canal, Nile River, Egypt. Egyptian Journal of Aquatic Research 40(3), 225-233.
[14] The standard of water quality of waterbody in Thailand. 1994. Ministry of Natural Resource and Environment. Thailand.

# Formal Knowledge Framework for Software Processes Architecture

Michael Alexander KOŠINÁR[a,1], Jan CZOPIK[a], and Jakub ŠTOLFA[a]
[a] *Department of Computer Science*
*VŠB – Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science*
*708 33 Ostrava – Poruba, Czech Republic*

**Abstract.** Last decades have introduced different improvements into software process modeling yet none has proven itself as a silver bullet; software development community has proposed various solutions from rigid prescriptive processes to agile methods, in the end, however, every good software process implementation require process modeling that can be used for different purposes like process auditing, analysis, and evaluation. This paper discusses application of explicit knowledge profiles based on process meta-model within software process modeling, alignment with visual process modeling, and further analysis with simulation and reverse engineering methods.

**Keywords.** Knowledge Modeling, OWL, Petri Nets, Process Mining, Reverse Engineering, Software Process, UML

## 1. Introduction

This paper is a part of ongoing research concerned in formal knowledge-based process modeling techniques. In this case the application of knowledge layer is in the software processes modeling. We develop a new process architecture method that meets following features that will be discussed in the paper:

- Modeling based on single process meta-model terminology defined in a formal way;
- Utilization of explicit reusable knowledge profiles;
- Intuitive visual modeling method based on diagrammatic language capable of software process modeling based on process patterns;
- Process simulation and analysis with extended process information derived from knowledge profile.

The paper is organized as follows: after the brief introduction we shall discuss existing work in this area, this section is then followed by details of techniques that have been selected and used for the software process modeling. Fourth and fifth sections describe the framework itself including the simulation and analysis of software

[1] Corresponding Author – michal.kosinar@vsb.cz

processes. The paper is concluded with summary that discusses benefits and disadvantages of the approach.

## 2. Existing Work

Software processes represent the core of any software house behavior [9, 16]. The modeling of processes in organization is motivated mostly by process improvement and re-engineering [15, 9] however models could (and apparently should) be used for simulation, process analysis, audit purposes, and automatized process execution [1, 13].

Advanced utilization of software processes, nonetheless, requires formal methods to be introduced not only to the modeling method but into the whole modeling framework as well [1, 6, 11, 12, 14] – a framework that comprise of modeling language, modeling methodology, and mechanics.

Some of existing projects and solutions utilize Petri Nets that are widely accepted in process modeling area for modeling of dynamic aspects of processes [1, 6, 7, 12]. Other projects that aim to improved process adaptability utilize system based on logics like OWL [4, 11, 13, 17] for modeling of process static aspects.

Our research, however, focus on a different way of software process modeling enhancement – we combine the modeling techniques with process meta-model [7, 12] to improve the intuitiveness of the modeling, utilize reusable common vocabulary, terminology alignment, process analysis, process simulation, exceptional states handling and reverse engineering.

The idea of one complex knowledge framework with specially adjusted supportive methods – like simulation, analysis, and reverse engineering – is what distinguishes our work from the others. Our work in the area of reverse engineering is based on process mining methods. Process mining aims for discovering, monitoring, and improving real process executions with extracting knowledge from event logs [3] – timestamps generated from process execution details. Process mining can be seen as a supportive method for the Business Process and Business process intelligence analysis and from the perspective of Business Process Management (BPM) can be used as a feedback to the BPM methods [2]. Process mining aims for being fill a gap between BPM and Workflow Management on the one hand and Data Mining, Business Intelligence, and Machine Learning on the other hand [3]. Process mining answers question how process execution was really executed, which variations were used and if there is any deviation or possibility to optimize next process executions or process models.

## 3. Software Processes

Our framework is able to model all aspects of software processes that are critical for process execution and further analysis – these aspects are usually referred to as aspects of *functional perspective*, *behavioral perspective*, and *structural perspective* – each concerns different scope (or viewpoint) of software houses. These perspectives can represent software processes but must be captured prior to process representation that is the primary goal of the process modeling. There are two basic categories of modeling techniques with the ability to capture the process (actually three but *informal* description has no value for our framework in the discussed context) [9]: *semi-formal* (unambiguous syntax, ambiguous semantics) and *formal* (unambiguous syntax and semantics).

Unfortunately most of process modeling techniques lack unified approach to specific domain terminology and our approach is based upon the so-called Unified Process Meta-Model (UPMM) that is, first of all, customizable and is able to map and match most software development domain terminology to meta-model types – before we shall progress with process modeling we shall introduce the meta-model.

### *3.1. Process Meta-Model*

There are various process meta-models available [12] however each had been designed to support different process aspects like context, goal (or product), activity, strategy, or automatization – our meta-model, on the contrary, unifies all base terminology necessary for complex process capture and representation. The Unified Process Meta-Model is depicted in figure 1 – its compatibility with other meta-models opens possibilities like pattern-based visual process modeling and explicit knowledge profiles creation.

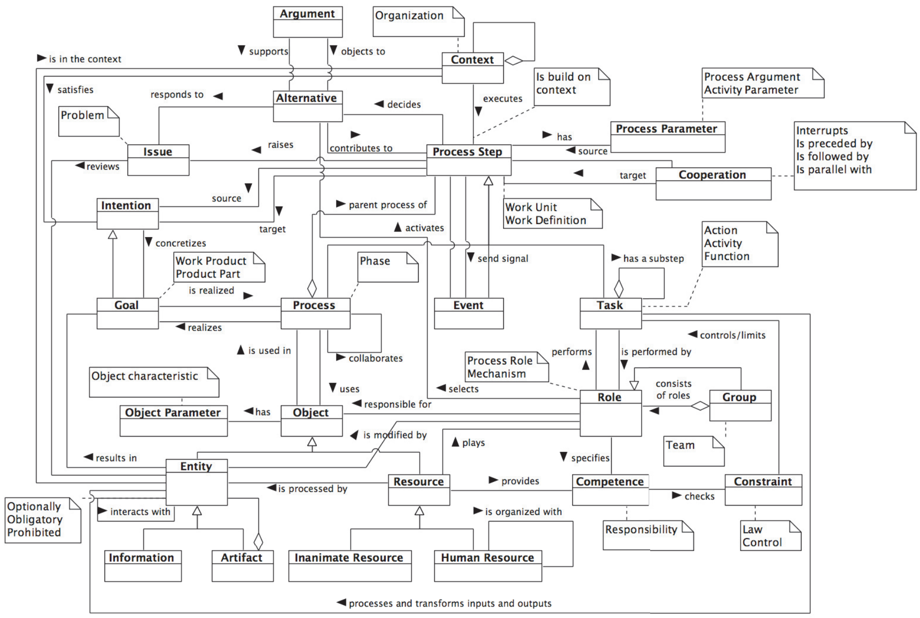



**Figure 1.** Unified Process Meta-Model

### *3.2. Explicit Knowledge Profile of Software Process*

Procedure of 'process details capture' naturally causes loss of information due to conversion of a the real-world state to generalized simplified model, typically visual and semi-formal – we call this side effect the Semantic Gap. Formal modeling methods are the best way for capturing the process models as they offer precise mathematical foundations that allow process specialists to create unambiguous models compatible with workflow systems and minimalize the Semantic Gap [1, 12].

Our framework and methodology combine two process formal representation techniques: *functional* and *descriptive*; both are essential – functional models dynamic aspects of the process and descriptive aims at the static aspects like types, relationships, properties, and constraints. Techniques we use are:

- Colored Petri Nets (CPN) – capturing the software house dynamics like activities, events, conditions, and concurrency.
- Web Ontology Language (OWL) – for meta-model creation, modeling of domain terminology, attributes, and visual modeling language mapping and alignment.

OWL has some nice characteristics suitable for process analysts and theorists [4, 17] – the rules and facts are separated in disjoint boxes known as TBox (classes/concepts), ABox (assertions/individuals), and RBox (object properties/roles/relationships); with this setup can be common vocabulary easily shared and reused, and its RDF/XML based schema is well prepared for further utilization.

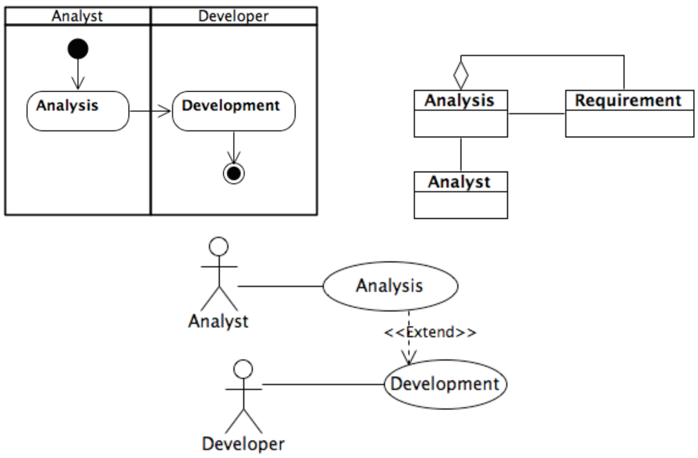



**Figure 2.** Visual process models examples using UML (upper left – activity diagram, upper right class diagram, bottom – use-case diagram)

### *3.3. Visual Process Modeling*

Our framework utilizes Unified Modeling Language [5, 9, 12] diagrams that are suitable for process representation – Activity Diagrams for behavioral modeling, Class Diagrams for structural modeling, and Use Case Diagrams for modeling functional aspects. Even though there are other modeling languages we decided to use UML because of its great characteristics that include ability to capture all necessary process aspects (in comparison with BPMN or EPC) and its great acceptance in software development community with great modeling tools support (for example IDEF is able to capture many process aspects but is not commonly used in the community). Example process diagrams are present in figure 2. So far so good – we have the terminology base – the process meta-model – and visual modeling language that is able to provide readable, intuitive and visual process representations. Only link we miss now is something that could connect those two elements – modeling mechanics [9, 12] and a domain knowledge that connect meta-model alignment and process creation phases.

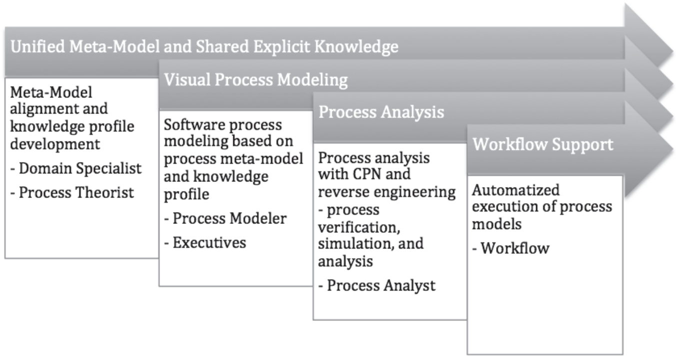



**Figure 3.** Modeling approach using formal knowledge process framework

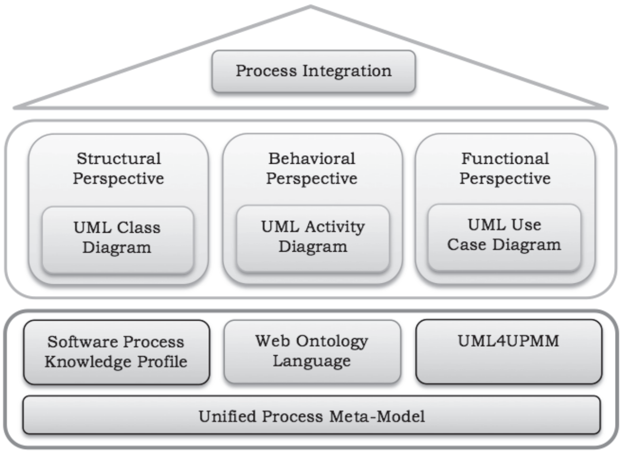



**Figure 4.** Software process modeling reference house

## 4. Process Knowledge Framework

Added value of our formal knowledge framework lies in the very combination of the techniques we discussed so far with a methodology (whole procedure is described in the figure 3) that is able to link them in iterative process modeling method that is based on phases: Process meta-model alignment, Common vocabulary capture (informal resources for process models), Software process explicit knowledge profile development (formalization using meta-model formal base), Consistency check (automatized with reasoners and knowledge validity with stakeholders), Pattern-based visual process modeling (based on reusable explicit profile elements), Process analysis and simulation (with Petri Nets based tools), Process optimization (utilizing reverse engineering). All aforementioned phases fit results into various software process perspectives of Formal Knowledge Process Modeling Framework reference house that is depicted in the figure 4:

- Level 0 of the house (basement) – Base Layer – meta-model and terminology and Middle Layer – explicit process profile (provides reusable elements for visual process modeling);
- Level 1 of the house – Top Layer – pattern based visual process modeling;
- Level 2 of the house (roof) – models integration, simulation, reengineering and optimization.

*4.1. Base Layer and Middle Layer – Meta-Model and Terminology Alignment*

The first step is dedicated to a meta-model consistency check – it is consulted with the customer and all terminology is aligned to match the one the organization is used to. This step is followed by reusing or creation of the very software process knowledge profile – the mechanics of this phase are based on conceptualization methods that allow rigid capture of real-world concepts and transform them into explicit knowledge profile with utilization of OWL language that supports many different semantic relationships including taxonomy, relation modeling, property modeling, annotation documentation, versioning, etc. [17]. It is crucial to realize that the Unified Process Meta-Model is completely independent on any process model and uses only basic language elements (in this case Concept, Object Property, Subclass of, Data Property, and Annotation). Captured terminology and relations are then categorized to unambiguous process perspectives with the support of process patterns, i.e. an approach we call the sandbox process modeling (modelers are presented with predefined process templates they can use and further manually extend in semi-formal modeling tier) – figure 5 depicts an example (navigable and oriented) graph structure that presents an extract of knowledge profile of Rational Unified Process. Every formal assertion of process model is defined as a 7-tuple – *Process Assertion (Identification, UPMM Type, Name, Public Label, Definition, Relationships, Extra Properties)*

- Process Assertion is an individual, an instance taken from an ABox (e.g. System Analyst);
- Identification is an URI (unified resource identifier) taken from OWL knowledge profile and is further used as a link between visual model elements and knowledge profile;
- UPMM Type is a class/type of the element according to customized and aligned (e.g. Role);
- Name is official human readable name in respective knowledge profile (e.g. System_Analyst);
- Public Label is a set of localizable human readable identifiers of specific process model element (e.g. {EN: System Analyst, FR: Analyste de Système, CZ: Systémový Analytik});
- Definition is a set of definitions in localized version (behaves similar to Public Label);
- Relationships is a set of object properties that exist in respective process model RBox (e.g. {Performs(System_Analyst, Develop_Vision), ResponsibleFor(System_Analyst, Vision)});
- Extra properties is a set of descriptive attributes that do not fit any other reserved slots (e.g. {AverageIncome:double}).

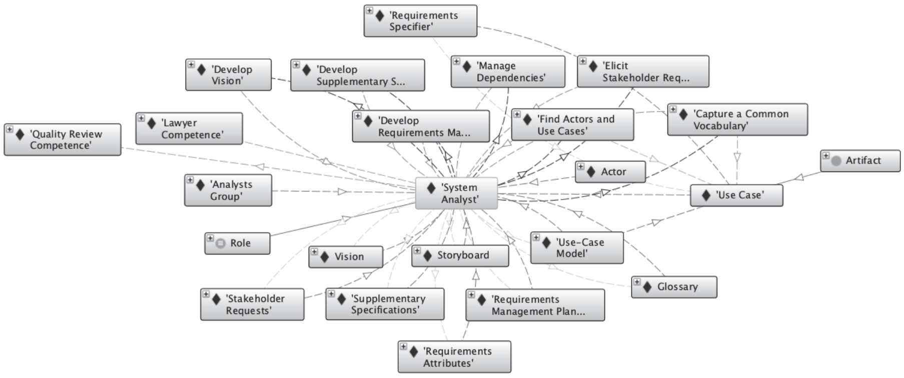



**Figure 5.** Extract of RUP knowledge profile

*4.2. Top Layer – Visual Process Modeling*

*Structural* perspective is captured and categorized with the combination of OWL and Class Diagrams. This step is usually performed by software house management and process auditors. Functional perspective is represented with Use Case diagrams and rules and constraints defined with OWL relations – necessary step of this phase is to capture various organizations scopes that are dependent on roles and competences (e.g. system analyst sees certain activities differently than project manager). *Behavioral* perspective combines the perspectives and viewpoints from the previous step. All available functions and responsibilities are generated from knowledge profile. The results then comprise of Process Meta-Model aligned with software house requirements and terminology ready for integration, Software process knowledge profile with rigid and most stable parts of structures and behavior, and Visual process models in UML linked to background knowledge profiles – figure 6 presents a whole modeling approach in two layers (process meta-model – upper and process model – lower).

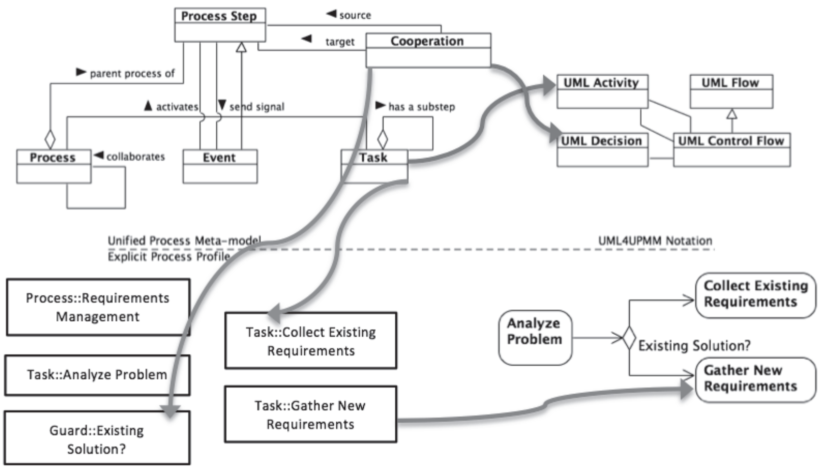


**Figure 6.** Software process modeling reference house

### *4.3. Execution Layer – Integration and Workflow*

Many modeling and tracing systems in sofwtare process domain are based on modules that bring different functionality to users. There is an analogy between functionality brought by automatized modules (e.g. storyboards modeling, requirements review, use-case modeling) and functionality captured in software process – both should be the same if done well (also using the same terms and rules preserving the integrity of activities performed in an organization). Chaining of the functionality on the other hand is a different way to describe process activities execution (sequence of functions done in right order under given conditions).

Set of models created in the framework with formal background and easy-to-read semi-formal foreground are easy to be transformed into an information system so the process execution could be automatized in a way described above. This approach allows improved configuration and adaptability of supporting modeling systems with various terminology including roles and responsibilities, competencies, functions, inputs, outputs, validations, pre- and post-conditions, events, context, issues, and exceptional states while preserving detailed model of software process behind the execution.

## 5. Software Processes Analysis

So far we have described process modeling that gives rigid terminology base in UPMM, explicit process profile describing stable part of software processes with knowledge background, sandbox modeling pool (gathering all elements of explicit profile available for visual modeling), and visual models created with reusable sandbox pool elements and extra model partitions captured and modeled above the original knowledge profile (undefined process elements can be easily distinguished from those taken from explicit profile as they miss link to respective knowledge element – there are several ways how to approach these elements but this is a problem of a methodology and shall not be discussed here).

The framework offers great possibilities in reusing and sharing process elements and provides great expressive power in its formal layer – this, combined with the modeling methodology, creates powerful and potent resource for further analysis. We support two major analytical tools that had been incorporated into the framework – process simulation with resources utilization and reverse engineering (modeling and optimization) methods.

### *5.1. Execution Optimization Layer – Process Simulation and Analysis*

We have introduced key parts of the framework in our previous papers [6, 7] but we have decided to refine the transformation rules and integrate it with the latest version of the Unified Process Meta-model (UPMM). Following diagram in figure 7 shows parts of the framework.

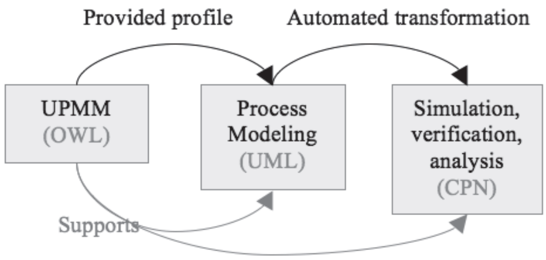



**Figure 7.** Framework parts and their relations

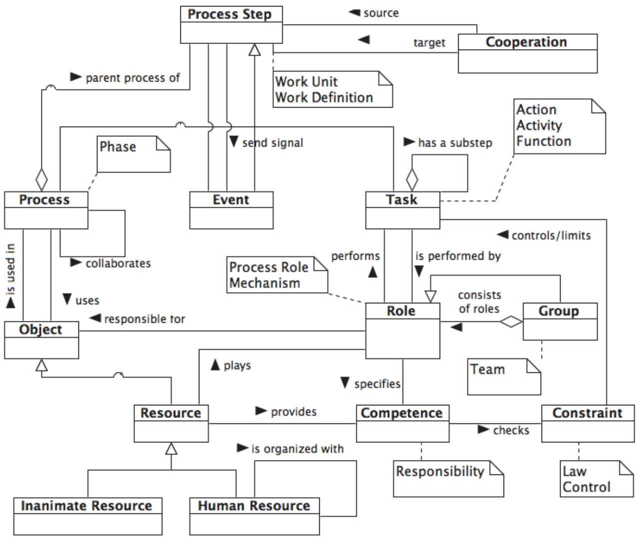



**Figure 8.** Subset of UPMM regarding resources and competences

Key part of the simulation layer is a small part of the UPMM shown in figure 8. It shows the relationship between Process step (whether it is an Event, a Task or a sub Process) and a Role that is played by Resource (Inanimate or Human) in Process Step execution. It also shows a relationship between Resource and Role through a Competence. The Competence is a skill, experience, or knowledge a Resource should posses. In our previous work, we haven't taken competences into account because we wanted to make the framework easy and compact but we have redesigned the framework to be extensible.

To achieve even more precise definition of software process and get more precise simulation and analysis reports regarding resource utilization, one cannot settle for specifying resources just as role based. As in the real world, resources (whether inanimate or human) can play several roles. For example resource *printer* may serve the role *printer* but may also serve as *coffee table*. To give you an example that is more likely to appear in Software process we shall consider a resource *employee* which is a *developer* but can also be a *technical leader*. Roles are important from constraint and security point of view. By giving resource a role, we can then use the role to limit access of other resources which are not in that role to specific activity. For example, only a resource under role *customer* can *create requirements* (activity) in our Software process, by stating this, we are excluding *developers*, *project managers* and others from being able to execute this activity. Roles can be also useful for grouping resources and thus creating an alias which can be used, instead of stating individual resources in the model.

We stated that using role based resources is not enough. We need to introduce a more precise mechanism. In the real world, each of us has a skill set which defines what we can do. We call each individual skill a competence, which can be view as a pair of skill and level of the skill (how good we are in that skill). So competences define what we can do and how good we can do it. The skill level can represent how fast we can do certain task or how precise the result will be. Now, the resource won't be defined just by its role, but will also have competences – the relation is depicted in figure 8.

In previous section we have described the relationships between resources, roles and competences. Now let's have a look on how are we going to incorporate it into Activity diagram and, of course, how are we going to transform it into CPN elements. Main element of every Activity diagram is an activity (activities); it represents a task (or tasks in case when activity is decomposed to further steps), i.e. a piece of work that is ought to be done. Activity is performed by a role that is played by a resource with certain competences. Activity can also consume (or utilize) passive resource (which is not performing the activity itself) and produce an artifact. The difference between consuming and utilizing is straightforward. When a resource is consumed, it is used and lost and cannot be reused later in the process (e.g. using paper in a printer). When a resource is utilized, it is used for certain amount of time but after that is free to be used again (e.g. printer).

A part of UML model on figure 9 shows how can we model an activity that is performed by a resource (represented by its role) with a competence (on certain level), consumes a passive resource and produces an artifact. The whole model is compliant with the UPMM. We have decided to use object and object flow instead of swim lane for modeling an activity being performed by a resource (as we did in [7]) because of a versatility (more than one roles can perform certain activity). You can also notice that UPMM uses one more element between a Competence and Task – Constraint. In example on figure 9 we are using an anonymous Constraint (not explicitly specified) represented by an object flow with a guard. The guard restricts a level of the competence and it reflects the data property Skill level in UPMM. It means that activity *Code integration into source control* can be performed only by a *Developer* with competence *Source control* on level greater than 70.

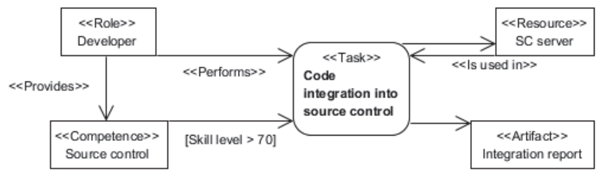



**Figure 9.** UML resource modeling compliant with UPMM

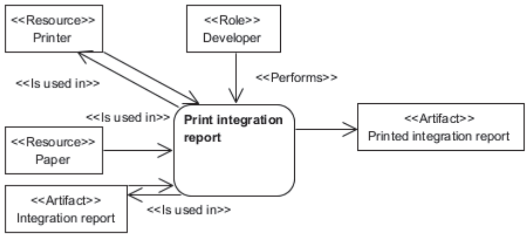



**Figure 10.** Difference between resource consumption and utilization

Figure 10 presents the difference in a UML model between resource consumption and utilization. It goes without saying that an Artifact can be also utilized (as input) in an activity. You can see the difference between consumption (only one output arc from object to activity) and utilization (two arcs between object and activity). You can also see that we can leave out the Competence if we feel it is not required.

When the models are ready, they are transformed into a CPN net using the information from both UML and OWL (figure 7). We can use two approaches to enhance our models via feedback from the results of the chosen approach. The first approach to improve (not only improve, but also to verify) the process models uses static analysis technique on generated net; the technique is called a state space analysis. The basic idea of state spaces is to calculate all reachable states and state changes of the CPN model and to represent these in a directed graph where the nodes correspond to the set of reachable markings and the arcs correspond to occurring binding elements [10]. The reachability graph is later used to examine properties of the CPN model and look for specific "quality" indicators. In that way we can expose the potential flaws (like a dead-lock situations or unnecessary activities) and fix or improve the process model.

The second approach is to use the CPN's natural ability for simulation and run a simulation (or more simulations) on the CPN model. Before we actually execute the simulation, we put markers on transition and places and mark the transitions and places- results are traced and logged when state changes. We are logging the number of executions in case of transition and marking changes in case of places. From the number of executions we can find potential bottlenecks in the process and marking changes give us an excellent overview of the resource utilization (detailed information can be found in [10]).

### *5.2. Models Reengineering Layer – Reverse Engineering and Models Optimization*

Reverse engineering provides an ability to discover, analyze or optimize process model from real executions of the software process. Reverse engineering is a supportive method of our knowledge framework. This approach gives us the possibilities to analyze the process from the bottom. It is possible to use data created by the process executions for discover, modeling and adjusting purposes. On the other hand, process modeling gives us view and process discovery from the top. It means that the domain of the process is discovered first and then, like aforementioned in this text, is process modeled and described in a knowledge profile [12].

Reverse engineering, how we call it, is technically a usage of the process mining in our framework with involvement of manual optimization of process models – the difference is that the reverse engineering is integrated method in our Formal Knowledge Process Framework. It provides a solution that is tightly related to captured processes and is especially adjusted in a way of usage in the knowledge framework. This allows various possibilities in the area of conformance checking and process model optimization. That is because of a process model preparation according to our framework rules – based on that we are able to make complex analysis and optimizations. Main benefit of the reverse engineering integration into the Knowledge Process Framework is that of our framework presents complex solution for the top-bottom and bottom-up process modeling.

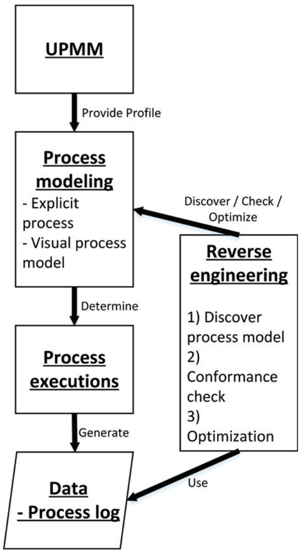



**Figure 10.** Reverse engineering approach

Figure 10 presents an idea of the reverse engineering. In the top of the picture is the UPMM that provides a terminology base for the Process modeling. We can choose which parts of the meta-model we need to use in process modeling activities. Results of all main modeling activities above our framework are firstly explicit knowledge profiles and visual process models that use knowledge elements in a way described in previous chapters. Behavioral visual models describe how the process executions should be executed, what activities should process execution perform, how are activities mutually cooperating, etc.

Specific process execution based on process models generates data to the log records. Logged data contain timestamps and detailed information about particular instances of the process execution. Assuming that the data are in a correct format and we are able to trace and collect all necessary execution details from the log, than we are able to use that information in the Reverse engineering phase; this approach offers further analysis of the data from the logs with techniques like review and discovery of the process model from the execution of the real software process (see [12] for more details). The scenario of the reverse engineering within our framework consists of three main parts:

- Conformance Check

- Optimization
- Process Model Discovery

Conformance Check includes control of the process model correctness. This check is done with comparison of the process model and the real process execution. The comparison procedure is based on the formal level, i.e. OWL and Petri Nets, in other words, we match traced data and process models to discover any form of deviations between proposed model and real process performance. So far we have recognized following types of unexpected states:

- Completely new and unknown activity had been executed and it does not exist in a process model;
- Activity that is in a process model had not been used ever;
- Activity is performed on regular basis and is captured in a process model but it is executed in different or strange context (followed or preceded by different activities).

With this type of analysis we are able to check whether the analyzed process model corresponds with the reality – if there are differences observed then it leads to the next step of the reverse engineering phase – Optimization.

Optimization focuses on the results analysis of the data extracted from the Conformance Check activity and use them to improve the model, e.g. if we find out that the activity had not been executed at all, we can adjust the process model.

The last, but certainly not least, method – Process Model Discovery – is used for gathering the knowledge about existing process that is performed in the reality and is not described in any process model. It is used in situation when process practitioners need to backtrack the process and create a process model based only on execution details. If there were some undocumented and undefined processes executed in observed software system, reverse engineering would help us to understand this domain and actual state of the software process hiding behind (from the information system perspective only, nonetheless). Discovery of the process model starts with the analysis of the data log – log record elements must be matched and mapped onto elements of the UPMM. If the mapping is correct, we shall be able to use process-mining algorithms and discover process model. Benefits of the reverse engineering technique used within the process framework are:

- Feedback on the process execution obtained from the reality – we are able to check if the process model is really followed;
- Continuous process optimization that goes on after models creation, integration, and automatization;
- Prevention of process model degradation – continuous conformity check avoids process model degradation;
- Deviations control and reduction – we are able to distinguish, discover and eliminate process deviations.

## 6. Conclusion

This paper discussed our results in knowledge framework for software processes and the techniques designed to support the framework. The benefits of the approach are following:

- Separation of semi-formal and formal methods brings the process modeling and understanding to less (mathematically) experienced staff;
- Formal (knowledge) definition linked to semi-formal models provides improved process documentation;
- Shared dynamic source of process knowledge brings modeling to the next level of modeling simplicity with process blocks pre-defined by domain specialist in explicit knowledge profile;
- Native framework support of Colored Petri Nets and reverse engineering methods provide powerful analytical abilities with features like simulation, exceptional situations handling, and process optimization and re-engineering.

Disadvantages of the framework lies in bigger base meta-model that is harder to adapt and train, also OWL process knowledge profiles modeling is not easy to learn and take some time to introduce into a system. It is obvious that our framework is suitable mostly for those software development environments that have rigid and unchanging processes (or the change is uncommon so it is less likely to introduce such an approach to agile environments) like those based on standards or reference processes like RUP. Future work will be dedicated to finalize the framework with complete meta-model and its full system integration that could be tested in a production environment.

## Acknowledgment

Work is partially supported by Grant of SP2015/85 - Knowledge modelling and its applications in software engineering, VŠB - Technical University of Ostrava, Czech Republic.

## References

[1] Aalst, W.M.P. van der, Hee, K.M., van, Houben, G.J. (1994). Modeling Workflow Management Systems with High-Level Petri-Nets, In G. De Michelis, C. Ellis, and G. Memmi, editors, Proceedings of the second Workshop on Computer-Supported Cooperative Work, Petri nets and related formalisms, pages 31-50, 1994.

[2] Aalst, W.M.P. van der, A.H.M. ter Hofstede, M. Weske, Business process management: a survey, Business Process Management (2003) 1–12.

[3] Aalst, W. M. P. van der 2014. *Process Mining in the Large: A Tutorial*. Lecture Notes in Business Information Processing. Vol. 172 LNBIP

[4] Allemang, D., Hendler, J., Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL, 2008.

[5] Booch, G., Jacobson, I., Rumbaugh, J. The Unified Modeling Language User Guide, Addison Wesley Longman, Inc., 1999.

[6] Czopik, J., Košinár, M., A., Štolfa, J., Štolfa, S. Formalization of Software Process Using Intuitive Mapping of UML Activity Diagram to CPN. Proceedings of IBICA 2014, VŠB-Technical university of Ostrava, Ostrava, 2014

[7] Czopik J, Kosinar A. M., Stolfa J, Stolfa S. Addition of Static Aspects to the Intuitive Mapping of UML Activity Diagram to CPN, Paper presented at the Proceedings of the 1st international conference on Afro-European Conference for Industrial Advancement, Addis Ababa, 2014.
[8] Du, Z. Yang, Y., Xu, J., Wang, J., Mapping UML Models to Colored Petri Nets Models based on Edged Graph Grammar, 2011.
[9] Dufresne, T., & Martin, J., Process modeling for e-business. Information Systems Department, George Mason University, L. Kerschberg (ed), 2003
[10] Jensen, K., Kristensen, L.M. Coloured Petri Nets: modelling and validation of concurrent systems. Dordrecht: Springer. ISBN 978-3-642-00283-0, 2009
[11] Kaufmann, M., Silver, G.A., Lacy, L.W., Miller, J.A., Ontology based representations of simulation models following the process interaction world view. Paper presented at the Proceedings of the 38th conference on Winter simulation, Monterey, California, 2006.
[12] Košinár, M., Štolfa, J., Štolfa, S. Knowledge Support for Software Processes (revisited). Paper presented at the 24th European-Japanese Conference on Information Modeling and Knowledge Bases, Kiel, Germany, 2014
[13] Narendra, N.C., Flexible support and management of adaptive workflow processes, Information Systems Frontiers 6(3), pp. 247–262, 2004.
[14] Scacchi W, Mi P., Process Life Cycle Engineering: A Knowledge-Based Approach and Environment. Intelligent Systems in Accounting, Finance, and Management 6:83--107-183—107, 1997
[15] Vergidis K., Tiwari A., Majeed B., Business Process Analysis and Optimization: Beyond Reengineering. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38 (1): 69-82, 2008.
[16] Workflow Management Coalition: http://www.wfmc.org/ WfMC Web, 1999
[17] W3C, OWL 2 Web Ontology Language. http://www.w3.org/TR/owl2-overview/, 2009.

# EVOLVING DATABASE CONCEPTUAL GRAPH PARTITIONING

**Elvira Immacolata LOCURATOLO**

*ISTI Consiglio Nazionale delle Ricerche*
*Via Alfieri, 1- Pisa*
*Telephone: +39 50 3152895 Fax: +39 50 3152810*
*E-mail: elvira.locuratolo@isti.cnr.it*

**Abstract:** Algorithms of graph partitioning exploited in conceptual database design were reused to define a methodology of database concept preservation. An algorithm, the *concept construction algorithm,* that relates concept theory to computer science was designed. This algorithm, however, is not suitable for implementation. In this paper, a relationship between conceptual graphs and concept generalization hierarchies is established at the boundary between concept theory and computer science. The algorithmic property of *class/concept completeness* is given and an algorithm achieving this property is introduced as a refinement step of the concept construction algorithm.

**Keywords:** database design, partitioning, conceptual modeling, algorithms, correctness, class/concept completeness.

## 1. Introduction

The connection between *concept theory* and *computer science* [3,4] is a novelty in the research field that takes its origin from algorithms of database conceptual graph partitioning [2,5,7].

Although a consolidated formal background is available in concept theory, only recently an algorithm, the *concept construction algorithm* [4], able to establish relationships between concept theory and computer science, has been designed. This algorithm is supported by a methodology which starts from concepts and goes towards classes/categories of objects. The first phase of the methodology consists in defining a formal background, which allows concepts and concepts operators to be introduced. The second phase is concerned with the input identification of the concept construction algorithm; finally, the last phase with the resulting concept structure, which is a structure *formalized in concept theory,* satisfying the following properties:

- Encloses all and only the concepts related to an initial concept structure.
- Encloses all and only the intensional inclusion relations between concepts.
- Results in leaves which are mapped to the classes resulting from algorithms of database conceptual partitioning.

The algorithms of database conceptual graph partitioning were designed to combine the conflicting quality desiderata of *flexibility* to modify a database schema supported by semantic data models with the *efficiency* of object data systems. As a consequence, the database applications can be specified with flexibility by referring to the conceptual schema, while the obtained implementations can exploit the efficiency provided by the object database systems [5].

In sections 2 and 3, a database partitioning algorithm, the *algorithm for the object classes*, and the *concept construction algorithm* [5] are respectively improved. The concept construction algorithm is formalized in concept theory, however, this notation is too expensive when compared with the database notations suitable for algorithmic implementations. In section 4, a relationship between conceptual graphs and concept generalization hierarchies is introduced at the boundary between concept theory and computer science, and the algorithmic property of *class/concept completeness* is given. An algorithm achieving this property, that can also be considered as a first refinement step of the concept construction algorithm, is introduced. The conclusions and the further developments are enclosed in section 5.

## 2. Database Conceptual Graph Partitioning

Algorithms of database graph partitioning [2,5,7] have been exploited in conceptual database design to map directed acyclic graph of classes supported by semantic data models, called *database conceptual graphs*, to classes supported by object systems, called *database object graphs*. They make it possible to achieve the coexistence of *flexibility* and *efficiency*, two conflicting quality desiderata. The following properties hold:

- *Database Conceptual Graphs*: each object instance can belong to any class of the graph, thus ensuring *flexibility*;
- *Database Object Graphs*: each object instance can belong to one and only one class of the graph, thus ensuring *efficiency*.

The links of a database conceptual graph are *is-a* relationships; whereas the links of a database object graph are *is-a*$_o$ relationships. In the following, the two definitions of *is-a* and *is-a*$_o$ are given.

- Class <Y> *is-a* class <X> *iff* the objects of class <Y> are enclosed into the objects of class <X> and class <Y> inherits attributes from class <X>. Class <Y> may have specific attributes.

- Class <Y> *is-a*$_o$ class <X> *iff* the objects of class <Y> are disjoint from the objects of class <X> and class <Y> inherits attributes from class <X>. Class <Y> may have specific attributes.

Various algorithms of partitioning can be applied to a database conceptual graph, all of them resulting in equivalent database object graphs and enjoying the properties of *correctness* and *completeness* [4]. *Correctness* is guaranteed by proving that root partitions of database conceptual graphs are provided step by step until all the disjoint classes composed in an object graph are obtained. *Completeness* is guaranteed by proving that the resulting disjoint classes define the finest partition of the graph root. In the following, the algorithm for the object classes is given that generates the maximum number of intermediate steps before reaching the object classes:

**Algorithm for the object classes (** $A_S$ )
**Begin**
If $A_s = <root>$ **then**
**Return** $A_O = A_s$
**Else**
If $A_s = <son> is-a <root>$ **then**
**Return**
$A_O = <son> is-a_O <root-son>$
**Else begin**
**Decompose** $A_S$ in $A_{S1}$ and $A_{S2}$
$A_{O1}$ = **Algorithm of maximum steps** ( $A_{S1}$ )
$A_{O2}$ = **Algorithm of maximum steps** ( $A_{S2}$ )
$A_O$ = **Merge** ( $A_{O1}, A_{O2}$ )
**Return** $A_O$
**End**

**Procedure Merge (** $A_{O1}, A_{O2}$ )
**Begin**
$A_O = <A_{O2}\ root>\ is-a_O <A_{O1}\ root>$
**End**

The **algorithm for the object classes** is applied to a graph of semantic classes $A_S$ and results in a database object graph $A_O$. In **Appendix A),** it is shown that the graph $A_O$ is equal to $<root>$ in *case* 1. and to $<son> is-a_O <root-son>$ in *case* 2. As the general case, the composition of two sub-problem solutions is given by the merge procedure, which links the root of $A_{O2}$ to the root of $A_{O1}$ through the *is-a$_o$* relationship. In *case* 3a. the given conceptual graph is decomposed into a conceptual graph of *case* 1., and a conceptual graph of *case* 2. The former resulting concept structure is the object graph $A_{O1}$, whereas the latter resulting concept structure is the object graph $A_{O2}$.These are evidenced through red circles in the global solution $A_O$. In *case* 3b. the given conceptual graph is decomposed into the two conceptual graphs of *case* 2. Each of them give rise to a resulting object graph, evidenced through red circles in the global solution $A_O$. Further, let us observe that the graph decomposition is performed with respect to the son at the left side of the root. Although the choice of a well established son is required for the graph decompositions, the choice of another son would lead to get the same disjoint classes. Various equivalent global solutions can result from the disjoint classes of case 3.b., also enclosing two cases of a node in multiple inheritance.

## 3. An intensional concept theory

An intensional concept theory [1,5], denoted by KC, is presented in a first-order language L that contains individual variables *u, v, z*, .... which range over the concepts, and one non-logical 2-place intensional containment relation, denoted by ≥. A concept *u* contains *intensionally* a concept *v* if the information content of *u* is greater or equal to the information content of concept *v*. As an example, the concept '*student*' contains intensionally the concept '*person*', formally *student* ≥ *person.* At the extensional/set theoretical level, set (*student*) ⊆ set (*person*). There are several non-identical concepts, which are co-extensional, and so we can infer from a concept to its extension, but not vice versa. Based on the *intensional inclusion* relation, the following relations of *compatibility* ⊥, *uncompatibility* T, *comparability* H, *uncomparability* I, and *intensional restricted negation* $^{\neg r}$ are introduced [5]. Concept *student* is compatible with concept *person* since $(\exists x)(x \geq student \wedge x \geq person)$; concept *student* is incompatible with concept *professor* since $\neg(\exists x)(x \geq student \wedge x \geq professor)$, concept *student* is comparable with concept *employee* since $(\exists x)(student \geq x \wedge employee \geq x)$.

Properties of these relations allow defining concept constructors, which have correspondence with the set-theory partitioning operators. If *u* is *compatible* with *v,* the least upper bound, denoted by ⊕, *exists*. Thus, the concept *person* ⊕ *student* can be defined. Correspondently, at set theoretical level, the following intersection of classes <*person*> ∩ <*student*> is obtained. If *u* is *not compatible* with *v,* then the two concepts *u* and *v* are incompatible. Correspondently, at the set theory level disjoint classes are defined. If *u* is *comparable* with *v,* the greatest lower bound, denoted by ⊗, exists; thus, the concept *student* ⊗ *employee* can be defined. Correspondently, at extensional/set theoretical level a class is obtained as union of the following classes: <*person*> ∪ <*employee*>.

The *Concept Construction Algorithm,* introduced to relate concept theory to computer science, is an algorithm of maximum step applied to the *initial concept structure,* defined through basic concepts directly intensionally linked to the general concept [4].

**Concept Construction Algorithm** ($\theta_n$)
**Begin**
If $\theta_1$ **then**
**Return** $R_{\theta 1}$
**Else**
If $\theta_2$ **then**
**Return** $R_{\theta 2}$
**Else begin**
**Decompose** $\theta_n$ in $\theta_{n1}$ and $\theta_{n2}$
$R_{\theta n1}$= **Concept Construction Algorithm** ($\theta_{n1}$)
$R_{\theta n2}$= **Concept Construction Algorithm** ($\theta_{n2}$)
$R_{\theta n}$ = **Merge** ($R_{\theta n1}$, $R_{\theta n2}$)
**Return** $R_{\theta n}$
**End**

**Procedure Merge** ($R_{\theta n1}$, $R_{\theta n2}$)
**Begin**
$R_{\theta n1} \geq R_{\theta n2}$
**End**

The structure resulting from the *concept construction algorithm* is formalized in concept theory. Following this formalization, all the database attributes are to be explicitly specified through information contents, which are primitive undefined concepts. An information content of a concept can be restricted to a given concept. As an example, [*name income matriculation*] $\geq$ [*name income*] and *student* $\geq$ *person* $\Rightarrow$ [$name|_s$ $income|_s$ $matriculation|_s$] $\geq$ [$name|_p$ $income|_p$]. Concept [$name|_p$ $income|_p$] is intensionally contained in concept [$name|_s$ $income|_s$ $matriculation|_s$]. Concept $matriculation|_s$ is specific to the concept [$name|_s$ $income|_s$ $matriculation|_s$]. In concept theory, the attribute inheritance of database conceptual graphs cannot be taken into consideration. The two concept theory operators $\oplus$ and $\Phi$ are applied to concepts defined through restricted information contents as follows:

- $[u_{j\backslash u}\ ,\ j \in J] \oplus [v_{i\backslash v}\ ,\ i \in I\ ] = [u_{j\backslash u\oplus v}\ ,\ j \in J]$
- $[u_{j\backslash u}, j \in J_1]\ \Phi\ [v_{i\backslash v}\ ,\ i \in I\ ]\ = [v_{i\backslash u\Phi v},\ i \in I\ ]$

The concept theory notation is too expensive when compared with the database notations, that exploiting the properties of is-a relationship, result to be suitable for implementations. In order to solve this problem, in the next section a new relationship is introduced at the boundary between concept theory and computer science.

## 4. Concept theory and computer science: the boundary

In order to introduce an algorithm for the refinement of the concept construction algorithm, the following definitions are given:

**Def:** *Database generalization hierarchy of concepts.*
Given a conceptual graph $A_S$, a *database generalization hierarchy of concepts* $A_c$ is a structure with general concept $[root]$ corresponding to the <root> node of the conceptual graph $A_S$ and remaining concepts linked through the following $is-a^C$:

Concept [X] $is-a^C$ concept [Y] *iff* class <Y> *is-a* class <X> *in* $A_S$

The generalization hierarchies of class/concepts are structures common to both computer science and concept theory. From the above definition of $is-a^C$, it results that the simplest concept structure located at the boundary between computer science and concept theory is the generalization hierarchy of concepts. Exploited in an intensional concept theory, the $is-a^C$ relationship can be replaced with the intensional inclusion relation $\geq$. Exploited in computer science, the intensional inclusion relation $\geq$ of generalization hierarchies can be replaced with the $is-a^C$ relationship.

The **algorithm for the integration of generalization hierarchies** is applied to a database conceptual graph $A_S$. Four initial cases and a general case compose this algorithm. The resulting concept structures of the initial case are enclosed in Appendix B. The general case consists in the recursive decompositions of the conceptual database graphs until elementary cases are reached. The algorithm for the integration has been designed to achieve the property of *class/concept completeness*, given in the following:

**Def:** *Class/concept completeness*
Given a conceptual graph $A_S$, for *class/concept completeness* we means the integration of n generalization hierarchies of classes/concepts, formalized through the set-theory and the $is-a^C$ relationship, one for each direct son of the given conceptual graph.

The details and the demonstrations of this algorithm are enclosed in [2].

### *4.1. Discussion*

Algorithms of database conceptual graph partitioning have been exploited in conceptual database design. The paper is oriented towards the evolution of the database conceptual graph partitioning research; at this purpose, a relationship is introduced at the boundary between concept theory and computer science. This relationship allows concept structures formalized in set theory through the $is-a^C$ relationship to be introduced. Differently, from the algorithm of concept construction, which is applied to an initial concept structure formalized in concept theory, the algorithm for the integration of generalization hierarchies is applied to a database conceptual graph formalized in *set-theory* through the *is-a relationship,* and thus all the properties of the is-a relationship which are suitable for the implementation can be exploited. In [8], differences between the extensional relationships and the intensional relationships are outlined. As results of research, the present paper adds further differences: at extensional level, the distinction between *is-a* and *is-a*$_O$ relationships; at the boundary between the extensional and the intensional levels, the $is-a^C$ relationship defined through the *is-a* relationship *in* a conceptual graph $A_S$.

As the conceptual graphs can also have only direct descendents, it can also be considered as a first refinement step of the concept construction algorithm. The algorithm for the object classes enjoys the property of *completeness* [4]; the concept construction algorithm enjoys the property of *concept completeness* [5]. The algorithm for the integration of generalization hierarchies enjoys the property of *class/concept completeness,* introduced in this paper. After the introduction of the new algorithm, the distinction between concept completeness and class completeness proposed in [5], has to be replaced with the distinction between concept completeness and completeness.

## 5. Conclusions and Further Developments

Generalization hierarchies of concepts are structures common in database and concept theory researches. In this paper, a relationship is introduced at the boundary between concept theory and computer science and the definition of class/concept completeness is given. An algorithm designed to achieve the class/concept completeness is introduced. This algorithm, that has a own autonomy, can also be considered as a first refinement step of the concept construction algorithm.

As further results of research, the present paper takes into consideration the following differences: at extensional level, the distinction between *is-a* and *is-*$a_O$ relationships; at the boundary between the extensional and the intensional levels, the $is-a^C$ relationship defined through the *is-a* relationship *in* a conceptual graph $A_S$. Further developments of this paper are concerned with advancements of the algorithm for the integration of generalization hierarchies.

**Acknowledgments**

The author thanks Prof. Bruno Codenotti, IIT-CNR, Pisa, for the availability and seriousness with which he followed her work, and Dr. Fausto Rabitti, ISTI-CNR, for the useful suggestions in improving this paper.

The author thanks her husband, Mr. Antonio Canonico, for his help during the preparation of this paper.

The connection between concept theory and computer science is a research undertaken by Dott.ssa Elvira I. Locuratolo and Prof. Jari J. Palomäki during periods of CNR short term mobility programs in Pisa and Pori.

## References

[1] Kauppi R. Einführung in die Theorie der Begriffssysteme*,* Acta Universitatis Tamperensis, Ser. A., 15 University of Tampere, Tampere, 1967.

[2] Locuratolo E. I. (2015) Evolving Database Conceptual Graph Partitioning TR- 003-2015

[3] Locuratolo E. I., Palomäki J. J. (2008). Extensional and Intensional Aspects of Conceptual Design, in: H. Jaakkola, Y. Kiyoki, T. Tokuda (Eds) Information Modelling and Knowledge Bases XIX, IOS Press, Amsterdam, Berlin, Oxford, Tokyo, Washington, pp.160-169

[4] Locuratolo E. I., Palomäki J. J. (2014). Relating concept theory to computer science. In: EJC 2014 – 24th International Conference on Information Modelling and knowledge Bases (Kiel, Germany, 3-6 June 2014). Proceedings, pp. 320 - 338. B. Thalheim, H. Jaakkola, Y. Kiyiki (eds.) (Kiel Computer Science Series (KCSS), vol. 4). Department of Computer Science, Faculty of Engineering, Kiel University.

[5] Locuratolo E. I., Rabitti F., (1998) Conceptual Classes and System Classes in Object Databases, Acta Informatica 35,181-210.

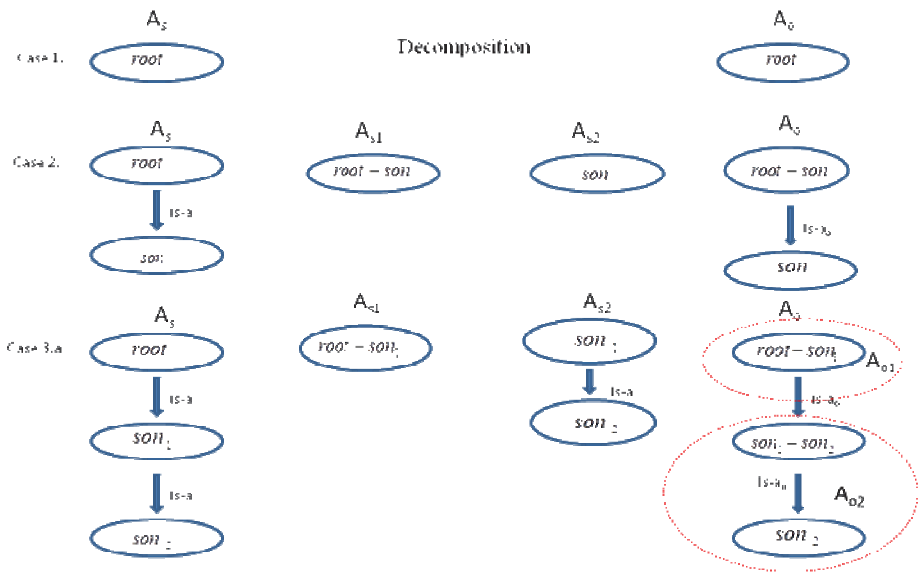
[6] Palomäki J.J, From Concepts to Concept Theory, Acta Universitatis Tamperensis, Ser. A. 416, University of Tampere, Tampere, 1994.

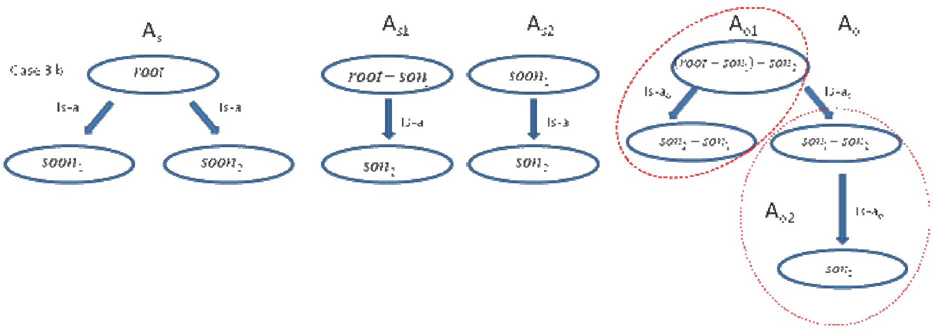
[7] Spagnolo, A. M., Incrementare la Qualità in Ambito Basi di Dati, Università degli Studi di Pisa, Tesi, 2000.

[8] Palomäki, J. and Kangassalo, H. (2012). That Is-IN Isn't Is-A: A Further Analysis of Taxonomic Links in Conceptual Models. In: G. Ramirez (Ed.). InTech: Rijeka, (pp. 3-18), *Advances in Knowledge Representation.* http://www.intechopen.com/books/advances-in-knowledge-representation/that-is-in-isn-t-is-a-a-further-analysis-of-taxonomic-links-in-conceptual-modelling

**Appendix A)**

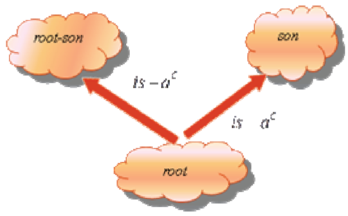
**Algorithm for the object classes ($A_s$)**



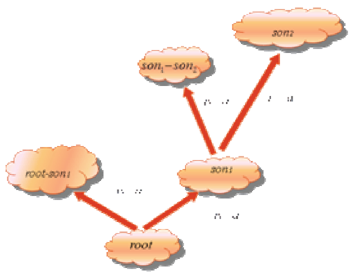



**Appendix B) ALGORITHM FOR INTEGRATION: Resulting concept structures**

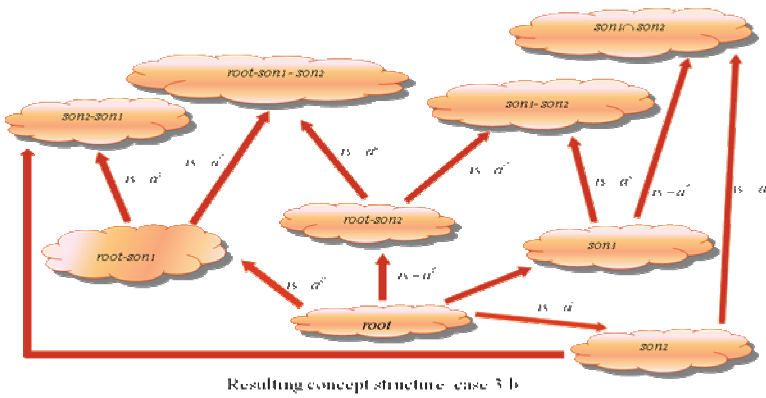

Resulting concept structure: Case 1.




Resulting concept structure: Case 2.




Resulting concept structure: case 3 a




Resulting concept structure: case 3 b

# Model Driven ActiveRecord with yEd

Matthias Sedlmeier, Martin Gogolla

**Abstract.** Since its release in 2004, Ruby on Rails has evolved into a widely used full stack model-view-controller (MVC) framework. But despite the fact, that Rails (short for Ruby on Rails) is also used for developing enterprise-scale applications like Github or scientific tools like QTREDS, there is no official support for graphical modelling. This paper introduces a proposal to fill this gap by suggesting a model driven approach using the free yEd diagram editor as well as a specifically developed transformation tool and ER dialect. The implementation is based on the Rails data abstraction layer ActiveRecord and its provided domain specific languages.

**Keywords.** model driven development, MDD, yEd, DSL, Ruby, Rails, ActiveRecord, graph, diagram, ER, modelling

## 1. Introduction

Since its release in 2004, Ruby on Rails [22] has become a widely used open source web application framework. Rails is purely implemented in Ruby, a programming language developed by Yukihiro Matsumoto in the mid-1990s [21]. Primarily extracted from a project management tool called Bootcamp, Rails evolved into a full stack model-view-controller (MVC) framework [11].

The framework essentially consists of 6 components, namely `ActionMailer`, `ActionPack`, `ActiveRecord`, `ActiveModel`, `ActiveSupport` and `Railties` as shown in Figure 1, where adjacent elements indicate an implementation or usage relation. While `ActionMailer` provides logic for email exchange, the `ActionPack` component is responsible for handling HTTP requests by providing controller code and view templates. This component roughly *implements* the view-controller part and makes usage of `ActiveSupport` functionality.

`ActiveRecord` [6] is responsible for mapping business objects to relational databases and for establishing connections between those objects carrying both persistent data and access logic. It brings its own domain specific languages for defining SQL schema migration and ORM (Object Relational Mapping) [6] model class files and hence *implements* the model part of the MVC triumvirate.

The `ActiveModel` component enables Rails to work with non-`ActiveRecord` models following the ORM principle, while `ActiveSupport` mainly delivers utility logic and extensions for the Ruby language. Finally, there is `Railties`, which glues together all components, handles the bootstrapping process and provides additional developer tools.

Rails development basically means creating text files containing Ruby code. While this is a common way to define business logic, it lacks of comfort when defining data models, because every single entity spawns 2 Ruby files. A schema containing 10 entities, for example, spreads at least over 20 separate files. The file count rises, if additional

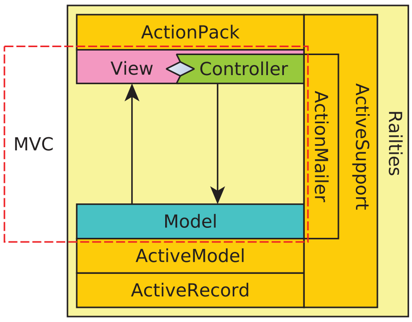



**Figure 1.** Ruby on Rails component structure

attributed associations are modelled. And despite the fact, that Rails is also used for developing enterprise-scale applications like Github [13] or scientific tools like QTREDS [18], there is no official support for graphical modelling although it brings advantages like clarity, comprehensibility and changeability.

This paper presents a proposal to fill this gap by introducing a model driven approach using the free yEd graph editor and a specifically developed transformation tool, which translates the yEd output into a valid textual `ActiveRecord` representation.

## 2. Textual Rails Modelling

To understand the advantages of the introduced approach, it is reasonable to look at the standard Rails data modelling process. The ActiveRecord DSL provides a clean abstract way to define models, their attributes and associations as well as restrictions. Each entity is represented as a Ruby model class depicting its table counterpart in a relational database. The necessary SQL schema is created by so-called migrations, which are also represented by Ruby classes.

A short minimal *Rental* example will illustrate this process in detail, while consciously avoiding any diagrammatic representations to reveal the significance of this work. An impatient reader might directly have a look at Figure 8.

We want to model the fact, that a person can rent a car and therefore extract Customer, Car and Rental as schema artefacts. Customer and Car are modelled as separate entities, while Rental is represented as a relationship between those entities. The Customer entity features a name, age, gender and customer number attribute. A Car instance has specific values for manufacturer, model, size and color. The Rental relationship keeps information about duration, free mileage and insurance coverage. A Customer may rent one or more cars at the same time, while one Car can only be attached to one renter at any given moment.

The formulated aspects are now implemented as an ActiveRecord data layer. In the first step the migration files as well as the empty model class files are created via special helper scripts, called generators (where g stands for generate), see Listing 1.

Listing 1: Command line generator calls

```
$ bin/rails g model Customer name:string age:integer
```

```
    gender:string number:integer

$ bin/rails g model Rental duration:integer
    mileage:integer insurance:string
    customer:references car:references

$ bin/rails g model Car manufacturer:string model:string
    size:integer color:string
```

### 2.1. Generated ActiveRecord Migration Files

The generated migration files are used to set up the SQL schema needed to save instances to a particular relational database (like SQLite [10], MySQL [23] or PostgreSQL [17]). As we modelled 2 entities and 1 attributed relationship, 3 migrations are generated.

Listing 2: Customer entity migration file

```ruby
class CreateCustomers < ActiveRecord::Migration
  def change
    create_table :customers do |t|
      t.string  :name
      t.integer :age
      t.string  :gender
      t.integer :number

      t.timestamps
    end
  end
end
```

The Customer migration in Listing 2 defines a change on the empty SQL schema in form of a table creation. The added relation is named `customers` and maintains four attribute columns, for which the data type is explicitly indicated. All migrations use a special `timestamps` directive, which creates additional columns for creation and update time.

Listing 3: Rental relationship migration file

```ruby
class CreateRentals < ActiveRecord::Migration
  def change
    create_table :rentals do |t|
      t.integer     :duration
      t.integer     :mileage
      t.string      :insurance
      t.references  :customer, index: true
      t.references  :car, index: true
```

```
      t.timestamps
    end
  end
end
```

Besides the description of the attribute columns, the Relation migration in Listing 3 introduces two foreign key columns. These columns will be used to join corresponding Customer and Car instances. The index option states, that the used database system defines an index on each of these fields for faster access. The last migration for Car is represented in Listing 4 and requires no additional remarks.

Listing 4: Car entity migration file

```
class CreateCars < ActiveRecord::Migration
  def change
    create_table :cars do |t|
      t.string  :manufacturer
      t.string  :model
      t.integer :size
      t.string  :color

      t.timestamps
    end
  end
end
```

The suggested solution models the Rental relationship explicitly in a separate join table. Due to the 1:n connection between Customer and Car this is not strictly necessary, because the Car table could save all the required information. There are also two different ways to place the foreign keys, assuming the Rental relationship is modelled explicitly. In the present example, Customer and Car instances are connected via two foreign keys in the Rental table. It is also possible to place a reference in the Car table pointing to the associated Rental instance, which again points to the corresponding Customer instance.

The current design decision is justified by the fact, that ActiveRecord offers additional support for 1:n connections when implemented as described.

### 2.2. *Generated ActiveRecord model class files*

Besides the migration files, also model class files are generated. These are incomplete yet and must be manually adjusted to fit the requirements. Here, we also use a ActiveRecord specific DSL to define associations and constraints. This way, we obtain 3 model class files, which are shortly described.

Listing 5: Customer entity ORM model class file

```
class Customer < ActiveRecord::Base
```

```
  has_many   :rentals
             :dependent => :destroy

  has_many   :cars,
             :through => :rentals

  validates_presence_of :name,
                        :age,
                        :gender,
                        :number

  validates_uniqueness_of :number

end
```

The first model class file represented in Listing 5 represents the Customer entity. A customer has one or more rentals, which is expressed by the `has_many` directive. ActiveRecord recognizes, that `rentals` is the lower case plural version of `Rental`, for which reason no additional information must be given. The ActiveRecord component now expects a foreign key named `customer_id` in the `Rental` table. The `dependent` option states, that associated `Rental` instances are removed, when the `Customer` is deleted. This prevents orphaned records in the rentals table. The second `has_many` directive enables direct navigation from `Customer` to `Car` instances via `Rental` records by using the `through` option. It expects a reference to `Car` in the `Rental` table, which is given by the corresponding foreign key.

Both directives implicitly establish so-called *association proxies* representing entity connections. Further, the ActiveRecord DSL allows the definition of so-called *validations*. The `Customer` model class checks, whether all attribute values are given, before an instance is saved. The customer number is additionally checked for its uniqueness.

Listing 6: Rental relationship ORM model class file

```
class Rental < ActiveRecord::Base

  belongs_to :customer
  belongs_to :car

  validates_uniqueness_of :car_id

  validates_presence_of :duration,
                        :mileage,
                        :insurance

  validates_numericality_of :duration,
                            :mileage

end
```

In the second model class file two `belongs_to` directives are used. These directives state, that each `Rental` instance is connected to one `Customer` and one `Car` instance by the corresponding foreign keys `customer_id` respectively `car_id`. To make sure each car can only be rented by one customer at the same time, the uniqueness of the `car_id` is checked. Furthermore, the `Rental` model ensures the presence of all its attribute values, of which `duration` and free `mileage` must be numeric.

Listing 7: Car entity ORM model class file

```
class Car < ActiveRecord::Base

  has_one :rental

  has_one :customer,
          :through => :rental

  validates_presence_of :manufacturer,
                        :model,
                        :size,
                        :color

  validates_numericality_of :size

end
```

The third model class file represents the `Car` entity and describes `has_one` connections to `Rental` and `Customer` via `Rental`. These directives can be seen as the counterpart of the `has_many` directives in the `Customer` model. Here, the ActiveRecord component expects a foreign key reference to `Car` and `Customer` in the `Rental` relation. As seen before, further attribute checks are declared.

It is worth to mention, that *structural constraints* or *value restrictions* are usually described and checked on application level within the Ruby code. However, it is possible to define them in the SQL schema using the described migration file mechanism as it is typically done when describing the column data types or allowing *null* values.

We have seen, that ActiveRecord is able to simplify the data modelling process by structuring the tasks necessary to implement a data layer with regard to SQL schema and ORM model class definitions. ActiveRecord also supports developers declaring associations and validations in a more abstract way by providing generators and domain specific language elements, which are basically self-explanatory using natural language. However, data modelling can be shaped more clearly and efficiently by introducing a model driven approach as we will see in the next part of this work.

In conclusion to this section, we give a short outline of the ActiveRecord migration and model class DSL fragments used in the examples above.

`create_table:`
: *defines a SQL table (t)*

`t.string`:
*defines a string column on table t*

`t.integer`:
*defines an integer column on table t*

`t.references`:
*defines a reference (id) column on table t*

`t.timestamps`:
*defines a create and update column on table t*

`has_one`:
*defines a one-to-one association to another model class;*
*the foreign key is expected to reside in the table of the other model class*

`has_many`:
*defines a one-to-many association to another model class;*
*the foreign key is expected to reside in the table of the other model class*

`belongs_to`:
*defines an association to another model class;*
*the foreign key resides in the current table;*
*usually used as counterpart to has_one and has_many directives*

`validates_presence_of`:
*checks the presence of an attribute or association*

`validates_uniqueness_of`:
*checks the uniqueness of an attribute value*

`validates_numericality_of`:
*checks, whether an attribute value is numeric*

## 3. Graphical Modelling with yEd

Using yEd for the graphical representation of ActiveRecord concepts, which are automatically transformed parsing the yEd GraphML [3] output, is a novelty in the area of Rails development. yEd is free of charge, although not open source, cross-platform and uses an XML-based format called GraphML for loading and saving created diagrams, which can be easily evaluated using XPath [7] expressions.

It ships with graphical language elements for a multitude of diagram types like flowcharts, UML diagrams, BPMN and so on. For this work, the provided Entity Relationship [4] palette is used to express ActiveRecord data models and to map graphical language elements to Rails DSL fragments. Basically, it is possible to adapt the concrete graphical yEd syntax with its nodes, edges and edge ends to fit the developer's requirements by creating custom palettes with even custom graphical language elements.

The developed transformation tool maps all graphical elements automatically to corresponding ActiveRecord migration and ORM model class files using the provided DSLs and releases the developer from manually creating any textual definitions in the best case.

Before translating the introduced textual example into a graphical representation, we give a short outline on the specifically developed ER dialect with its language elements.

Domain specific languages [5], no matter if textual or graphical, are specific, meaning, that they focus on certain main aspects of the modelled target domain. The graphical DSL introduced in this work especially aims for fast and easy description of entities, their attributes and notably their relationships also with regard to agile development processes with frequently evolving requirements. The assumption is, that restraining the expressiveness helps developers to model faster and less error-prone. Another aim is to offer a kind of standard pattern catalogue for structural composition of artefact connections, i.e. a default mapping for explicit 1:n relationships as seen in the given example.

### *3.1. Basic Language Description*

Based on the ER notation, the proposed language supports the Entity artefact. An Entity represents an independent concept and is drawn as a rectangle labelled with its name. In contrast to popular ER dialects entities cannot be weak. In UML class diagrams they would be expressed as classes. As usual for object relational mapping approaches, an Entity is generally mapped to a single table, provided that relational database systems are used, see Figure 2.

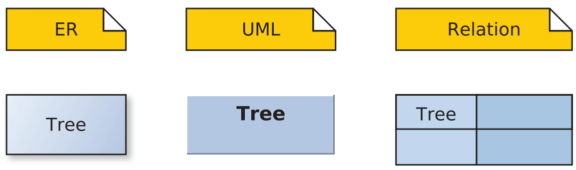



**Figure 2.** The entity artefact

Entities usually express their uniqueness in form of an attribute set. Attributes are drawn in circles, which are connected to entities via plain edges. Those circles are labelled with the attribute name followed by a colon and the data type. Figure 3 shows a comparative example. In some ER dialects also multi-valued attributes and derived attributes are supported. The introduced DSL does it without.

One of the main modelling tasks concerns the mapping of relationships between entities. The proposed language generally differentiates 4 basic *multiplicity types*, where type 3 represents the inverse of type 2 with swapped source and target.

1. one-to-one (1:1)
2. one-to-many (1:n)
3. many-to-one (m:1)
4. many-to-many (m:n)

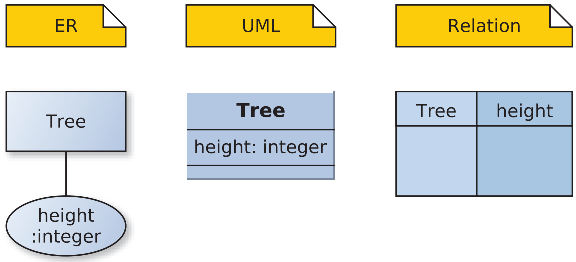



**Figure 3.** The entity artefact with attributes

Each basic type has multiple variations derived by its rendering, shape, presence, multiplicity and connection type. Another special case handled respects the potential self reference of entities.

The *many-to-one* multiplicity type was introduced to enhance modelling semantics in cases, where the relationship direction mismatches the intended foreign key placement. Figure 4 shows a corresponding annotated schema example. We assume, that a person has a `name` and a `gender`, while the latter one is not literally saved, but represented by an `GenderEnumeration` instance. Therefore, a foreign key column is placed in the `Person` table. Using a *one-to-many relationship* forces us to draw the edge from `GenderEnumeration` as source to `Person` as target, but that does not match our wording. We phrased, that *a person has a gender* and not, that *a gender has many persons*, which is indeed also true, but does not match the drawing direction implied by the word order, which causes confusion to no purpose.

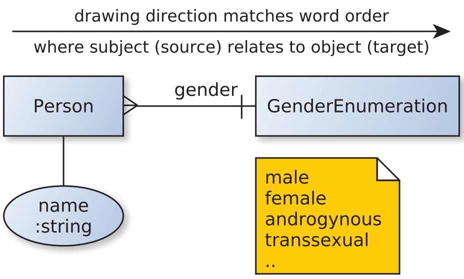



**Figure 4.** Exemplary usage of the many-to-one relationship

The *connection type* can be set to *association*, *aggregation* or *composition*. An association relationship does not demand any existence constraints. If A associates B and A is removed, then B remains. An aggregation relationship demands, that if A and C aggregate B and A is removed, then B remains, because C is still referencing it. If C is also removed, then B will also be destroyed. You can say, that B is *shared* between A and C. A composition relationship however demands, that if A composes B and A is removed, then B is also immediately destroyed. In this case B may not be shared between instances.

This paper will only give examples for association and composition relationships, because they can be directly mapped to ActiveRecord DSL fragments under some limitations. Supporting aggregation requires some more custom logic. It should be made clear, that some combinations of relationship and connection types make no sense and are thus not supported. A many-to-many relationship, for example, will only have the connection type association, because it usually does not (or rather should not) model a whole-part aspect expressed as aggregation or composition.

Relationships have different *presence types* denoting, whether B instances must exist. A one-to-one relationship, for example, is this way specialized to one-to-zero-or-one (B is *optional*) or one-to-exactly-one (B is *required*).

The *rendering type* specifies, whether a relationship is modelled as a simple edge (*implicit*) or as an *explicit* relationship entity (represented by a diamond) carrying further attributes. This latter concept is known from the UML association class.

The *shape type* specifies, whether a relationship allows only one target type (*uniform*) or accepts participating entities of different types (*polymorphic*).

This paper will not discuss every reasonable combination, but will rather present selected examples to reveal the principles.

### *3.2. Model Examples*

Figure 5 shows several 1:1 relationship variations. The upper example models the fact, that one `Citizen` maybe owns one `IdentityCard`. The relationship is expressed by an edge connecting the given entities, while the edge ends take different shapes depending on the desired configuration. In the first case the source end is clean and the target end is zero-or-one, denoted by a small circle and a single vertical line. Thus, we see an *implicit uniform optional one-to-one association relationship.*

The second example is an *implicit uniform required one-to-one association relationship.* The additional single vertical line drawn at the source end is redundant and it is the developer's choice to use it. Besides, role names are introduced at this point.

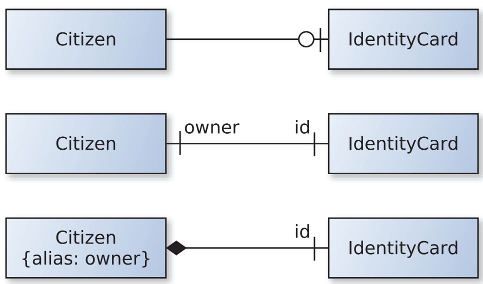



**Figure 5.** Implicit 1:1 relationship variants

The last case models the connection between `Citizen` and `IdentityCard` as an *implicit uniform required one-to-one composition relationship* and introduces the *alias* modifier. This modifier defines an alternative default role name, which is used instead of the real entity name. If one wants to navigate from entity A to entity B, then B is accessible by its role name (in the context of a specific relationship) respectively alias, if given, or by its primary name.

Figure 6 shows a model, which connects planets with its (astronomic) satellites. In the first version an implicit composition relationship is modelled. The second version however uses association and is explicit, since the relationship is actually visible. The `Orbit` relationship is drawn as a diamond and introduces a `distance` attribute. In both cases target ends with circles and tripods indicate, that zero-or-more (optional) connections are allowed. The special *participation edge* is also introduced linking the diamond (`Orbit`) to its finally destiny (`Satellite`). Thus, the second case represents an *explicit uniform optional one-to-many association relationship*.

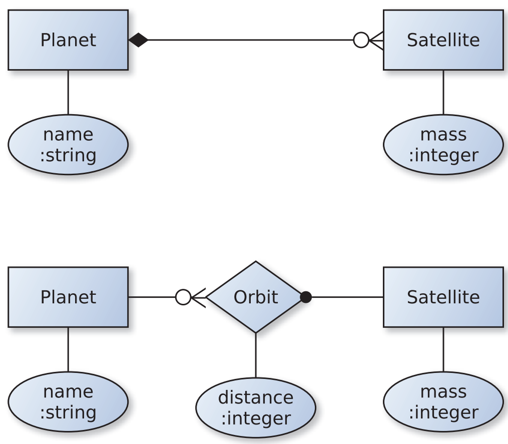



**Figure 6.** Implicit and explicit 1:n relationships

Figure 7 introduces implicit and explicit polymorphism. Both examples model an `Article`, `Gallery` and `Profile` entity, which are connected to the `Image` entity in different ways. The first connection (from left) denotes, that an `Article` instance is allowed to compose zero-or-more `Image` instances when a state is built. If a specific `Article` instance is deleted, then all referenced images are removed, too. The second connection depicts, that a `Gallery` associates at least one `Image` instance without any existence constraints. The last connection is a composition, which states, that a `Profile` instance composes exactly one `Image` instance, which is existence dependent from that specific profile.

Usually these connections exist without interdependencies, that means, that in a potential SQL migration, the `Image` relation would hold 2 foreign key columns for `Article` as well as `Gallery` and the `Profile` would reference `Image`. But as all connections have the same source role name `imageable`, the transformation tool derives an *implicit polymorphic relationship*. Naturally speaking, an image is able to reference an article *or* a gallery *or* a profile, which is realized by an additional type column in the `Image` SQL representation and directly supported by ActiveRecord.

This mechanism is also used in the second example, which introduces an explicit version. Here, the `Has` relationship enables the developer to attach additional information to the specific connection, a `caption` attribute in this case.

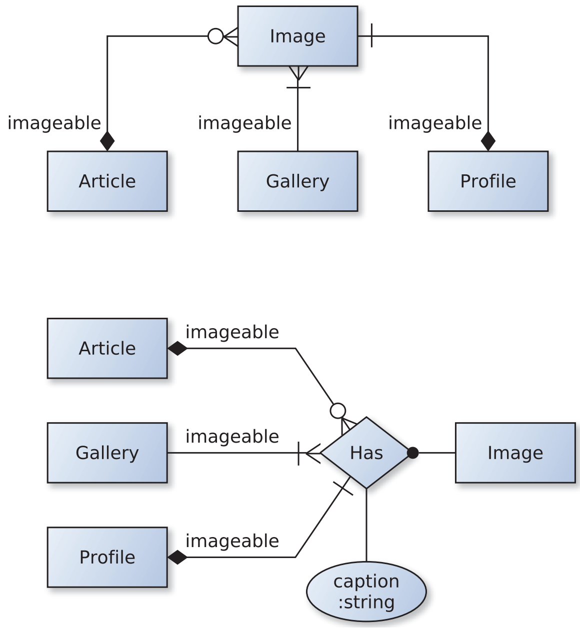



**Figure 7.** Implicit and explicit polymorphic relationships

### *3.3. Rental Example Translation*

The right tools at hand, we can now translate the example introduced in the last paragraph into a graphical representation shown in Figure 8.

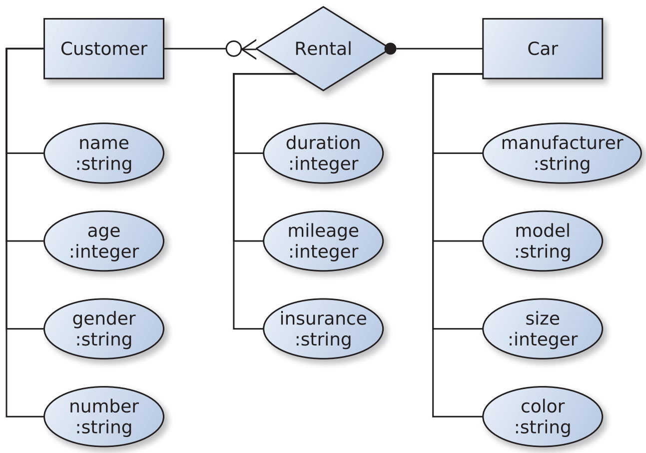



**Figure 8.** Rental example as diagram

This model contains 2 entities named `Customer` and `Car` as well as an *explicit uniform optional one-to-many association relationship* called `Rental`. All types shown are connected with their specific attributes by plain edges.

This model can now serve as input for the transformation tool to generate the migration and model class files considered above – thus completing the circle. By means of a real world example, the next section shows, how the generated code can be used to establish a runtime ActiveRecord data layer. It will also give some query examples.

## 4. A Real World Application

In this chapter we show the everyday usability of the described approach by introducing a domain model representing the political administrative structure of Germany based on the census data provided by the German Federal Statistical Office (Destatis) [2]. This data can be officially obtained on the Destatis web portal and contains one monolithic ASCII file as well as human readable record descriptions required to interpret the provided data line by line.

The different ASCII records contain information about federal states (German: Bundesland), districts (German: Bezirk), counties (Kreis), municipalities associations (Gemeindeverband) as well as municipalities (Gemeinde). Further, the hierarchical connections between those units are given as well as the places of administration. On the lowest level of the municipalities additional geographical position information was collected from public data of the Open Street Map project [1].

The domain model in Figure 9 was derived by interpreting the Destatis record descriptions and by examining examples of the instance data given. It was afterwards transformed into valid ActiveRecord migration and model class representations. The result of this translation can be found in the appendix of this work. This appendix will not be included in later versions of the paper and will be replaced by a reference to an appropriate document on the web. Furthermore, all existing data was extracted from the ASCII file, converted and used to build a corresponding state in an SQLite database.

After a short discussion on the rendered domain model and the ActiveRecord initialization process, we show some example queries formulated in ActiveRecord linguistics.

### *4.1. The Destatis Domain Model*

The model contains elements introduced before, like entities, attributes and *implicit uniform optional* and *required one-to-many association relationships*. Furthermore, it makes use of so-called *Base Entity Types*, which are used to generalize standard entities. In the current example, `BaseUnit` as well as `UnitType` represent abstractions of concrete concepts passing common attributes. This way, entities like `FederalState` or `Municipality` inherit additional features while specializing their conceptual ancestors. Generalization is indicated by an edge between the concerned artefacts ending with a large white arrow and is *virtual*, because base types are not directly transformed into ActiveRecord representations.

The actual model also reveals different attribute variations. The `area` attribute of `Municipality` is drawn with a dashed line indicating, that this attribute is optional. Besides, the `zip_code` attribute shows a double lined border stating, that the value range must be unique. The underlined name and data type express, that this attribute must be indexed on database level, i.e. for faster access.

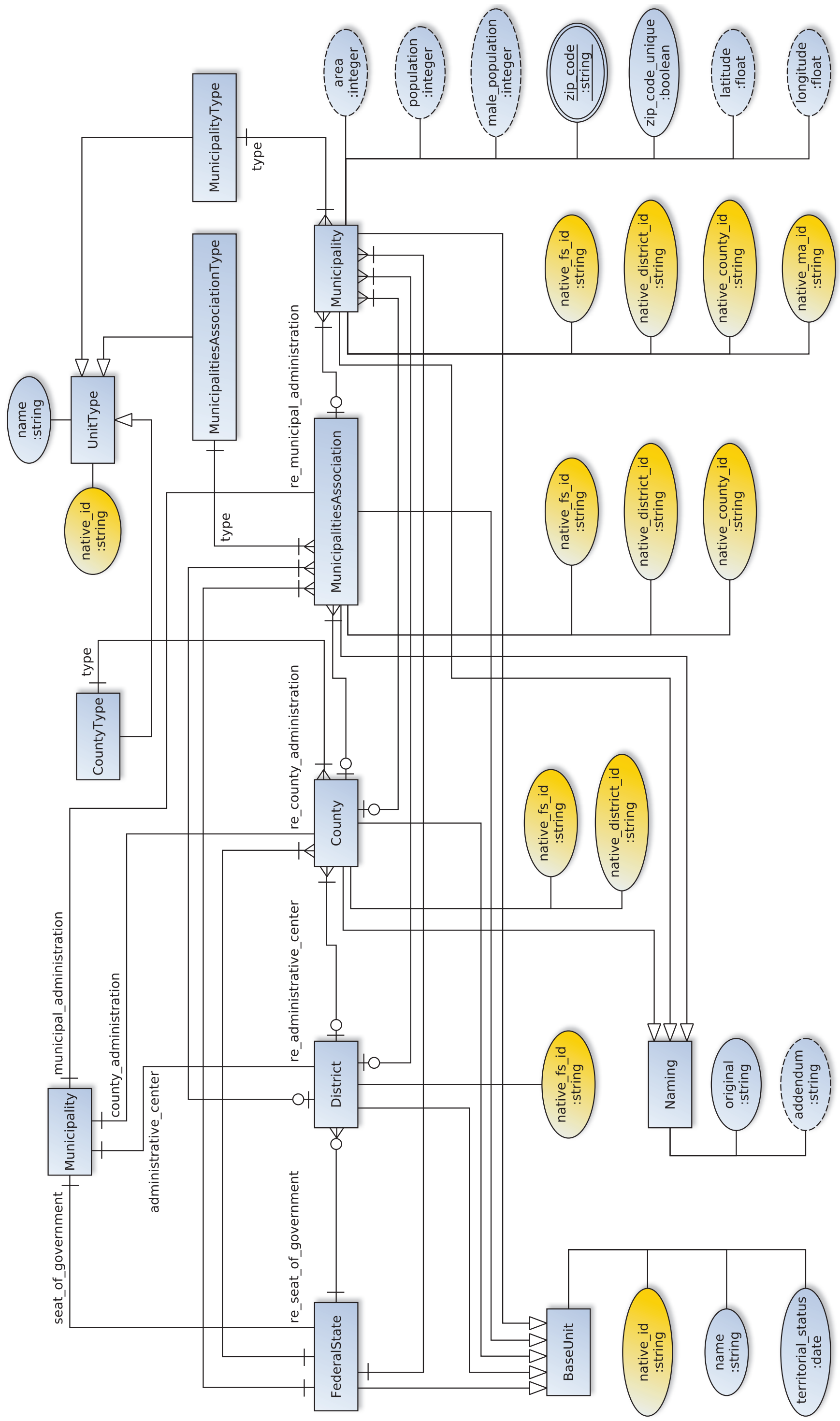



**Figure 9.** Model of Germany's administrative structure

The different attribute coloring has no syntactical meaning. It just highlights those attributes, which represent a reference in the native ASCII dataset and are used to load the original Destatis data.

Another noteworthy point is the double emergence of the `Municipality` entity. The transformation tool allows the multiple reference of artefacts enabling the developer to expand attribute or relationship definitions in several places in aid for clear diagram structuring even across multiple GraphML model files. Heavy connected items can be represented much easier this way as our example demonstrates.

### 4.2. *Initializing the ActiveRecord Layer*

Prior to presenting some example queries, we will shortly describe how the ActiveRecord data layer is made available at Rails runtime (see Listing 8). The transformation tool generates 16 migration and 16 model class files as described before (line 1). Afterwards, the migration files are executed to establish the SQL schema via the Ruby task management and build automation tool *rake* (line 2). Finally, another task is started, which creates the database state according to the given ASCII dataset. This process is called *seeding* and realized by separate Ruby seeding task files, which were also specifically created for this paper.

Listing 8: Rake commands

```
1 $ rake tibet:output
2 $ rake db:migrate
3 $ rake db:seed
```

When the Rails environment is booted, i.e. for a live console or in the context of a web application server instance, the model class files are loaded automatically and establish connections to the SQL database holding the according schema. The transformation tool implemented supports a special mode, which does not generate a textual representation of the model class files at all, but loads them dynamically when the Rails environment emerges. Therefore, it keeps a representation of the model in the memory, which exists during the execution time of the particular Rails process and injects the classes via metaprogramming.

The generated model class files can be customized by creating specifically named Ruby modules. If, for example, the `Municipality` model class has to be extended beyond its basic definition, the developer can create a corresponding Ruby module named `MunicipalityCustom`, which is automatically included. At this point, it is important to mention, that all additional functionality presented in this work is unobtrusive towards the Rails framework, meaning that the framework components are not changed in some way and that only the official Rails API is used. The whole approach is modelled as an additional layer upon the already existing ones.

### 4.3. *Example Queries*

As announced, we give some query examples executed in the Rails development console. We will at first specify each query in natural language and then present the ActiveRecord translation as well as the results. Internally, the so-called *ActiveRecord Query Interface* is used rendering corresponding SQL queries.

The first query takes the first `Municipality` instance and reads the `longitude` attribute value.

Listing 9: Example query 1

```
> Municipality.find(1).longitude
= 9.43333
```

The second example returns the names of all municipalities, whose latitude value ranges from 51 to 51.5 and whose longitude value lies between 10.48 and 10.5.

Listing 10: Example query 2

```
> Municipality.where(latitude: 51..51.5)
  .where(longitude: 10.48..10.5).map(&:name)
= ["Flarchheim", "Kammerforst", "Oppershausen",
   "Vogtei"]
```

Example 3 returns the zip codes of all municipalities lying in all counties belonging to the federal state of Schleswig-Holstein.

Listing 11: Example query 3

```
> FederalState.where(name: 'Schleswig-Holstein')
  .first.counties.map(&:municipalities)
  .flatten.map(&:zip_code)
= ["25770", "21493", "23898", "25917", "25853",
   "25938", "25860", "23730", "23738", "23758", ..]
```

Query 4 sums up the population of all municipalities, which do not belong to a county. These municipalities are linked to a corresponding `MunicipalityType` instance.

Listing 12: Example query 4

```
> MunicipalityType.find(2).municipalities
  .sum(:population)
= 23561333
```

The following query example returns the count of all municipalities belonging to the district Oberbayern, which have a male population between 1000 and 10000 and whose area ranges between 1000 and 2000 units.

Listing 13: Example query 5

```
> District.where(name: 'Oberbayern')
```

```
  .first.municipalities
  .where(male_population: 1000..10000)
  .where(area: 1000..2000).count
= 6
```

And finally, the last query 6 returns the original names of all municipalities starting with the letters “Be”. The original name is the name taken from the Destatis dataset before normalization.

Listing 14: Example query 6

```
> Municipality.where('name LIKE ?', 'Bas%')
  .map(&:original)
= ["Basedow", "Basthorst", "Bassum, Stadt",
   "Basdahl", "Bassenheim", "Basberg",
   "Bastheim", "Basedow", "Bastorf", "Basdorf"]
```

## 5. Related Work

There are several contributions that can be related to our present work. The importance of visual model representation is particularly emphasized in [9], where drawing sketches are used to design complex dynamic systems. Using available graph editors to generate software in modelling environments is also done in [8], where Microsoft Visio is employed. The freely obtainable yEd diagram editor introduced in this work also finds application in [20], where it helps creating schemas for a prototypical graph database. Using yEd as modelling environment is also considered in [14] discussing example-driven meta-model development. And in [12], the GraphML format underlying yEd is employed to describe UML diagrams characterizing existing PHP code.

In [16], ActiveRecord models are transformed to Alloy formal data model specifications and afterwards checked for errors using bounded verification techniques. [15] uses Ruby on Rails to show, that the developed security tool is an effective aid in the implementation of secure web application features like authentication and authorization.

Our approach is the only one that employs an everyday drawing tool for the ER-based design of an information system on the basis of Ruby on Rails and that is validated by a complex case study.

## 6. Conclusion and Future Work

The practical experiences made with the described approach reveals its capability. There is actually no need to define migration and model class files by hand anymore. Further, the developer does not have to cope with the question, how to represent specific artefact connections like relationships on the relational level. This makes establishing a Rails data layer significantly more comfortable and efficient. Given the graphical abstraction of the domain, the developer can immediately start working on specific application code, like controllers, views or additional model class code.

The presented approach satisfies the requirements for comprehensibility as the domain structure becomes immediately receivable. It brings clarity, because artefact definitions can be spread over multiple places, even files, and thus prevents overloaded views due to heavy connected items. Because of yEd's sophisticated drawing interface, developers can easily change the model. In combination with the entirely automated transformation process, especially agile approaches can benefit from the presented work. Consequently, developers save time and routine work.

However, there is certainly some more work to do. Currently, every model evolution causes a complete schema reset. That means, if a model *M1* is changed to model *M2*, all migration files are recreated and the database is rebuilt. Although it is possible to work with that restraint, the transformation tool should adapt to the Rails migration workflow, which allows successive schema evolutions by altering existing versions. It is possible to implement such a feature by calculating model deltas and to create migrations only for the changed parts.

Furthermore, it should be analyzed if using the ActiveRecord model inheritance mechanism brings any advantages compared to the current implementation. And besides the checking of multiplicity constraints and the uniqueness of ActiveRecord collections, one could think about realizing validations based upon OCL expressions [19].

## References


[1] OpenStreetMap (OSM). `http://www.openstreetmap.org/`, accessed: 2015-01-18

[2] Statistisches Bundesamt (Destatis). `https://www.destatis.de/DE/Startseite.html`, accessed: 2015-01-18

[3] Brandes, U., Pich, C.: GraphML Transformation. In: Pach, J. (ed.) Graph Drawing, 12th International Symposium, GD 2004, New York, NY, USA, September 29 - October 2, 2004, Revised Selected Papers. Lecture Notes in Computer Science, vol. 3383, pp. 89–99. Springer (2004), `http://dx.doi.org/10.1007/978-3-540-31843-9_11`

[4] Chen, P.: The Entity-Relationship Model: Toward a Unified View of Data. ACM Transactions on Database Systems 1, 9–36 (1976)

[5] Consel, C.: Domain-Specific Languages: What, Why, How. Electr. Notes Theor. Comput. Sci. 65(3), 1 (2002), `http://dx.doi.org/10.1016/S1571-0661(04)80422-5`

[6] Fowler, M.: Patterns of Enterprise Application Architecture. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2002)

[7] Gottlob, G., Koch, C., Pichler, R.: XPath processing in a nutshell. SIGMOD Record 32(2), 21–27 (2003), `http://doi.acm.org/10.1145/776985.776988`

[8] Horkoff, J., Yu, Y., Yu, E.S.K.: OpenOME: An Open-source Goal and Agent-Oriented Model Drawing and Analysis Tool. In: de Castro, J.B., Franch, X., Mylopoulos, J., Yu, E.S.K. (eds.) Proceedings of the 5$^{th}$ International *i** Workshop 2011, Trento, Italy, August 28-29, 2011. CEUR Workshop Proceedings, vol. 766, pp. 154–156. CEUR-WS.org (2011), `http://ceur-ws.org/Vol-766/paper27.pdf`

[9] van Joolingen, W.R., Bollen, L.: Interactive drawing tools to support modeling of dynamic systems. In: Goldman, S.R., Pellegrino, J., Gomez, K., Lyons, L., Radinsky, J. (eds.) Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences, ICLS '10, Chicago, IL, USA, June 29 - July 2, 2010, Volume 2. pp. 169–171. International Society of the Learning Sciences / ACM DL (2010), `http://dl.acm.org/citation.cfm?id=1854589`

[10] Kreibich, J.A.: Using SQLite - Small. Fast. Reliable. Choose any Three. O'Reilly (2010), `http://www.oreilly.de/catalog/9780596521189/index.html`

[11] Leff, A., Rayfield, J.T.: Web-Application Development Using the Model/View/Controller Design Pattern. In: 5th International Enterprise Distributed Object Computing Conference (EDOC 2001), 4-7 September 2001, Seattle, WA, USA, Proceedings. pp. 118–127. IEEE Computer Society (2001), `http://computer.org/proceedings/edoc/1345/13450118abs.htm`

[12] Lemos, M.: GraphML Generator: Generate UML diagrams from PHP code using GraphML. http://www.phpclasses.org/package/6025-PHP-Generate-UML-diagrams-from-PHP-code-using-GraphML.html, accessed: 2015-01-18

[13] Lima, A., Rossi, L., Musolesi, M.: Coding Together at Scale: GitHub as a Collaborative Social Network. CoRR abs/1407.2535 (2014), http://arxiv.org/abs/1407.2535

[14] Lopez-Ferndandez, J.J., Cuadrado, J.S., Guerra, E., de Lara, J.: Example-Driven Meta-Model Development. Software and System Modeling (2015)

[15] Munetoh, S., Yoshioka, N.: Model-Assisted Access Control Implementation for Code-centric Ruby-on-Rails Web Application Development. In: 2013 International Conference on Availability, Reliability and Security, ARES 2013, Regensburg, Germany, September 2-6, 2013. pp. 350–359. IEEE Computer Society (2013), http://dx.doi.org/10.1109/ARES.2013.47

[16] Nijjar, J., Bultan, T.: Bounded verification of Ruby on Rails data models. In: Dwyer, M.B., Tip, F. (eds.) Proceedings of the 20th International Symposium on Software Testing and Analysis, ISSTA 2011, Toronto, ON, Canada, July 17-21, 2011. pp. 67–77. ACM (2011), http://doi.acm.org/10.1145/2001420.2001429

[17] Obe, R., Hsu, L.: PostgreSQL - Up and Running: a Practical Guide to the Advanced Open Source Database. O'Reilly (2012), http://www.oreilly.de/catalog/9781449326333/index.html

[18] Palla, P., Frau, G., Vargiu, L., Rodriguez-Tomé, P.: QTREDS: a Ruby on Rails-based platform for omics laboratories. BMC Bioinformatics 15(S-1), S13 (2014), http://dx.doi.org/10.1186/1471-2105-15-S1-S13

[19] Richters, M., Gogolla, M.: On Formalizing the UML Object Constraint Language OCL. In: Ling, T.W., Ram, S., Lee, M. (eds.) Conceptual Modeling - ER '98, 17th International Conference on Conceptual Modeling, Singapore, November 16-19, 1998, Proceedings. Lecture Notes in Computer Science, vol. 1507, pp. 449–464. Springer (1998), http://dx.doi.org/10.1007/978-3-540-49524-6_35

[20] Sedlmeier, M., Gogolla, M.: Design and Prototypical Implementation of an Integrated Graph-Based Conceptual Data Model. In: B. Thalheim, H. Jaakkola, Y.K. (ed.) Proceedings of the International Conference on Information Modelling and Knowledge Bases (EJC 2014). KCSS, vol. 2014/4, pp. 376–395. Department of Computer Science, Faculty of Engineering, Kiel University (June 2014), http://ebooks.iospress.nl/volumearticle/38513

[21] Ueno, K., Fukasawa, Y., Morihata, A., Ohori, A.: The Essence of Ruby. In: Garrigue, J. (ed.) Programming Languages and Systems - 12th Asian Symposium, APLAS 2014, Singapore, November 17-19, 2014, Proceedings. Lecture Notes in Computer Science, vol. 8858, pp. 78–98. Springer (2014), http://dx.doi.org/10.1007/978-3-319-12736-1_5

[22] Viswanathan, V.: Rapid Web Application Development: A Ruby on Rails Tutorial. IEEE Software 25(6), 98–106 (2008), http://doi.ieeecomputersociety.org/10.1109/MS.2008.156

[23] Widenius, M., Axmark, D.: MySQL reference manual - documentation from the source. O'Reilly (2002), http://www.oreilly.de/catalog/mysqlref/index.html

## A. Complete Transformation Output of Destatis Model

Please see next pages.

CreateFederalState.rb

```ruby

class CreateFederalStates < ActiveRecord::Migration

  def change
    create_table :federal_states do |t|
      t.string :native_id, null: false
      t.string :name, null: false
      t.date :territorial_status, null: false
      t.references :seat_of_government, null: false
      t.timestamps null: false
    end
  end

end

```

CreateDistrict.rb

```ruby

class CreateDistricts < ActiveRecord::Migration

  def change
    create_table :districts do |t|
      t.string :native_fs_id, null: false
      t.string :native_id, null: false
      t.string :name, null: false
      t.date :territorial_status, null: false
      t.references :federal_state, null: false
      t.references :administrative_center, null: false
      t.timestamps null: false
    end
  end

end

```

CreateCountyType.rb

```ruby

class CreateCountyTypes < ActiveRecord::Migration

  def change
    create_table :county_types do |t|
      t.string :native_id, null: false
      t.string :name, null: false
      t.timestamps null: false
    end
  end

end

```

CreateCounty.rb

```ruby

class CreateCounties < ActiveRecord::Migration

  def change
    create_table :counties do |t|
      t.string :native_fs_id, null: false
      t.string :native_district_id, null: false
      t.string :native_id, null: false
      t.string :name, null: false
      t.date :territorial_status, null: false
      t.string :original, null: false
      t.string :addendum
      t.references :district
      t.references :federal_state, null: false
      t.references :type, null: false
      t.references :county_administration, null: false
      t.timestamps null: false
    end
  end

end
```

CreateMunicipalitiesAssociationType.rb

```ruby

class CreateMunicipalitiesAssociationTypes < ActiveRecord::Migration

  def change
    create_table :municipalities_association_types do |t|
      t.string :native_id, null: false
      t.string :name, null: false
      t.timestamps null: false
    end
  end

end

```

CreateMunicipalitiesAssociation.rb

```ruby

class CreateMunicipalitiesAssociations < ActiveRecord::Migration

  def change
    create_table :municipalities_associations do |t|
      t.string :native_fs_id, null: false
      t.string :native_district_id, null: false
      t.string :native_county_id, null: false
      t.string :native_id, null: false
      t.string :name, null: false
      t.date :territorial_status, null: false
      t.string :original, null: false
      t.string :addendum
      t.references :type, null: false
      t.references :county
      t.references :federal_state, null: false
      t.references :district
      t.references :municipal_administration, null: false
      t.timestamps null: false
    end
  end

end

```

```ruby

class CreateMunicipalityTypes < ActiveRecord::Migration

  def change
    create_table :municipality_types do |t|
      t.string :native_id, null: false
      t.string :name, null: false
      t.timestamps null: false
    end
  end

end

```

CreateMunicipality.rb

```ruby

class CreateMunicipalities < ActiveRecord::Migration

  def change
    create_table :municipalities do |t|
      t.integer :area
      t.integer :population
      t.integer :male_population
      t.string :zip_code, null: false, unique: true, index: true
      t.boolean :zip_code_unique, null: false
      t.string :native_fs_id, null: false
      t.string :native_district_id, null: false
      t.string :native_county_id, null: false
      t.string :native_ma_id, null: false
      t.float :latitude
      t.float :longitude
      t.string :native_id, null: false
      t.string :name, null: false
      t.date :territorial_status, null: false
      t.string :original, null: false
      t.string :addendum
      t.references :type, null: false
      t.references :municipalities_association
      t.references :federal_state, null: false
      t.references :district
      t.references :county
      t.timestamps null: false
    end
  end

end

```

```ruby

class FederalState < ActiveRecord::Base

  belongs_to :seat_of_government,
             class_name: :Municipality

  has_many :municipalities_associations,
           class_name: :MunicipalitiesAssociation,
           inverse_of: :federal_state,
           foreign_key: :federal_state_id

  has_many :counties,
           class_name: :County,
           inverse_of: :federal_state,
           foreign_key: :federal_state_id

  has_many :districts,
           class_name: :District,
           inverse_of: :federal_state,
           foreign_key: :federal_state_id

  has_many :municipalities,
           class_name: :Municipality,
           inverse_of: :federal_state,
           foreign_key: :federal_state_id

  validates_associated :municipalities

  validates_associated :municipalities_associations

  validates_associated :districts

  validates_associated :counties

  validates_presence_of :territorial_status

  validates_presence_of :name

  validates_presence_of :seat_of_government

  validates_presence_of :native_id

end

```

District.rb

```
class District < ActiveRecord::Base

  belongs_to :federal_state,
             class_name: :FederalState

  belongs_to :administrative_center,
             class_name: :Municipality

  has_many :municipalities_associations,
           class_name: :MunicipalitiesAssociation,
           inverse_of: :district,
           foreign_key: :district_id

  has_many :counties,
           class_name: :County,
           inverse_of: :district,
           foreign_key: :district_id

  has_many :municipalities,
           class_name: :Municipality,
           inverse_of: :district,
           foreign_key: :district_id

  validates_associated :municipalities

  validates_associated :municipalities_associations

  validates_associated :counties

  validates_presence_of :federal_state

  validates_presence_of :territorial_status

  validates_presence_of :name

  validates_presence_of :native_id

  validates_presence_of :administrative_center

  validates_presence_of :native_fs_id

end
```

CountyType.rb

```ruby

class CountyType < ActiveRecord::Base

  has_many :counties,
           class_name: :County,
           inverse_of: :type,
           foreign_key: :type_id

  validates_associated :counties

  validates_presence_of :name

  validates_presence_of :native_id

end

```

```ruby

class County < ActiveRecord::Base

  belongs_to :district,
             class_name: :District

  belongs_to :federal_state,
             class_name: :FederalState

  belongs_to :type,
             class_name: :CountyType

  belongs_to :county_administration,
             class_name: :Municipality

  has_many :municipalities_associations,
           class_name: :MunicipalitiesAssociation,
           inverse_of: :county,
           foreign_key: :county_id

  has_many :municipalities,
           class_name: :Municipality,
           inverse_of: :county,
           foreign_key: :county_id

  validates_associated :municipalities

  validates_associated :municipalities_associations

  validates_presence_of :district

  validates_presence_of :name

  validates_presence_of :territorial_status

  validates_presence_of :original

  validates_presence_of :native_id

  validates_presence_of :federal_state

  validates_presence_of :type

  validates_presence_of :native_district_id

  validates_presence_of :county_administration

  validates_presence_of :native_fs_id

end

```

MunicipalitiesAssociationType.rb

```ruby

class MunicipalitiesAssociationType < ActiveRecord::Base

  has_many :municipalities_associations,
           class_name: :MunicipalitiesAssociation,
           inverse_of: :type,
           foreign_key: :type_id

  validates_associated :municipalities_associations

  validates_presence_of :name

  validates_presence_of :native_id

end

```

MunicipalitiesAssociation.rb

```ruby

class MunicipalitiesAssociation < ActiveRecord::Base

  belongs_to :type,
             class_name: :MunicipalitiesAssociationType

  belongs_to :county,
             class_name: :County

  belongs_to :federal_state,
             class_name: :FederalState

  belongs_to :municipal_administration,
             class_name: :Municipality

  belongs_to :district,
             class_name: :District

  has_many :municipalities,
           class_name: :Municipality,
           inverse_of: :municipalities_association,
           foreign_key: :municipalities_association_id

  validates_associated :municipalities

  validates_presence_of :original

  validates_presence_of :native_county_id

  validates_presence_of :native_id

  validates_presence_of :name

  validates_presence_of :territorial_status

  validates_presence_of :native_district_id

  validates_presence_of :type

  validates_presence_of :county

  validates_presence_of :federal_state

  validates_presence_of :district

  validates_presence_of :native_fs_id

  validates_presence_of :municipal_administration

end

```

MunicipalityType.rb

```ruby

class MunicipalityType < ActiveRecord::Base

  has_many :municipalities,
           class_name: :Municipality,
           inverse_of: :type,
           foreign_key: :type_id

  validates_associated :municipalities

  validates_presence_of :name

  validates_presence_of :native_id

end

```

Municipality.rb

```ruby

class Municipality < ActiveRecord::Base

  include MunicipalityCustom

  belongs_to :type,
             class_name: :MunicipalityType

  belongs_to :municipalities_association,
             class_name: :MunicipalitiesAssociation

  belongs_to :district,
             class_name: :District

  belongs_to :federal_state,
             class_name: :FederalState

  belongs_to :county,
             class_name: :County

  has_one :re_municipal_administration,
          class_name: :MunicipalitiesAssociation,
          inverse_of: :municipal_administration,
          foreign_key: :municipal_administration_id

  has_one :re_administrative_center,
          class_name: :District,
          inverse_of: :administrative_center,
          foreign_key: :administrative_center_id

  has_one :re_seat_of_government,
          class_name: :FederalState,
          inverse_of: :seat_of_government,
          foreign_key: :seat_of_government_id

  has_one :re_county_administration,
          class_name: :County,
          inverse_of: :county_administration,
          foreign_key: :county_administration_id

  validates_associated :re_county_administration

  validates_associated :re_seat_of_government

  validates_associated :re_administrative_center

  validates_associated :re_municipal_administration

  validates_numericality_of :area,
                            :allow_nil => :true,
                            :allow_blank => :true,
                            :only_integer => :true

  validates_numericality_of :male_population,
                            :allow_nil => :true,
                            :allow_blank => :true,
                            :only_integer => :true

  validates_numericality_of :population,
                            :allow_nil => :true,
                            :allow_blank => :true,
```

```ruby
                                :only_integer => :true

  validates_numericality_of :longitude,
                                :allow_nil => :true,
                                :allow_blank => :true

  validates_numericality_of :latitude,
                                :allow_nil => :true,
                                :allow_blank => :true

  validates_presence_of :native_id

  validates_presence_of :native_county_id

  validates_presence_of :native_district_id

  validates_presence_of :native_ma_id

  validates_presence_of :name

  validates_presence_of :territorial_status

  validates_presence_of :original

  validates_presence_of :type

  validates_presence_of :district

  validates_presence_of :native_fs_id

  validates_presence_of :zip_code_unique

  validates_presence_of :county

  validates_presence_of :zip_code

  validates_presence_of :federal_state

  validates_presence_of :municipalities_association

  validates_uniqueness_of :zip_code

end
```

# Culture-Aware Web Information System Development

Hannu JAAKKOLA [a,1] and Bernhard THALHEIM [b,2]

[a] *Tampere University of Technology, P.O.Box 300, FI-28101 Pori, Finland*
[b] *Christian-Albrechts-University Kiel, Computer Science Institute, 24098 Kiel, Germany*

**Abstract.** Globalisation has a strong impact on information systems both from a development and usage point of view. Development is done by geographically distributed teams, with team members representing different cultures. In addition, the information systems are targeted at widening markets to get new clients and business - either as they are or as localized variations. In the case of traditional information systems, the needs related to globalisation are more or less manageable, because the clients are at least to some extent known, as well as their needs and preferences. Cloud technology and cloud-based solutions are replacing traditional information systems at an accelerating speed. In some cases, the question is only one change to an execution platform, but more often also opening up the information system usage to a "faceless" mass of users over the Internet - the information system (IS) becomes an SaaS-based WIS (Web Information System). In addition to SaaS-based solutions, the WIS category covers a wide variety of web services available in a more or less open manner; as a consequence the ability to react to the needs of a multi-cultural set of users causes new challenges to the developers and service implementation. Our paper initiates discussion on the challenges related to a WIS in a multi-cultural context. The complexity is structured by recognition of six concerns (viewpoints), which are handled in an interrelated manner. The foundation is built by analysis of cultural differences, which are used to clarify and explain the culture-based differences in information system structure and usage. Culture-related aspects of storyboards and database schemas are dealt with and evidence is derived from selected existing information systems.

**Keywords.** Information systems, web information systems, culture-aware systems, adaptation of systems to users, culture, adapted culture, extended culture, multicultural, culture stereotypes, storyboards, adaptive data base structure

## 1. Introduction

Web information systems are nowadays widely used and are becoming one of the major realisations of information systems. E-business, edutainment, infotainment, community, and identity web systems are data and information intensive websites. They integrate a variety of database, workflow and other processing, communication and presentation

[1] hannu.jaakkola@tut.fi
[2] thalheim@is.informatik.uni-kiel.de

systems. Due to their worldwide usage, web information systems have to take different cultures into consideration.

At the same time, systems development has changed from general-purpose systems toward user-oriented systems. User-oriented web systems currently mainly reflect the local culture of a group of users. Worldwide applications must however also support a large variety of cultures, as mentioned above. Moreover, the classical approach to technical systems is no longer appropriate. Therefore, we must distinguish between the socio-cultural and the technical view for such systems. Since these systems are becoming more complex, we have to separate concerns and be able to derive requirements for the systems.

Modern web information systems reflect at least six different concerns: intention, usage, content, functionality, context, and presentation.

Modern web information systems reflect at least six different concerns: intention, usage, content, functionality, context, and presentation.

**Intention:** The intention aspect is a very general one, centered around a mission statement for the system. The primary question is: what is the purpose of the system? Which users will use the system? Which skills and capabilities can be expected? What kind of support is needed? Users typically have their own specific behavior.

**Usage and the resulting story space:** Once some clarity with respect to the intentions of the web-based system has been obtained, the question arises of how the system will be used and by whom? As web-based systems are open systems, it is important to anticipate the behaviour of the users. We model usage through storyboards that allow the specification of the stories that users might use with the web information system. The story space is the combination of all the stories supported.

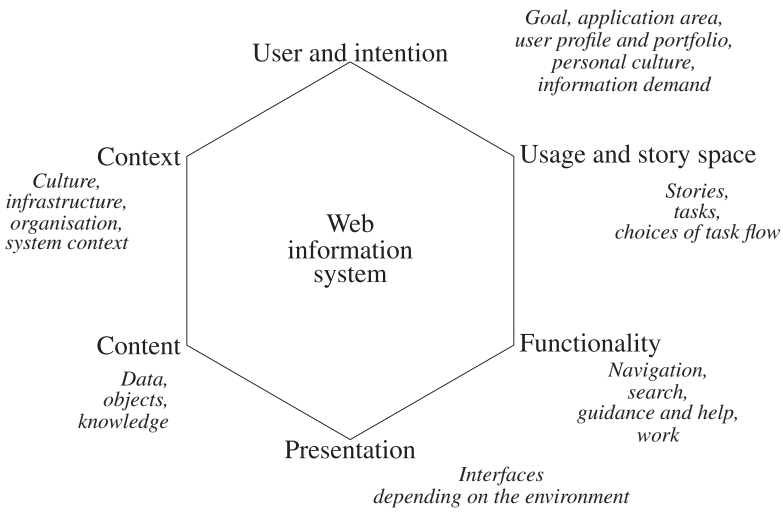



**Figure 1.** The six concerns for web information system and the conceptual facilities

**Content:** The content aspect concerns the question: Which information should be provided to which user at which stage of the work? It is coupled with the problem of designing an adequate database. However, the organization of the data presented to the user via a website is significantly different from the organization of data in a database.

**Functionality:** The functionality aspect is coupled with the question of whether the site should be passive or active. A passive site only allows a user to navigate through the pages without any action or to search for corresponding content. It can be combined with guidance and help support. In an active site, input is also required from the user. Specific functions allow the processing of user input and the provision of features such as searching, printing, marking, and extraction.

**Context:** The context aspect deals with the context of the web information system with respect to society, time, expected users, the history of utilisation, and the paths of these users through the system. One major element of the context is culture in all its peculiarities, e.g. national and regional, organisational, website provider, and work environment cultures.

**Presentation:** The presentation aspect concerns the final realisation in web pages. The presentation can be separated into layout and playout. It follows principles of screenography [19] and is based on principles of visuality such as visual communication, visual cognition and visual design [34]. Design also depends on the support of technical end-devices such as computer screens, television, cell phones, etc., and the set layout preferences.

The concerns can directly be represented by separate conceptual structures. For instance, the intention dimension can be specified on the basis of life cases [27] and solutions to them. It is typically combined with a user model that has a profile describing the user properties and with a portfolio of tasks that the user has to complete. Users can be stereotyped. The stereotype must then reflect the specific personal culture of a group of users. Figure 1 displays the concerns with their specific elements.

The adaptation possibilities discussed in [23] are: (1) information density, (2) navigation, (3) accessibility of functions, (4) guidance, (5) structure, (6) colourfulness, (7) saturation, and (8) support. Information density is a special property of content; functionality must reflect (2), (3) (4), and (8); presentation also considers (5), (6), and (7).

Our separation into six concerns goes beyond these adaptation facilities. Depending on the given culture and on the given application, we can develop specific stereotypes that appropriately support the user and can be refined depending on the chosen system environment. The following section develops these stereotypes.

We can now develop a framework for culture adaptation depending on the six concerns in Figure 1. In our first paper [11], we developed an adaptation approach based on the development of specific culture-dependent profiles and portfolio of users. Based on these profiles and portfolio, a number of adaptation facilities were derived for the user interface. The adaptation of storyboards and content can now be tackled. Figure 2 depicts the plan for this paper. The adaptation of functionality to cultures is the topic of a forthcoming paper.

The paper is structured in the following way. Section 2 introduces our approach to the concept of culture; two commonly used frameworks - the Hofstede model and the Lewis model are discussed. Section 3 extends this discussion with elements of cultural stereotypes; new concepts of adapted culture and extended culture are presented, based on the work of Reinecke. Section 4 introduces an approach to generic and culture-adaptive storyboarding based on the mini-story approach [5,36]. Section 5 shows that

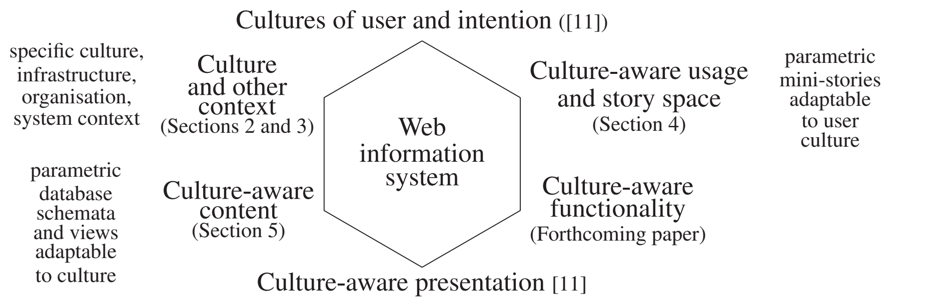



**Figure 2.** Our research plan

database technology is fit for adaptation of content to the user culture. We introduce parametric database schemata and generalise the solutions presented in [1,11,17,18,33]. Section 6 concludes the paper.

## 2. Culture Models

Hofstede (2010) defines the term culture as "a collective phenomenon, which is shared with people who live or lived within the same social environment, which is where it was learned; culture consists of the unwritten rules of the social game; it is the collective programming of the mind that separates the member of one group or category of people from others." Hofstede's definition refers to national cultures, which he sees as a layered (pyramid) structure including inherited and learned properties finally adapted in the level of personality into the individual behaviour. Values (of a national culture) form the foundation for a national culture, having culture-dependent practices as a manifestation. The "collective programming of the mind" in this definition indicates the individual's learning capacity, in which the topics learned are organized according to the inherited part of the "program". Human beings are learning intensive and adaptable, which means that ultimately their culture consists of elements recorded from lifeline experiences having their source in regional culture (in addition to national culture), organizational cultures (e.g. based on work experience), professional culture (based on education) etc. The learning aspect is also seen also in Reinecke's studies [23]. She defines the term "*user's extended national culture*" as storage of the influence of foreign cultures on the national one, when a person is living in a foreign culture. These experiences change and enrich the original culture. In addition, the national culture also adapts to external influence. The heart remains the same but the outer layers are adaptive.

The paper [11] reports that national culture dominates the behavior of individuals: language, educational tradition, religion, beliefs, attitudes, and social context are important parts of cultural identity. In addition, different levels of organization-related cultures (organization, work, team, project, site), professional culture, regional culture etc. represent collective similarities of groups of people having the same national inheritance and provide the means for variations inside a national culture. These may cover similarities in behavior, interaction, decision-making, organizational structure, and goals; people represent different roles more than themselves, which are parallel from an individual's point of view.

The most common approach in analyzing differences between national cultures is based on cultural stereotypes. Stereotypes[3] provide generalized knowledge of factors related to national cultures. The frameworks most commonly used and referred to were published by Geert Hofstede [6,9,10,8] and Richard Lewis [13]. An additional approach supporting our goals is related to information system usage and has been handled by Reinecke (& Bernstein) [21,23] in several papers. These are introduced briefly in sub-section 3.2.

## 3. Culture and Other Contexts

### *3.1. Cultural stereotypes*

Cultural stereotypes define patterns typical to national cultures. The Hofstede model uses six dimensions (indices) to define cultural behavior (adapted from [11]):

- Power Distance (PDI): the extent to which power differences are accepted
- Individualism / Collectivism (IDV): the extent to which a society emphasizes the individual or the group
- Masculinity / Femininity (MAS): refers to the general values in the society - hard / soft values
- Uncertainty avoidance (UAI): refers to the extent that individuals in a culture are comfortable (or uncomfortable) with unstructured situations
- Long-term / Short term orientation (LTO): refers to the extent to which the delayed gratification of material, social, and emotional needs is accepted.
- Indulgence / Restraint (IVR): acceptance of enjoying life and having fun vs. controlling life by strict social norms.

Every national culture has an index value for each dimension, which indicates its tendency in this dimension. Comparison of index values between cultures provides means of identifying the need for adaptation in a collaboration context. Hofstede's resource pages [7] provide a tool for the comparison of selected national cultures. The idea to provide easy to notice differences between selected cultures. Hofstede's work is applied in several studies to indicate people's preferences in a certain context; e.g. Reinecke's work applies it in the user interface structure and the look-and-feel of it (discussed later in this sub-section). In our work Hofstede's dimensions are used in the context of Web Information Systems analysis. The culture model of Trompenaars [35] is a variation of Hofstede. Therefore, this model is not discussed in detail.

Richard Lewis is a business consultant who has long (since 1950s) experience of training organizations for successful interaction in a global context. He has synthesized his experience in a triangular model [13,15] that classifies national cultures in three basic categories. A *linear-active culture* is task-oriented and value is given to technical competence. They are cool, factual and decisive planners. A *multi-active culture* is extrovert and human force is seen as an inspirational factor. They are warm, emotional, loquacious and impulsive. A *reactive culture* is people-oriented and dominated by knowledge, pa-

[3] Stereotypes should satisfy at least five properties: (1) they must be accurate; (2) the quality of the stereotype allows it to be used consciously; (3) they should be descriptive, not evaluative; (4) they should be flexible so that they can be modified from time to time; (5) they can be used as a first "best guess".

**Table 1.** Characteristics of cultural stereotypes according to Lewis' model, applied in WIS context

| Linear-active | Multi-Active | Reactive |
|---|---|---|
| introvert | extrovert | introvert |
| patient | impatient | patient |
| plans ahead | plans grand outline | looks at general principles |
| does one thing at a time | does several things at once | reacts |
| punctual | not punctual | punctual |
| compartmentalises activities | one activity influences another | sees the whole picture |
| sticks to plans | changes plans | makes slight changes |
| sticks to facts | juggles facts | statements are promises |
| gets information from official sources | prefers oral information | information from official and oral sources |
| follows correct procedures | pulls strings | networks |
| completes action changes | completes human transactions | react to partners |
| likes fixed agendas | interrelates everything | thoughtful |
| uses memos | rarely writes memos | plans slowly |
| dislikes losing face | has ready excuses | must not lose face |

tience and silent control. They are courteous, amiable, accommodating, compromisers, and good listeners. National cultures representing clean basic categories are located on the angles of the triangle and cultures having mixed influence of two basic behaviors are on the sides (distance from the corner indicates its weight in the mix).

Lewis has collected the common traits of the three basic categories [13](pp. 33-34). Some (those relevant for our purposes) of these characteristics are shown in Table 1.

Lewis' original list indicates differences between the basic culture categories according to thirty overall properties. In his detailed model, Lewis [13,15] introduces national cultures using five main factors: general facts (geography, history, politics and economy), culture (general classification, values, cultural black holes, concept of time, concept of space, self-image), communication (communication pattern, body language, listening habits, audience expectation), and interaction (concept of status, gender issues, leadership, management, motivation factors, meetings, negotiating, contracts & commitments, manners and taboos, how to empathise).

Reinecke ([20]; & Bernstein [21,23]) has studied user interface preferences in a multi-cultural context. Her research problem focuses on the structure and look-and-feel of the user interface in web-based systems. The test tool (MOCCA) is a web interface that recognizes a variety of (task and user interface related) dimensions (information density, navigation, and accessibility of functions, guidance, structure, colors, saturation, and support). MOCCA builds a user model based on the user's demographic data. It covers national culture-based knowledge of the user's national (birth) culture according to Hofstede's five first dimensions, parameters to assign changes to it as a consequence of time spent under the influence of a foreign culture, reference to the parents' culture (used to take parents' dominance into account), and a set of other parameters (birth year, political orientation, social structure, religion, educational level, and computer literacy).

The look-and-feel and functionality of the information system are adapted according to the user model parameter values. MOCCA provides 115,000 different culture-dependent user interface versions to follow the presumed expectations of the user. The "user's extended national culture" is a changed and enriched modification of the culture by birth; a model is used to recalculate the parameter values of the user model as a consequence of external influence; we use the term "*adapted culture*" for this modified culture, because it is not a question only of extending the cultural base, but also "overwriting" (replacing), deleting, and modifying it.

MOCCA has been validated by multi-cultural users; some of whom represented their original national cultures and some had an international career, covering periods of living in several countries. The results are reported in a journal article by Reinecke & Bernstein [23]. The goal was to compare the system's user model and the users' real behavior in task execution. The results were promising (better than chance), but not completely satisfactory. The prediction accuracy of the model improved when learning from users' majority choices was added to the prediction. This indicates the importance of feedback, which forms a kind of "usage context". General findings cover the following, for example: high computer literacy does not supersede cultural preferences and cultural exchange does not lead to an adoption of foreign values – instead it seemed to enhance one's own cultural identity; parents' nationality does not have major importance in prediction accuracy.

### *3.2. Cultural Stereotypes and their Utilisation for System Development*

Sub-section 3.1 introduced our approach to use cultural stereotypes to support culture-aware web information system development. The stereotype models (Hofstede, Lewis, and others) partially overlap and indicate the same behavior from different viewpoints. Reinecke's work combines user interface design with Hofstede's dimensions. Certain user interface properties and interaction process features have a Hofstede index value; the layout of the web page, color selection, guiding the interaction process etc. are carefully bound in the interface properties (Table 2).

Reinecke's original table is based on collected information from several sources. It also seems to include one minor error in the PDI part, in which the last items are in the wrong order (not relevant for our table and not included in it). In MOCCA the central role is played by the information system's understanding of the user in the terms of user model. The user interface adaptation is based on the best fit between the user model and interface content / interaction structure. This is worth generalizing and being used in other contexts, which is one of the aims of our paper. The work of Lewis is not recognized by Reinecke. Our aim is to combine his ideas in our analysis also.

### *3.3. An Example: Culture-Aware Search*

Search support is one of the central facilities of each web information system. It is currently based either on *query-answer* approaches for a *property-based* search [2] or based on a keyword search. *Keyword-based* search is more culture-dependent than is often acknowledged. According to ReturnOnNow [24] statistics, Google is most popular search engine in 25 out of 31 countries; the exceptions are Czech Republic (Seznam), China (Baidu; partially because of political restrictions), Hong Kong and Japan (Yahoo), Russia (Yandex) and South Korea (Naver). The same statistics one year later [25] showed that the Czech Republic had also turned to Google. The most recent statistics [28] show that currently in Western and Middle Europe there are no alternatives to Google and also in other non-Google countries Google is winning marketshare: Czech Republic (Seznam:

**Table 2.** Relationships between Hofstede Dimensions and UI Design (modified of Reinecke [22,23])

| Index | Low | High |
|---|---|---|
| **PDI** | · access variations, non-linear navigation | · linear navigation, few links, minimal number of alternatives |
| | · non-structured data allowed | · structured data expected |
| | · most information at interface level, deep hierarchy not accepted | · data in hierarchy, little data on top level |
| | · friendly messages | · strict messages |
| | · need for support is low | · support must be available |
| **IDV** | · traditional colors and images | · colors used to encode information |
| | · image to text direction | · text to image direction |
| | · high multimodality | · low multimodality |
| | · colorful expression | · monotonous color map |
| **MAS** | · little saturation, pastel colors | · high contrast, bright colors |
| | · exploration and different navigation paths allowed | · restricted navigation possibilities |
| | · personal content presentation and friendly communication | · encouraging words in communication |
| **UAI** | · most information at interface level, complex interfaces | · organize information hierarchically |
| | · nonlinear navigation accepted | · linear navigation paths, guided |
| | · colors, typography and sound to maximize information | · redundancy and cues to avoid ambiguity |
| **LTO** | · reduced information density | · most information at interface level |
| | · content structured into small units | · content can be arranged around a focal area |

approx. 24%) [29], Hong Kong (Google: > 70%, Yahoo: approx. 20%), Japan (Google - 64%, Yahoo - 32%) [30]. Even in Russia Google is outperforming Yandex [31]. Simultaneously, however, the role of Google is decreasing in USA in its most significant market [32].

The simple straightforward American kind Google look-and-feel is accepted by nationalities having wide diversity in their cultures; on the other hand, some Asian cultures prefer a more informative, colorful, and complex front page structure.

In spite of using search engine selection as an example of culture-related preferences, we understand that there are a lot of other arguments on the background, not only look-and-feel of the user interface. These statistics are also the basis of the research question discussed in this paper: *How to make information systems adapt to the cultural differences of the users?*

## 4. Culture-Aware Storyboards

We have now developed a framework and a realization strategy for the automatic adaptation of storyboards to users. Mini-stories are the main constructive component of storyboards. A storyboard is a composition of mini-stories. They are generalized into parametric mini-stories. These parametric stories can be adapted to suit different cultures.

### 4.1. Mini-Stories

On a high level of abstraction the storyboard of a web information system specifies who will be using the system, in which way and for which goals. Storyboard pragmatics deals with the question of what the storyboard means for its users. One part of pragmatics is concerned with usage analysis by means of life cases, user models and contexts. We also addressed another part of pragmatics that complements usage analysis by WIS portfolios.

These comprise two parts: the information portfolio and the utilization portfolio. The former is concerned with the information consumed and produced by the WIS users, which leads to content chunks. The latter captures functionality requirements, which depend on the specific category the WIS belongs to.

The storyboard language we are going to use extends SiteLang [3] with the explicit introduction of mini-stories. The conceptual model of storyboarding (see e.g. [26]) takes this up by providing an integrated model comprising the story space capturing the stories and the plot, actors, and tasks. Inspired by approaches in theatre and film, the story space comprises scenes and the actions in these scenes, and the plot describes the details of the action scheme. Furthermore, the model describes the actors in these scenes, i.e. groups of users, which leads to roles, profiles, goals, preferences, obligations, and rights. The actors are linked to the story space by means of tasks.

A *mini-story* [5,36] typically captures a small, self-contained, tightly connected set of scenes similar to a movie clip. It may be characterised through a (semantic) word field [4]. In B2C e-business, typical mini-stories are (1) advertise, (2) quote, (3) request, (4) response, (5) select and collect, (6) bargain, (7) contract, (8) requisition and order, (9) deliver, (10) invoice, (11) pay, and (12) return. Mini-stories have an input, an output, associated actors within certain roles and plays, controls for their playout, features to be used within the story, and conditions for their usage.

### 4.2. Parametric Mini-Stories

Similar to situations in applications, tasks may be specified on the basis of a general description of a possible way for satisfaction and completion. These activities may vary in their style and flow of scenes or activities. In traditional approaches, an activity is specified supported by its own mini-story, e.g. a contract according to a specific set of regulations, a contract in a private form or depending on the parties, historical contracts, contract proposals, a contract on the basis of other contracts, contracts according to company styles, a contract as an extension, etc. The traditional approach leads to a huge number of mini-stories that have a similar behavior.

We will extend mini-stories in order to become adaptable. The mini-story is independent of the parties involved, the application domain, the ends, the supporting means, the style and the pattern of applying the actions, the spacial and temporal context as well as other contexts, the content and the functionality to be used, the concrete intention or target of its utilisation, its benefits and values in applications, and the utilisation constraints. The dependence is however expressed through an instantiation of corresponding parameters.

Parametric actions can also be annotated by verb word fields that declare what kind of action is going to be used and what are the object and valencies of objects within this action. Typical general parameters are: *wherefore, whereof, wherewith, worthiness, why,* *whereto, for when, for what reason, by whom, to whom, whatever, wherein, where, for what, wherefrom, whence, what, how, whereat, whereabouts, whither, when, why, what properties, what scenario, which restrictions*. These parameters generalise the rhetoric frame (who, what, when, where, why, in what way, by what means) that can be traced back to Hermagoras of Temnos [2]. These general parameters form an extensible *parameter space*.

### 4.3. Adaptation of Parametric Mini-Stories to Culture of Users

Parametric mini-stories can be adapted to the current situation, to the current user, to the systems that might be used, to the flow of parametric actions that seem to be appropriate, to the environment that is going to be used, to the styles and pattern, and to the conditions within an application. Other contexts are typically cultural contexts, e.g. the layered structure [11,12] (universal human nature, culture that is specific for a group, personality that has been inherited or learned) and multidimensional aspects of cultures [11] (national, regional, or specific cultures versus organisational, project, team cultures).

The Hofstede [10] and Trompenaars [35] dimensions govern the instantiation of parameters within the parameter space. Here we follow the research on user interfaces (e.g. [23], Table 2) and presentation adaptation in [11] and generalise these results to stories.

(I) People with a high PDI score (or power close people) are used to being driven. Therefore, the story may also be the driving force. Linear stories are better acknowledged. In cultures with low PDI values (or better "power distant people" [4]), people prefer story spaces with many opportunities. Stories must not be entirely structured.
(II) People with a high Individualism score also accept interleaved and multimodal stories. Less individual and more collective people prefer collaborative stories. At the same time, multimodality is lower.
(III) Masculinity also accepts jumps in stories. Decisions must however be clear. Femininity requires more smooth stories with clear steps. Some decisions may also be delayed.
(IV) Uncertainty avoiding people need a preview for the next actions and following mini-stories. Stories should not be interwoven. So, we derive a story tree from our mini-story. Stories should be unambiguous. In contrast, uncertainty tolerant people also accept story forests, i.e. interwoven, concurrent story trees that might reuse parts of other stories.
(V) Long-term orientation requires that the storyboards developed must be made stable before delivery to the user. They should not often be changed. Short-term oriented people prefer shorter stories. So, mini-stories can kept separate without integration into larger ones.
(VI) Indulgence allows integration of storyboards with others. Conservative or restrained people cope better with a flow of events they are used to.

The Lewis [11,13,14] separation into linear, reactive and multi-active cultures guides the adaptation rules and the parameters as well. The linear-active culture is better supported by a sequential story flow whereas the multi-active culture allows multi-tasking, and parallel flows of action. Stories for reactive people are better given in a fully-fledged form with zoom facilities for current activities.

[4] In Hofstede's terminology: low score in the power distance index.

## *4.4. Case Study of Stories*

S. Liermann [16] made an analysis of popular, commercially successful and well-accepted e-commerce portals in China, Finland, Germany, Japan, Spain, and the US. She analysed checkout procedures for portals such as Hobbyhall.fi and Rakuten.com[5].

Typical mini-stories in a business-to-customer e-commerce setting are

- *advertise and quote goods and services in various representations, variations, and collections;*
- *request and response based on communication and procurement procedures;*
- *select and collect choice with various versions;*
- *bargain and contract with dependence on conditions and business profiles of companies;*
- *requisition and order depending on the choices the customer made with a number of sub-mini-stories such as*
  - ∗ *checkout including completion and payment;*
- *deliver and invoice depending on the logistics or the company;*
- *and potentially return goods.*

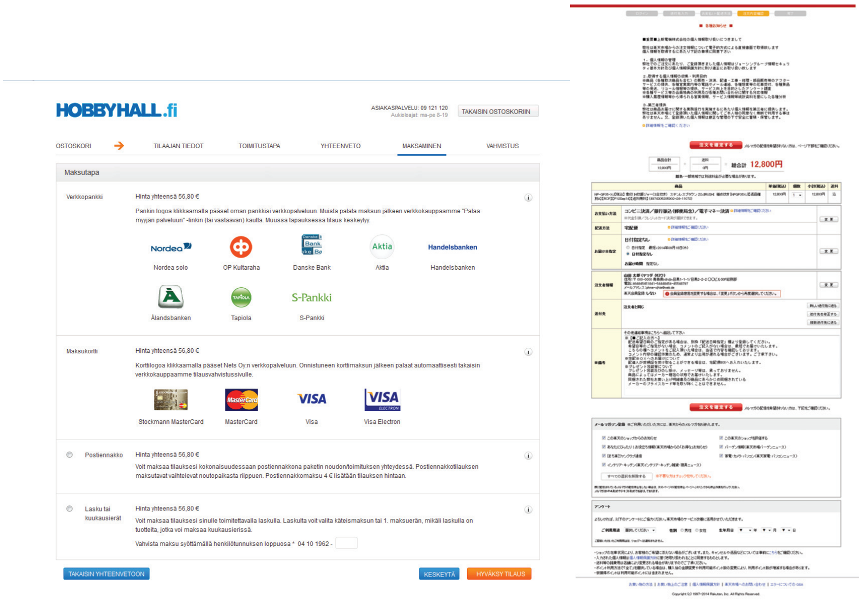


**Figure 3.** Two typical checkouts in Finland (Hobbyhall) and Japan (Rakuten) websites

The checkout procedures in Figure 3 represent typical behaviour in the two countries. We can model these checkout procedures directly using mini-stories in Figures 4

[5]The study included all concerns for websites in detail. In the forthcoming paper we shall generalise the findings and demonstrate the adaptation of parametric functionality in dependence of the given culture.

and 5. The storyboard representation is based on the website specification language SiteLang [3,4,26]. A mini-story is composed of scenes. Each scene belongs to a general activity. The usage of a scene in a mini-story is controlled by pre-conditions for entering the scene, by accepting conditions for leaving the scene, and by events that opens the scene. The transitions among scenes are represented by curves. The story space thus consists of scenes, their control, actors either involved in the story space or enabled for specific scenes, context, tasks, and transitions among scenes according to the specification space that has been introduced in Figure 1.

We represent basic scenes here with ellipses. Each of these scenes might be represented by a web page. Often web pages combine a number of scenes. This combination results in large pages that require repeated scrolling back and forth. The user must then memorise the parts that are not visible on the current screen.

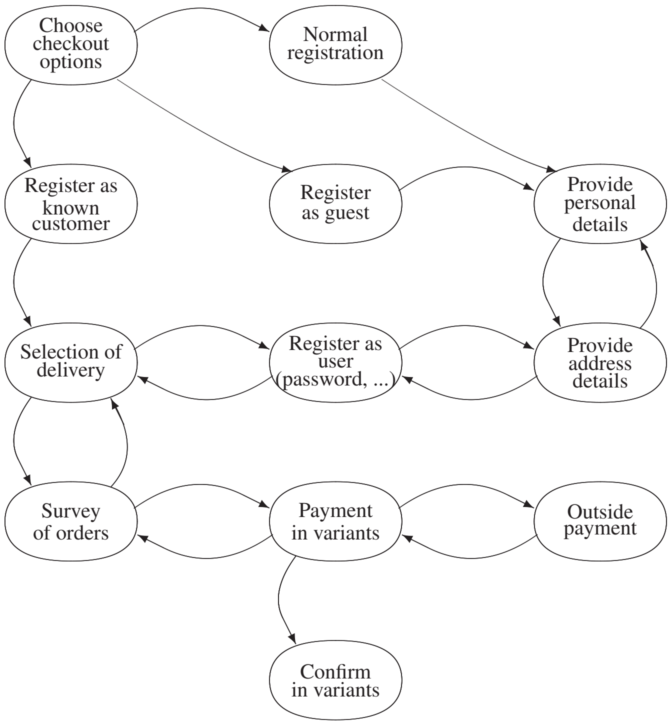



**Figure 4.** Mini-story for checkout in Finland (see Figure 3)

A mini-story must often represent the large variety of usage that users prefer. Therefore, mini-stories are far more flexible than the workflows to which the stories are mapped. The concrete deployment of a web information system by users is a view of this story space, i.e. a *scenario* or path in this space for the given user by abstracting from those scenes for which the user is enabled. The workflow specification may use a suite of such scenarios. Each scenario is then supported by one workflow.

The mini-story in Figure 4 results in a good number of such scenarios. Each of the potential paths from the choice of checkout options towards the confirmation would have been supported by one workflow. Here the potential and capacity of our approach can be understood. Our proposed solution is economical and simple. Website behavior is modelled on the basis of parametric mini-stories. The adaptation procedure for parametric mini-stories can be mapped to an adaptation procedure for generic workflows.

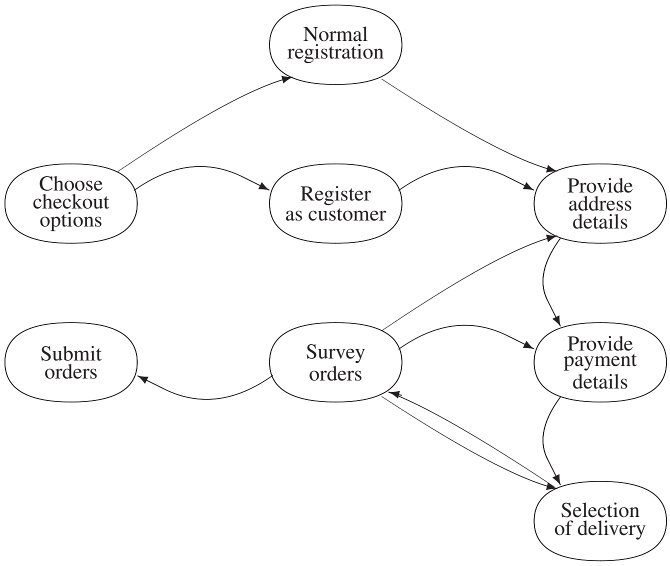



**Figure 5.** Mini-story for checkout in Japan (see Figure 3)

These mini-stories vary a lot depending on the culture. Let us compare two mini-stories for Finland and Japan in Figures 4 and 5. The two mini-stories provide an inside into the different cultures. According to [8], Finish people are used to a flat and hierarchical organization, are individual-oriented, prefer soft values, are more used to unorganised situations, and are interested in short-term results. They are linear-active and data-oriented. They are used to the linear execution of actions and prefer independence. Older generations of Finnish people also possess reactive tendencies, especially in the area of communication. Time is an essential issue in Finland. Japanese people are used to hierarchy and bureaucracy, know the importance of grouping, prefer hard values, are organized, and long-term oriented. They are more oriented towards multitasking and prioritisation. Japanese culture is reactive and a listening culture. There is no guest registration for the analysed portal.

The communication habits are partially similar (silence, listening, minimal speech) and partially different (ways to find a conclusion, to harmonise). The differences in revising actions are based on these communication habits. The importance of payment information varies. Payment systems vary from country to country.

## 5. Culture-Aware Content in Web Information Systems

We have developed a framework and an implementation strategy for the automatic adaptation of content to user cultures. Classically information systems are built on top of a database system that is structurally specified on the basis of a conceptual, logical, or physical database schema. Views provide a facility for derivation of user-specific data. We have generalized this approach by introducing parametric database schemata based on a pattern from [1] and by developing general views. These views can be adapted to the user culture through generation of specific views for each culture stereotype. The concept of view towers [11] has already been used for the generation of interfaces [11].

### *5.1. Parametric Database Schemata*

Database schemata can be represented conceptually by diagrams in an extended entity-relationship model [33]. However, these schemata reflect some specific viewpoint and some culture. They are dependent on assumptions, paradigms, and postulations that drive the development of such diagrams. The linear-active culture is neatly supported by development tactics such as Salami slice or class-oriented schemata. The reactive culture is better supported by schemata that reflect all aspects of a given thing from reality, e.g. by XML schemata (networks). The multi-active culture, however, is not reflected at all. In order to support this culture, a large number of views must be developed. Therefore, we need a more flexible way to represent a database structure.

A *parametric database schema* consists of a collection of structure patterns together in a composition that binds some of the parameters of the pattern to each other. The patterns are parameterized. Parametrisation also uses a parameter space similar to the one for mini-stories. The composition of patterns follows the approach developed in [17,18]. In general, it is an open database schema with parameters that can be adapted according to the specific form of the web information system.

A structure pattern [1] consists of the name of the pattern $N$, the schema of the pattern $S$, deployment conditions $\Psi$, *parametric* integrity constraints $\Sigma$, the parameters $p_i$ of the pattern with their pre- and post-conditions $\gamma_i, \delta_i$, the controls $\mathfrak{C}$ for all functions that can be applied to the pattern, and the supports $\mathfrak{U}$ for all functions that can be applied to the pattern, i.e.

$\mathcal{SP} \; = \; (N, S, \Psi, \Sigma, \mathfrak{P}, \mathfrak{C}, \mathfrak{U}) \quad .$

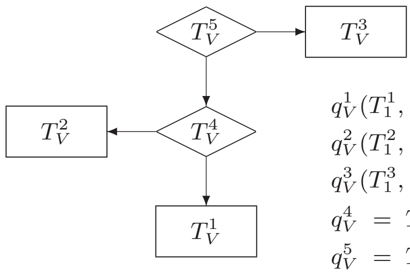
We can now develop a number of *stereotypes for database schemata*:

(a) strictly hierarchical (ER-like) database schemata,
(b) schemata with local viewpoints that reflect the needs of some stakeholders (local-as-view approach),
(c) variants of XML-schemata, Bachman diagrams,
(d) sets of local database schemata with the requirement that the corresponding database schemata is simply the union of the set (global-as-view based on local viewpoints),
(e) sets of personalised views based on local database schemata with some kind of coherence constraint among all views (rigid global-as-view) etc.

This list of schema stereotypes is not exhaustive. It demonstrates, however, that different schemata can be given for the same application. Which database schema is appropriate depends on the utilization of the database schema and on the community of practice and its culture.

### 5.2. Parametric Views

Similar to parametric database schemata, parametric views can be derived. The set of views can form a view tower [11]. Views are schemata on their own[6], e.g. like the view in Figure 6.




$$
\begin{aligned}
&q_V^1(T_1^1, ..., T_{k_1}^1)\\
&q_V^2(T_1^2, ..., T_{k_2}^2)\\
&q_V^3(T_1^3, ..., T_{k_3}^3))\\
&q_V^4 \;=\; T_V^1 \bowtie T_V^2\\
&q_V^5 \;=\; T_V^4 \bowtie T_V^3
\end{aligned}
$$

**Figure 6.** An abstract example of a view in the higher-order entity-relationship model with five types $T_V^i$ that are defined five queries $q_V^i$ which use base types $T_1^1, ..., T_{k_1}^1, T_1^2, ..., T_{k_2}^2, T_1^3, ..., T_{k_3}^3$ where the two relationship types are defined through joins of their component types

We use the algebra of the higher-order entity-relationship model [33] for view specification, i.e. union, difference, intersection, projection, join, selection, renaming, filter, aggregation, variation for presentation, data presentation (sorting, grouping, etc.), nesting, and unnesting operations. Additionally [11], views are associated

- with actions and users (e.g. for authorisation),
- with enforcement control for modification of data through views, and
- with data maintenance policies (either virtual views as a specific computation formula with query expansion techniques or as materialized views with an enforcement policy for cases where data in the base database tables are modified).

A view inherits the parametrisation of the types that are used in their query expressions. There is no need to unify the parameters since these parameters are identifiable by the query that uses them.

### 5.3. Content Schemata

A *content schema* of a web information system consists of a database schema and a collection of views. Views can be enhanced by functionality. The database schema and the views are typically bound by data enforcement strategies. Traditional enforcement is based on a rigid local-as-virtual-view approach, i.e. data for views are virtual and generated only when needed. Data warehouses follow the partial local-as-materialized-view approach, i.e. some of the views are materialized and will be changed whenever the (global) database updates correspondingly. We may also use the global-as-view approach where the local databases are the master and the global database is the slave that follows the local ones.

[6]Traditional object-relational approaches only support singleton table views. In this case we must define a complex view as a collection of views that are associated through integrity constraints - mainly (pairwise) (generalised) inclusion constraints.

A content schema is *integrated* with parametric mini-stories as each mini-story action only consumes data according to the content schema (e.g. one of its views) and produces only data that can be used for a modification of the content database (i.e. either for one of its updateable views or for the database itself).

### *5.4. Adaptation of Content Schemata to the Culture of the Users*

Database schema stereotypes may be classified in a form similar to the classification of database schemata discussed above ((a)-(e)). Cultures are stereotyped as a combination of [10,13,14,35]. The kind of diagram used also depends on the educational and organisational cultures within the community of practice.

Database schema stereotypes can be associated with cultural stereotypes, as shown in Table 3. We do not intend to present all possible aspects but use it as an example. Our table is based on [23] (and Table 2) but concentrates on structures instead of interfaces. The adaptation mechanism that can be used for this flexibility uses the database schema translation approach [33]. We start with a database schema and general views. The translation rules depend on the stereotype for the content schema.

**Table 3.** Cultural stereotypes, kinds of database schemata that are potentially preferred, and potentially useful database schema stereotypes

| **Cultural stereotype** | **Preferences** | **Schema** |
|---|---|---|
| High Power Distance | completely specified and well-formed, easy to understand and persistent database schema | (a) |
| Low Power Distance | freely configurable database schemata that are adaptable to current needs and preferences | (d) |
| Individualism | my own database schema according to my and only my preferences (work profile, education profile, personality profile, security profile) | (e) |
| Collectivism | commonly agreed database schema reflecting all elements within a group according to the collaboration style | (b) |
| Masculinity | restriction to essential elements and only those, strict structuring | (a) |
| Femininity | schema with additional and optional elements, with exploration opportunities, personalised schemata | (e) |
| Uncertainty avoidance | complete schema with all elements, hierarchical structuring, more linear, well-scoped sub-schemata with simple reference to main schema | (a),(d) |
| Uncertainty tolerance | extensible schema, flexible schema style, web-like schemata | (c),(e) |
| Long-term culture | all potential elements are reflected as well as all viewpoints, focused (oil stain) schemata | (a), (b) |
| Short-term culture | handy schemata depending on current use and their smooth integration, decomposable schemata | (e) |
| Indulgence | schema with a central part containing all necessary elements and further elements that might be of use in future | (e),(c) |
| Restraint | puritanical schemata without any non-essential elements | (a) |
| Linear-active culture | schemata with step-wise exploration of all its aspects | (b) |
| Multi-active culture | different variants of the global schema for parallel integrated work | (d),(c) |
| Reactive culture | completely fledged schemata with all details and views for later work | (d) |

## 6. Conclusion

The paper dealt with the challenges related to information systems used by a wide variety of users representing different cultural backgrounds. The complexity of the problem is structured by a model of six concerns, which also dictates the structure of the paper. We have used an analysis based on well-known cultural stereotype models, which are applied in user interface design and the functionality of information systems. The application of these models in interaction and information management principles provides new means in the field - to be used both to guide developers and to analyze the quality of existing web information systems; understanding the challenges is the first step towards better information systems. The work has connections to our earlier studies (references included), but also to plans for future work (also specified in the paper). We are grateful for the support for our research work (see Acknowledgements); it has provided us with an opportunity to collaborate as a multi-cultural team with experience of information system development in different cultural contexts. This experience is built into our paper, as well as in our future work.

## References


[1] B. AlBdaiwi, R. Noack, and B. Thalheim. Pattern-based conceptual data modelling. In *Information Modelling and Knowledge Bases*, volume XXVI of *Frontiers in Artificial Intelligence and Applications, 272*, pages 1–20. IOS Press, 2014.

[2] M. Berg, A. Düsterhöft, and B. Thalheim. Query and answer forms for sophisticated database interfaces. In *Information Modelling and Knowledge Bases*, volume XXIV, pages 161–180. IOS Press, 2013.

[3] A. Düsterhöft and B. Thalheim. Conceptual modeling of internet sites. In *Proc. ER'01, LNCS 2224*, pages 179–192. Springer, 2001.

[4] A. Düsterhöft and B. Thalheim. Integrating retrieval functionality in websites based on storyboard design and word fields. In *NLDB'2002*, volume 2553 of *LNCS*, pages 52–63. Springer, 2002.

[5] F. Förster and B. Thalheim. An effectual approach for a data and information management for humanists. *Qualitative and Quantitative Methods in Libraries (QQML)*, (2):121–128, 2012.

[6] G. Hofstede. Cultural dimensions. http://www.geert-hofstede.com, 2003.

[7] G. Hofstede. Geert hofstede resource pages. http://www.geert-hofstede.com, Retrieved December 4th, 2014.

[8] G. Hofstede. Cultural dimensions - WWW. http://www.geert-hofstede.com, Retrieved November 20th, 2013.

[9] G. Hofstede and G.J. Hofstede. *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. McGraw-Hill, New York, 2004.

[10] G. Hofstede, G.J. Hofstede, and M. Minkow. *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. McGraw-Hill, New York, 2010.

[11] H. Jaakkola and B. Thalheim. Multicultural adaptive systems. In *Information Modelling and Knowledge Bases*, volume XXVI of *Frontiers in Artificial Intelligence and Applications, 272*, pages 172–191. IOS Press, 2014.

[12] W.R. King. A research agenda for the relationships between culture and knowledge management. *Knowledge and Process Management*, 14(3):226 – 236, 2007.

[13] R.D. Lewis. *When Cultures Collide. Managing Successfully Across Cultures*. Nicholas Brealey, London, 3rd edition, 2011.

[14] R.D. Lewis. Richard Lewis resource pages - Cross-culture. http://www.crossculture.com/services/cross-culture/ & http://www.cultureactive.com, 2013. Retrieved November 20th, 2013.

[15] R.D. Lewis. Richard Lewis resource pages - Cross-culture. http://www.crossculture.com/services/cross-culture/ & http://www.cultureactive.com, 2014. Retrieved December 4th, 2014.

[16] S. Liermann. Kulturabhängige E-Commerce Scenarios. Bachelor thesis, Christian-Albrechts-Universität zu Kiel, Department of Computer Science, Kiel, September 2014.

[17] Hiu Ma, R. Noack, K.-D. Schewe, and B. Thalheim. Using meta-structures in database design. *Informatica*, 34:387–403, 2010.

[18] Hui Ma, K.-D. Schewe, and B. Thalheim. Modelling and maintenance of very large database schemata using meta-structures. In *UNISCON*, volume 20 of *Lecture Notes in Business Information Processing*, pages 17–28. Springer, 2009.

[19] T. Moritz, R. Noack, K.-D. Schewe, and B. Thalheim. Intention-driven screenography. In *ISTA 2007*, volume LNI 107, pages 128 – 139, 2007.

[20] K. Reinecke. Automatic adaptation of user interfaces to cultural preferences. *it - Information Technology*, 54(2):96–101, 2012.

[21] K. Reinecke and A. Bernstein. Tell me where you've lived, and i'll tell you what you like: Adapting interfaces to cultural preferences. In *User Modeling, Adaptation, and Personalization (UMAP)*, volume 5535 of *Lecture Notes in Computer Science*, pages 185–196. Springer, 2009.

[22] K. Reinecke and A. Bernstein. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. *ACM Transaction on Computer-Human Interaction*, 18(2):8, 2011.

[23] K. Reinecke and A. Bernstein. Knowing what a user likes: A design science approach to interfaces that automatically adapt to the culture. *MIS Quarterly*, 37(2):427–453, 2013.

[24] ReturnOnNow. Search engine market share by country. http://returnonnow.com/2012/06/search-engine-market-share-country/, 2012. Retrieved in April 4, 2015.

[25] ReturnOnNow. Search engine market share by country. http://returnonnow.com/internet-marketing-resources/2013-search-engine-market-share-by-country/, 2013. Retrieved April 4th, 2015.

[26] K.-D. Schewe and B. Thalheim. Conceptual modelling of web information systems. *Data and Knowledge Engineering*, 54:147–188, 2005.

[27] K.-D. Schewe and B. Thalheim. Life cases: A kernel element for web information systems engineering. In *WEBIST 2007, LNBIP 8*, pages 139–156, Berlin, 2008. Springer.

[28] StatCounter. Search engine market share by country. http://gs.statcounter.com/#search_engine-eu-monthly-201402-201502, 2015. Retrieved April 4th, 2015.

[29] StatCounter. Search engine market share by country: Czech. http://gs.statcounter.com/#search_engine-CZ-monthly-201402-201502, 2015. Retrieved April 4th, 2015.

[30] StatCounter. Search engine market share by country: Japan. http://gs.statcounter.com/#search_engine-JP-monthly-201402-201502, 2015. Retrieved April 4th, 2015.

[31] StatCounter. Search engine market share by country: Russia. http://gs.statcounter.com/#search_engine-RU-monthly-201402-201502, 2015. Retrieved April 4th, 2015.

[32] StatCounter. Search engine market share by country: Us. http://gs.statcounter.com/yahoo-gains-further, 2015. Retrieved April 4th, 2015.

[33] B. Thalheim. *Entity-relationship modeling – Foundations of database technology*. Springer, Berlin, 2000.

[34] B. Thalheim. Syntax, semantics and pragmatics of conceptual modelling. In *NLDB*, volume 7337 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2012.

[35] F. Trompenaars and C. Hampden-Turner. *Riding the waves of culture*. McGraw-Hill, New York, 1998.

[36] M. Tropmann and B. Thalheim. Mini story composition for generic workflows in support of disaster management. In *DEXA 2013*, pages 36–40. IEEE Computer Society, 2013.

**Acknowledgment.** We would like to thank the Academy of Finland and the German Academic Exchange Service (DAAD) for their support of this research.
We are very grateful to our reviewers for their detailed, fruitful, and challenging proposals, remarks, and requests.

# Global Communication with Icons: Hotel Safety as an Environmental Context

Anneli HEIMBÜRGER[a1] and Sukanya KHANOM[b]
*anneli.a.heimburger@jyu.fi*
*University of Jyväskylä*
*[a]Faculty of Information Technology*
*[b]Department of Mathematical Information Technology*
*Finland*

**Abstract.** Icons are small signs with fixed meanings. Icons are usually context specific. For example in the context of a hotel, the client can often find icons in hotel room books and safety guides. Scandic Hotel chain, for example, currently provides the manual for its safety system in 14 languages. There are at least two major shortcomings of this system: (1) in emergency or panic situations, it is very difficult to find your own language from the leaflet, and (2) there are no Asian languages. There is an obvious need for a global icon-based hotel safety language. In our paper, we introduce an icon-based model, language and mobile application for hotel safety.

**Keywords.** Global communication, icons, environmental context, hotel safety.

## Introduction

The focus of our study is on visual communication. Visual communication is one of the oldest disciplines in the humanities. It began with stick drawings on the walls of caves. We are living in cultural spaces: Japanese in Japanese cultural space and Finns in Finnish cultural space. In these spaces people face different kinds of situations and pictorial symbols or icons in their everyday life and environments.

Icons are small and isolated signs with fixed meanings and very simple graphics. The collections of icons are usually context specific. For example in the context of a hotel, you can often find icons in hotel room books. However, there is an obvious need for a global icon-based hotel safety language. Scandic Hotel, for example, currently provides the manual for its safety system in 14 languages. There are at least two major shortcomings of this system: (1) in emergency or panic situations, it is very difficult to find your own language from the leaflet, and (2) there are no Asian languages. In our paper, we introduce an icon-based model, language and mobile application for hotel safety. This study applies the design science research approach.

This paper is organized as follows. In Section 1, we discuss iconic information, and in Section 2 we provide some examples of environmental icons. The main phases of the design science research approach and the modelling icon design process are described in Section 3. An icon-based model, language and mobile application for hotel

---

[1]Corresponding Author: anneli.a.heimburger@jyu.fi

safety are introduced in Section 4. Section 5 is reserved for conclusions and a discussion of issues requiring further study.

## 1. Iconic Information

Icons are small images that perceptually resemble a particular function or condition, and that are used as main components for nonverbal communication [1, 2]. The appropriate use of icons can mitigate complexity and encourage users to interact with a system in easy and effective ways [3]. Principally, icons are self-explanatory [4, 5], depending upon the representation, the intended meaning, and the referred objects. As an example, Figure 1 illustrates an environmental context with the following components:

- **Representation:** an arbitrary icon (a crossed-out red circle) and a concrete icon (a cigarette).
- **Object:** the cigarette icon generally reflects the real world object, and the crossed-out red circle semantically refers to the "prohibited" meaning.
- **Interpretation:** "no smoking" is the intended meaning.

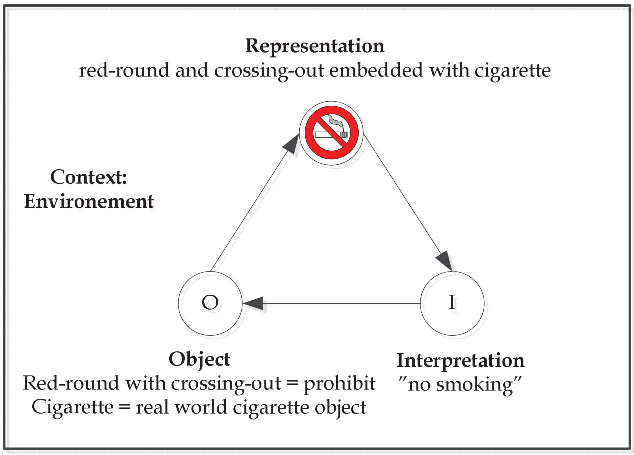



**Figure 1.** Components relevant to an icons' interpretation.

Icons can typically be classified into three genres depending on their styles and usages [6, 7]: concrete, abstract and arbitrary.

Concrete icons are also known as direct representational icons. Their design explores visual similarity between an icon and its reference. For instance, the camera icon in Figure 2 does not include each component of a real camera, but rather it carries the relevant qualities: the assembly shape and structure of a camera. Using this design style to reference real-world objects has the advantage of being easily perceived, learned and taught.

Abstract icons, also known as indirect representations, attempt to illustrate a concept or an idea that is underlined upon the prosperity of icons to convey the intended meaning. Unlike concrete icons, indirect representational icons explore semantic relationships between an icon and its reference. The clacking glass in Figure 2 is an example of an abstract icon. The intention of this style is to communicate the concept of “fragile”, not to directly represent a glass object. The understanding of abstract icons tends to be more difficult than that of concrete ones.

Neither a concrete nor an abstract icon, arbitrary icons are established by a social convention. The meanings of these icons are the most complex and must be learned (for example traffic signs). The meanings of these icons are somewhat complicated to discover from context alone. The radiation warning icon shown in Figure 2 is a common example of an arbitrary icon. Normally, it is designed using geometric shapes and colours to help ease the learning process and allow rapid recognition.

| Concrete | Abstract | Arbitrary |
|---|---|---|
|  |  |  |

**Figure 2.** Examples of concrete, abstract and arbitrary icons.

## 2. Environmental Icons

In our physical environments, icons are often used in a wide variety of ways to inform people about particular conditions. Examples of common icon uses are healthcare facilities, crisis situations, and public places [5, 8, 9, 10].

In the healthcare context, icons have been introduced with the goal of allowing reliable and accurate interactions between users and the system in cases of emergency [2, 11]. Comprehension of medical instructions can often be particularly challenging for patients who have reading disabilities [12]. Communication through icons may also help physicians or nurses avoid critical errors, especially in multi-cultural environments [2]. A medical icon-based language offers a set of medical icons that can be used to identify signs, diseases, physiological states, life habits, drugs and tests [13]. Examples of the icons that are typically used in healthcare are provided in Figure 3.

**Figure 3.** Examples of icons in the medical field.

Icons have also been used successfully in crisis management as a form of auxiliary communication among various users [13, 14, 15, 16]. For instance, in a disaster management system, stress is frequently placed on tackling the shortcomings caused by communication breakdowns. In addition to many other advantages, icons permit people to communicate with each other in a relatively uncomplicated way. In research conducted by Fitrianie et al. [13, 14], the authors devised iconic communication tools to represent concepts and functions in crisis environments. These tools were generally composed of geometric figures, such as arrows, lines, ellipses, rectangles, triangles or icon strings. An example of icons developed for use in crisis management [13, 14] is presented in Figure 4. These icons reflect the situations they represent. The right hand figure shows a way in which iconic sentences are generated and converted into natural language sentences based on the semantic context.

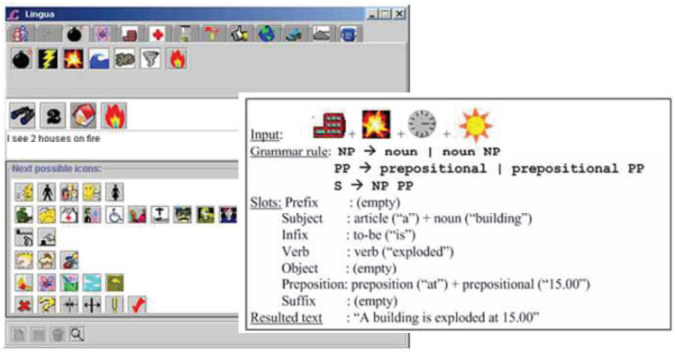



**Figure 4.** Iconic interface and natural language sentence in crisis situations according to Fitrianie et al.

In particular, icons are widely used in public places [5, 9, 17] to express specific meanings. Icons in Figure 5 have been designed to reinforce comprehensibility for users based on their surroundings. For instance, concrete equipment, gestures, and circumstances are used to design icons in recreation areas, whereas abstract icons are used to represent places (e.g., using the letter "P" to represent parking, and using an anchor to represent marine areas). These icons represent a simple language stemming from the need to communicate among people with different backgrounds.


**Figure 5.** Examples of icons in public places.

When seeking an icon to convey a message, it is plausible that the representation is not distinctively unique to one culture. Likewise, when individual reviewers understand icons, they do so personally; each reviewer interprets icons based on his or her own culture, knowledge, and familiarity with the icons. In the context of traffic signs, different countries tend to design their icons based upon their cultural preferences (Table 1).

**Table 1.** Cultural characteristics of icons for traffic signs in Finland, Japan and Thailand [18, 19, 20].

| Meaning | Finland | Japan | Thailand |
|---|---|---|---|
| No entry for bicycles |  |  |  |
| Stop |   |   |   |
| Be aware of animal |  |  |  |

## 3. Modelling the Icon Design Process

We have applied the design science research (DSR) approach [21] to develop and evaluate iconic communication for hotel safety environments. We have first reviewed the state-of-the-art in relevant fields to gain insightful ideas on the utilization of icons in this environmental context. The findings have shown that icons have been adapted for a wide range of areas because they provide an intuitive and effective means of communication, which allows people to interact with each other in a comparatively uncomplicated way. In order to propose a possible solution, we have to define a concrete objective. In this study, the intention is to build an icon design process that can be used as the protocol for the creation of new icons.

At the design and build stage, we infer the requirements for our artefact by following theoretical foundations in iconic communication, human-computer interaction and cognitive science. We further combine knowledge and techniques from the research field of information modelling.

Based on the existing theoretical assessment, our proposal must be tested to confirm whether it can be applied in practice. Usability measurement, including satisfaction and aesthetics, is an outstanding way to justify a new concept. The final activity involves contributing new knowledge obtained by this study to the body of scientific evidence. In our paper we are reporting the three first main phases of the DSR process: environment, design and build artefact. The first iteration of the evaluation phase is under construction for assessing and developing the proposed approach and model.

The design of icons may appear to be the esoteric work of graphic artists or designers; however, there are some guidelines that provide basic structures to the design process [8, 9]. Figure 6 illustrates a model for icon design process, which can be divided into three distinct steps or phases: defining an object to depict, rendering the design and creating an icon, and testing the icon.

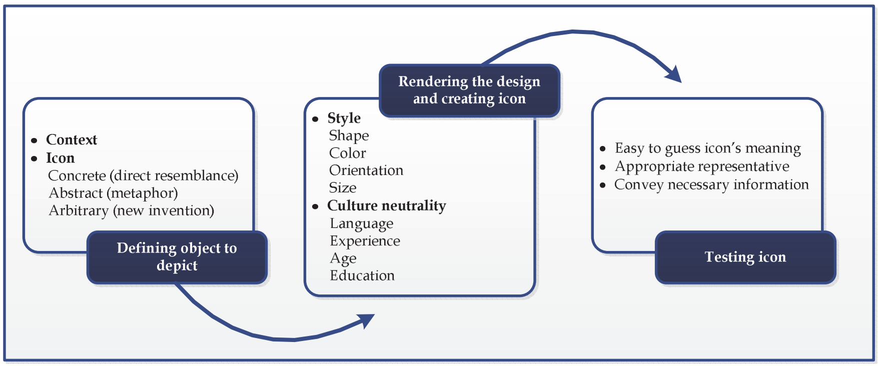



**Figure 6.** Icon design process.

Beginning with defining an object to be depicted by icons, context definition is essential to ensure that icons will be successfully interpreted. In a broad sense, the context is the message or boundary—what is to be communicated. Ideally, the context must be initially defined in an explicit manner that can later be sufficiently depicted by an icon. Next, the icon type—concrete, abstract or arbitrary—must be determined to depict the defined context.

The last and most difficult phase of this process is the design and rendering of icons. Self-evident icons rely on their style and individual culture. In this regard, there are four chief elements that must be considered and countered during the design process: shape, colour, orientation and size. Colour can be used to highlight information that could further improve discernibility. The dominance of shape plays a particular role in discriminating between different iconic constructions. Orientation is more visually expressive when the icon is either too complex or designed with a high density. The size of the icon is another essential factor in determining whether recognition will be clear enough. To make icons more universal, cultural aspects have to be taken into account during the design process [22, 23]. Cultural preference may include influences from a users' background, including spoken language, experience, age and education.

During the icon design process, a critical concern is the question of how to validate whether the icons, both their design and usage, are reasonable to distribute and use. Three characteristics have been proposed to test new icons. (1) Reviewers should be able to guess the intended meaning of icons on the first encounter. (2) Context or messages should be appropriately affected by the resembling icons. (3) Icons should be designed to contain adequate information to convey the intended message, not too overwhelming or too simple.

## 4. Icon-Based Language for Hotel Safety

Creating a standard icon-based language for hotel safety and security guidelines may ensure positive outcomes by helping in emergency situations or preventing an accident/crime before it occurs. Associated stakeholders such as travel agencies or detectives emphasize to offer safety lessons that are likely to protect traveller's property [24, 25, 26, 27]. Nevertheless, many end-users of hotel safety systems encounter usability problems such as learnability and flexibility. For example, the Scandic Hotel safety system [28] provides its safety manual in 14 languages (Figure 7). This may provide an obstacle for users if they are not capable of quickly finding their own language or reading and understanding other languages. Because Scandic Hotel is an international chain, they should also support Asian languages in addition to the 14 existing languages.

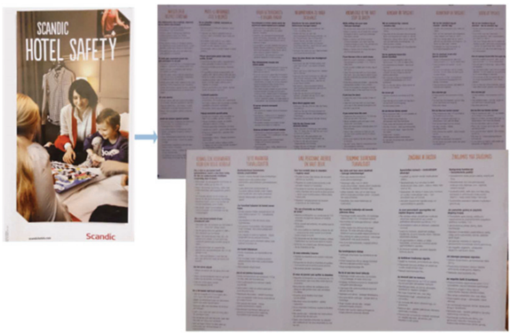



**Figure 7.** Scandic Hotel safety booklet which includes 14 languages.

We introduce a new model in Figure 8 called the *HotelSafetyModel*. The foremost idea behind this model is to develop generality, which can be realized by icons to speed up recognition and to improve understanding of hotel safety topics among all users. The main class, *HotelSafetyMode,* links to the following subclasses: *AlarmCase*, *SettlingCase*, *SmellSmokeCas*e and *StruggleCase*. All of these classes are interconnected through datatype properties. Each of these knowledge classes contains the possible situation and instruction relevant to their context.

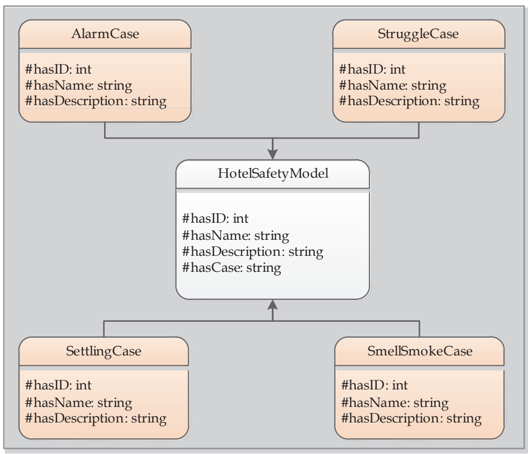


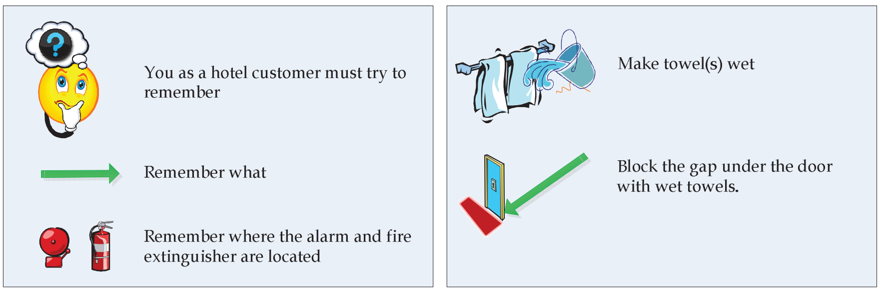

**Figure 8.** High-level hotel safety model.

Icons should be self-explanatory and able to convey the concept of hotel safety conditions to people who speak different languages. However, icons must also provide suitable connections between the concept and the representation that enhance a reader's interpretation. When constructing a sentence, rule-based principles are employed, for example:

- A subject is regularly represented by a noun form (concrete or abstract icons) such as a "smiley face" introduced to represent the actor of the sentence.
- A verb is signified by gesture such as a person looking at a book to signify "study", which is an action of the sentence.
- An object is designated by a noun form (concrete or abstract icons) such as an extinguisher icon used to represent an extinguisher object in the real world.
- An adjective is portrayed by arbitrary icons (geometric shapes, arrows, and lines).

We have used the MS Visio icon collection for our proof-of-concept study. Interpretation is the process of reconstructing the meaning of icons, as shown in Figure 9. Figure 9(a) demonstrates the iconic sentence that can be perceived from left to right, whereas Figure 9(b) portrays the iconic sentence when reading from right to left.

A smiley icon is normally treated as an actor of the sentence. In this context, the smiley icon means "you" as a hotel customer. A line and arrow together, such as a forward arrow, indicate that the next item is the object or the specific condition that must be executed. In contrast, a backward arrow indicates that what happens before the action must be performed.

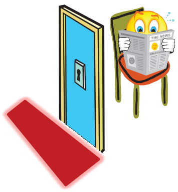
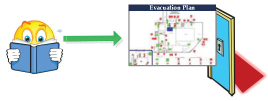
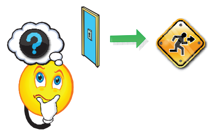
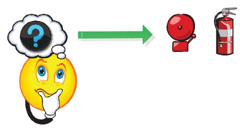
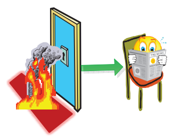
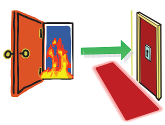



(a) (b)

**Figure 9.** Examples of how to interpret icons using left to right direction (a) and right to left direction (b).

In accordance with the four classes in *HotelSafetyModel* (Figure 8), this study attempts to propose icons that are representative of safety instructions. As portrayed in Table 2, concrete, abstract and arbitrary icons have been amalgamated.

**Table 2.** An example set of icon-based hotel safety instructions.

| Class | Safety Instruction | Resemble Icons |
|---|---|---|
| *SettleCase* (While settling in the room)  | Study the evacuation plan on the door. |  |
| | Make a mental note of where the nearest emergency exit is, such as a staircase that leads outsides. |  |
| | Make a mental note of the number of doors between your room and the nearest emergency exit. This will help you to find the exit even in the dark. |  |
| | Make a mental note of where the alarm and fire extinguisher are. |  |
| *SmellSmokeCase* (Discovering a fire or smelling smoke)  | Alert the hotel's reception or telephone the emergency number - 112 - from a safe place. |  |
| | If the corridor is filled with smoke, stay in the room. |  |
| | Take the room key with you if you leave the room. You may need to return if escape is blocked by smoke or another obstacle. |  |
| | If the fire is small and limited - use the fire extinguisher located in the corridor. |  |
| | If it is not possible to extinguish the fire close the door of the room that is on fire. |  |

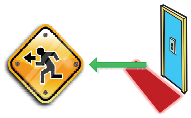
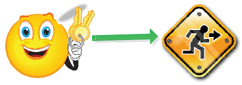
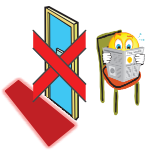
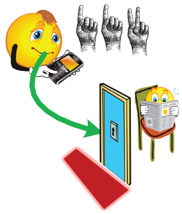
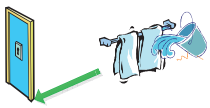
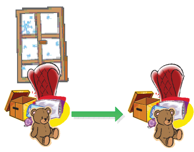
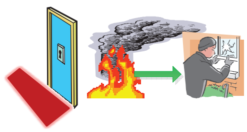
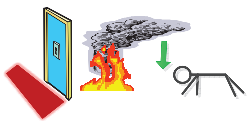
| | | |
|---|---|---|
| *AlarmCase* (Hearing the alarm)  | Leave the room if the corridor is free of smoke. |  |
| | Take the room key with you and make your way to the nearest emergency exit. |  |
| | Do NOT use the lift. |  |
| *StruggleCase* (Cannot leave the room)  | Telephone the reception or 112 and say that you are still in the room. |  |
| | Block ventilation grills and gaps under doors with wet towels. |  |
| | Move all inflammable materials away from the windows. |  |
| | Stand near a window so that you can be seen from outside and try to attract attention. |  |
| | If the room starts to fill with smoke, open a window. |  |
| | In a smoke-filled room you will be able to see and breathe better if you crouch or lie down on the floo |  |

Figure 10 illustrates the mobile client application for icon-based hotel safety systems. The application provides the presentation layer by means of iconic interfaces associated with safety information, which allows individual users to understand all safety instructions in the form of icons. The data layer is connected to the databases that store relevant data, and to the hotel's booking system for retrieving booking information such as check-in and check-out period. The application is activated when customers check-in to the hotel and deactivated immediately after customers check-out.

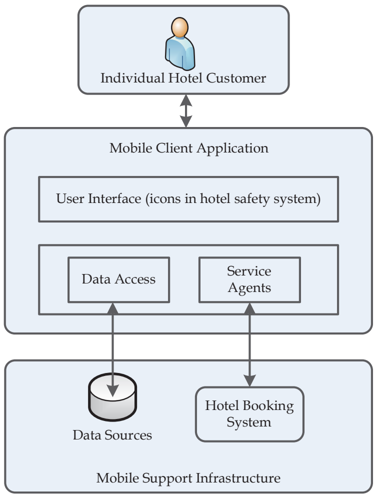



**Figure 10.** Mobile application design for hotel safety system.

## 5. Conclusions

In this paper, we introduced a first iteration of the icon-based model, language and mobile application for hotel safety. From more formal point of view, a new iconic language definition should be based on visual notations and on validated syntactic and semantic construction rules. Our next steps include implementing the prototype and a first iteration evaluation in a hotel. Interpretation of icons is an essential and culture-sensitive issue. The challenge is to minimize misinterpretations. The interpretation and the use of icons are not only task-dependent but also culturally determined beside the context. Icons are therefore not easy to handle whenever different cultures are concerned. They might be self-explanatory in one environment and completely misleading in another one. The evaluation of hotel safety icon language should be carried out side-by-side with the hotel safety booklet. Globally speaking, the evaluation results can give us very valuable information about cultural interpretations of icons.

## References


[1] Goonetilleke, R. S., Martins Shih, H., Kai On, H. & Fritsch, J. Effects of training and representational characteristics in icon design. *International Journal of Human-Computer Studies* **55**(2001), 741-760.

[2] Salman, Y. B., Cheng, H. & Patterson, P. E. Icon and user interface design for emergency medical information systems: A case study. *International journal of medical informatics* **81**(2012), 29-35.

[3] Cheng, H. & Patterson, P. E. Iconic hyperlinks on e-commerce websites. *Applied Ergonomics* **38**(2007), 65-69.

[4] Nakamura, C. & Zeng-Treitler, Q. A taxonomy of representation strategies in iconic communication. *International Journal of Human-Computer Studies* **70** (2012), 535-551.

[5] Marcus, A. Icons, symbols, and signs: Visible languages to facilitate communication. *Interactions* **10**(2003), 37-43.

[6] McDougall, S. J. P. & Curry, M. B. *More than just a picture: Icon interpretation in context*. Coping with Complexity Workshop, University of Bath, 2007.

[7] Microsoft. *Design Guidance exploration icons and symbology*, 2008.

[8] Abdullah, R. & Hübner, R. *Pictograms, icons & signs: A guide to information graphics*. Singapore: Thames & Hudson, 2007.

[9] Aykin, N. *Usability and internationalization of information technology*. Lawrence Erlbaum Association, Inc., Publisher, 2005.

[10] Heimbürger, A. & Kiyoki, Y. *Pictorial symbols in context - A means for visual communication in cross-cultural environments*. Proceedings of the IADIS International Conferences, IADIS Press, 463, 2010.

[11] Galesic, M., Garcia-Retamero, R. & Gigerenzer, G. Using icon arrays to communicate medical risks: Overcoming Low Numeracy. *Health Psychology* **28**(2009), 210-216.

[12] Kim, H., Nakamura, C. & Zeng, Q. Assessment of pictographs developed through a participatory design process using an online survey tool. *Journal of Medical Internet Research* **11**(2009).

[13] Lamy, J., Duclos, C., Bar-Hen, A., Ouvrard, P. & Venot, A. An iconic language for the graphical representation of medical concepts. *BMC Medical Informatics and Decision Making* **8**(2008), 1-12.

[14] Fitrianie, S., Datcu, D. & Rothkrantz, L. J. M. (2007). Human Communication Based on Icons in Crisis Environments*. Usability and Internationalization, Springer-Verlag Berlin Heidelberg, 57-66.

[15] Fitrianie, S. & Rothkrantz, L. J. M. A Visual communication language for crisis management. *International Journal of Intelligent Control and Systems* **12**(2007), 208-216.

[16] Tatomir, B. & Rothkrantz, L. (2005). *Crisis management using mobile ad-hoc wireless networks*. Proceedings of the 2nd International Crisis, 147.

[17] Heimbürger, A., Kiyoki, Y. & Kohtala, S. *Intelligent icons for cross-cultural knowledge searching*. In Henno, J., Kiyoki, Y., Tokuda, T., Jaakkola, H. & Yoshida, N. (Eds.). Information Modeling and Knowledge Bases XXIII. ISO Press, Amsterdam, 2012, 77-89.

[18] Department of Land Transport, *Thailand traffic signs*. Available in: http://www.dlt.go.th/th/sara/knowledgedrl/sign11.htm. Accessed: 11 January 2015.

[19] Finnish Transport Agency, *Road signs and other safety devices*. Available in: http://www.expat-finland.com/pdf/finnish_traffic_signs.pdf. Accessed: 11 January 2015.

[20] Japan Driver's License, Japanese road signs. Available in: http://www.japandriverslicense.com/japanese-road-signs.asp. Accessed: 11 January 2015.

[21] Hevner, A. R., March, S. T., Park, J. & Ram, S. Design science in information systems research. *MIS Quarterly* **28**(2004), 74-105.

[22] Reinecke, K. & Bernstein, A. Knowing what a user likes: A design science approach to interfaces that automatically adapt to culture. *MIS Quarterly* **37**(2013), 427-453.

[23] Khanom, S., Heimbürger, A. & Kärkkäinen, T. Icons: Visual Representation to Enrich Requirements Engineering Work. *Journal of Software Engineering and Applications* **6**(2013), 610-622.

[24] American Hotels & Lodging Association Guests Safety Trips. Available in: https://www.ahla.com/uploadedFiles/AHLA/information_center/travelers_tips_1.pdf. Accessed: 11 January 2015.

[25] Coffey, K. Hotel Safety Tips For Travelers. Available in: http://www.kevincoffey.com/hotel/hotel_safety.htm. Accessed: 11 January 2015.

[26] Chan, E. S. W. & Lam, D. (2013). Hotel safety and security systems: Bridging the gap between managers and guests. *International Journal of Hospitality Management* **32**(2013), 202-216.

[27] Feickert, J., Verma, R., Plaschka, G. & Dev, S. C. Safeguarding your customers: The guest's view of hotel security. *The School of Hotel Administration Collection at the Scholarly Commons* **47**(2006), 224-244.

[28] Scandic Hotel Safety. Available in: http://www.scandichotels.com/. Accessed: 11 January 2015.

# Information and Information Security

Jaak HENNO
*Tallinn University of Technology*

**Abstract.** News about hacking have become so common that most people do not follow them anymore and many do not even care, considering hacking an inevitable vice. But situation is rapidly deteriorating and many people(even in academy) are not aware or do not take seriously the emerging future, where Internet is a 'free wild West' where everyone can freely break everything and steel whatever seems to be valuable. Here is presented overview of current situation, analyzed reasons for this and presented some ideas about future and an overview of progress made in Estonia in secure use of e-tools.

**Keywords**. Internet, Information, Security, Passwords

## Introduction

The annual damage to the World economy from cyber crime in 2014 is estimated to be 445 USD [1]. This is already a measurable part of GDP of many developed countries, e.g. 1.6% of GDP of Germany, 1.5% of GDP of Netherlands and is greater than GDP of many other countries. The actual damage may be substantially higher, since online crime costs are hard to measure - companies, banks and governments often don't report damages from intrusions or the reports are rather ambiguous.

Many individuals, companies and governments greatly underestimate the risk they face from cybercrime and how quickly the cost of cybercrime increases. With more and more business functions performed online and more companies and consumers connecting to the Internet grows also cybercrime. Thieves are rapidly becoming more sophisticated and inventing new tools and methods. According to intelligence officials, "Cybercrime remains a growth industry" [2], therefore also Cyber Security jobs are growing three times faster than any other information technology jobs [3],[4].

It is estimated, that Internet economy annually generates \$2..\$3 trillion USD and this share of the global economy is rapidly growing, but cybercrime steals 15..20% of the value created by the Internet [5],[6].

Cybercrime concerns not only governments and companies, but also individuals. According to study of Pew Research [ 7], 18 percent of Internet users in USA announced that their personal information was stolen in 2013. The Identity Theft Resource Center (ITRC) claims, that in USA in every 3 seconds there is a new victim [8].

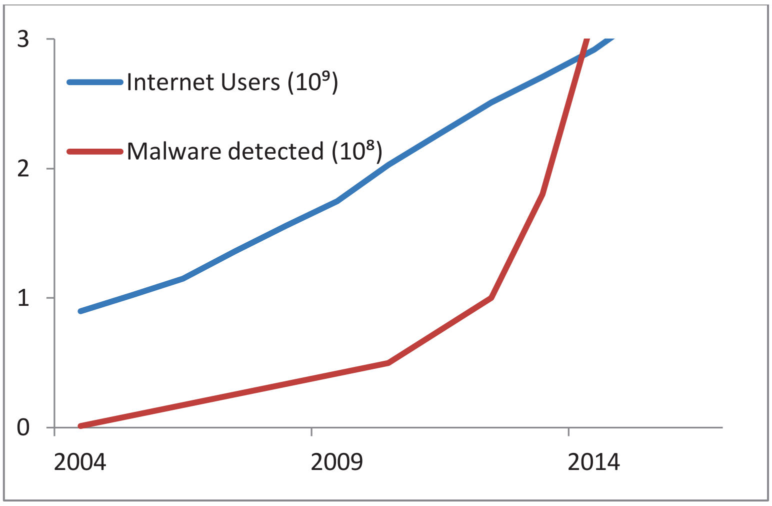



**Figure 1.** Growth of number of Internet users (data from [9], [10]) and detected malware (data from [11]). The exponential increase of malware ensures, that soon there is something for every Internet user.

Malware growth is far more rapid than change of any other Internet statistics. And this is only the visible part of the iceberg - the established anti-virus (AV) products are rather weak protection against all the time rapidly increasing threats. According to recent report from threat protection company Damballa [12], only 4% of the almost 17,000 weekly malware alerts are investigated. Inside the first hour of submission, AV products missed nearly 70% of malware, only 66% were identified after 24 hours, 72% after a full week and it took more than six months for AV products to create signatures for 100% of the malicious files used in the study.

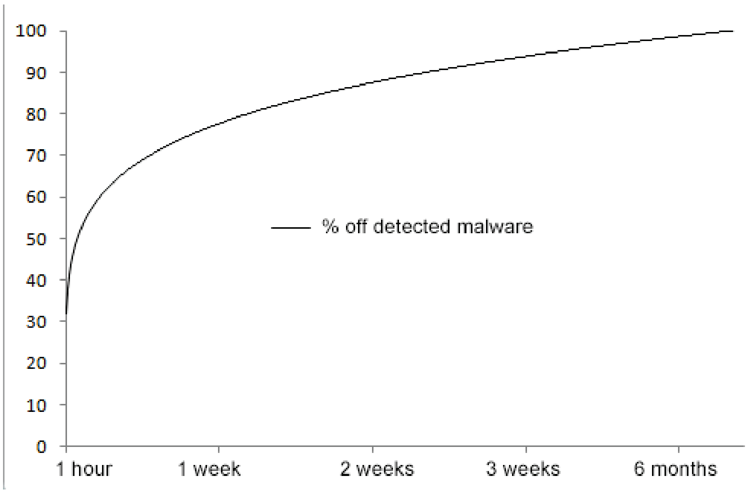



**Figure 2.** Percentage of detected malware after hours of discovery [12].

Malware is a computer program and computer programs are quick. Malware completes the main part of its malicious action in the first thirty minutes after the beginning of the infection [13]. If a malware is detected only after six months of discovery it has had time to do lot of damage.

Especially bad is situation with mobile and tablet devices. Recent report [14] from Alcatel-Lucent's Security Labs claims that 16 million mobile devices worldwide have been infected by malware, which is used by cybercriminals for corporate and personal espionage, information theft, denial of service attacks on businesses and governments, and banking and advertising scams. Growth of infections of mobile devices is rapidly accelerating with an increase of 25% in 2014 compared with 20% in 2013.

Looking at these prospects - are all Internet users (nearly half of the global mankind) turning into criminals? September 11th 2001 changed forever the way air transportation deals with security. The International Civil Aviation Organization (ICAO) initiated strict security measures which concern the whole World.

Something similar should happen for Internet, but what and when? Increasing penalties [15],[16] has not been very effective measure and it is difficult to believe that it will be in the future. The virtual world of Internet does not know borders and hackers who have been announced criminals in one country may be national heroes in other.

## Theft and value

'Theft' is unlawful change of ownership of some value. But categories 'value' and 'owning a value' should be considered more precisely.

The Merriam-Webster on-line dictionary [17] (the first source provided by Google on query 'theft definition') states:

"**Full Definition of THEFT**

1 *a* **:** the act of stealing; *specifically* **:** the felonious taking and removing of personal property with intent to deprive the rightful owner of it

*b* **:** an unlawful taking (as by embezzlement or burglary) of property "

This is a totally circular definition: the noun 'theft' is explained using verb 'stealing', concept 'felonious' is explained as[18] "very evil : villainous, relating to, or having the nature of a felony", 'unlawful' - "not lawful : illegal, not morally right or conventional"[19].

Better definitions are provided by professionals, who have to deal with this, e.g. the UK Legislation states [20]: "A person is guilty of theft if he dishonestly appropriates property belonging to another with the intention of permanently depriving the other of it; and 'thief' and 'steal' shall be construed accordingly...It is immaterial whether the appropriation is made with a view to gain, or is made for the thief's own benefit" and the used above concepts 'Dishonestly', 'Appropriates', 'Property' etc are further explained in less circular terms.

But whatever a definition, all begins from concept "value" - nobody steals a thing which does not have value. In Internet everything is data/information, thus we have to consider values, which are based on Information.

## What is Information?

Information is always syntax - a structure of a signal and this signal becomes information when received by a system which can decode this syntax – an Information Processing System (IPS). There is no information without some IPS receiving this signal understanding it, i.e. for whom encoded in the signal information has some meaning [21],[22]. Information acts like a program - it changes the future behaviour of receiving IPS, is becomes semantics for the receiver.

The (semantic) quantity of information in a signal is determined by change of behaviour of receiving IPS - the greater the change, the more there was information in the signal, the greater the value of the information.

Values are not universal. They depend on Information Processing System who receives/uses Information. There are several types/levels of IPS: living systems (especially humans), businesses, governments, but also the whole Internet behaves in many ways more and more as a huge super-IPS. Correspondingly there are also different values.

The utmost (informational) value for humans is our identity: our name, our relations with others, our possessions, our memories, beliefs, experience, skills and knowledge. We share our values in society with our friends and acquaintances, but we have many items, which we want to guard, do not want to give unlimited access. And our estimates and valuation changes with time – we often want at later time to restrict access to some personal facts what we earlier distributed or considered insignificant. Human values have been extensively studied and categorized (see e.g. [23]), but current rapid changes in society have made many older value systems less popular[24] which with increasing anxiety, stress and unhappiness has evoked many people to search for new values: slow money[25], slow technology[26], right to be forgotten[27]. The 'mainstream' value in the current 'Age of Information' is the value of knowledge, human curiosity, which is the base for growing our knowledge and independence, ability to be by themselves, rely less on others and common beliefs.

All businesses have also their 'private' information, business secrets, which they do not want to share. Although businesses should first of all produce some goods for customers, currently this is often seen as a secondary goal, for businesses (and for many people) the utmost value has become money. This deteriorating of business values is vividly exposed in the study [28] - many businesses just copy-paste value statements for their corporate websites from the websites of leaders in the field, they do not have any personal additions.

The utmost value for government is taking care for their citizens - organizing infrastructure, advancing social structures, education and business, but again in many cases the utmost value seems to be collecting as much information about their citizens and their businesses in order to maximize taxes and increasing health of government officials and politicians.

## Online Values and Passwords

In our 'Age of information' the main storage of informational values is Internet. Internet contains many resources, which should be secured and not be accessible to everyone.

The simplest method to discover informational values on Internet is to use search engines – 'googling'. When Internet was young and search engines just appeared was proposed a method to hide some pages using a special file: *robots.txt*[29]; the idea was, that ban in this file should stop crawlers visiting the site. The proposal did not become a standard and currently Google ignores it - it still lists the page, only adds remark "A description for this result is not available because of this site's robots.txt".

The prevailing method to restrict access is with passwords. Password-checking function is a 'black box', finite automaton, which opens only when the correct password is provided. For brute force methods password security depends on password length $n$ and the size $m$ of the alphabet, i.e. number of symbols which can be used in the password string.

If only lower-case ASCII letters are used in a password, $m = 26$. Thus the time required for brute-force crack of a password with length $6$ symbols is proportional to $26^6 = 308915776 \approx 3*10^8$. With a computer performing 500000 character comparison operations per sec this time would be ca 10 min (the same estimate is provided in [30]).

But if the length of a password is 8, and the vocabulary uses both lower- and uppercase characters, numbers and printable special symbols from the ASCII table, altogether 96 symbols [31], the brute-force cracking time becomes 463 years.

These estimates hold only if the cracker knows the length of the password. Most password checking functions (e.g. logging in to Windows) require the input to be of right length, i.e. if your password is *ich@myself£* and you enter *ich@myself£1* and then press 'Enter', you will not get in and should start over. Thus actually the estimation for time to crack even a 6-character password if all 96 printable ASCII symbols could be used becomes

$$\sum_1^6 96^i / (500000*60*60) = 18.3 \text{ days}$$

and with 8-character password - 462.3 years.

Thus if your password is 8+ characters long, contains both capital and lower-case letters and at least 1 number, then it will be almost impossible to break the password with brute force.

Another often cited method for password cracking is not syntactical, by characters, but semantic, using word lists. Currently every Internet user has to remember tens of passwords. It is often stressed, that these should be different, otherwise revealing a user password in one badly designed website makes user vulnerable on all other sites where the same password is used. It is very difficult to remember passwords created from arbitrary symbols, e.g. *K9%#@hjM9x<* (a suggestion from a webmaster), thus usually users create passwords from words with some semantics, e.g. something like *password.*

Native speakers (and usually we use words from our native language) know in average about 20000 words [32], [33]. Suppose that all words known to user can occur in a password with equal probability and password is created from 3 words (all from his mother language).

The average length of a word in 20000-word English dictionary is 8 [34], thus expected brute-force cracking time would be $5.449*10^9$ years.

However, not all words are essential. Several researches of frequency of use of words (Zip-s law [ 35 ] or Herdan–Heaps's Law [ 36 ]) indicate, that frequency distribution of use of words is close to power law. Only 100 words are needed to understand 50% of any text, 1000 words are sufficient to understand 75% of all the words used in common, every day English [37] and native speakers use 7000 words for 90% of what they say and write. Assuming that words used in passwords are selected with the same frequency distribution and taking into account, that the most frequent words are mostly short (average length of the 1000 most frequent words in English language is 5.449, calculated for the list [38]) reduces the above estimate of time needed to break 75% of passwords created from three frequently-used words to 3 days, which is already sufficient in many cases.

But Internet uses do not select passwords the same way as they speak. Hackers have many times published long lists of passwords from hacked sites, e.g. in December 2014 they published 13000 user passwords from Amazon, Walmart and Brazzier sites [39]. These lists show, that users are extremely reckless when selecting passwords - e.g. in 2013 the most frequent passwords were '123456', 'password' '1234567' etc [40]; the list for 2014 is nearly the same[41].

Hackers use first of all these lists of published often-used passwords. And if they can break even only in some computers, they can load to those computers trojans which make them zombies which assist breaking other sites, making hacking activity parallel, from more and more computers simultaneously.

Thus the passwords as a security technology is not the culprit. Culprits are users, who still do not take passwords seriously.

Password strength can be measured using entropy of the distribution of symbols in the password. How people create passwords depends on many variables - age, culture, skill of computer use, general awareness of dangers when using weak passwords. Comparing passwords for participants of the course "Games and Virtual World Programming" in Tallinn University of Technology during 2006..2015 (mainly first year IT students) indicates, that in spite of repeated warnings and suggestions for using strong passwords the situation remained nearly the same; improvement appeared only in current year when the registering program did not accept weak passwords - all passwords had to satisfy regular expression

(?=^.{8,}$)(?=.*\d)(?=.*[!@#$%^&*]+)(?![.\n])(?=.*[A-Z])(?=.*[a-z]).*$

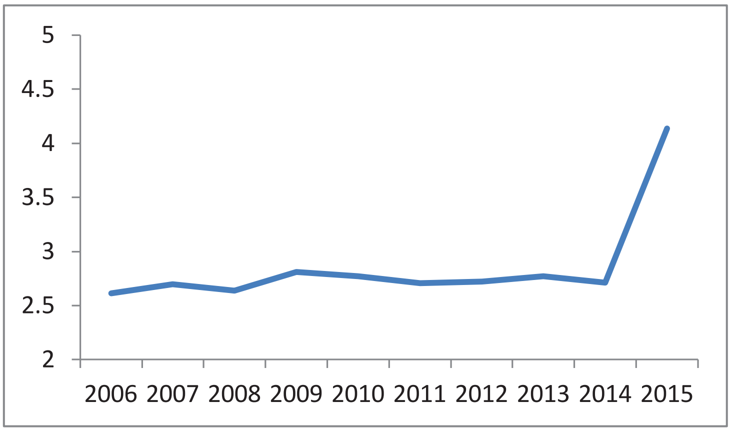



**Figure 3.** Average entropy (for the whole group, 30-60 students) of passwords used by the first year IT students in Tallinn University of Technology

## Locks and keys

Use of locks/keys in real life and in virtual life (Internet) is diametrically opposite.

In real life, if you have to pass a door, which is not always open, you will get a key. In virtual life usually you have to create the key (password) yourself - it is not provided by authorities who can lock the door, i.e. restrict access to some Internet resources.

Since there are many users with very different understanding of needs for security in Internet, inevitably appear some weak passwords. When computers with weak passwords are discovered (and they always will be discovered by crawlers), they get infected with trojans, which start attempting to break into other computers. As a result, infections of computers on network propagate like infectious of human diseases [42].

Internet were much safer, if passwords (the keys to a door) were also created not by users (persons who want to pass the door), but by authorities (owners of the door) - like it is in real life.

We investigated use of passwords, which were created be a program beforehand and provided to users who have right to use resources.

Such passwords should satisfy two main conditions.

They should be strong: length > 8 characters (and for ease of remembering - less than 16), contain both upper- and lowercase characters, numbers and some special characters - !@#&. Currently are created passwords checked with regular expression, which agrees with Microsoft requirements for strong password [43] :

```
(?=^.{6,10}$)(?=.*\d)(?=.*[a-z])(?=.*[A-
Z])(?=.*[!@#$%^&amp;*()_+}{&quot;:;'?/&gt;.&lt;,])(?!.*\s).*$
```

And passwords should have semantics - it should be easy (for humans) to reduce them to some meaningful phrase, which is easy to remember (the idea is similar to cryptographic keys built from environmental data [44]).

To satisfy these requirements, passwords were created with the following procedure (it is still experimental).

A piece of text in users mother language (for Estonian students - in Estonian, for Russian students - in Russian, for foreign students - in English, but they may also provide their own text in whatever language) is used, e.g. an item from the morning news - something, what has meaning for the user and is thus easy to remember. The item is converted into list of words (all spaces, punctuation marks, newline characters removed) and from this list are randomly selected some consecutive words so that the total number of characters were >8; these words are converted to a string of consecutive characters, the result may be like 'oneyearborrowing'. Next some randomly selected characters are converted to uppercase and some numbers and special characters inserted in random positions. If for the seed text was used a news item from a newspaper, the final result can be e.g. 'atHen1@1swill5nowhave', 'Ce9NT31@RalBankhasincreased', 's@harP3lyo9nthenews' etc.

The method was tested with students of Tallinn University of Technology. The common complaint was – "it is too complex, impossible to remember!", thus the function will be simplified. Instead of several uppercase characters and several digits it is sufficient to insert only one of both – this still makes sure, that a brute-force attack using only lowercase letters will not be successful.

## Stronger methods

Compared with information used for identification in 'real life', passwords contain miserable amount of personal identification information.

If we appear personally, our acquaintances recognize us by our visual image, by our voice, by our gait (our walk alone is enough to recognize us [45]). In digital form (if digitized), this is at least several gigabytes of data.

For strangers (e.g. when boarding a plane and/or in border control) we present passport. Nowadays in more and more countries this is e-passport or some other electronic travel document, which store information about carrier in embedded EPROM (Erasable Programmable Read Only Memory, a memory chip that retains its data also when its power supply is switched) or RFID chip (Radio-Frequency Identification, similar to previous, but can be accessed from distance, without card reader and functionally a contactless smartcard, often used in shops to prevent stealing). These chips contain lot of biometric information - facial or eye maps and fingerprints as primary identification features accepted by ICAO.

The ICAO has also described in ICAO Doc 9303 Part 1 Volume 2 e-passport standards as well as in associated ISO/IEC 19794 biometrics standards the standardized file formats and communication practices that should be used to access information stored in e-passports.

Another stronger method of secure access with keys provided by resource owner (not generated by user) is achieved with electronic ID (eID) cards. These are already used in nearly twenty countries and several other countries are contemplating their introduction; all EU countries are either already using or are introducing eID cards [46].

In Estonia the electronic ID card was introduced already in 2000, when Estonian parliament passed the Digital Signature Act (DSA) which stated, that the ID-card is a new type of personal identification card, which is mandatory for all Estonian residents [47]. In the card is embedded chip with at least 8 KB of EEPROM memory and processor supporting the elliptic curve cryptosystem using 2048-bit public key encryption; the card is regularly renewed so that its safety corresponds to developments in computing technology. Access to the card is guarded with two passwords: the first one - to activate/read the card, the second - to create digital signature for a document or bank transaction.

The created by card electronic identity enables on-line electronic authentication, digital signing of documents and all other activities which require establishing person's identity, e.g. managing bank accounts, creating business enterprises and participation in elections (on-line). Digital signatures are equivalent to handwritten ones and all public sector organizations must accept digitally signed documents. In December 2014 Estonia made Estonian ID Card available also to non-residents [48]; these (without owner's image) cards are intended for electronic identification, speeding up many identification procedures, e.g. on borders, in registering new enterprises etc.


**Figure 4.** Reverse side (with the EPROM chip) of the Estonian Id-card. The chip contains personal certificates which are encoded using the SHA-256 hash algorithm.

On-line voting has been used in Estonia starting from 2005 in six elections (municipal, national and European). The system was accused [ 49 ] in many shortcomings, but the practice (six elections) has not revealed any incidents which have influenced the outcome, thus the accusations were soon strongly refuted [50],[51],[52]. The accusations in the paper consider procedures, i.e. acts of humans involved in organization of e-elections, the technical side (cryptographic procedures, communication protocols) are not even mentioned. Humans are in every system the weakest link, and elections as a politically very sensitive operation have been also e.g. in USA strongly criticized [53],[54]. In general elections in Estonia (spring 2015) more than 30% of voters used e-voting (53% of them were women) and the procedure was used from 116 countries. Already in the first day of elections the number of e-voters exceeded twice the corresponding number from previous elections [55], thus Estonians have embraced the procedure. And for Estonians who were at the time of elections abroad this was the easiest way to participate - it is very difficult to believe, that voters (Estonians abroad) from 116 countries could participate if traditional methods of voting (traveling to some Estonian embassy or consulate) were used.

In 2007 the Estonian Certification Centre and Estonian Mobile operator EMT launched an innovative Mobile ID service which is used for identity verification and digital signing using the SIM card in mobile phone. The Certification Centre and the State also created the basic software needed for using the card and developed the DigiDoc software, which provides the digital signature, eSignature validity verification and data encryption.

In all 11 banks operating currently in Tallinn, e-Id-card based full remote access is available in all of them; mobile-Id based remote access - in 9.

**Why is the current state of security in Internet so miserable?**

Internet is young - only 25 years old. It is still in its adolescence, growing and changing all the time.

Many of us, especially older generations, baby boomers and millennials have difficulties with adapting to these rapid changes. Many consider internet as a heavenly gift and spend hours in clicking around in Facebook, Tumbr or in some darker corners , e.g. Pastebins. And if antivirus programs (if they are installed and updated) begin giving warnings, they are switched off. The main reason for massive containment of mobile devices is the willingness of device owners to check out every link, to click everywhere in social network sites. The social networks have become for many a real cacoethes and forces them e.g. to create Facebook profile even for unborn child [56]. They see in Internet only the excitement of NEW and do not consider dangers. They do not follow and/or know rules of conduct, which often do not even exist yet. Their understanding of Internet has not yet achieved the 'Peak of Inflated Expectations' of Gartner's technology development cycles [ ].

But the generation of 'digital natives' [58] (currently students and younger) does not even know life without computers. They have grown with computers and computer games, with Internet and iPads and tablets, with Facebook, Twitter and YouTube, with Blogs and Wikis.

Video/digital games have become an essential part of our culture and life, especially for the younger generation. The whole our planet spends 3 billion hours a week playing videogames; more than half a billion people worldwide are playing computer and videogames at least an hour a day [59]; 97% of teens in the USA play video games, and sales of games are growing. Massive technology adoption, everyday use of internet and playing video games have already changed behavior, attitudes, skills and styles of the younger generation. Older generations have to adapt to these changes.

Video games are great schools of hacking. With every new game player has to discover - how does this work, what happens, if I do this or that? While elders are often afraid to push unknown buttons or try out unknown devices ("Where is a manual for this ?"), youngsters press every button, try every device and never ask for any manual - this is the behavior what they have learned playing games. Games present to player constantly new challenges, new problems which he/she must solve. The player cannot get through with what they already have or know, he must find new combinations and incorporate old skills with new skills to overcome obstacles.

A game is more interesting if there are living opponents. Massive Multiplayer On-line games (MMOG), where players compete not only with a (static) computer program, but with other players - living, intelligent opponents - are one of the fastest growing sectors of digital games.

And for some of them even this is not enough. Some of 'digital natives' have grown to 'digital hooligans', who are seeking excitement (and money) by hacking in Internet. In November-December 2014 were seriously attacked several big companies - Sony Pictures Entertainment, on Christmas Day a group under the name Lizard Squad took down Xbox Live and PlayStation Network services. Later one of originators of the attack confessed that he and his team of hackers considered the attack a "sort of a game", and partly did it for their own amusement [60].

The world of MMOGs is a good indicator of future - trends in behavior of the generation of 'digital natives'. Security problems in MMOGs are becoming increasingly critical. Cheating, virtual frauds, and other security attacks are widespread in the virtual worlds of MMOGs. Gamers have developed illegal methods to obtain virtual wealth and convert it into real wealth in the real world using Real Money Trading, which allows to buy using real money game values (trophies, achievements). A million-dollar business of developing cheating tools has appeared: bots for (half) automated gaming, cheats, trainers and walkthroughs and web-sites, where one could get for monthly fee super-human abilities in some game, e.g. see other players through walls [61]. It is difficult to tell exact percentage of cheating players, but in study involving one on-line racing game [62] was collected information about gamers on the Steam Community global gaming network: Steam has an anti-cheating system which marks cheaters public profile; from more than 12 million analyzed gamers over 700 thousand (ca 6 %) had their profiles flagged as cheaters.

If earlier cheating was inevitably considered as a despicable member of society, scumbags, currently the attitude has changed - the study discovered, that "cheaters are well embedded in the social and interaction networks: their network position is largely indistinguishable from that of fair players". And seeing fellow players succeed through cheating creates the idea that it might be worth the risk also to try.

While youth is changing quickly and have fun from hacking and/or cheating in games older people and big companies are slow to adapt with this brave new world. In spite of deep financial resources Sony Pictures Entertainment has after more than two months after break not been able to secure its network [ 63 ],[ 64 ]; USA State Department email servers are after three months of discovery still accessible to remote access [65].

We do not (yet) have traditions and moral positions related to our all the time changing digital environment.

One of the great achievements of ancient nations was creation of Ten Commandments, which originate from Hittite and Mesopotamian laws and treaties. People understood, that following the prohibitions of idolatry, blasphemy, murder, theft, dishonesty, adultery make their life better and these prohibitions have become basics of human morality and values like honesty, trust, caring, fairness, which (most of us) follow in our everyday, 'real' life. Unfortunately we do not yet have something analogous for our virtual, digital life, the "Ten Commandments of Computer Ethics" [66] introduced in 1992 [67] have not become a set of standards to guide and instruct people in the ethical use of computers.

But the 'real' Commandments were created during centuries and the Internet is only 25 years old.

The current dreadful situation with security in Internet is based also on economical organization of Internet. The important reason why ICAO become after Sept11, 2001 the leader in security measures which concern the whole World was money - airplanes are expensive and aviation companies could not allow that they are used as sledgehammers to attack skyscrapers. But in Internet attacks suffer Internet users, not companies providing Internet services or domain name registrars. For them every user and every domain registration means income, so they are very reluctant to restrict use or cancel a domain. Soviet Union has died, but the .su internet suffix assigned to the USSR in 1990 is still alive and has turned into a haven for hackers who've flocked to the domain space to send spam and steal money. When in 2011 administrators of Russia's .ru domain space toughened their rules, scammers began to move to .su. In 2013 were registered 125299 .su domains [68]. Russian IT exports consider that more than half of cybercriminals in Russia and former USSR use the .su domain to send spam, launch attacks against websites and control botnets - networks of hijacked computers used to empty bank accounts.

Nobody is directly responsible for the state of the Internet Network. In Jan 23, 2014 at the World Economic Forum in Davos was announced Global Commission on Internet Governance, leaded by the former Swedish Prime Minister; the task of the new body was to consider and propose a strategic vision for the future of Internet Governance [69]. The results after one year after the establishment of this high-level organ are not very impressive: their website [70] mentions 9 meetings (three - virtual) and 9 publications. Assessment for their work from the web watchdog 'The Register' was: " the result is somewhat of a soggy mess" [71].

The rapidly growing malware problem is in great deal psychological. Computers and Internet are intellectual tools and we have not yet learned to use them. We learn use of, real-life tools early in childhood experience (when hammering with something and hit a finger) or later in school or courses (driving). And we are not allowed to use complex machinery or drive a car without proper credentials about our abilities. To use a computer (usually) no credentials are required, but computer can be far more dangerous than e.g. a car, since when using computers we behave differently than in real life. In real life we do not ask strangers into our house, we do not open strange packages left somewhere. In airports a warning "unattended luggage will be removed and destroyed" is repeated routinely. But in virtual life many people invite everybody into their virtual houses collecting Facebook 'likes', open every webpage and mail attachment. The result is constant flow of news about hackings and thefts on Internet.

Constant need of 'being connected' has in many developed FOMO - Fear Of Missing Out [72] syndrome. They constantly scan email or Facebook and stare their iPad in hope that something happens - somebody calls, tweets, posts a picture. More than half of social networkers are experiencing FOMO and more than 60% of age 18-34 have multiple networking accounts [73]. They want desperately to be part of WORLD, thus they click every link and open every attachment and the result - new infections appear daily. They are 'Alone Together' [74]. FOMO drives them to decisions they don't otherwise want to make - instead of staying home and resting they cannot miss a dinner party, cannot bear a thought that they have not been invited to (usually dull) meeting on workplace, can't rest for the thought that something important might happen while they are not present. Are they willingly delegating their ability to think to Network ?

Soon we will be connected not only with computers and (possibly) humans behind these computers, but with all kind of 'things' with IP-s - cars, doors, TV-s - the whole network of interconnected devices and sensors, which will dissolve the current Internet [75]. It is expected that instead of seven billion devices in 2010 there will be 50 billion devices connected to the internet by 2020. Security of these devices with (comparatively) little memory and slow processors will raise additional concerns, there are already reports of hacking into baby webcams [76] and taking down trones [77].

All these devices can be seen as black-box finite automata. Since they have finite number of states, they inevitably go into cycle, start repeating their responses after sufficiently many inputs. This allows to identify (decode, decipher) the functionality of a black box without any knowledge of the internal structure except an upper bound $k$ on the number of states, but this can be easily obtained from the size of memory of the device (usually freely available from technical documentation). Already in 1973 was proved the following result [78, p.211]:

Let size of the input alphabet of finite automaton (black box) $\mathfrak{M}$ be $m$ and the size of the output alphabet - $n$ . For any natural number $k$ one can effectively construct an input word $d(k)$ of length $|\mathrm{d(k)}|=]4k^2(\ln nk)m^{2k}[$ which residually distinguishes all automata with $k$ states, i.e. any two automata with $k$ states after getting this input either produce different outputs (they are recognized to be different) or the automata will afterwards act the same way. Since there is only small number (length is small-power polynomial) of such words, it follows, that if automaton $\mathfrak{M}$ is connected to another (bigger) automaton $\mathfrak{M}_1$ with number of states $c_1 * |d(k)| + c_2$, where constant $c_1$ depends on encoding of words of length $|d(k)|$, constant $c_2$ - encoding of search algorithm, then automaton $\mathfrak{M}_1$ can analyze behavior of automaton $\mathfrak{M}$ . Thus $\mathfrak{M}_1$ will know beforehand what is the response of $\mathfrak{M}$ , thus make $\mathfrak{M}$ to do everything what $\mathfrak{M}_1$ wants, i.e accomplish C&C (Control&Command). For the case where an upper bound on the number of states is not known, in [79] is described a polynomial-time probabilistic inference procedure; in tests this procedure was able to analyze automaton based on Rubic's Cube ($10^{19}$ states). For developing a botnet it is sufficient, that probability of success of C&C of a computer from network is greater than 0.5.

Most of arriving 'things' with IP-s will be always on and connected, thus a beautiful playfield for hackers having a bit more computing power. Member of the Lizard Squad confessed after the attack against Microsoft and Sony that their botnet sent out 1.2 terabits per second [80]. If this was enough to take down servers of big commercial companies then it is certainly enough to e.g. to lock all doors of all self-driving cars, set all ATM to spit out money (this has already happened [81],[82]) or set all traffic lights to red. A week after attack Lizard Squad launched a service called Lizard Stresser that allows customers to attack any target, including large websites or Internet services. Prices range from $6 to $500, depending on the length of the attack, so everyone can try "just for fun".

**Who is in Control ?**

The coming Internet of things is sometimes characterized as “universal global neural network”. It is estimated, that in 2020 the Internet contains 44 zetabytes (zetabyte - $10^{21}$ bytes) [83] of information. Capacity of human brain is estimated only ca 2.5 petabytes (petabyte - $10^{15}$ bytes) [84]. Thus if to this universal global network is added sufficient AI, e.g. Google's Spanner [85], [86], it is able to hack anyone of us.

There are already autonomous, intelligent programs in the wild, able to switch on street lights, manage telephone networks and viruses and worms. The source code of the very sophisticated worm Stuxnet which severely damaged Iranian nuclear program (not only) is now freely available on Github [87]; everyone can modify this superworm for new targets. The recently discovered banking worm Carbanak [88] which made ATM in Kiev to dispense cash at seemingly random times of day exposed very sophisticated evasive and stealthy behavior [89] which hide it for years.

Our commercial culture has created tendency to turn everything what we use into a service. Once upon a time we got personal car - the T-model ('whatever color, if it is black'). For a long time cars remained products which we proudly owned and managed, own car provided for young adults the ultimate sense of freedom.

Nowadays we do not have any more total ownership of our cars. In modern personal cars great deal of functions are performed by electronics which (in many countries) we do not own and which e.g. in USA is covered by DMCA (Digital Millennium Copyright Act). The Section 1201 of the DMCA prohibits unlocking ‘access controls’ of the software and car companies can threaten anyone who needs to get around those restrictions. So we can recall the time when we our cars had individuality only from Jeremy Clarkson's shows. Instead of owning we are now supposed to rent or lease, and young generation isn't interested in cars any more [90].

Soon we just ask and from somewhere appears a driverless car or drone which takes us to another location. Instead of moving we are moved, consuming services of movers industry (and meanwhile adding some fat to our consuming body). Once we bought records and DVD-s to listen the music or watch cinemas, we were music/movie listeners/lovers, nowadays we should consume streamed (what you can not store!) services of entertainment industry. Similar process is currently underway with personal computers. We are supposed not to own any programs, but to use/consume SaS - Software as Service. In social networks we are not even consumers, we are consumed by advertizing industry, by Google and other collectors of 'Big Data'. It is very difficult to create anything for iPad/Android, you should consume services of apps. Looking at persons who all the time stare their iPad/Android displays: "Who is in control ?"

With desisting ownership of things (cars, records, programs) we desist also control. Do we really need all those embedded sensors and connectivity everywhere, to turn everything into (anonymous) service - we have already spamming refrigerators [91]. Maybe it were good to design some things to be difficult to use (like currently computer games) [92] - managing difficulties gives us great emotional pleasure and satisfaction.

## Conclusions

We - the Mankind - have learned after the Ten Commandments many other truths, e.g. slowly, but surely we (most of us) have understood, that Democracy, Freedom of Speech and free access to all kind of information are cornerstones for our future development. Hopefully we will also manage all problems of our relations with technology and Internet.

Attitudes of younger generation, the 'digital nations' are far more realistic and practical, they are not 'information objectrs' to be consumed by Facebook, Google, Microsoft. First year IT specialty students in Tallinn University of Technology do not actively upload photos/selfies and collect 'likes' on Facebook (even if they have a account - ca 50% do not); they use their mobile phones to read e-mail and pay parking fees, but do not keep their mobiles constantly open; instead of Facebook many have accounts in LinkedIn, Stackoverflow or some other professional forums, where they often already participate in professional conversations.

Currently our relationship with rapidly advancing information technology is still in its infancy, we're still searching correct ways to interact mindfully, meaningfully with it. Often it is difficult to separate innovations aimed to improve our life from attempts to earn more and more money, creating bigger and bigger divide among us. We should learn to manage technology, otherwise technology will soon manage us. We should craft our technology to serve us and make the best of possibilities for social connections enabled by technology, not to be only sources for more information to Network.

## References

[1] Net Losses: Estimating the Global Cost of Cybercrime. Economic impact of cybercrime II. Center for Strategic and International Studies, June 2014, http://www.mcafee.com/ca/resources/reports/rp-economic-impact-cybercrime2.pdf

[2] Bloomberg. Cybercrime Remains Growth Industry. http://www.bloomberg.com/news/2014-06-09/cybercrime-remains-growth-industry-with-445-billion-lost.html

[3] http://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm

[4] http://www.cybrary.it/cyber-security/

[5] Net Losses: Estimating the Global Cost of Cybercrime. Economic impact of cybercrime II. Center for Strategic and International Studies, June 2014, http://www.mcafee.com/ca/resources/reports/rp-economic-impact-cybercrime2.pdf

[6] D. Bartlett. Talking Points - Cyber Crime: A Growing Threat to Global Companies. http://rsmi.com/publications/talking-points/691-talking-points-cyber-crime-a-growing-threat-to-global-companies.html

[7] Public Perceptions of Privacy and Security in the Post-Snowden Era. http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/7

[8] http://www.idtheftcenter.org/id-theft/data-breaches.html

[9] http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

[10] http://www.internetlivestats.com/internet-users/
[11] Malware Statistics. http://www.av-test.org/en/statistics/
[12] Damballa State of Infections Report Q4 2014. https://www.damballa.com/state-infections-report-q4-2014/
[13] E. Filiol. Strong Cryptography Armoured Computer Viruses Forbidding Code Analysis: the Bradley Virus. http://vxheaven.org/lib/pdf/Strong%20Cryptography%20Armoured%20Computer%20Viruses%20Forbidding%20Code%20Analysis:%20the%20Bradley%20Virus.pdf
[14] Motive Security Labs report. https://resources.alcatel-lucent.com/asset/184652
[15] EU increases penalties for cybercriminals and hackers. http://www.cnet.com/news/eu-increases-penalties-for-cybercriminals-and-hackers/
[16] Obama's proposed changes to the computer hacking statute: A deep dive. http://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/01/14/obamas-proposed-changes-to-the-computer-hacking-statute-a-deep-dive/
[17] http://www.merriam-webster.com/dictionary/theft
[18] http://www.merriam-webster.com/dictionary/felonious
[19] http://www.merriam-webster.com/dictionary/unlawful
[20] http://www.legislation.gov.uk/ukpga/1968/60/crossheading/definition-of-theft
[21] J.Henno. Emergence of Information, Communication and Language. Information Modelling and Knowledge Bases XXIV, IOS Press, 2013, pp 277-299
[22] J. Henno. Grounded Multi-level Computations. Information Modelling and Knowledge Bases XXVI, IOS Press, 2014, pp 140-151
[23] A. Maslow. New Knowledge In Human Values. Joanna Cotler Books, 1959, http://www.amazon.com/Knowledge-Human-Values-Abraham-Maslow/dp/0060341408
[24] International Association for Human Values. http://www.iahv.org/us-en/
[25] https://slowmoney.org/
[26] http://slowtech.org/
[27] EU regulators to Google: "Right to forget" needs to go worldwide. http://arstechnica.com/tech-policy/2014/07/24/eu-regulators-to-google-right-to-forget-needs-to-go-worldwide/
[28] S. Roth. Common values? Fifty-two cases of value semantics copying on corporate websites. Human Systems Management, Vol. 32 No. 4, pp. 249-265, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2368340
[29] http://www.robotstxt.org/
[30] The Problem with Passwords. http://www.bloomberg.com/bw/magazine/content/11_06/b4214036460585.htm
[31] http://www.ascii-code.com/
[32] http://testyourvocab.com/
[33] http://www.economist.com/blogs/johnson/2013/05/vocabulary-size
[34] J. Henrick, On Word-lenth and dictionary size. http://digitalcommons.butler.edu/cgi/viewcontent.cgi?article=1116&context=wordways
[35] C.J.Sorell. Zipf's Law and Vocabulary. http://www.academia.edu/550703/Zipfs_Law_and_Vocabulary
[36] L. Egghe. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. Journal of the American Society for Information Science and Technology, 03 (2007), 58 pp 702-709
[37] Oxford Dictionaries. The OEC: Facts about the language. http://www.oxforddictionaries.com/words/the-oec-facts-about-the-language
[38] 1000 Most Common Words. http://www.giwersworld.org/computers/linux/common-words.phtml
[39] Hackers leak 13,000 Passwords Of Amazon, Walmart and Brazzers Users. http://thehackernews.com/2014/12/password-hacking-data-breach.html
[40] http://splashdata.com/press/worstpasswords2013.htm
[41] The 25 Most Popular Passwords of 2014: We're All Doomed http://gizmodo.com/the-25-most-popular-passwords-of-2014-were-all-doomed-1680596951
[42] B. K. Mishra. A. K. Singh. SIjRS E-Epidemic Model With Multiple Groups of Infection In Computer Network. International Journal of Nonlinear Science, Vol.13(2012) No.3, pp.357-362
[43] https://technet.microsoft.com/en-us/library/cc786468%28v=ws.10%29.aspx
[44] J.Riordan,B.Schneier. Environmental Key Generation towards Clueless Agents. https://www.schneier.com/paper-clueless-agents.pdf
[45] Jim Giles. Cameras know you by your walk. New Scientist, vol 215:2883, Sept 22, 2012, pp 17
[46] https://www.tractis.com/help/?p=3670
[47] Electronic ID Card. https://e-estonia.com/component/electronic-id-card/
[48] E-residency in Estonia for entrepreneurs. http://1office.ee/en/e-residency/

[49] http://www.theguardian.com/technology/2014/may/12/estonian-e-voting-security-warning-european-elections-research
[50] https://www.ria.ee/e-voting-is-too-secure/
[51] http://estonianworld.com/security/attacks-estonias-e-voting-political-rather-technical/
[52] http://www.vvk.ee/valimiste-korraldamine/vvk-uudised/vabariigi-valimiskomisjoni-vastulause-the-guardianis-ilmunud-artiklile/
[53] A. Liptak. Lost Votes, Problem Ballots, Long Waits? Flaws Are Widespread, Study Finds. http://www.nytimes.com/2013/02/06/us/politics/us-voting-flaws-are-widespread-study-shows.html?pagewanted=all&_r=0
[54] J. Kaminsky . Voting machine, ID problems crop up in U.S. elections. http://www.reuters.com/article/2014/11/05/us-usa-elections-irregularities-idUSKBN0IP06M20141105
[55] 30.5% of all votes for Estonian parliamentary election submitted online. http://www.baltictimes.com/30_5__of_all_votes_for_estonian_parliamentary_election_submitted_online/
[56] Parents Make Facebook Page for Unborn Child; Becomes Online Journal. http://abcnews.go.com/Technology/facebook-page-unborn-baby-marriah-greene-texas-page/story?id=13734579
[57] Hype Cycle. http://en.wikipedia.org/wiki/Hype_cycle
[58] Marc Prensky. Digital Natives, Digital Immigrants. http://www.marcprensky.com/writing/Prensky%20-%20Digital%20Natives,%20Digital%20Immigrants%20-%20Part1.pdf
[59] http://www.ted.com/conversations/44/we_spend_3_billion_hours_a_wee.html
[60] Xbox 'Hacker' Reveals Why He Attacked Consoles http://news.sky.com/story/1398435/xbox-hacker-reveals-why-he-attacked-consoles
[61] Maiberg, E. 2014. Hacks! An investigation into aimbot dealers, wallhack users, and the million-dollar business of video game cheating. PCGamer. (2014).
[62] J. Blackburn, A.I. N. Kourtellis J. Skvoretz M. Ripeanu. Cheating in Online Games: A Social Network Perspective. ACM Transactions on Internet Technology. (2014).
[63] S. Peters. Report: Russian Hacker Broke Into Sony & Is Still There. http://www.darkreading.com/report-russian-hacker-broke-into-sony-and-is-still-there/d/d-id/1318964
[64] S.Biddle. Russians Hacked Sony, and They're Still Inside: Report. http://internet.gawker.com/russians-hacked-sony-and-theyre-still-inside-report-1683977374
[65] Wall Street Journal. Three Months Later, State Department Hasn't Rooted Out Hackers. http://www.wsj.com/articles/three-months-later-state-department-hasnt-rooted-out-hackers-1424391453
[66] Ten Commandments of Computer Ethics. http://courses.cs.vt.edu/professionalism/WorldCodes/10.Commandments.html
[67] R. C. Barquin "In Pursuit of a 'Ten Commandments' for Computer Ethics", http://computerethicsinstitute.org/barquinpursuit1992.html
[68] SU TLD: Registration and Delegation Statistics. http://stat.nic.ru/en_su/2014/01/01/titul-20140101.shtml
[69] T. Brewster. Davos: Two-Year Investigation Into Future Of The Internet Announced. http://www.techweekeurope.co.uk/workspace/davos-internet-future-commission-136935#cg4hEfcqWqvcYwlm.99
[70] Global Commission on internet Governance. https://www.ourinternet.org/#home
[71] The Register. Global Commission on Internet Governance wobbles into IANA debate. http://www.theregister.co.uk/2014/11/27/gcig_wobbles_into_iana/
[72] Oxford Dictionaries. FOMO. http://www.oxforddictionaries.com/definition/english/FOMO?searchDictCode=all
[73] Report: 56% of Social Media Users Suffer From FOMO. http://mashable.com/2013/07/09/fear-of-missing-out/
[74] S. Turkle. Alone Together: Why We Expect More from Technology and Less from Each Other. http://www.amazon.com/Alone-Together-Expect-Technology-Other/dp/0465010210/?tag=psychcentral
[75] Google Chairman Expects Internet To 'Disappear' Soon. http://www.techweekeurope.co.uk/e-innovation/eric-schmidt-google-internet-disappear-160126#UjAgqIeDdGhqtitb.99
[76] Man spooks nanny after hacking into baby monitor. http://kxan.com/2015/01/28/man-spooks-nanny-after-hacking-into-baby-monitor/
[77] Maldrone: Watch Malware That Wants To Spread Its Wings Kill A Drone Mid-Flight. http://www.forbes.com/sites/thomasbrewster/2015/01/27/malware-takes-down-drone/

[78] B.A. Trakhtenbrot, Ya.M. Bardzdin', "Finite automata. Behaviour and synthesis" , North-Holland (1973)
[79] R.L. Rivest, R.E. Schapiro. Diversity-Based Inference of Finite Automata. Journal of the ACM, Vol 41 No 3, May 1994, pp555-58
[80] Meet Lizard Squad, the hackers who ruined Christmas for gamers worldwide. http://www.dailydot.com/technology/lizard-squad-hackers/
[81] How to Make an ATM Spit Out Money. http://video.mit.edu/watch/how-to-make-an-atm-spit-out-money-147/
[82] Hackers Can Force ATMs to Spit Out Money With a Text Message. http://gizmodo.com/hackers-can-force-atms-to-spit-out-money-with-a-text-me-1551119933
[83] The EMC Digital Universe study 2014. http://www.emc.com/leadership/digital-universe/index.htm?pid=home-dig-uni-090414
[84] What Is the Memory Capacity of the Human Brain? http://www.scientificamerican.com/article/what-is-the-memory-capacity/
[85] Exclusive: Inside Google Spanner, the Largest Single Database on Earth. http://www.wired.com/2012/11/google-spanner-time/all/
[86] Spanner: Google's Globally Distributed Database. ACM Transactions on Computer Systems Vol. 31:3, August 2013, http://dl.acm.org/citation.cfm?id=2491245
[87] https://github.com/Laurelai/decompile-dump
[88] Bank Hackers Steal Millions via Malware. http://www.nytimes.com/2015/02/15/world/bank-hackers-steal-millions-via-malware.html?_r=0
[89] Ninety Five Percent of Carbanak Malware Exhibits Stealthy or Evasive Behaviors. http://labs.lastline.com/ninety-five-percent-of-carbanak-malware-exhibits-stealthy-or-evasive-behaviors
[90] Millennials Don't Care About Owning Cars, And Car Makers Can't Figure Out Why. http://www.fastcoexist.com/3027876/millennials-dont-care-about-owning-cars-and-car-makers-cant-figure-out-why
[91] Internet of things-connected fridge implicated in botnet spam email attack. http://www.itwire.com/business-it-news/security/67205-internet-of-things-connected-fridge-implicated-in-botnet-spam-email-attack
[92] Why We Should Design Some Things to Be Difficult to Use. http://www.wired.com/2015/02/on-the-joy-of-mastery/

# Ontology Design Pattern Extractions for Ontology Visualization

Pavel LOMOV [a], Maxim SHISHAEV [b]
[a] *Institute for Informatics of Kola Science Center of RAS*
[b] *Kola Branch of Petrozavodsk State University*

**Abstract**. This work continues investigation of a problem of ontology visualization focused on effective ontological knowledge transmission to an user. For achievement of this purpose the authors had offered to form on the basis of an ontology fragments special structures - cognitive frames (CFs) which allowed to represent concepts of ontology in a compact and complete visual image aimed at perception by user not familiar with ontological modeling. The general ways of OWL axioms interpretation as a sets of SKOS model elements for their adaptation for visualization by means of CFs are presented.

**Keywords.** ontology, semantic web, ontology visualization, cognitive frame.

## 1. Introduction

Visualization of ontologies become an important aspect of their practical use. Our investigation oriened on the ontology visualisation for knowledge exchange between users not familliar with ontology engineering. This work is continuation of the research devoted to ontologies visualization on the base of cognitive frames [1,2,3].

A cognitive frame (CF) is defined as a combination of a content - some ontology fragment representing a certain viewpoint [4] on a concept and a visual image constructed on the basis of the contents taking into account laws of human perception. Important task in this case is a formation of CFs content which is similar to a problem of viewpoint extraction from knowledge base. The solution of this task has some problems concerning ontologies described by OWL [5] on which visualization our research is directed.

An OWL ontology is a set of logical expressions (axioms) describing conceptual system. These axioms often have a difficult internal structure which isn't corresponding triplets (object - property - value). This fact makes difficult their parsing to define their accordance to viewpoints. Also it is necessary to make some transformations of such axioms before their addition to CFs contents to make them adapted to the creation of visual image fragment. It is also necessary to consider such circumstance that some axioms need to be analyzed jointly as they represent a realization of some ontological modeling trick.

For the solution of the given problems it is offered to extract viewpoints on the concepts drawing on ontology content design patterns (CDPs), which represent one king of the ontology design patterns (ODP) [6,7]. CDPs represent small ontology fragments for describing objects of subject domains - for example: "Task execution", "Description-Situation", "Time interval", "Sequence" and other. In our case sets of axiom corresponded to CDPs (pattern's instantiation) which were used for definition of some concept will be included in the contents of its CFs.

The paper is organized as follows: in section 2 the place of this work in the context of our research is considered, in section 3, the general procedure of extraction and interpretation of pattern's instantiations for their following visualization within CF is presented, the results and the directions of further researches are given in section 4.

## 2. Background

Our first work [8] was devoted to the problem of automatic obtaining a OWL-ontology representation adopted for the simplified visualization in node-link diagram form. Such representation was called User Presentation Ontology (UPO) and based on the SKOS model [9].

However it left open a question of what fragment of UPO should be showed to the end-user when he choose some concept for examination. Since showing of a whole set of relations and neighboring concepts often complicates perception. Along with that a node-link diagram, though is a universal way of visualization, but not always provide simple meaning interpretation of concept by an user

Therefore this circumstance forced us to refer to the notion of viewpoint [4]. And next paper [1] was devoted to visualization of UPO, corresponding to some domain ontology, based on CFs. The following general definition of a cognitive frame was given:

$$CF(t) = \langle CT, VS \rangle, \tag{1}$$

where $t$ - a target concept of cognitive frame; $CT$ - a content of the frame; $VS$ - a visual image formed on the basis of the content. The content is consist of a set of triplets in form "concept-relation-concept", which reflect a meaning of the target concept.

Also this work considered a CF generation procedure, which is based on such basis relations as "taxonomy", "partonomy" and "dependence". However such way of visualization is useful for visualization of ontologies with simple structure. So in the next paper [3] we considered formation of a CF (mainly its visual image) with using ODPs and principles of Gestalt psychology [10,11]. And that way of building visualization showed promised results during empirical evaluation, but a problem of ODPs extraction from an OWL-ontology to form CFs content has appeared to be more complex than we expected. So therefore in this work we examine it in more detail.

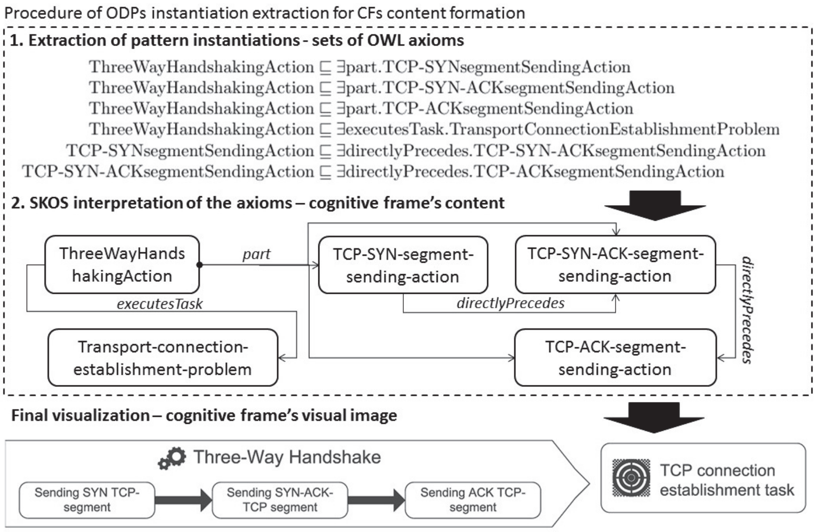



**Figure 1.** A general scheme of an ontology concepts visualization by means of pattern-based cognitive frames.

## 3. Procedure of ODP's instantiation extraction for CFs content formation

The main objective of a CF content formation is to obtain a set of pattern's instantiations related to some concept from ontology and their their representation in form adapted for building visual image (fig. 2). As a result of execution of such procedure axioms, belonging to the various pattern's instantiations , will be presented as a group of interconnected elements of the SKOS-model, underlying UPO. Under the notion "pattern's instantiation" we imply a set of axioms from an ontology corresponding to some ODP.

Further we will define the main types of axioms, in accordance to their syntactic structure and meaning for ODPs. Then we will consider the main stages of the procedure of their extraction from ontology and interpretation in form of a set of SKOS elements in UPO.

### *3.1. The main types of axioms in ODP's instantiations*

Pattern's instantiation can consist of axioms which may refer to the following types in accordingly a role which the axioms play:

- **Structural pattern axioms** - define relations between concepts in pattern's instantiation;
- **Semantic pattern axioms** - enrich a concept definition by formal semantics, which can be considered by a reasoning engine.

A group of structural axioms usually consists of **Simple class axioms** containing single classes on left and right sides (e.g. $InformationTransfer \sqsubseteq Process$) and **Simple restrictions class axioms**, containing a single property restriction, which consists of an object property and a named class (e.g. $Activity \sqsubseteq$

$\exists produces.Outcome$) or a datatype property and an elementary datatype (e.g. $Activity \sqsubseteq \exists hasStart.xsd{:}dateTime$). Such axioms can be simply parsed and interpreted as edges of a graph structure.

Semantic axioms, as a rule, have a complex class expression from right side. Therefore their syntactic analysis is more difficult, and their interpretation may require taking into account a meaning of a pattern.

### *3.2. Definition of used ODPs*

At this stage, detection of patterns used in an ontology for concepts descriptions is performed. During that an analysis of simple class axioms and simple restrictions class axioms is carried out. Any axiom which consists of elements which have labels or IRIs typical for some patterns, is linked to corresponding patterns name and the concept.

As a result a set of used patterns is defined as follows:

$$UsedPatternsSet = \{\langle Concept, \\ \{\langle `PatternName', PatternsInstantiation\,(`PatternName')\rangle\}\rangle\}\,, \quad (2)$$

where $PatternsInstantiation(`PatternName') = \{Axiom\}$.

The important subtask is taking into account inheritance of patterns. It assumes a linking with some pattern and its instantiation used for definition of a concept successors of the concept as well.

Notice that in this case for each axiom from inherited instantiation an existence of more specific axiom describing concepts-successor is checked. If it is true, then in a final inherited instantiation will be included the more specific axiom and not axiom from the initial inherited instantiation.

### *3.3. Interpretation and representation of pattern's instantiations in the UPO*

At this stage identifications of semantic axioms of pattern and a replenishment by them the instantiations created at the first stage is carried out. The main complexity is that semantic axioms of a pattern are used only for enriching a concept definition by formal semantics. Thus their syntactic contents only indirectly point on the concept described by them. Therefore for their detection and correct interpretation it is necessary to take into consideration a known structure and a semantic meaning of the patterns detected at the previous stage.

After that an interpretation of axioms contained in the obtained instantiations and their representation in the UPO are performed. A general interpretation of a pattern instantiation is a set of SKOS concepts connected by relations. The SKOS concepts and relations could have text annotations characterizing them. A formal definition of the general interpretation is as follows:

$$Interpretation\,(PatternInstantiation) = \langle\{InstConceptSet\}\,, \\ \{InstRelationSet\,(ConceptA, ConceptB)\}\,, \{AnnotSet\}\,, \rangle\,, \quad (3)$$

where $InstConceptSet \subseteq ConceptSet$, $ConceptSet$ - a set of all SKOS concepts in the UPO corresponding to classes of initial OWL ontology, $InstRelationSet = \{Relation_1 (ConceptA_1, ConceptB_1), \ldots Relation_n (ConceptA_k, ConceptB_l)\}$ - a set of the relations defined between SKOS concepts in some pattern instantiation, $ConceptA, ConceptB \in InstConceptSet$,
$AnnotSet = \{Annotation_1(SkosElement_1), \ldots Annotation_m(SkosElement_j)\}$, - a set of annotations for elements of the pattern instantiation, $SkosElement \in InstConceptSet \cup InstRelationSet$.

Further we will consider the general ways to interpretation of structural and semantic axioms of a pattern instantiation, and also some special cases.

### *3.3.1. Interpretation of structural axioms*

Structural axioms define relations between two objects of a subject domain presented by ontology classes or, in case of a restriction on datatype property, set some attribute value of a concept. Their interpretations are represented by named arches of a graph structure.

However it is often possible to meet more complicated ways of an attribute description. It allows to represent a subject domain in more detail. One of such ways is creation of a object property corresponding to some attribute, and a class defining its range. For example, for an attribute "weight" object property "has weight" and the class "Weight" are created. An interpretation in this case will mainly correspond to the general interpretation for structural axioms eq. (3). The main difference will be an addition of the annotation $DefinitionAnnotation$ specifying that this SKOS concept is a range. Subsequently this fact will be taken into account during a visualization.

More difficult way of an attribute representation is its definition in the form of a class, but not a property, and connection by means of some object property with other class representing a range:

$$Class \sqsubseteq \exists hasAttribute. (AttributeClass \sqcap RangeClass) \tag{4}$$

A concrete name of the object property pointed on association of the attribute (property - *hasAttribute* ) is depend on concrete ontology. It is used for an identification of such axioms at the pattern detection stage. An interpretation of such type of axioms corresponds to the general one - eq. (3). The difference is that the property of affiliation attribute is ignored, and the attribute class is represented as the relation between SKOS concepts - the attribute carrier and the range class.

Besides the considered nuances there is one more model trick which is connected with using of object properties. It is called "Covering and closure axioms"[12]. It allows to restrict a range of some attribute in ontology to only one class. For this purpose two axioms for one object property are defined. One of them contain an existential ($\exists$) restriction and the second contains an universal one ($\forall$):

$$Class \sqsubseteq \exists Property.RangeClassA \tag{5}$$

$$Class \sqsubseteq \forall Property.RangeClassB \tag{6}$$

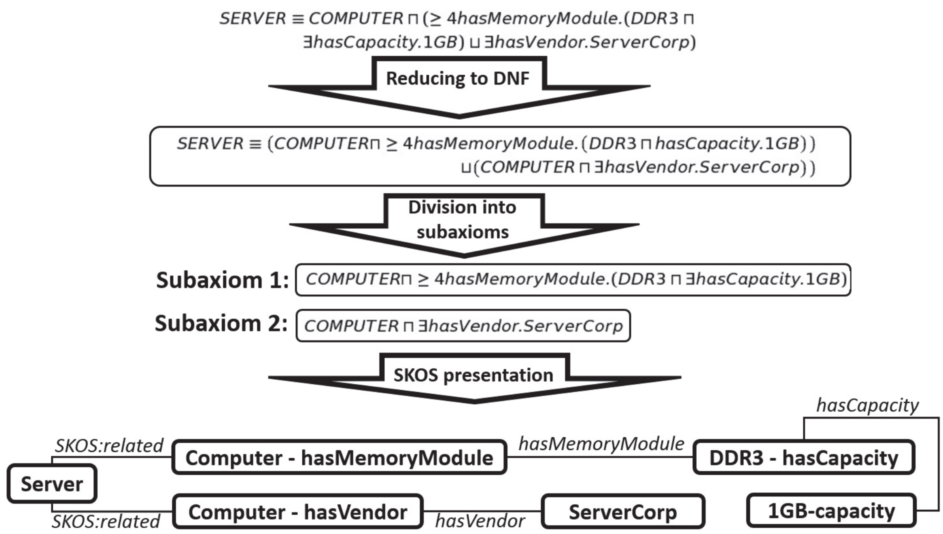



**Figure 2.** A general interpretation of a semantic axiom.

In this case the axiom with the existential restriction is interpreted according to the general interpretation eq. (3), and other axiom is ignored. And the narrowest class from restrictions became a range of the property. That is if it is true that $RangeClassA \sqsubseteq RangeClassB$, then $RangeClassA$ goes to the interpretation as a range and vice versa. Notice that usually both axioms are defined for the one class, however in more difficult cases one of them can be inherited from super classes.

### *3.3.2. Interpretation of semantic axioms*

Further we will consider ways for semantic axioms interpretation. It should be noted that in some cases each such axiom can be considered alone and in another - they may form a group so that is necessary to interpret them together. In the latter case such axiom combination could be regarded to a logical design pattern which is used within a ODP.

A general way for semantic axioms interpretation is presented in Fig. 2. At first a right side of an axiom containing complex class expression reduced to a disjunctive normal form. Then each conjunct, called "Subaxiom", is reduced to a negation normal form [13]. Such transformations allow to simplify the following syntactic analysis and construction of annotations. The obtained subaxioms are represented as a set of SKOS concepts and are linked by means of the $SKOS$ : *related* relation with the SKOS concept, designated a class from the left side of the axiom.

Thus if the analyzed subaxioms consist of restrictions on properties, then their fillers are also transformed and broken down into subaxioms. For example, the restriction - $4 \geq hasMemoryModule.(DDR3 \sqcap \exists hasCapacity.1GB)$ in the subaxiom 1 (fig. 2), broken down into one subaxiom - $DDR3 \sqcap \exists hasCapacity.1GB$. As a result the obtained subaxioms are also presented as a set of SKOS concepts, which are linked with SKOS concept of the parental subaxiom by a relation from the restriction. Further for the obtained subaxioms (in case of internal complex restrictions) the process will be repeats recursively. As a result a set of related SKOS elements which reflects a structure of an axiom will be obtained.

Among semantic axiom there are so called general concept inclusion axioms (GCI axioms). These axioms are characterized by complex expressions from the left side. Therefore they have no obvious relation to some class. GCI axioms usually are used to achieve necessary results of automatic classification of concepts. From the point of view of a visualization in this case they present same information as structural axioms but in more strict form, that allows to influence process of logical inference. However there are situations in which these axioms could brings new elements in a visual image of a concept.In this case it is necessary to carry out their analysis and interpretation.

One of such situations is existence of GCI axiom which includes a class expression on the left side and one named class on the right side (e.g. $\exists hasStart.\top \sqcap \exists hasEnd.\top \sqsupseteq FixedActivity$). It allows an inference engine to classify individuals and classes satisfying to the expression from the left side as members and subclasses of the class from the right side. Such GCI axiom could be considered as a set of **formal** sufficient conditions to be class member. In such case GCI axioms may have the same meaning as structural ones but sometime they could contain new conditions which need to be taken into account.

In this case an interpretation is made according to the general scheme for semantic axioms. The differences consist that the left side of an axiom is exposed to the analysis and at breakdown on the subaxioms, which are compared to the right side of structural axioms. If they bring new information (logically don't follow from the right side expressions of the structural axioms), their analysis and/or interpretation as a set of SKOS concepts are executed. An interpretation of GCI axioms having another structure need to be considered within concrete patterns.

## 4. Conclusion

In this work we considered a problem of ODP's instantiations extraction from OWL-ontologies and their interpretation and representation in the UPO. For this purpose the procedure was offered. It includes detection of patterns used in ontology, parsing of an OWL axioms and their interpretation as a set of connected SKOS concepts. Also the special cases of an axioms interpretation caused by some widespread methods of ontological modeling were also considered. Pattern's instantiations presented as a result in the UPO form CF's content for concepts of an initial OWL ontology

The main directions of further researches are definition of the graphical notation for visualization of widespread ODPs and development of the software tool for execution of automatic formation of a set of CFs for some domain ontology and its visual representation.

### *4.1. Acknowledgments*

The work was supported by the Russian fund of basic researches (RFFI) - grant 15-07-03321A.

## References

[1] P. Lomov and M. Shishaev. Technology of ontology visualization based on cognitive frames for graphical user interface. In Pavel Klinov and Dmitry Mouromtsev, editors, *Knowledge Engineering and the Semantic Web*, volume 394 of *Communications in Computer and Information Science*, pages 54–68. Springer Berlin Heidelberg, 2013.

[2] P. Lomov and M. Shishaev. Visualization of ontologies on the basis of cognitive frames for knowledge transmission. In Y. Kiyoki. B. Thalheim, H. Jaakkola, editor, *Proceedings of the International Conference on Information Modelling and Knowledge Bases (EJC 2014)*, volume Number 2014/2 of *Kiel Computer Science Series*, pages 339–347. Kiel University, 2014.

[3] P. Lomov and M. Shishaev. Creating cognitive frames based on ontology design patterns for ontology visualization. In Pavel Klinov and Dmitry Mouromtsev, editors, *Knowledge Engineering and the Semantic Web*, volume 468 of *Communications in Computer and Information Science*, pages 90–104. Springer International Publishing, 2014.

[4] Porter B. Acker, L. Extracting viewpoints from knowledge bases. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 547–552. San Francisco, CA: Morgan Kaufmann, 1994.

[5] Ian H. Boris M. Parsia B. Peter F.P. Sattler U. Bernardo, C.G. Owl 2: The next step for owl. *Web Semant.*, 6(4):309–322, November 2008.

[6] A. Gangemi. Ontology design patterns for semantic web content. In Yolanda Gil, Enrico Motta, V.Richard Benjamins, and MarkA. Musen, editors, *The Semantic Web, ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 262–276. Springer Berlin Heidelberg, 2005.

[7] Presutti V. Gangemi, A. Ontology design for interaction in a reasonable enterprise. In P. Rittgen, editor, *Handbook of Ontologies for Business Interaction*, 2007.

[8] Dikovitskiy V.V Lomov, P.A. Shishaev M.G. Owl-ontology transformation for visualization and use as a basis of the user interface,. *Design Ontology*, pages 49–61, 2012.

[9] Skos simple knowledge organization system reference, 2009.

[10] K. Koffka. *Principles of Gestalt psychology.* Harcourt, New York, 1935.

[11] W. Kohler. *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology.* Liveright Publishing Corporation, 1992.

[12] R. Stevens. Closing down the open world: Covering axioms and closure axioms., 2011.

[13] Voronkov A. Robinson, A. *Handbook of Automated Reasoning.* Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, 2001.

# Abstract handling of Information and Knowledge

Enn TYUGU
*Institute of Cybernetics of the Tallinn University of Technology*

**Abstract.** We discuss main formalisations of information and knowledge available in literature. Knowledge is considered as a meaning of data for an actor who can understand and use it. The actor can be a computer, a robot, a human being or even a virus. We make an assumption that the usage of knowledge depends on material data carrier only as much as the latter allows one to perform operations and restricts knowledge processing through capacity and time. This permits one to consider knowledge abstractly. Knowledge handling requires some tools that are called knowledge systems. These systems have different forms and they can be connected in various ways as it follows from the analysed publications. A metrics of knowledge is discussed that reflects the capability of solving problems.

**Keywords.** Knowledge, knowledge system, modularity of knowledge, metrics of knowledge.

## 1. Introduction

We present here a survey of formal results obtained on fundamental level about information and knowledge. These works have not been in the focus of the researchers and practitioners that work in the field of knowledge-based systems. We expect that this survey will be useful for researchers and software devlopers, in particular, facilitating reasoning about software on architectural and knowledge level.

It has been difficult to extend the concept of information so that it would cover also the essence of knowledge. Beside that, the words *information* and *knowledge* are being often used as synonyms, and this adds extra difficulties to reasoning about knowledge and information. The present work is an analysis of literature on definitions and properties of knowledge, as well as on some mathematical tools for handling it.

For both concepts, we have to deal with finding a meaning for a sign. The signs are objects or phenomena that are rather easy to observe, but meanings can largely vary, and may be hidden from the observer. For instance, listening to music can be considered as an act of communication, where a musical piece as a sign may have emotions as its meaning.

The case of extracting knowledge is complicated. Beside the probability of getting a message with a particular information content, an important role belongs to background knowledge of the receiver. Also the ability of the receiver to “understand the content of the message,” i.e. the ability to assign a meaning to the received text, and to use the received meaning in a useful way is important. It is obvious that the concept of knowledge is related to semiotics, language semantics and mathematical logic. The results in knowledge understanding could be useful also in psychology, artificial intelligence and data mining.

One of the first essential writings about the knowledge belongs to Gottlob Frege from the nineteenth century [1]. He is known as the author of the first formal logic system, and a pioneer of the research in semantics of natural languages.

The present paper is composed as follows. The following two sections introduce a common conceptual basis, and present main assumptions for the discussion. The fourth section discusses briefly a role of background knowledge. The fifth and the sixth section constitute the central part of the paper, where a survey of various knowledge systems and their connections are discussed. Deductive and connectionist approaches to knowledge handling and metrics of knowledge is discussed in sections 7 and 8, followed by a brief summary of the paper.

## 2. Information, knowledge and data

We will try to create a common conceptual basis for the phenomena related to knowledge in the first place. A central concept is the concept of data. Independently of how information and knowledge are defined, we have a good reason to think that both can exist only on a material carrier. More precisely, one can say that, in the real world, information and knowledge are contained in data on some data carrier, where the data can occur in a form of any objects which can be distinguished as notations or signs of something. The information may come from data occurring as texts, pictures, music, electrical signals and so forth. One can always make an abstraction, and encode data in a binary form — as a bitstring, using as many bits as required by the formula for information content given in the beginning of this text. Transformation from physical phenomena to abstract data in the form of bitstrings may be a rather mysterious process that we will mainly ignore in this text. Let us note that a source of information can be some social process or a biological phenomenon, and this makes the process of creation of abstract data even more difficult to understand. We will consider data from the moment when they are explicitly representable as bitstrings. We can assume in particular that states of neurons and signals in a neural network belong to this kind of data as well[1]. One has to bear in mind that this is only a theoretical model. The amount of required resources may prevent even approximate simulation of the brain processes today — constraints of the material world determine the limits of information processes.

## 3. Information processes depend on the material world only through the constraints

It is highly probable that the laws of information processing (as well as of any kind of knowledge handling) do not depend in principle on the material carrier. The latter determines constraints on the speed, capacity and applicable operations, but this is all the influence. In other words, as soon as we fix the set of possible operations and will not violate the limits on the processing speed and amount of information, we can handle information processes without taking into account the material world, i.e. the informa-

[1] This is a very strong statement. Its consequence is that there is no need to use a help of any kind of fluidum or a spirit in explaining knowledge, thinking and consciousness. However, from this statement does not follow any real possibility of explaining complex psychological phenomena without using higher level concepts that must come from bio- and neurosciences.

tion carrier and means of performing the operations[2]. Possible operations play the central role. For instance, it is expected that quantum computing will provide a set of very powerful operations on quantum information carriers — qbits. As soon as the quantum computing becomes a reality, todays encryption methods will be obsolete, because quantum computers will be able to break the encrypted code easily.

A set of operations considered in theoretical computer science, and especially in logic, is very restricted. On the other side, one can not even determine the set of applicable operations in neural science today. One can rather speak of the use of operations there: understanding speech, processing visual information, etc. There exists a principal difference between the computer science and software science from one side and the neural science from the other side. A set of possible operations is well determined for the first, and rules of composing new operations are also determined, but it is not the case in neural science. Even a single neuron performs complex processes, and the neurons perform in parallel to a large extent. This makes understanding of complex operations in neural nets very difficult.

## 4. Background knowledge and beliefs

Although a receiver must be always present in communication, its role is restricted to acquiring data and extracting information from the data. In the case of knowledge usage, a knowledge handler has a more complex role. This person or system receives, stores and uses knowledge. It is assumed that it may have *background knowledge* obtained earlier. The background knowledge is needed both for extracting knowledge from data and using it. A role of background knowledge and its usage in acquisition of new knowledge has been considered by Fred Dretske [2]. First of all, Dretske distinguishes knowledge and *belief*. Let us have a statement $p$, for instance "it rains." Dretske writes that a person knows that $p$, if he believes that $p$, and $p$ is true in the reality. Then this belief bears information that $p$. Dretske describes the growth of knowledge in the following way. Let a person have some background knowledge $k$, and he or she receives a message that $r$ is $F$. This includes knowledge that $s$ is $G$ only, if the probability that $s$ is $G$ equals 1, and was less than one in the case of knowledge $k$ only. Dretske postulated also the Xerox principle (transitivity) of knowledge: if $r$ is $F$ bears knowledge that $s$ is $G$, and $s$ is $G$ bears knowledge that $t$ is $H$, then r is $F$ bears knowledge that $t$ is $H$.

## 5. Knowledge and knowledge systems

In order to do something with knowledge, one has to have some tools. In this way we come to an abstract concept of a knowledge system that includes tools for knowledge handling. We can call this system anthropomorphically also a knower or a thinker who knows how to use knowledge.

---

[2] We have made here an assumption that the information processes consist always only of operations from a well defined set, and the processes can be performed concurrently in time.

*5.1. Notation-denotation relation.*

The simplest concept of knowledge is presented by Peeter Lorents [3]. He defines an element of knowledge as a pair *(notation, denotation),* or symbol and its meaning in other words. A knower is not explicitly present in this case, but a *notation-denotation relation* is used that binds a denotation, i.e. a meaning with each notation. A knower is implicitly present as a bearer of the notation-denotation relation in this case. One can also say that a knower understands the language whose syntax determines possible notations and semantics provides their meaning. A denotation may be a notation of another piece of knowledge, therefore, one can build hierarchies of knowledge. In the decreasing direction, these hierarchies sooner or later end with meanings that are not denotations of anything else. These are *fundamental concepts*, not describable by anything else. This approach has a drawback that it does not include any operations with knowledge.

*5.2. Infomorphisms.*

John Barwise and Jerry Seligman [4] introduce a similar concept, but they add also means for working with knowledge. They are well-known logicians from the Stanford University. They present knowledge abstractly by means of classifications. A classification $\boldsymbol{A}$ is a pair of sets, where one is a set $\Sigma_A$ of *types*, the other is a set $M_A$ of *tokens*. Tokens and types are bound by a classifying relation $\vDash_A$, that provides types for tokens. Hence a *classification* is presented as ($M_A$, $\Sigma_A$, $\vDash_A$). Tokens have the role of meaning, and types have the role of notations, although this analogy is not precise. Classifications are bound by mappings that constitute *infomorphisms*. An infomorphism between two classifications $\boldsymbol{A}$ and $\boldsymbol{B}$ is a pair of functions $f' : \Sigma_A \rightarrow \Sigma_B$ , $f'' : M_B \rightarrow M_A$, that conform to the following constraints: $f''(b) \vDash_A t_a$ iff $b \vDash_B f'(t_a)$ for all tokens $b$ of $\boldsymbol{B}$ and all types $t_a$ of $\boldsymbol{A}$, see Fig. 1. This infomorphism is denoted by $\boldsymbol{A} \rightarrow \boldsymbol{B}$.

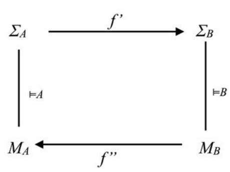



**Figure 1.** Properties of infomorphism

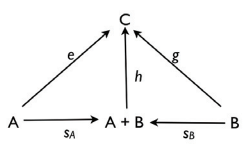



**Figure 2.** Example diagram of infomorphisms

Classifications with infomorphisms constitute networks that can be visualised as diagrams where infomorphisms are arrows between nodes denoting classifications. Fig. 2 represents an example from the book [4]. It describes infomorphisms of a sum of classifications and visualises a property of the sum that, for any infomorphisms $e$ and $g$ with domains $\boldsymbol{A}$ and $\boldsymbol{B}$, and codomain $\boldsymbol{C}$, there exists a unique infomorphism $h$ such that the diagram in Fig. 2a commutes. Infomorphisms allow us to describe information flow in larger systems. In essence, a language of categories is used here.

A classification defines a language, for instance, it can be a language of a logic. In this case, types are formulas of the logic, and tokens are semantical objects (models). One can bind a *local logic* with a classification. This enables one to formalise local handling of knowledge in more detail. It is essential that infomorphisms between classifications create infomorphisms also between the local logics.

*5.3. State transition systems and deductive systems.*

One more representation of knowledge was given by Sergei Maslov [5] even before the work of Barwise and Seligman. Maslov assumes that one works with knowledge by making inferences step by step, expanding gradually what is known. In the most general setting this looks as follows. The whole amount of knowledge is considered as a *state* of a system that works with knowledge. For instance, it can be a state of a neural net. Using knowledge, one applies *rules* from a given set of rules. A rule enables one to derive a new piece of knowledge from some given pieces. One can continue this process as long as the required result is obtained, or nothing new can be derived. Sometimes this process just does not stop, and time or space resources will be exhausted. This very general scheme can be made more detailed in various ways, depending on the specific features of a knower.

It is natural that Maslov as a logician uses deductive systems as knowledge systems. In this case, knowledge is represented by texts (words), and new knowledge appears as new words derived according to inference rules from given words. Meaning of a word is given by interpretation given for a deductive system. When a new word is derived, its meaning is also constructed. The most general form of a deductive system is *canonical Post's system* [6]. Post's systems are in some sense minimal and universal deductive systems. From the other side, any system of formal logic can be easily encoded, i.e. presented as a canonical Post's system.

The formalisation of a knowledge system as a deductive system has a sense, if we assume that knowledge can be encoded in the form of texts, these can be bitstrings, as we have already earlier agreed. There are several examples of deductive systems for handling knowledge in economy, technology etc. in the Maslow's book.

Applicability of knowledge languages can be also described by means of Maslov's formalism. One defines a language for representing knowledge, and develops a program to handle knowledge written in this language. This constitutes a deductive system. Some failures of knowledge languages can be explained by the fact that one tries to develop as expressive language as possible[3], but one will be unable to develop a program that could handle knowledge sufficiently well. This was especially true for early knowledge languages like KRL [7], FRL [8]. The efficiency of automated handling has been taken into account in the versions of description logic [9] used for representing knowledge about web services.

## 6. Modularity of knowledge and networks of knowledge systems

Marvin Minsky was probably the first to present the idea of modularity of knowledge [10], and this idea is being used in multi-agent systems, that are systems with some intelligence. If one considers local logics of Barwise and Seligman as knowledge systems, then one can see that even these constitute networks of knowledge systems.

The idea to use a network of knowledge systems for describing architecture of intelligent software (of artificial knowledge based systems) is presented in [11], and implemented as a set of graphical notations for describing a knowledge architecture. One assumes in this case that knowledge systems are local, but they can be connected in

---

[3] Strictly speaking, every language has a purpose to represent knowledge. Here we still speak of languages that are intended for presenting knowledge to the programs that we consider as knowledge based software (see, for instance, Tyugu,Yamaguci [12]).

various ways: hierarchically, operationally, semantically, etc. Let us consider a hierarchical connection. We use a set $S$ of notations and a set $M$ of meanings of a knowledge system, together with a notation-denotation relation between these sets, like in [3]. Beside that, rules must be given for manipulating knowledge in a system, like in [5]. Let us have two knowledge systems with the sets of notations and meanings denoted as $(S,M)$ and $(S',M')$. A *hierarchical connection* between these systems is provided by mappings $h: M \rightarrow S'$ and $h': S' \rightarrow M$ between the meanings $M$ of one system and notations $S'$ of the other system. If $h$ and $h'$ are one-to-one mappings, we speak about a *strong hierarchical connection*. An example of hierarchical conection of knowledge suystems is deductive program synthesis. The upper system is logic and the lower system is the knowledge system of computations that is constituted by a computer and its system software[4]. The connection means that computational meanings of formulas of logic of upper system are transformed in programs that are notations of computations in the lower system. These computations are the meanings in the lower system.

Knowledge systems are connected semantically if their sets of meanings have a common part. This is so in the case of translation from one language in another. Syntax of the languages, i.e. the notations of meanings are different, but sets of meanings must sufficiently overlap, a translation is possible only in such a case. Even machine translation must use the semantic connectedness, otherwise the translation will be very poor.

## 7. Deductive and connectionist knowledge systems

Knowledge systems can be divided in two large classes. All knowledge systems considered above belong to *deductive knowledge systems*. They have well defined syntax for texts that denote meanings, and their derivation rules are explicit as well. Their strength is a possibility of computational realization in a straightforward way. A well known proponent of these knowledge systems has been John McCarthy from the Stanford University.

Another class of knowledge systems are *connectionist knowledge systems,* where knowledge is represented by networks of concepts and entities. These systems are suitable for approximate representation of complex concepts. A concept is described by a semantic net, where nodes are other concepts or objects, and they are bound by associations — arcs representing relations between them. As concepts in the nodes of a semantic net are described by semantic nets as well, it is possible to move along the associations far away, and to make interesting conclusions on the way. One can derive new knowledge in this way. A proponent of this approach has been Marvin Minsky.

It is easy to show how to present any connectionist knowledge symbolically, as a text. Every arc of a net, let us say — the one that goes from a node $x$ to a node $y$ and represents a relationship $f$, can be denoted by a triplet $(x,f,y)$. A collection of triplets of all arcs of a semantic net contains precisely the same information that the net. A knowledge representation language RDF (Resource Definition Format) [13] is in essence a language of triplets from above. RDF permits to include additionally hyperlinks in triplets that provide information about their implementation. In order to use a semantic net in a computer, one has to write a program that works with the net. One can consid-

---

[4] Various abstract knowledge systems of comuptations have been investigated in the first half of the 20th century: Turing machines, Church's lambda calculus, Post's machines, Markov's normal algorithms. They all are universal, and in this sense equivalent systems. A computer with system software is in principle also a universal system of computations, but in this case one has to take into account the material restrictions, as we have discussed it above.

er this program as a representation of rules of knowledge handling that adds a deductive flavour to the implementation.

An example of knowledge representation that is on the border between the connectionist and deductive knowledge systems is given by ontologies. Generally speaking, an ontology means knowledge about some application domain represented as a set of concepts and relations between them. Ontology has to be presented in an ontology language that is a knowledge representation language, hence we have here the problem of contradicting requirements of expressiveness and efficiency of realisation as we have discussed it above. An ontology together with respective software constitutes a knowledge system in our sense.

One can specify connections between the ontologies and build networks of ontologies. The paper [14] describes operations on knowledge systems, lattices of knowledge systems and their relatedness to ontologies. First, knowledge representation is described by a semantic system -- an algebraic system with notation-denotation relation. Semantic systems are closely related to ontologies. They are pre-ontologies, i.e. ontologies without an inference mechanism. A central concept of the work [14] -- the concept of knowledge system has been defined in two different ways. First, following the ideas of S. Maslov, it has been defined as a free deductive system with interpretation. Second, it has been defined as a semantic system supplied with derivation rules, i.e. as an ontology. Both definitions give the same result. A lattice of consistent semantic systems is described in [14]. Knowledge systems generated by consistent semantic systems also constitute lattices that are related to lattices of goals defined on the knowledge systems through solvability of goals in a rather natural way: if a goal $g$ is solvable on a knowledge system $K$, then it is solvable on any knowledge system $K'$ such that $K < K'$; if a goal $g$ is solvable on a knowledge system $K$, then any goal $g' < g$ is solvable on $K$.

## 8. Measuring knowledge

Knowledge has a role in informatics analogous to the role of energy in physics — the more one has it the more one can do. Therefore one should be able to measure knowledge as well, like one does it with energy. Energy can be measured using calories, electron-volts, kilogram-meters etc., and transformations between these units are well known. However, we don't have a generally accepted measure of knowledge. Information can be measured with bits and bytes, but these units are not suitable for measuring knowledge, because they express only the diversity, but not the capability that should be a meaning of metrics of knowledge. Still, we are not quite correct when we say that one does not measure the knowledge. In particular, one measures knowledge of students at school. Unfortunately, this metrics is not universal, and it is quite imprecise.

Let us make an attempt to introduce a metrics of knowledge, based on the work [14]. We introduce a concept of a set $P$ of atomic problems. An atomic problem can be any problem that is expected to have a solution representable in the form of data, and can not be divided into smaller problems. We define a problem as a finite set of atomic problems. One can define partial order between the problems by using the set inclusion. Even more — problems defined this way constitute a boolean lattice. We can take the set of solvable problems as a measure of knowledge. The more problems are solvable the more knowledge one has. Now we have a metrics for knowledge, but a possibility to use it depends on how we can define the set of atomic problems, and how we can determine the set of solvable problems. The work [14] contains analysis of ordering of knowledge systems depending on their sets of notations, meanings and rules. It demonstrates the dependency of solvability of problems on this order. More precisely A relation between between a lattice of problems (called goals there) and a lattice of knowledge systems is presented, see the section above. Only relative capability of problem solving can be estimated on the basis of solvable problems, because the atomic problems can be of different complexity, and this is not taken into account. Not all problems are comparable, because only a partial oder exists. Still, in the case of finite set of atomic problems this approach may be useful. In the case of infinite set of atomic problems one can try to introduce a measure on the set of problems, find a finite approximation of this set and use it for measuring the knowledge.

## 9. Summary

We have considered knowledge as a meaning of data under the assumption that the usage of knowledge depends on material data carrier only as much as it allows one to perform operations and restricts knowledge handling through capacity and time. This permits us to consider knowledge abstractly, and to give broad definitions applicable both for biological and technical systems. In the most general case, knowledge is useful and valid information. Knowledge usage requires some tools that we call knowledge systems. These systems have different forms and they can be connected in various ways. The abstract works on knowledge come mainly from logicians and computer scientists. An interesting task is measuring knowledge. We have made only the first steps in this direction taking into account usefulness of knowledge.


### Acknowledgements

This research has been supported by the project No. 3.2.1201.13-0026 Model-based Java software development technology and the institutional research grant IUT33-13.


## References


1. G. Frege. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*. NF 100, 1892, 22 - 43.
2. F. Dretske, *Knowledge and the Flow of Information.* MIT Press, Cambridge, Mass. 1981.
3. P. Lorents. Formalisation of data and knowledge based on the fundamental notation-denotation relation, In: *Proc. IC-AI´2001, v. III.* CSREA Press, 2002, 1297 – 1301.
4. Barwise, J. Seligman. *Information Flow: The Logic of Distributed Systems*. Cambridge University Press, Cambridge, 1997.
5. S. Maslov. *Theory of Deductive Systems and its Applications.* MIT Press, 1987.
6. E. Post. Formal Reduction of the General Combinatorial Decision Problem. *American Journal of Mathematics.* 65, No. 2. 1943, 197 - 215.
7. D. Bobrow, T. Winograd. An Overview of KRL: A Knowledge Representation Language. *Cognitive Science* 1. 1977, 3–46.
8. R. Roberts, I. Goldstein. *The FRL Primer*, 1977
9. F. Baader et al. *The Description Logic HandbookTheory, Implementation and Applications.* Cambridge University Press, Cambrige, 2010.
10. M Minski. *The Society of Mind.* Simon & Shuster Paperbacks, New York, 1986.
11. E. Tyugu. Understanding Knowledge Architectures. *Knowledge-Based Systems*. 2004, No. 3, 1 - 7.
12. E. Tyugu, T. Yamaguchi (editors). *Knowledge-Based Software Engineering.* Frontiers in Artificial Intelligence and Applications, vol. 140. IOS Press, 2006.
13. M. Rouse. Resource Description Framework (RDF). *http://searchsoa.techtarget.com/definition/Resource-Description-Framework*
14. P. Lorents, E. Tyugu. Lattices of Knowledge Systems. In: Proc. *IC-AI´2001, v. III.* CSREA Press. 2009.

# Privacy-Enhancing Technologies: Privacy Wallet

Frank KRAMER [a] Oliver JÄGER [b] Bernhard THALHEIM [a]
[a] *Information System Engineering, Christian-Albrechts-University at Kiel, Germany*
[b] *NTT DATA Deutschland GmbH, Hammerbrookstrasse 89, 20097 Hamburg, Germany*

**Abstract.** Sharing private data between various mobile users is a daily occurrence nowadays. The owner has the right of a self-determined and policed usage of his private data by foreign holders. To ensure and improve the privacy of our personal data, we need privacy-enhancing technologies that secure a privacy-friendly and lawful processing of personal data. We develop an approach that allows an owner to store and share his private data on his smartphone and personal computer to gain full control over the usage of his private data. Therefore, we develop and implement a flexible system we call *privacy wallet*. This system uses modern encryption standards and a peer-to-peer architecture to realize data privacy.

**Keywords.** privacy, communication, security, data storage, data management

## 1. Introduction

Today, the self-determination of personal data must be taken under consideration of privacy. For example, the scientific article [1] shows what happens, if we abandon our data privacy. It describes how easy it is for someone, to get information about a person only by scanning an image of this person. To preserve and improve the privacy of our personal data we need privacy-enhancing technologies (PET). Beside the minimization and avoidance of identifiable data of a user Fischer-Hübner describes in [6] that PET can also secure a privacy-friendly and lawful processing of personal data. This includes the control of the personal data that covers, on the one hand, the encryption of the data, and on the other hand, the best possible logging of all usage of the data. Furthermore, a system for controlling personal data must be easy to use for people who want to protect their privacy. Since every user has his own personal demand for data control and privacy our system gives the user only the technology to protect his data. The user still has to decide which persons and organisations get access and usage rights to his personal data.

To evolve such a system to secure private data, it is important to distinguish between the owner and the holder of data. A person is the owner of data, if he has created this data. Therefore, an owner can read, change and delete his data at will. Furthermore, only with his permission other persons can use his data and the owner is informed about how his data is used and changed from other persons. A person who uses the data of an owner is a holder of data. He can only use the data with the permission the owner has given him for the data. Moreover, the owner must inform the holder, if he has changed something on his data. Hence, a system is needed that allows an owner permanent control of his private data against a holder.

This paper presents an approach to achieve such a system for data control. Therefore, we develop and implement a concept called *privacy wallet*. Every owner gets his own wallet to store his private data inside. To secure the data in the wallet, all data is only stored encrypted and solely the owner knows the key. Furthermore, the *privacy wallets* are connected within a network. The owner of the data can exchange his data with other users within the network. For every exchanged data the owner defines the permissions a holder can get. Additionally, all usage of the holder is logged for the owner. Consequently, an owner is informed permanently which holder uses his data in what way. In our approach, section 2 will first present the basic technologies of the *privacy wallet*. This includes Peer-to-Peer networks and secure communication. Then, section 3 will present the functionality of our *privacy wallet*. We will present our six main processes that are realized in our approach. Section 4 will then describe, how we realize our *privacy wallet*. This covers solutions for user management, user groups, ground functionality, encryption and data storage. Moreover, we also present our prototype for a smartphone application. In section 5, related work will be presented. We will take a look at other existing solutions to secure private data. Finally, a conclusion and a short outlook on future research will be given in section 6.

## 2. Technologies of the Privacy Wallet

Before we describe our approach of a *privacy wallet*, we will take a short look at the design of peer-to-peer systems and the possibilities for a secure communication and data transfer in such a system. Both technologies are core components within our *privacy wallet* concept.

### *2.1. Peer-to-Peer Communication*

This section describes the characteristics of a peer-to-peer system and the advantages over a client-server system. In [10] a peer-to-peer system is described as a self-organizing system that is composed of autonomous units called peers. All peers have equal rights in the system and all peers use their resources and services completely decentralized. This brief description can be extended by the following characteristics of a peer-to-peer network as described in [15]:

- All relevant operating capital is evenly distributed and used from the peers in the network.
- A set of peers use the operating capital that the peers within the group offer to give up.
- Peers are spread throughout the world and are connected through a network.
- Peers have a variable connectivity. They can not be found on the same address in the network.
- Peers interact with other peers directly to use their operating capital.
- The usage and transfer of operating capital takes place directly between the peers.
- Peers have both client and server functionality.
- A peer has the complete autonomy on his operating capital.
- Finding resources does not need a centralized service.

An ideal system has all these characteristics. But in most cases only a subset of these characteristics can be found in a peer-to-peer system. Using a peer-to-peer system offers some advantages compared with a client-server system. First, it has a very good scalability because the number of peers depends on the network capacity and not on the capacity of a server. Moreover, a peer-to-peer system is a reliable and secure system. Servers are more predestined for attacks, such as distributed denial of service attacks or unauthorized access on data because a server is a single point of failure. Such attacks on a peer-to-peer system are more complex or bring only a small yield for an attacker. Last, a peer-to-peer system is more flexible for integrating new services in the system. Furthermore, these new services have high quality of service. On a server a new service must be installed and tested over time and the server is not available during this period. In a peer-to-peer system, a peer can offer a new service and other peers can use it without loss of availability of the system. For our approach, we will use such a peer-to-peer system for the communication within the *privacy wallet*. This will be shown in section 4. In a next step we want to show the existing possibilities for a secure communication.

### *2.2. Securing Communciaction*

In this section, we take a look at how communication can be secured. As described in [13] there are two different approaches for secure communication that are shown in figure 1.

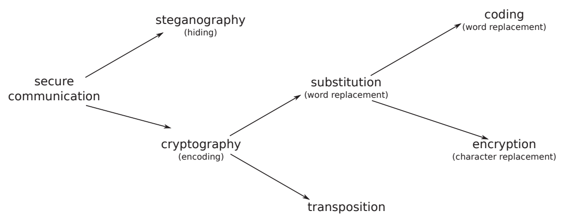



**Figure 1.** Secure Communication Methods [13]

Firstly, steganographic methods can be used to hide information from an attacker. Therefore, the information is hidden within a picture or a text. Only if a person knows how the information is hidden, the person can extract the information from the picture or text. Secondly, information can be encoded with a special key. If one piece of information is encoded, only persons who know the right key can decode it. The most popular methods to encode information are symmetric and asymmetric encryption methods. A symmetric encryption uses a single key that a sender as well as a receiver know. The sender uses the encryption algorithm with the key to encrypt his information. The receiver can decrypt the information by using a decryption algorithm and the same key as the sender. Today, the AES algorithm is an example for such a symmetric encryption [3]. Symmetric encryption pose a problem. Both, sender and receiver must know the key. So, they have to interchange it in a secure way. To avoid this problem, asymmetric encryption can be used. For this, sender and receiver have a public and a private key. The public keys are published to everyone who wants to know it. The private key is only known by the owner. Hence, the sender can encode the information with the public key of the receiver. The receiver can use his private key to decode the information. The RSA algorithm is an example for an asymmetric encryption [11]. The problem with using asymmetric encryption is the maturity of the algorithms. Therefore, we will use a hybrid of asymmetric and symmetric encryption for the *privacy wallet* to minimize the disadvantages of the encryption methods. After describing the relevant technologies for our approach, we will present the range of functionality of our *privacy wallet* in the next section.

## 3. The Privacy Wallet Functionality

This section presents the functionality of our *privacy wallet* approach. There are six different business use cases for the *privacy wallet*. Each business use case can be regarded as an independent process. All transfers between pairs within a process are asynchronous transfers. We will describe each business use case. All these use cases can be modelled with a business process modelling language, for example BPMN 2.0[9]. Due to the limitations of this paper, we present only one process model as an example. We refer to the Bachelor thesis of [8] that presents the other process models too.

### *3.1. Request-a-Key*

The first functionality we want to present is *Request-a-Key*. With *Request-a-Key*, the key for decrypting a specific document is requested by a user to the system of the key owner. Figure 2 presents this use case as BPMN model.

In the owner's system, a check is carried out in the keystore for whether the specific document and the specific user, who requests the key, exist. If no key exists in the keystore, the fact that there was an incorrect request from this user for this document is logged. Reasons could be that the user has been denied rights to the document or the user received this document without authorization. If a

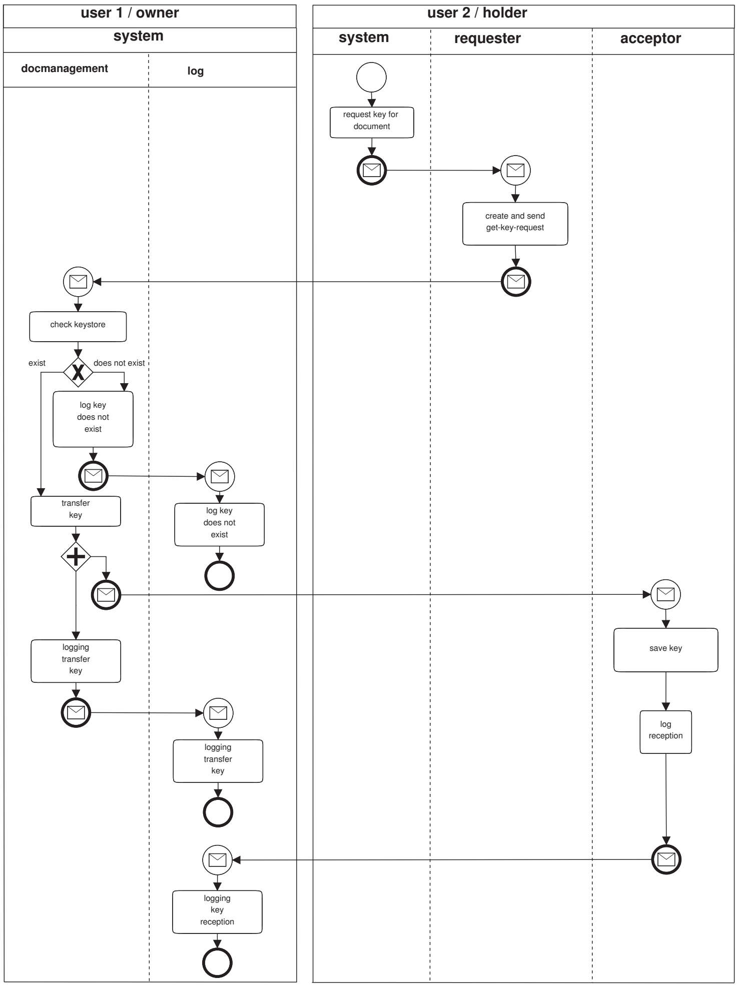



**Figure 2.** BPMN request-a-key

key exists in the keystore, this key is transferred and the transfer of the key is logged in the system of the owner. The user's wallet receives the key and stores the key to the document. The reception of the key is confirmed. The wallet of the owner stores the receipt confirmation in a log.

### *3.2. Request-a-Document*

With *Request-a-Document*, a particular document in the owner's system is requested. First, a connection is established to the desired participant. This connects both *privacy wallets*. The requester identifies itself on the basis of authorization features. The *privacy wallet* receiving the request now shows to the requesting *privacy wallet* all documents to which the requesting *privacy wallet* has permission to access. It does not matter whether the requesting *privacy wallet* has full access or read-only access to the document. The requester can now select from this document list the documents that are to be transferred into his own *privacy wallet*. These documents are then sent by the owner to the requester. Each receipt must be confirmed. This receipt confirmation is then stored by the wallet which did send the documents.

### *3.3. Response-a-Document*

With *Response-a-Document* a particular document is send to a particular person from the owner's system. To do this, the owner chooses the document to be send in his own wallet. The document is encrypted for the selected person and sent to him. The key is also stored in the local keystore of the owner. The document send is automatically written to the wallet of the recipient. Once this has happened, the document is displayed in the document list and a receipt confirmation is send to the sender. The sender, that is, the owner of the document, receives this receipt confirmation and the system automatically stores it.

### *3.4. Synchronize*

With *Synchronize*, updated data are send to other users. These data cover new files, people, person objects, or altered document objects, whereby altered document objects can only be synchronized from the owners privacy wallet. Holders can not synchronize any changes that they have done on documents where they are not the owner. Therefore, the wallet of the user that has changed the data starts the synchronization. The other wallets receive these changes and update the data in the wallet in the background on a fully automatic basis. Thereby all wallets only synchronize from the wallet that has started the synchronization. The user himself does not need to take any action. If a wallet cannot be reached, synchronization is repeated later. Thus, the synchronizing wallet checks within a periodical time, if the missed wallet is online, and then sends the changes. This ensures that each participant of the group receives all updates needed.

### *3.5. Ask-for-Extended-Rights*

With *Ask-for-Extended-Rights* a holder of a document requests from the owner an extension of the rights to this document. In our case, extended rights mean an extension of existing rights. For example, if a holder has only read rights on a document, he can ask for edit rights for this document. The owner must actively decide whether to grant the requested rights or not. If the owner does not grant the requested rights, the request is logged with the negative decision. If the owner grants the requested rights, the document is re-encrypted and transferred to the holder. In addition, the file transfer is logged and the new key is stored in the keystore. The wallet of the holder automatically receives the file and stores it in the filestore. In the process, the old file is overwritten and the old key, if it exists, is erased. After a successful receipt of the document, a receipt confirmation is send to the owner. The wallet of the owner saves the receipt confirmation in form of a log.

### *3.6. Request-a-Person*

If a new user joins a group, it is necessary that all members of the group know this new user. In the peer-to-peer network, the new user is known in the form of his email address if he is online, but other information is missing, in particular the public key. If a new user now participates in the group, his complete data is automatically distributed by the system to all participants who are online. To reduce administrative effort, that is, the management of the information regarding which users have been informed and which have not, the principle of an obligation to provide information is replaced by an obligation to retrieve information in this case. If a *privacy wallet* notices that an 'unknown' user is in the group, that is, in the group online, then this *privacy wallet* automatically requests the data of the 'unknown' user from the *privacy wallet* in the background. This data is then send back to the requesting *privacy wallet* on a fully automatic basis.

In contrast to the other processes, a check for whether the data actually arrived is not performed with this process. This is not necessary, as no security and control mechanisms are disrupted or undermined if the data is not present. Sending of the documents can only take place if the personal data is locally available. If problems occur with the transmission of personal data, the wallet simply tries again. To make sure that there was really a receipt confirmation in the operations *Request-a-Document, Response-a-Document, Request-a-Key* and *Ask-for-Extended-Rights*, a check is carried out. After an appropriate period of time, the system checks whether there has been a receipt confirmation. If this is the case, then everything is fine. If this is not the case, this document is blocked for the receiver. This happens when a delete command is send to the receiver wallet and at the same time the entry is removed from the keystore. This process is logged.

After presenting the functionality, the next section will describe the concrete realization of the *privacy wallet*.

## 4. Privacy Wallet Realization

This section describes the realization of the *privacy wallet*. The *privacy wallet* serves as an intelligent interface. This interface can be placed over a document management system (DMS), for example, but it can also work on a standalone basis. This makes it possible to use it in a very flexible manner. A set of wallets communicate in a peer-to-peer network. The whole communication is secured with a hybrid encryption approach and an additional usage of steganographic functions. Figure 3 shows this general structure of such a *privacy wallet* peer-to-peer network.

The next sections will take a closer look at the construction of such a *privacy wallet*. Moreover, we also present our prototype of a mobile application for smartphones. Also, there is an implementation for personal computer. For this implementation we again refer back to the Bachelor thesis of [8].

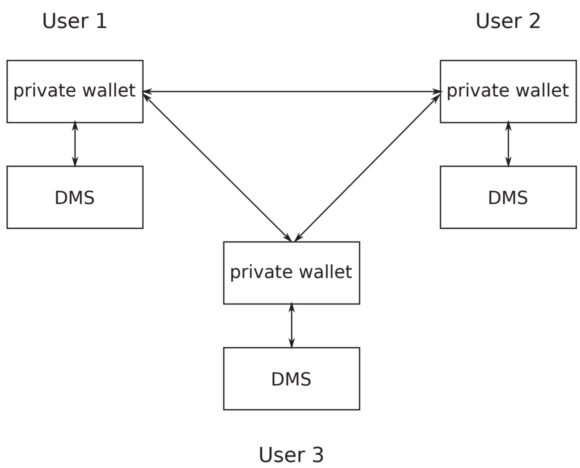



**Figure 3.** Peer-To-Peer *privacy wallet* Model

### *4.1. User Management*

The *privacy wallet* requires its own user management because only if users are registered in the *privacy wallet* network they can exchange data. Therefore, every user has to login into the *privacy wallet* network as shown in figure 4.

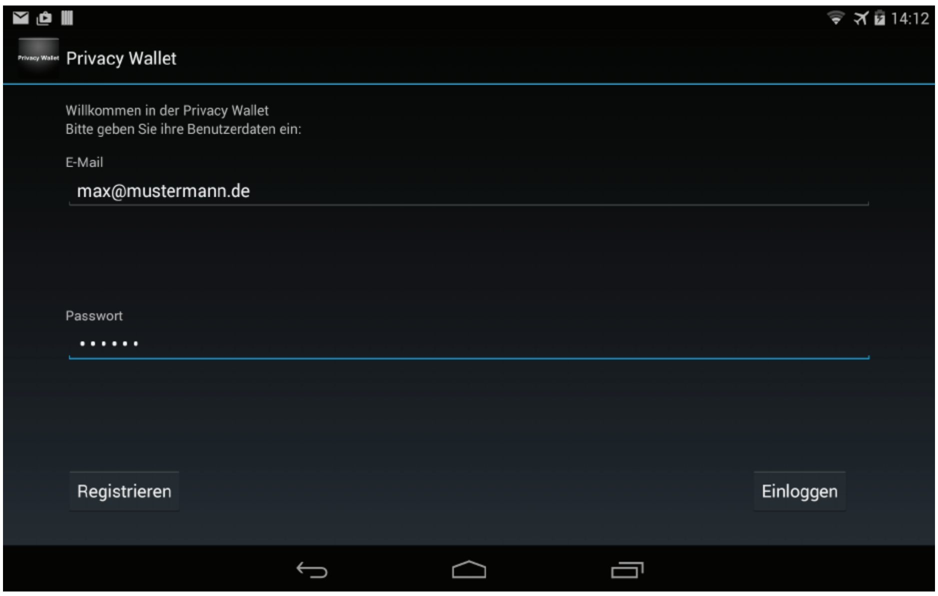



**Figure 4.** *Privacy wallet* Login

To enter the *privacy wallet* a valid email address and a password must be entered into the login screen. If a user wants to create a new account, he can register within the network. Therefore, he must enter his email address, a password and his first and last name. Figure 5 shows how this registration looks like.

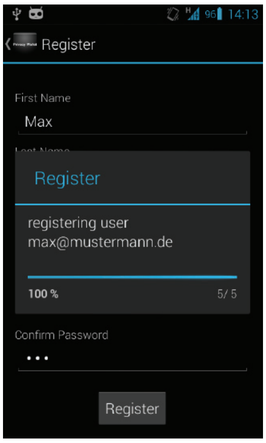



**Figure 5.** Registration Screen

After the login the user gets into his main screen that is shown in figure 6. There, he can access information about groups he has created, which groups he has joined and to which groups he is invited. We will describe the group management in part 4.2. Furthermore, he can access his imported documents and open the settings for his *privacy wallet*.

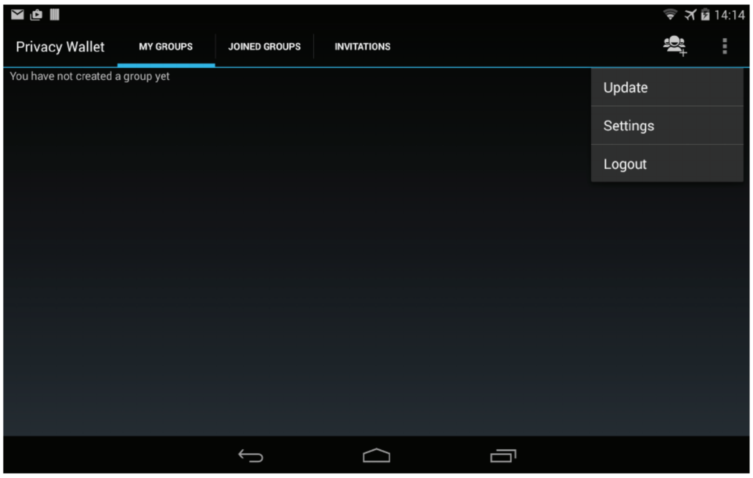



**Figure 6.** Main Menu

The settings define for each user the used default encryption, how the loggs should be displayed, which entries from the loggs are displayed and the management for the notifications of encryption, decryption and invitation. Figure 7 shows how the settings looks like.

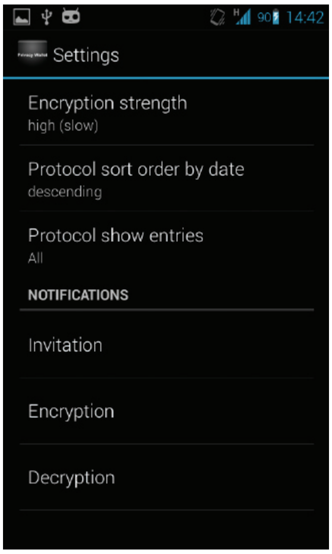



**Figure 7.** *Privacy wallet* Settings

*4.2. Groups*

The aim of the *privacy wallet* is to securely exchange data and to keep as much control as possible over the exchanged data. This makes it necessary to distribute data. These data are documents or document data, keys, and log entries. In our *privacy wallet* app document exchange is only possible between users in the same group. A group offers the possibility of an easy assignment of keys and rights to documents to several users and allows the option to automate certain approvals for the use of documents. As figure 8 shows a user only has to enter a group name to create a group. After this the group is displayed in his *My Groups* section of the app as figure 9 shows.

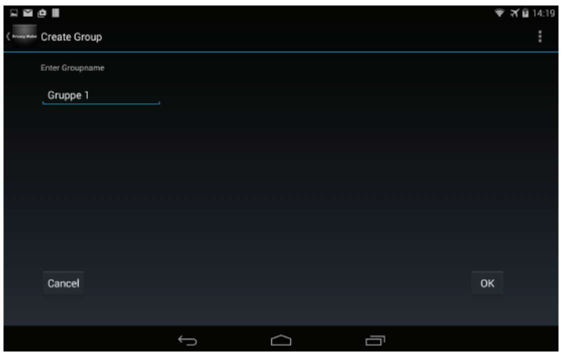



**Figure 8.** Group creation

After creating a group, a user can invite other users of the *privacy wallet* network in his group. For this, he must either enter the email address of the user

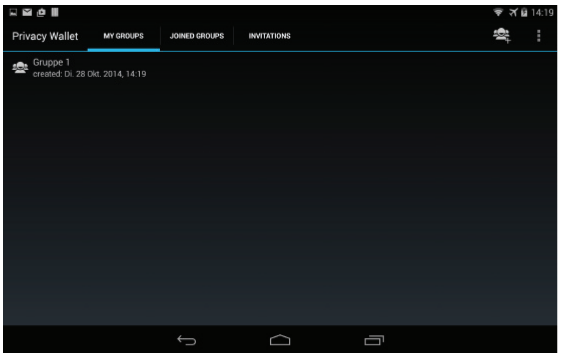
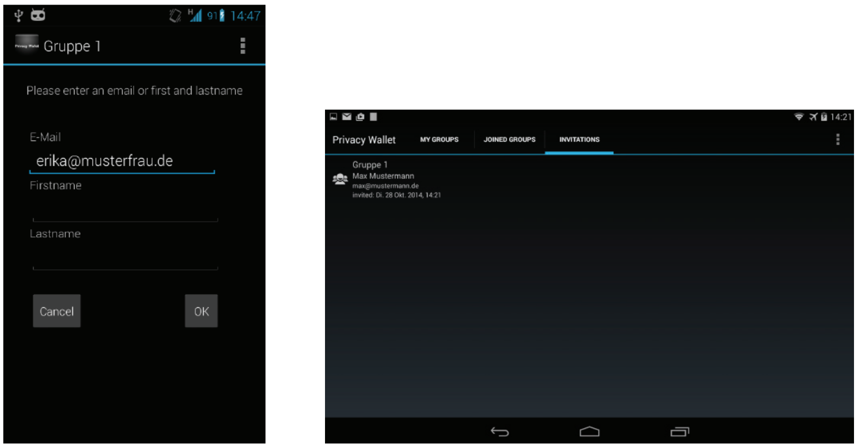



**Figure 9.** Display own groups

or his first and last name, as shown in figure 10(a). If a user is invited, he will get a notification about this in his *Invitations* screen, as shown in figure 10(b).




(a) Invite a user

(b) Show my invitations

**Figure 10.** Invite a user into a group

There a user can see, who has invited him in which group on which date. A user can now accept or cancel the invitation as shown in figure 11. In this special screen the user gets once again all information about the invitation and he has to actively accept or cancel the invitation. After this decision, the inviting user is informed about the user's choice. If he accepts the invitation, the data exchange within the group can start.

### *4.3. Modules and task areas*

The implementation of the *privacy wallet* is modular. This allows an easy replacement of individual components, such as another encryption or another commu-

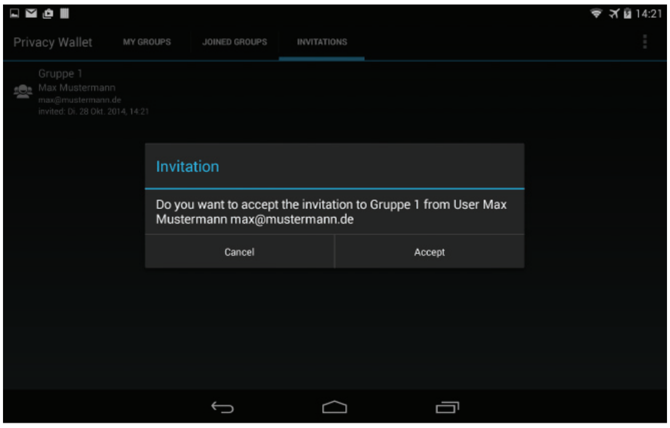



**Figure 11.** Confirm an invitation

nication framework. The following basic functional components are implemented for the *privacy wallet*.

*4.3.1. Importing documents*

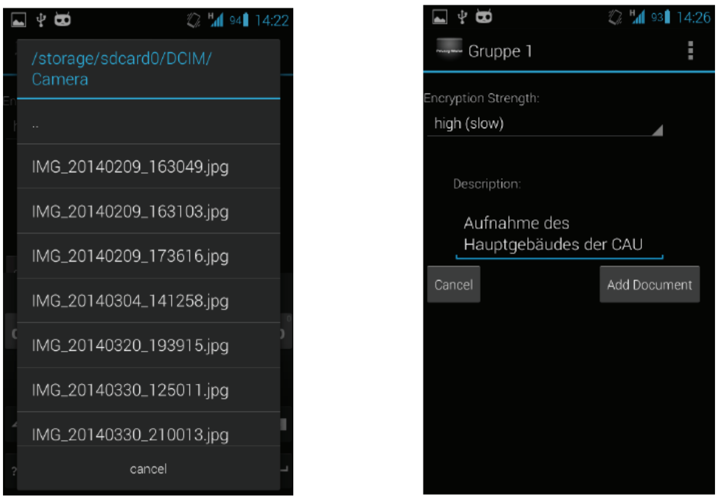
The *privacy wallet* is a closed system. Only within this closed system can the built-in security mechanisms be effective. Therefore, a function must be provided that allows you to import data into the system. For this purpose, the *privacy wallet* makes the *Import* function available. A user can select any files stored on his smartphone that should be added to the *privacy wallet*. Figure 12(a) displays how photos can be added from a smartphone to the *privacy wallet*.

Data such as a document ID, the creation date, and the author are generated by the system and added automatically to the files. Other data must be entered by the user. This includes a description of the stored file that allows all users within the group to see the content of the file and the encryption level that is applied on the stored file to secure it. Figure 12(b) shows how these information are added to the imported picture.

Depending on the level of the encryption the app points out, how long the encryption of the file can take. The file encryption depends to a large extent on the hardware of the smartphone. Therefore, we can only give a relative time as for example >> *slow* << for the encryption. After this step the file is encrypted immediately and stored in the filestore. The associated metadata are stored in the database. From this point on, the document is available in the *privacy wallet* as figure 13 shows.

*4.3.2. Communication*

A major component of the *privacy wallet* is the exchange among the participants. In this approach, the decision was made in favour of a peer-to-peer network. In [10] a peer-to-peer system is described as a self-organizing system that is composed of autonomous units called peers. All peers have equal rights in the system and




(a) Adding files to the *privacy wallet* app

(b) Import file description

**Figure 12.** Document import

all peers use their resources and services completely decentralized. This brief description can be extended by a set of characteristics of a peer-to-peer network as described in [15]. An ideal system has all these characteristics. But in most cases only a subset of these characteristics can be found in a peer-to-peer system. Using a peer-to-peer system offers some advantages over a client-server system. Through the omission of a server, the vulnerability of the entire system is reduced. Another advantage of a peer-to-peer network is simple distribution of information. It offers the possibility of sending a message to individual members of the group or to all members at the same time. Communication is always encrypted. This ensures that even if an outsider infiltrates the group, the information is secure.

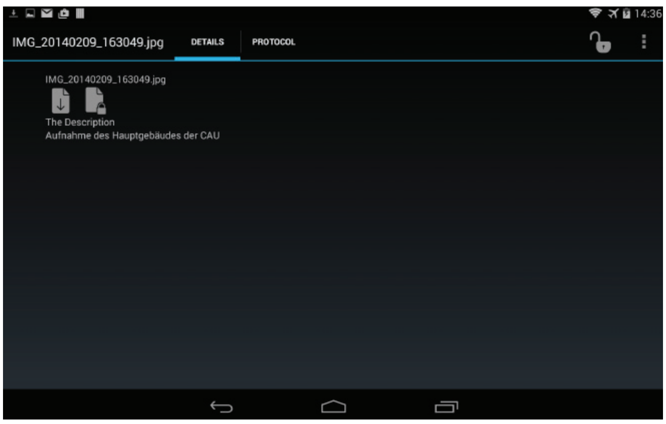



**Figure 13.** Available imported documents

*4.3.3. Logging*

Any action on or with a document triggers notification of the owner and author. To ensure this, the information about such an action is kept in a log object and send to the *privacy wallet* of the owner and author. Figure 14 shows how such a log looks like within the app. The owner and author stores this log object in his overall log. If one of the two participating *privacy wallets* is offline or unreachable, this log object is stored in the sender wallet and marked as not sent. As long as the sender wallet is turned on, the system attempts again and again to transfer non-transferred log objects at appropriate intervals until a transfer is successful. A log entry contains the document ID in your own *privacy wallet*, timestamps of when an action took place and when this log entry reached your own *privacy wallet*, the type of use of the document, the status, how many times this action was tried, and the user and his IP address. But our logging is limited. If a holder copies a document or takes a screenshot, we have no logging about what he is doing with the document any more. So at this point, we lose control over an owner's document.

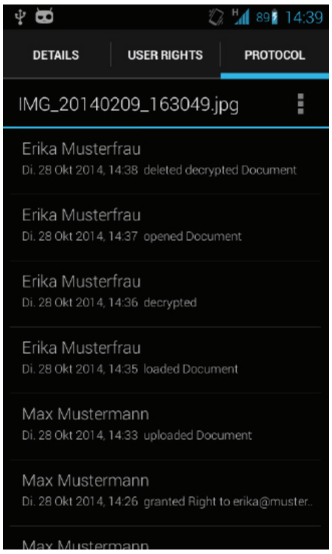



**Figure 14.** The *privacy wallet* log

*4.4. Encryption*

As described in [13] there are two different approaches for secure communication, steganography and cryptography. In our implementation a hybrid of two types of encryption is selected for cryptography. Symmetrical encryption is chosen for the encryption of the document. This way, it is possible to benefit from the speed advantage over asymmetrical encryption. The downside in terms of the effort and lower security with the key exchange is balanced out by asymmetrical encryption of the key of the symmetrical document encryption. A 128-bit key with the AES cryptosystem [3] is used for document encryption. A 192-bit or 256-bit key is also possible. For the RSA encryption [11], 2048 bit is used. Adequate security is already provided with a smaller number of bits. Because RSA encryption is only used to encrypt the AES key, meaning the text to be encrypted is very short, there are no performance drawbacks, but a higher level of security.

Furthermore, this *privacy wallet* offers the possibility of increasing the security of the encryption through steganography. The encrypted key is hidden in an image. Hence, an owner can hide the key by using a special function from the *privacy wallet*. A holder can only reveal the key from the image by using the same function in his *privacy wallet*. Due to the combination of RSA encryption and steganography, the security can be classified as very high. The method used in this *privacy wallet* for steganography is very simple. The method can be exchanged by every other known steganography algorithm. Here, every image is composed of colour points or pixels. A pixel has a red value, a green value, and a blue value. These values are between 0 and 255. To hide a message in the image, in each pixel each colour channel is changed in the last bit. The decision whether the last bit is set to 1 or 0 depends on the message to be hidden. For this, the letters of the message are converted into the associated ASCII values. These ASCII values are converted into binary form. Take for example that DB is hidden. The ASCII value for D is 68, and the value for B is 66. In binary terms, the binary representation for decimal 68 is 01000100 and for decimal 66 is 01000010. With the binary values, it is necessary to ensure that the leading zeros (always eight digits) are taken into account. Considering that it is always just the last bit of a colour channel that is changed and a pixel has three colour channels, eight pixels are needed to hide two letters and the message end character (ASCII 0). The number of pixels required is given by the following formula:

$$\lceil numberofpixels \rceil = (numberofcharacters + 1) * 8/3$$

The bit sequence of the letter combination DB results in 0100010001000010, which is 16 bit values. Now the the bit values are entered in sequence into the most recent bit of a colour channel in each case. If the last bit of a colour channel is identical to the bit entered no change takes place. In this example, the last bit of the first colour channel of the first pixel receives the value 0, the last bit of the second colour channel receives the value 1, and the last bit of the third colour channel receives the value 0. Since all three colour channels have now been used, the next pixel is used. This is repeated until all bits of the bit sequence and the message end character have been entered. Since only the last bit of a colour channel is manipulated, the overall colour of a pixel is changed by 1/256. This is only 0.39 percent. As a result, the colour change for each pixel is from 0 to 0.39 percent. On average the change is about 0.2 percent. This change is not discernible to the human eye. Therefore, the message has become invisible.

### 4.5. Data Transfer

There are various possibilities for transferring the data between the *privacy wallets* within the network. Using TCP and UDP, data can be transferred from both computers and mobile devices such as smartphones or tablets. This works both in a local network and over the internet. The advantage of this transfer option is that all systems have this possibility of use.

Transfer via Bluetooth or NFC is only suitable for mobile devices such as smartphones or tablets. However, this type of transfer offers great advantages in terms of security. Since NFC[1] or Bluetooth[2] only work over short to very short distances, the probability that such a connection will be 'overheard' is close to 0, since it is immediately apparent if someone is hanging around in the vicinity.

Transfers over a local network or the internet are suitable due to the high transfer rates both for the exchange of documents and for the exchange of keys. Networks have transfer rates up to 1 gigabit/s, and the internet provides transfer rates of between 1 and 100 Mbit/s per second for private households.

Bluetooth, then again, only has transfer rates of 57.6 kbit/s to 732.2 kbit/s. Transfer of documents is possible here, but the advantages of Bluetooth come with smaller data packets, such as a key, because the security is significantly higher due to the proximity to the exchange partner. The same applies to NFC. With NFC the data transfer rate is only up to 424 kbit/s.

### 4.6. Data Management

To store the data, every *privacy wallet* gets its own small database. Every database gets the same database schema to store the data. Figure 15 shows this database schema as Higher-Order Entity-Relationship Model (HERM)[16].

The central element is the document. A document can have one owner and multiple holders. The owner is the person who owns the document. This is in most cases the author of the document. The holder is the person who has control over the document. A document can have any number of keys with the corresponding rights. Behind each key is a holder with the appropriate rights. For each document, there are metadata. Information is stored for every person who participates in the *privacy wallet*. The log contains any number of log entries. A log entry provides information about the affected document, the point in time of the activity, the actual process, the status (the action was successful or not successful), the number of attempts, and the IP address of the device on which this action was performed. In the *privacy wallet*, users can define the document types in any required way. This ensures a high degree of flexibility.

---

[1] range: up to 10 cm.

[2] range: class 3 approx. 10 m outdoors, class 2 approx. 20 m outdoors, class 1 approx. 100 m outdoors.

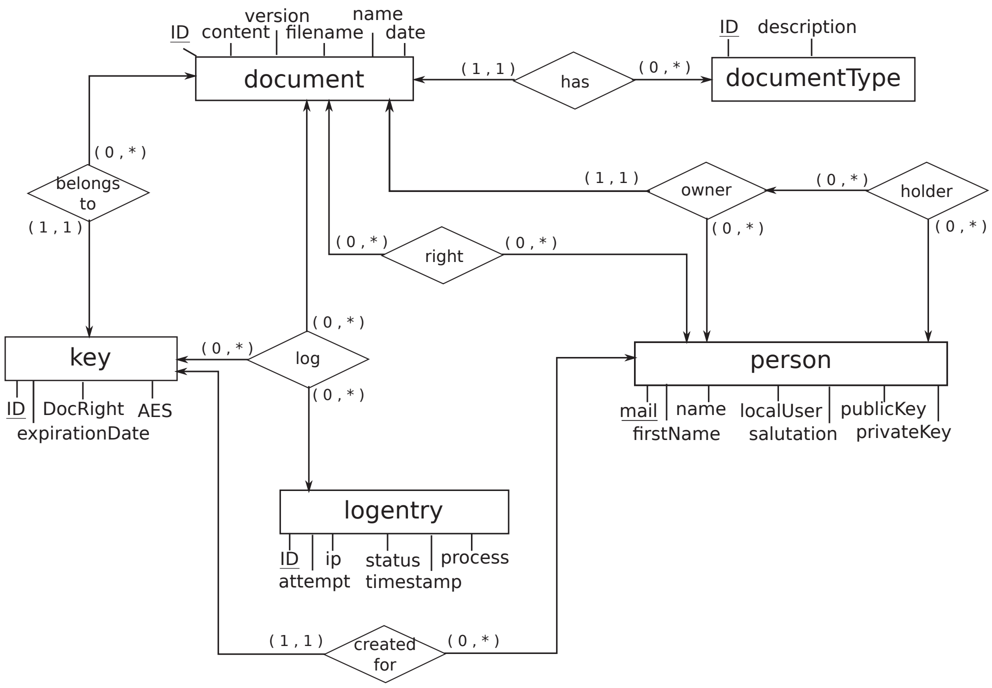



**Figure 15.** HERM-Schema of the *privacy wallet*

Throughout the application, work is carried out only on data objects that are already stored in the database. The decisive factor for such an approach was the objective of not losing any data. This is particularly important for guaranteed logging of actions on documents. For some operations, it is necessary to update existing data in the database. Only the following personal data can be changed: first name, last name, the marker, whether it is a local user or a group member, the password for local login to the *privacy wallet*, the public and private key, the document data rights to a document, the document key, and the expiration date. Figure 16 shows the displayed document informations. All other data, for example the email or the salutation, may not be altered under any circumstances.

Data which are stored in the database are deleted only by setting a marker which indicates that it is a deleted data record. The only exception is temporary data. Temporary data in this *privacy wallet* consist of log entries that have not yet been transmitted to the author of the document. They are removed from the database. Deleting data records only logically offers the advantage that no data inconsistency can occur. For documents marked as deleted, the physical file is deleted in the file system.

## 5. Related Work

There exists a large quantity of implementations in the field of secure data storage and exchange. Here, we want to describe some of these implementations that

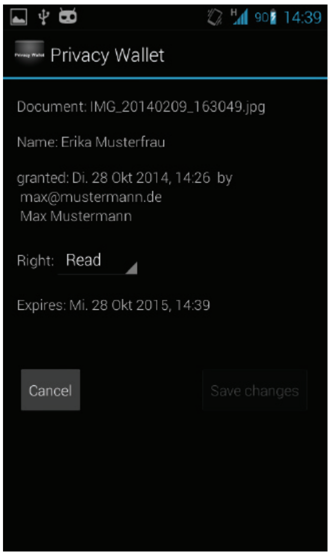



**Figure 16.** Displayed document informations

track concepts similar to our *privacy wallet* approach. First, we look at cloud services like Dropbox[5] or Apple iCloud [7]. There, a user can store data on a foreign server and can share these data with other users of the cloud. The main problem is that all data that are saved in these cloud services are stored on a third party foreign server. At this point, an owner does not have full control over the usage of his data by this third party. He must trust in the service that there is no illegal access to the data by the third party. Furthermore, an owner is dependent on the security of the cloud service system. In our approach, only holders who have the owner's permission can get access to data. If an owner shares data, this data are transferred from the owner's system to the holder's system. Thus, no third party server system for storage and transfer is used. Additionally, in most of the cloud services our data are not stored encrypted within the cloud. We must install special programs for encryption before storing the data into the cloud. In our *privacy wallet* approach every document is encrypted automatically when it is send to a holder. Only if the owner gives the holder the key permission, the holder can decrypt and use a document.

Another mobile application for smart phones and tablets that can be used for a secure exchange of chat massages with attached videos or images between users is Snapchat [14]. Before a user can exchange a message with another one, he must define an expiration date for the message he wants to send. The receiver of the message can look at it on his device until the expiration date is reached. After this, the message is deleted from the receiver device. This should guarantee the securtiy of the videos and images that are send via Snapchat. But there are some problems with Snapchat. On the one hand, the receiver of a Snapchat message can take a screenshot of the received object. Then he has permanent access to the object without the control of the owner. On the other hand, user have to trust

Snapchat that the message is fully deleted. There are hints that this is not done by Snapchat [2]. Despite everything, Snapchat is a first approach for a secure data exchange where the user has control over how long another user has access to his data.

Next, we want to take a look at the data exchange service Mega [12]. Mega can be used to exchange data with other users through the internet. Therefore, a user can import any digital information to the Mega server. The information is end-to-end encrypted with AES 128 before it is stored on the server. Only the user gets the key to decode the data. Hence, no other user or Mega itself can decode the private data. If a user wants to share the information with another user, he has to exchange the key with this user, so that the other user can get access to the data too. How the users exchange the keys is not the task of Mega. Unlike our *privacy wallet*, the focus on the encryption of information is not a user's self-determination about his own data, but it is a form of legal protection for Mega. Thus, Mega can assert that they never know what users send over their filesharing system.

The last service we want to present is an application that is close to our approach. The *Deutsche Post AG* provides a software called *DocWallet* [4]. This software can be used for secure personal document management. Here, a user can import documents into the *DocWallet* system. Within the system, the document is end-to-end encrypted with AES 256 and synchronized with every system on which the *DocWallet* software is installed. Hence, a user has access to his documents on all his systems, including mobile devices like smartphones or tablets. All this is free of charge for the user. Additionally, one can pay for a premium account. Then the data within the *DocWallet* is synchronized with a cloud server in Germany that backups all imported files in the cloud. Unlike our *privacy wallet*, there is no data exchange between various *DocWallet* users possible. The system is only for single user data management.

## 6. Conclusion and Outlook

Defining the personal demand of data control and protection by the user is one way to realize privacy-enhancing technologies. In this paper, we present our approach of a system in which an owner of data can store and exchange his data with other users without loss of control. Hence, we implement a system that is based on a peer-to-peer network to exchange data between users of this network. All exchanged data is encrypted with modern algorithms and only the owner of the data can decontrol the usage for different holders. To increase the security of the private data once again, we use the storage of private data within a database and steganographic algorithms. With the logging of all actions performed by the owner and the holder we gain the control over the usage of the data.

We have only outlined a first approach of our *privacy wallet*. In a subsequent step, we have to finalize our mobile prototype to a stable version for release. This includes also the key exchange by steganographical methods. At the moment, our first version is only usable for personal computers. After the final completion of our mobile version, we can extend our system with additional features. For example, we can report and analyse statistical informations, implementing a versioning of documents and encrypt the passwords and keys that are stored in the local database that is installed on the smartphone or personal computer. Furthermore, we have no recovery strategies for media break while using the *privacy wallet*. Finally, we have to examine whether users accept the usage of the *privacy wallet* to protect their personal data or if our system will be rejected because the effort is to high in the daily usage.

## 7. Acknowledgement

We want to thank Marten Göttsche, Jan Hoffmann and Christian Schielke for their support in the development and implementation of the *privacy wallet* concept.

## References

[1] A. Acquisti, R. Gross, and F. Stutzman. Faces of facebook: Privacy in the age of augmented reality. *BlackHat USA*, 2011.
[2] CHIP. Snapchat: Sexting-App löscht Videos nicht. http://www.chip.de/news/Snapchat-Sexting-App-loescht-Videos-nicht_63702600.html.
[3] J. Daemen and V. Rijmen. *The Design of Rijndael: AES - The Advanced Encryption Standard*. Springer Verlag, 2002.
[4] Deutsche Post. Docwallet. https://docwallet.de/.
[5] Dropbox. https://www.dropbox.com/.
[6] S. Fischer-Hübner. Privacy-enhancing technologies. In L. LIU and M. ZSU, editors, *Encyclopedia of Database Systems*, pages 2142–2147. Springer US, 2009.
[7] iCloude. https://www.icloud.com/.
[8] O. Jäger. Technologien für das Privacy Wallet., April 2013.
[9] OMG. Business Process Model and Notation (BPMN) Version 2.0. http://www.omg.org/spec/BPMN/2.0/PDF.
[10] A. Oram. *Peer-to-Peer - Harnessing the power of disruptive technologies*. OŔeillt, 2001.
[11] R. Rivest, A. Shamir, and L. Adleman. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM*, 21:120–126, 1978.
[12] K. Schmitz. Mega. https://mega.co.nz/.
[13] S. Singh. *Geheime Botschaften Die Kunst der Verschlüsselung von der Antike bis in die Zeiten des Internet*. Carl Hanser Verlag München Wien, 2000.
[14] Snapchat. http://www.snapchat.com/.
[15] R. Steinmetz and K. Wehrle. Peer-to-peer-networking & -computing. *Informatik Spektrum*, 27(1):5154, 2004.
[16] B. Thalheim. *Entity-Relationship Modeling: Foundations of Database Technology*. Springer-Verlag, 2000.

# A Conceptual Framework for Modelling Human Involvement in Cyber-Physical Social Systems: Leveraging STT Information in Heterogeneous Data

Sulayman K. SOWE and Koji ZETTSU
*Information Service Platform Laboratory,*
*National Institute of Information and Communications Technology (NICT),*
*3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan.*
*Email:sowe@nict.go.jp, zettsu@nict.go.jp*

**Abstract.** Cyber-Physical Social Systems (CPSS) are transforming how we live and interact with the physical world by semantically linking devices, data, and people. However, while tremendous progress have been made in modelling various CPSS physical components such as sensors and actuators, the involvement of humans in CPSS poses additional challenges for conceptual modelling experts and systems designers. Part of the problem is due to the fact that, even though humans sense, actuate, and process information like other CPSS component, they do these things differently. Furthermore, it is difficult to exactly know, in advance, how the human entity will interact with a CPSS. In this paper, we present a framework for modelling human's involvement in CPSS. Use cases, lessons learnt, and challenges we face in using the framework to develop a CPSS for research and development in the environmental sciences are also discussed.

**Keywords.** Cyber-Physical social systems, Conceptual modelling, Information modelling, human factor, Participatory sensing, Heterogeneous data fusion, System design

## 1. Introduction

In this information revolution era where everybody, everything, and anything are networked, Cyber-Physical Systems (CPS), also known as Internet of Things (IoT) [1], are having profound impacts in all aspects of human life. They are transforming how we live and interact with things in the physical world at a rate that is unprecedented in the history of mankind. The technological impacts of CPS is evident in many fields, including transportation, disaster response and recovery [2], healthcare [3], power grid [4], smart cities [5], to mention a few. CPS are generally regarded as physical and engineered systems whose operations are monitored, coordinated, controlled and integrated by a computing and communication core. A CPS can also be a system of collaborating computational elements controlling physical entities, or as an integrations of computation and physical processes [6]. From systems engineering point of view, however, a CPS is more than computational elements, processes, or devices. It can be an integrated composite that consists of, among others, people that provides a capability to satisfy a stated need or objective (ISO/IEC 15026:1998, 3.18). Furthermore,

[7] noted that CPS not only connects the virtual world with the physical world, but they have the ability to add more intelligence to social life by dynamically responding to real world scenarios. Essentially, CPS enrich people-to-people, people-to-device, and device-to-device interaction. This tight relationship between CPS and social entities such as people and their social networks is what Cyber-Physical social systems or CPSS are all about. According to [8], CPSS integrate various resources (which should interact efficiently) from physical, cyber, and social worlds.

In many CPSS, humans perform the same functions as the more traditional CPS components such as sensors, actuators, or actuators. However, the way humans perform these functions is fundamentally different. The concept of information is represented differently in a human, compared to a digital system or device. The fundamental building blocks for the human information system are different. Because of this, not only do humans function differently, but the modelling of humans in CPSS is fundamentally different from modelling other engineered systems.

### *1.1. Research Challenges*

Since conceptual modelling entails describing the semantics of an object, humans in our case, at a high level of abstraction, a clear understanding of the behavious or capabilities of the object is essential. Once we have this understanding, we can then conceptualize and sketch a model that can be used to define CPSS data and resource requirements and how, in practice, the model or framework can effectively information sharing between the CPSS components, including people. There are at least three challenges associated with developing a conceptual framework for modelling human involvement in CPSS.

First, human observations are difficult to interpret, quantify, and model. Our observations are context or situation specific and could be influenced by a host of factors. This makes it difficult for other humans or even a system to actually know which observations are reliable, how they should be interpreted, stored, processed, and presented so that other humans or systems can understand.

Second, as opposed to a mechanical device or component, humans may not do exactly what they are told to do. For instance, as long as the system is up and running, an air quality sensor in a CPSS will always report the air pollution values from a given geographical location. On the other hand, a human acting as an air pollution sensor might be reporting pollution values from a given geographical location, but change the location at any time, stop sending readings without notice, or even give false readings. This leads to uncertainties and unreliability when using the human sensor.

Third, a CPSS system designer knows in advance the components in his system, what data or information they provide or what predefined task or subtasks each component performs. The situation is completely different with a human component. The challenge for modelling is to identify what tasks the human can perform, whether he/she has the required knowledge and experience to perform that task, when and where will he be at a given time or his availability, how he is going to interface with other components in the CPSS, etc.

### *1.2. Problem definition and research contribution*

Giving the above challenges, the problem for conceptual modellers is how to capture and represent these human characteristics, describe the relationship between the human as a service component in CPSS and other CPSS physical components. Furthermore, we are also confronted with the problem of developing a conceptual model that captures all the

CPSS complexities and yet it is easy to understand, so that CPSS system designers or engineers can use the model to build a system that is in sync with the human

These challenges and the problems it poses, perhaps justifies why we need CPCC conceptual models or frameworks. [9] defined conceptual modelling as "the abstraction of a simulation model from the part of the real world it is representing (the real system)". The author went on to argue that the real system or CPSS may or may not exist, but the abstraction should simplified at the right level. [21] agrued that conceptual modelling has a significant importance in eliciting and clarifying system structure, behavior, and functionality.

In order to address the above challenges, we present a framework for modelling human's involvement in CPSS and discuss how we used the framework to develop a CPSS that allow users to send data to the system and the system can combine this data with other available data and send alert messages to the users so that they can take action about a certain event. The main objective of the framework is help us model human involvement in CPSS. In the framework we represent, at a high-level, the information necessary to describe a human component. The framework also defines the service interface between the human as a component and other CPS components, describes some general human capabilities that are needed to perform a particular service (e.g. sensing or actuating) within a CPSS. With these descriptions of capabilities, we hope to make it easy for the CPSS system designer to integrate human components based on the requirements of the CPSS, and to have enough information about the human component(s) so that he can assemble them together to create a complete system that integrates the human.

The rest of the paper is structured as follows. The background and work related to this research is presented in section 2. In section 3 we present CPSS real-life scenarios to help use determine the roles of humans in a CPSS. In section 4, we present the framework for modelling human involvement in CPSS. Concluding remarks, and future research are presented in section 5.

## 2. Background and Related Work

Cyber-physical systems are some of the most important and exciting emerging technologies of our times. They are increasingly being used for a wide range of utilities necessary for daily living. Governments and research institutions worldwide are investing heavily in CPS research and development. For example, the National Science Foundation and the European Union are heavily investing in developing CPS as an integral components of their cyber-infrastructure [10]. Large-scale projects like the Open Science Grid and the TeraGrid are few other examples of large CPS scale system of systems. Furthermore, a compendium of researchers are investigating the challenges [6, 11, 12] and social impacts of CPS. However, some of the most challenging research issues in CPS comes from the need to interface the modeling, control, and adaptation of CPS with humans. Furthermore, [13] argued that most contemporary CPS research often leaves out a key component, that is, the ordinary user. [26] also highlighted the important role humans play in cyber-physical cloud computing systems.

There is a compendium of initiative, projects (e.g. OpenSense [14], Common Sense [15]), and research literature discussing various aspects of human involvement in CPSS, primarily focusing on participatory sensing or community sensing. In [16] the authors presented a survey of mobile phone sensing applications and systems and discussed a number of challenges (e.g. infrastructure, privacy) emerging from mobile sensing. [17] presented a method and described how community sensing applications can make contact with users, using traffic monitoring as a case study. With regards to modelling, [18] used spatio-temporal data and applied location based social networks to model user activity preferences. However, while this kind of modeling is very useful in helping us understand the serendipity of the users, there is no information to help us understand the relationship between the users (humans) and other components in a CPSS sensing platform.

In addressing the scientific foundation of CPSS, [19] argued that we must address the human and social dimension of CPSS. This pervasive intelligence, CPSS, have reached a point where these two dimensions must be [modelled] or considered an integral part of any CPSS design and operation, the author concluded. Furthermore, [20] posits that CPSS (consisting of cyber and physical spaces, human knowledge, sociocultural elements) regards human factors as part of the system instead of placing them outside the system boundary. In order to model CPSS entities or agents, [21] introduced the concept of physical-informatical essence duality (PIED) to explain how CPS interact with real world entities. The authors discussed modelling the PIED process can be applied to threat management such as safety and counter-terrorism.

However, despite the abundance of literature discussing conceptual or information modelling in CPSS, there is little or no detailed discussion on how to model human's involvement in CPSS. Even at the device or sensor level, [22] argued that there is no common data modeling for heterogeneous devices. Therefore, the conceptual modeling framework we present in this paper may have the potential of contributing to the expanding body of knowledge in conceptual and/or information modelling in CPSS.

## 3. CPSS Use Cases Involving Humans

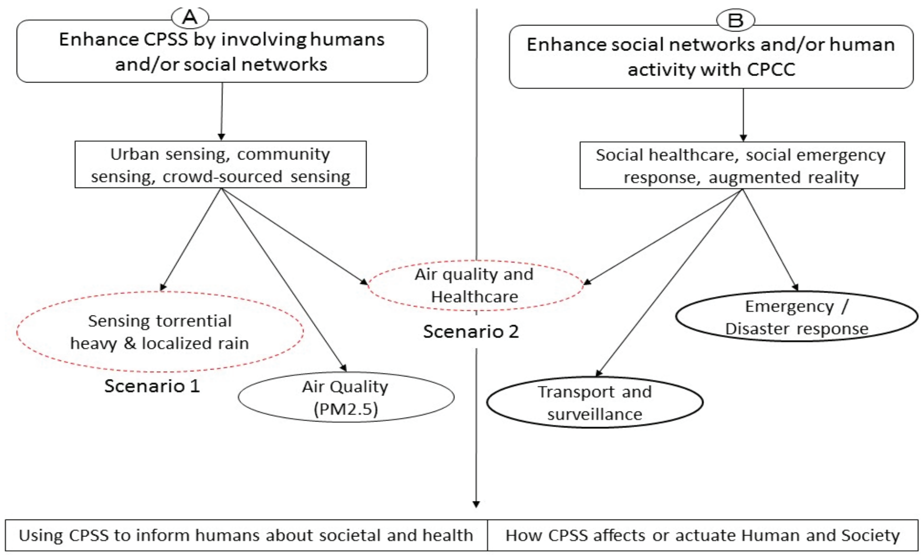



**Figure 1**: Classifying CPSS Use Cases Involving Humans

To help us describe CPSS use cases involving humans, we classify top-level CPSS use cases into two main categories (**A** and **B**), as shown in Figure 1. Use cases in category **A** are used to describe and improve our understanding about the role of humans in extending the functionalities or capabilities of a conventional CPSS. For instance, as

exemplified in urban and opportunistic sensing [23], community sensing [24], crowd-source sensing [23] and mobile phone [16] sensing. Use cases in category **B** are used to describe and understand how CPSS affect our lives, both at the individual and societal level. The trend in these modes of sensing stem from the recognition that humans are not bystanders in the sensing and communication processes, passively waiting on the fringe of the network for data to appear, but form an integral part of CPSS.
Below is a brief description of two use cases (in red dotted ovals in Figure 1) currently implemented by our CPSS.

### *3.1. Scenario 1: Sensing torrential heavy and localized rain*

The objective of this use case is to detect extreme or heavy rain in a localized area such as a city and predict how it impact people's social life. The weather in Japan is being monitored by a network of physical sensors that include precipitation (rain), temperature and wind sensors, to name but a few. Despite the wide country coverage, these sensors cannot report all the extreme weather conditions throughout the country or in some cases, even fail to report some heavy rains or hailstorms in some localized areas in the country. For example, heavy rain and hailstones hit Tokyo and the surrounding areas on June 24, 2014, with hailstones of up to 3 centimeters wide falling for about 30 minutes, according to some reports. Unfortunately, weather sensors could not report this natural disaster in time. On the other hands, the amount of Japanese tweets talking about "hailstorm" in Tokyo peaked sharply just after this disaster within a short time period.
Therefore, immense societal benefit can be realized if we are able to integrate the information from physical sensors with that reported by social human sensors. Logically, the role of the human and social networks, in this case, is crucial in supplementing or enhancing the lack of sufficient information in CPSS. Besides, the impacts of torrential or heavy rain in a small area on social life cannot be discovered by using only weather sensors. Furthermore, physical sensors can only give us quantified values. It is left to the Human Sensor to report the emotional, social, economic, and destructive nature of a disaster. Human sensors, with some degree of bias and reliability, can report exactly what is impacting them and their surroundings.

### *3.2. Scenario 2: Air quality and Healthcare*

One of the greatest application areas of CPSS is in the field of healthcare, where a host of applications have been developed to address specific health scenarios [7]. Our focus in this scenario is on how asthma can be aggravated by air pollutants such as fine particulate matter ($PM_{2.5}$), a case of enhancing or improving human health with CPSS (use case category B). EventShop [25] is a CPSS applications that can detect asthma attacks by integrating and analyzing information reported by physical sensors (e.g. air pollution or particulate matter ($PM_{2.5}$) levels)) and data from human patients. The system has medical records, telephones, emails, and each person life events or Personicle (jogging, shopping, etc.). "Knowing" what a person is doing or where a person is at any time, the system can send alert message to warn people not to approach a particular area because the $PM_{2.5}$ or Carbon monoxide levels are very high in that area. Users can, in turn, propagate or tweet the message they got through their social network or send feedback to the EventShop system to improve the information and health related events processing.
These scenarios demonstrates the strategic importance of the human within a CPSS. The human can be conceived as service component who can interact with other CPS sensors,

processors, and actuators. For example, for a human contributing air quality data with spatio-temporal and thematic information, Table 1 shows how this human sensor capability could be described and the kind of information it can contained.

**Table 1:** Description of a Human Social Sensor for Sensing Air Quality

| **Human Sensor Component** |
|---|
| **Service** |
| Type: *human service.* Name [<<**Thematic**>>]: *air quality ($PM_{2.5}$) data from Japan.* Function: *provide hourly $PM_{2.5}$ readings from Tokyo city.* |
| **Qualification [<<Constraints>>]** |
| Training: *Environmental scientist.* Experience: *5yrs working with air pollution data.* Certificate: *MSc in environmental science* |
| **Tasks** [<<**Tasks**>>] |
| PerformedTask: *Contributed $CO_2$, $PM_{2.5}$, data to AirQualityEgg project.* PotentialTask: *Contribute $CO_2$, $NOxPM_{2.5}$* |
| **Spatio-Temporal** |
| Location [<<**Location**>>]: 35° 41′ 22.22″ N, 139° 41′ 30.12″ E (Tokyo). Time [<<**Time**>>]: 12noon -1pm |

## 4. A Framework for Modelling Human's Involvement in CPSS

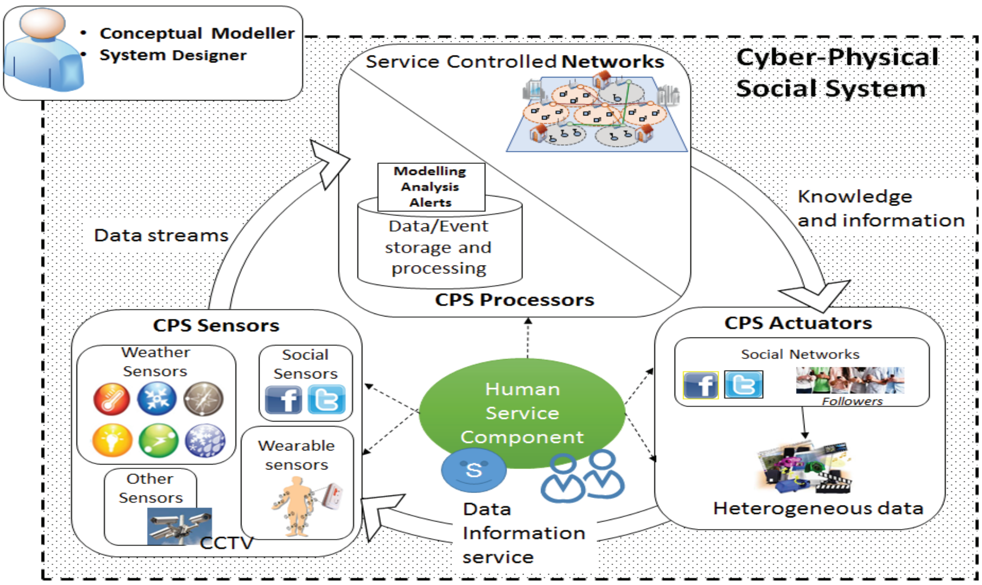



**Figure 2**: A Conceptual Framework for Modelling Human's Involvement in Cyber-Physical Social Systems

Figure 2 shows a high level diagram of the framework for modelling human's involvement in CPSS, as well as external actors (conceptual modeler and CPSS system designer) who might be the beneficiary or users of this framework.

The framework is described in terms of the CPSS entities; humans and physical components. At the highest level, the framework consist of components that convert data and information from the physical world into digital data for use in the cyber world. A

*Sensor* measures some property of the physical world and an *Actuator* effect a change in the physical world based on output from processing in the cyber world. *Sensors* monitor physical properties (e.g., rain, temperature) and generate digital representation of the measurement corresponding to physical events or phenomena being measured. Additionally, location and timing information must be associated with the measurement so that what is being measured can be integrated with other physical data. *Actuators* convert digital information into an action in the physical world.

The framework also shows traditional CPSS components that process, store, and transmit data and produce new information: *Processor*, *Networks,* and *Data Storage*. A *Processor* takes sensor input and/or input from other non-sensor data sources and processes it according to a defined algorithm. The output of a *Processor* may be fed directly to an actuator, to another *Processor* (allowing multiple *Processor*s to be combined to perform a function), or to a stored data service. The *Data store* component class stores data. The service controlled *Network* class connects other services by exchanging data between services.

The *Human* service component, at the center of the framework, represents the human and describes the human's sensor, actuator, processor, and data store capabilities. *Humans* exhibit all the main characteristics of CPS services, but the capabilities are manifested in a completely different manner. The *Social* class, represented as "*S"*, represents groups of humans whose capabilities are combined. An example of a *Social* component is a social network used as a virtual sensor to provide data aggregated from *Human* observations. These *Social* sensors must communicate and exchange information about the sensor availability and capabilities and to provision the *Sensor* for use within the CPS. Furthermore, *Social* groups may process data or perform actions or some complex combination of these. However, not only can *Humans* participate as a part of CPS, but a *Human* can be one of the physical entity that a cyber-physical system is sensing and acting on. In some cases, a *Human* is both a component of a cyber-physical system and the physical entity being sensed or acted upon. A healthcare monitoring system is a good example of this. A *Human* may have connected monitoring devices (e.g., heart rate monitor, blood pressure monitor) that are remotely read by a physician. At the same time, the *Human* may use an application on a smart phone to send the physician specific observations (such as "I feel light headed or "My arm feels numb."), which can be combined with the monitoring devices to give the physician a more complete picture.

## 5. Summary and Conclusion

In this research, we presented a framework for modelling human's involvement in CPSS and used two scenarios to describe what the human component could do in a CPSS. In the framework, the essential CPSS entities or components (such as sensors, data stores, networks, processors, actuators) and how they exchange data, knowledge and information, services were presented. At the center of the exchange was the human service component and other social artefacts.

CPSS are becoming more and more complex as we continue to rely more and more on them. This complexity means that the number of abstraction levels of information needed to generate a good conceptual model of CPSS will also increase. While our conceptual framework may not have captured all the features that may be need to model humans' involvement in CPSS in specific scenarios such as sensing torrential heavy rain and air and health quality, we posits that we have laid the foundation for a debate that might be of great interest to the modelling expert community.

The future direction of this research is to evaluate the framework by leveraging a number of information modelling evaluation techniques. For instance evaluation can be done in many ways [26], including; literature surveys, domain experts' reviews, and state-of-the-art assessment.

## References


[1] Antonio J. Jara, Dominique Genoud, and Yann Bocchi. Big Data for Cyber Physical Systems: An Analysis of Challenges, Solutions and Opportunities. In *Proceedings of the 2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing* (IMIS '14), 2014, 376-380.

[2] Eric D. Simmon, Kyoung-sook Kim, Eswaran Subrahmanian, Ryong Lee, Frederic J. de Vaulx, Yohei Murakami, Koji Zettsu, Ram D. Sriram. A Vision of Cyber-Physical Cloud Computing for Smart Networked Systems. NIST Interagency/Internal Report (NISTIR) – 7951.

[3] S. Soegijoko. A Brief Review on Existing Cyber-Physical Systems for Healthcare Applications and their Prospective National Developments. In *3rd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering*, pp: 2–2, 2013.

[4] S. Khaitan and J. McCalley. Cyber physical system approach for design of power grids: A survey. In *Power and Energy Society General Meeting (PES)*, 2013 IEEE, pages 1–5, July 2013.

[5] L. Gurgen, O. Gunalp, Y. Benazzouz, and M. Gallissot. Self-aware cyber-physical systems and applications in smart buildings and cities. In *Design, Automation Test in Europe Conference Exhibition (DATE)*, 2013, pp: 1149–1154.

[6] E. A. Lee. Cyber physical systems: Design challenges. *Technical Report UCB/EECS-2008-8*, Electrical Engineering and Computer Sciences, University of California at Berkeley, January 23 2008.

[7] Shah Ahsanul Haque, Syed Mahfuzul Aziz, Mustafizur Rahman. Review of cyber-physical system in healthcare. *International Journal of Distributed Sensor Networks*, 2014(Article ID 217415), 20 pages.

[8] Smirnov, A.; Levashova, T.; Shilov, N.; Sandkuhl, K. Ontology for cyber-physical-social systems self-organisation. In *16th Conference of Open Innovations Associatio*n (FRUCT16), 27-31 Oct. 2014, pp: 101-107.

[9] R. Stewart. Conceptual Modelling: Who Neets It? SCS M&S Magazeine, No 2, April, 2010, pp: 1-7

[10] V. Welch, A. Walsh, W. Barnett, and C. A. Stewart. A Roadmap for Using NSF Cyberinfrastructure with Incommon. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, 2011, pp: 28:1–28:2.

[11] A. Klesh, J. Cutler, and E. Atkins. Cyber-physical Challenges for Space Systems. In *IEEE/ACM Third International Conference on Cyber-Physical Systems (ICCPS)*, 2012, pages 45–52.

[12] Siegel, Neil G. The Challenges of Emerging Software Eco-systems (keynote). In *Proceedings of the 2013 Int'l Conference on Software and System Process (ICSSP2013)*, 2013, pp: 1–8.

[13] A. D. Wood and J. A. Stankovic. Human in the loop: Distributed Data Streams for Immersive Cyber-Physical Systems. *SIGBED Rev*., 5(1), 2008, pp: 20:1–20:2.

[14] OpenSense Project. Available at http://www.opensense.ethz.ch/trac/, January 25, 2015

[15] Common Sense Project. Available at http://www.communitysensing.org/index.php, January 25, 2015

[16] N. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. Campbell. A Survey of Mobile Phone Sensing, *IEEE Communications Magazine*, 48(9), 2010, pp: 140–150.

[17] Krause, A.; Horvitz, E.; Kansal, A.; Feng Zhao, Toward Community Sensing, *International Conference on Information Processing in Sensor Networks (IPSN '08), 2008*, pp.481-492.

[18] Dingqi Yang; Daqing Zhang; Zheng, V.W.; Zhiyong Yu, Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs, *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol.45, no.1, 2015, pp.129-142.

[19] Fei-Yue Wang, The Emergence of Intelligent Enterprises: From CPS to CPSS, *Intelligent Systems, IEEE*, vol.25, no.4, 2010, pp.85-88.

[20] hong Liu; Dong-Sheng Yang; Ding Wen; Wei-Ming Zhang; Wenji Mao, Cyber-Physical-Social Systems for Command and Control, *IEEE Intelligent Systems,* vol.26, no.4, 2011, pp.92-96.

[21] Mordecai, Y.; Chapman, C.; Dori, D., Conceptual Modeling Semantics for the Physical-Informatical Essence Duality Problem, *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* 2012, pp.4505- 4510.

[22] Xing Gao; Shaoyong Guo; Lanlan Rui; Kai Zhu; Xuesong Qiu, Novel Information Modeling for Management of Heterogeneous Device, *2010 3rd IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT),* 2010, pp.247-251.

[23] G. S. Tuncay, G. Benincasa, and A. Helmy. Participant Recruitment and Data Collection Framework for Opportunistic Sensing: A Comparative Analysis. In *Proceedings of the 8th ACM MobiCom Workshop on Challenged Networks*, 2013, pp: 25–30.

[24] Z. Zhang. Feeling the Sense of Community in Social Networking Usage. *IEEE Transactions on Engineering Management*, 57(2), 2010, pp: 225–239

[25] S. Pongpaichet, V. K. Singh, M. Gao, and R. Jain. Eventshop: Recognizing Situations in Web Data Streams. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, Geneva, Switzerland, 2013, pp: 1359–1368.

[26] Moody, D.L.; Sindre, G.; Brasethvik, T.; Solvberg, A., Evaluating the quality of information models: empirical testing of a conceptual model quality framework, In *Proceedings 25th International Conference on Software Engineering, 2003*, pp.295-305.

# An Enhanced Application Benchmark for Smart Cities

Eiman ALSABI, Layla AlRWAIS , Reem ALOSIMAI, Ruba ALRWAIS, and Ajantha DAHANAYAKE [b]

*Prince Sultan University – College for Women, King Abdullah Road, Riyadh 11586 Saudi Arabia*
[b] *Tel.: +966-11-494-8319, E-mail address: ADahanayake@pscw.psu.edu.sa*

**Abstract.** Some cities are called smart because these cities make use of ICT in an effective ways to manage their resources to achieve high satisfaction levels for their citizens. The ICT services make a city more competitive on local and global levels. The smart city applications immensely depend on the achievement of characteristics, factors, and indicators to measure the degree of ICT that contribute to provide services. Therefore, the degree of ICT permits the judgment of the level of smartness of a city. This research proposes an enhanced framework to evaluate the smartness of a city based on three functional layers comprising of three smartness measuring factors (SMF). The smart measuring factors include smart infrastructure, smart citizen, smart governance, smart mobility, smart economy, smart lifestyle, smart technology adaptation tendency, smart service integration level, smart response feedback model and ICT Maturity comprise of the umbrella activity. The proposed framework computes the level of efficiency in utilizing city's resources and thus fills the research gap presenting a way to the evaluation of the smart cities.

**Keywords:** Smart City, Smartness Benchmark, Smartness Measuring Factors, Service Integration Level, Smart Mobility, Smart Economy, Layered architecture

## 1 Introduction

In today's globalized world, there have been unprecedented advancements in technology. This has had a lasting impact on the cities that we inhabit thus transforming them into smart cities. Smart cities are cities that rely on sophisticated technology to enhance the wellbeing of its citizens. This is achieved through the efficient management of resources, provision of advanced infrastructure, implementation of advanced communication networks to stimulate sustainable economic growth. Therefore, this results in high quality of life [1]. Researchers have employed different terminologies to frame a smart city as: a digital city, an intelligent city, an informative city, a creative city, and a knowledge city [2, 3] are some of those examples. The smart city concept has been surfaced at the end of twentieth century by a small group of countries, and within a short time it captured the entire globe.

The term "smart city" refers to different things to different people and is used in many different contexts. Therefore, the smart city has been defined in a number of following ways. For instance, the concept of smart city is defined by IBM as, the use of information and communication technology to sense, analyze and integrate the key information of core systems in running cities. At the same time, smart city can make intelligent response to different kinds of needs, including daily livelihood, environmental protection, public safety and city services, industrial and commercial activities [12]. A Smart City according to [1] is a City performing in a forward-looking way in economy, people, governance, mobility, environment, and living, built on the smart combination of endowments and activities of self-decisive Independent and aware citizens [1]. This definition is based on the traditional regional and neoclassical theories of urban growth and development.

Caragliu has explained about smart in [3] as: a city is considered to be smart when investments in human and social capital and traditional (transport) and modern (ICT) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources, through participatory governance. Whereas, [13] defines a smart city as: a city that monitors and integrates conditions of all of its critical infrastructures including roads, bridges, tunnels, railways, subways, airports, seaports, communications, water, power, even major buildings, can better optimize its resources, plan its preventive maintenance activities, and monitor security aspects while maximizing services to its citizens. Furthermore a city that monitors and integrates conditions of all of its critical infrastructure including roads, bridges, tunnels, railways, subways, airports, seaports, communications, water, power, major buildings and a city that has the ability to optimize its resources. This results in planning maintenance activities, and monitoring security aspects therefore maximizing services to its citizens [2].

Based on [4], a city combining ICT and Web 2.0 technology with other organizational, design and planning efforts to dematerialize and speed up bureaucratic processes and help to identify new, innovative solutions to city management complexity, in order to improve sustainability and livability is a smart city.

Barcelona's is ranked at the top of smart city rankings. Its transformation into a smart city is not a result of sole contribution of the IT department, rather it was the joint effort among the town planning department, local councils and IT infrastructure providers. The organizational transformation goes far beyond employees' training in ICT or an information systems upgrade. In Fact, people are the key asset of cities in the transformation into smart cities as in the case of Barcelona [8].

The focus of this research paper is to propose a framework to evaluate the smartness of a city based on three functional layered approach in which each layer comprising of three smartness measuring factors (SMF). The smart measuring factors include smart infrastructure, smart citizen, smart governance, smart mobility, smart economy, smart lifestyle, smart technology adaptation tendency, smart service integration level, and smart response feedback model. The umbrella activity is defined as the ICT Maturity. The proposed framework computes how efficient are cities in utilizing their resources and fills the research gap presenting the aforementioned evaluation of the smart cities. To analyze the efficacy of the proposed scheme, we have designed a case study which collected data from five different cities against the proposed smartness parameters and displays a summarized view indicating their value of smartness. Based on the case study and observations, the framework has displayed favorable results for measuring a city's smartness.

## 2 Related Works

In literature, there are several articles in which researchers have explained the concept, implementation and evaluation criteria about a smart city.

It is explained in [1] that a smart city is based on nine characteristics namely: smart infrastructure, smart citizen, smart governance, smart mobility, smart economy, smart lifestyle, smart technology adaptation, smart service integration, and smart feedback response. An in depth analysis of each characteristic can be further explored on the basis of 31 factors and each factor is measured on a 1-4 index. This approach is rather abstract as it defines a smart city measurement system, but it does not provide any sound information about its measurement architecture. The six characteristics are explored with the help of the conceptual framework of smart cities and then evaluated on prominent smart cities such as Amsterdam and Barcelona. The framework in [1] further explores the context at which characteristics of smart cities have been implemented in the aforementioned cities.

In [3], the author summarizes the characteristics that are proper to smart cities. The first one is the utilization of networked infrastructure to improve economic and political efficiency and enable social, cultural and urban development. Second characteristic include underlying emphasis on business-led urban development. Third factor focuses on the aim to achieve the social inclusion of various urban residents in public services. The fourth characteristic stresses on the crucial role of high-tech and creative industries in relation to the established urban growth. The fifth, factor has a profound focus on the role of the social and relational capital in urban development. The sixth factor then focuses on social and environmental sustainability as a major strategic component of smart cities. Characteristic five and six are the most interesting and promising ones.

IBM researchers in [6] recognize cities as a system with seven subsystems, with various urban core elements such as city services, citizens, business, transport, communication, sanitation, and energy sources.

The author in [7] has illustrated a comprehensive set of factors for understanding a smart city. These characteristics, brought together to identify a smart cities framework, can be used to study and determine the smartness of a city. These eight factors include: management and organization, technology, governance, policy, people and communities, economy, infrastructure, and the natural environment. The author developed an integrative framework to illustrate the relationships and influences between these factors and the smart city. In addition, author represented two different levels of influence. The first level is outer factors which includes; governance, people and communities, natural environment, infrastructure, and economy and the second level is inner factors. The inner factors include technology, management, and policy. The outer factors influence is heavier than inner factors influence.

The city of Barcelona has taken many initiatives to observe the concept of smart cities. Barcelona city in collaboration with research centers and industrial partners has been testing the smart city concept with the dual purpose to contribute to the creation of standards and providing an interim solution to cope with the heterogeneity of providers, in particular from the wireless sensor part. This work has been recognized and structured as a Barcelona Intelligent City project [9].

The research work from Netherlands [10], where Amsterdam as a case study, is focusing on a number of ambitious environmental targets for the city which includes a 40% reduction in $CO_2$ emissions by 2025, and 25% reduction in energy use by 2025. In addition, Amsterdam has introduced and evaluated a lot of projects, divided into four areas: sustainable public space, sustainable mobility, sustainable living, and sustainable working.

Furthermore, three elements of smart cities including technologies-suppliers-customers are also defined [4]. The smart cities can also be studied in terms of the following characteristics like city administration, education, healthcare, public safety, real estate (energy efficient and of high performance buildings), transportation (traffic and parking management and public transportation), utilities (resource capitalization, environmental services).

Various smart cities have depicted different concepts and characteristics during the last few years. A lot of research has been conducted on the topic of smart cities but a framework of characteristics to identify the smartness of a city has never been formulated. Since such a framework has not been developed yet, this research will introduce a smart city framework that can efficiently identify the level of smartness of a city.

## 3 The Proposed Framework

To evaluate the smartness of a city, we have proposed a three-layered framework (figure 1). In which each layer represents a set of key parameters. The layered architecture will make the overall smartness evaluation process expeditious. One can easily discard a city without calculating the advanced parameters, if it a city fails to satisfy the initial parameters. For instance, if we take a city to evaluate and the city does not fulfill the criteria of smart infrastructure. Then it is obvious that the city does not hold a smart environment then how it can implement a smart governance there. Moreover, the new parameters in the third layer make the selection process more robust and result oriented. In the section below, we explain the underlying components in each layer.

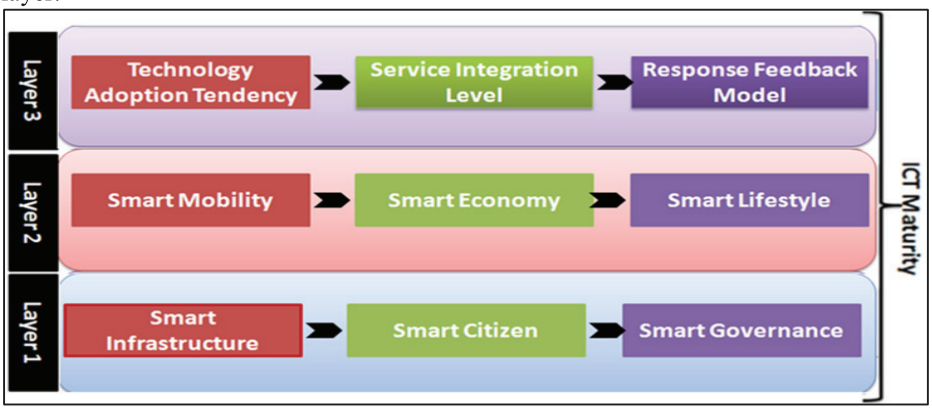



**Fig. 1.** The proposed framework to evaluate city smartness

### 3.1 Smart Infrastructure

Smart Infrastructure refers to the well deployed, interoperable and secured ICT services. Any new device can automatically get its configurations from server and will start a collaborative work with other device and appliances working in smart environment. There are some factors that can help to identify if the environment is being smartly managed or not. The infrastructure can be smarty managed in a number of ways. Firstly, it includes security and trust. Safety and security is achieved through discovering new ways to protect populations in public and private zones by using a new technology of information and communication systems [2]. Secondly, the concept of culture and identity is used. The smart cities use new technology for promoting and transporting their culture to other areas to publicize their cultural identity. Likewise, new technology is applied to record the cultural traditions. The use of initiatives for the digitization of heritage assets is an indicator to measure the utilized smart infrastructure [2]. Finally, natural environment, which is using a technology to protect and manage natural recourse such as climate, green space, waterways, sewers etc., can effectively be used to manage the infrastructure smartly in smart cities.

### 3.2 Smart Citizens

Smart people refer to the level of qualification or education of citizens as the social capital, flexibility, creativity, tolerance, and cosmopolitanism. It also includes the quality of social interactions regarding integration and public life and social relations with others that assists cities to be open and connect with the outside world [5]. People are considered as main element of progress in smart cities. They have different skills, educational levels, and high quality of social interaction. There are some factors that can be used to improve and make people smarter such as education and training for promoting the population's higher education with college degrees. Implementation of E-learning programs with the help of ICT in education includes plans for digital development in classrooms, research and development, innovation in human capital [6].

### 3.3 Smart Governance

Smart Governance refers to using ICT in management of city and using e-government or e-democracy to improve citizens' participation in public decision-making. There are some factors that must be taken into consideration when applying smart governance in cities. Initiatives can be taken such as local public spending and investment in the field of ICT to achieve the intelligent management of the city. Website availability, which is the major channel to communicate with public through Internet, can also be implemented in the process. Online public services can help in improving the quality and productivity of the services. Other factors include transparent governance by making the budgets and regulations, municipal plans available to the public, e-democracy which includes increasing the opportunities for public to interact with government by citizen participation or electronic voting, promoting ICT and innovation through open dialogue between experts to discuss how cities can face their challenges and can improve their usage of ICT [3] [6].

### 3.4 Smart Mobility

Smart mobility refers to the ability to provide every user to share and process any information instantly from anywhere locally and/or internationally by good broadband infrastructure and high speed internet access in order to ensure ICT use in homes, Internet usage, broadband coverage, broadband usage, mobile phone usage, mobile Internet usage penetration, public Internet Access, Wi-Fi hotspots in cities, Public Internet access centers, Promotion deals with ISPs. The use of new technologies in all means of transportation is also considered smart mobility. [3] [2].

### 3.5 Smart Economy

Smart economy refers to increase e-business and e-commerce, improve productivity, ICT-enabled and advanced manufacturing and delivery of services. Smart economy depends on technology and information in the production process to achieve maximum profit. The smart economy incorporates ICT by utilizing factors such as financial, promotion, retaining, attracting talent and promoting creativity, support for entrepreneurship, development of business spaces and internationalization [2].

### 3.6 Smart Lifestyle

Smart lifestyle refers to improving citizens' quality of life with the help of cultural facilities, health conditions, individual safety, housing quality, touristic attractivity, education facilities, and social cohesion. Smart lifestyle means a more harmonious life that implies the development of each of these aspects, especially through e-health (alarm systems for patients, electronic health card, remote home control, on-line medical services) and e-inclusion and accessibility [3] [2].

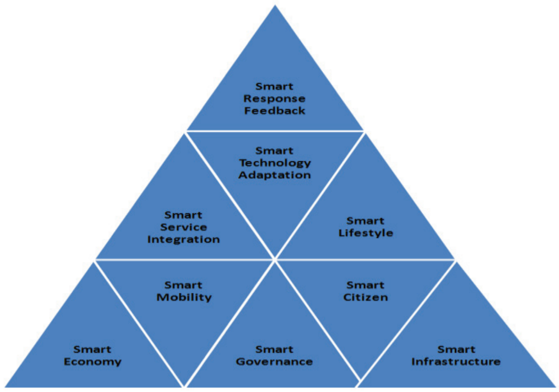



**Fig. 2.** Pyramid of smart city dimesions

### 3.7 Smart Technology Adoption

In order for the smart cities to foster and survive, smart technology needs to be adopted by various stakeholders. This is an integral part of the success of smart cities. The adoption of smart technology takes place with the help of proper education that is provided from the respective governments to its citizens. Smart technology adaptation is supported by increase in knowledge base of the local population. Furthermore, the adoption of technology is also a function of easy availability [11]. Once technology is widely available, it is easy to adopt. An example that can be taken in the real world includes the wide availability of technological innovations in the current Era. It can be easily seen that previous generations were not adoptable to technological innovation but the current generation is highly adaptable to technological sophistication. The major reason behind that is the availability of technology. Therefore smart cities incorporate high technology adoption within its residence base in order to function as successful smart cities [12].

### 3.8 Smart Service Integration Level

Smart service integration level refers to an integrated environment in which all the aforementioned factors could work collaboratively. The concept of smart city would not be achieved unless all the smartness factors work with each other while sharing the information. For instance, if a smart city achieves a smart economy but do not provide ICT services to its citizen's then cycle of smartness will remain incomplete [13].

### 3.9 Smart Response Feedback Model

The smart cities using the ICT incorporate a very successful citizen feedback model in order to record response of the people. Citizen response is a critical element in the development and proper functioning of the smart cities. A very comprehensive response feedback communication model is very important because smart cities use real time communication to adopt quickly in order to meet citizen needs. An example can be of a resident requiring an urgent ambulance service. The smart response feedback model will be built in such a way that the resident's urgent need of ambulance is met through feedback model's response communication [14]. Smart cities continuously meet the needs of its citizens with these response models. Furthermore, the citizens also give their feedback regarding the services of the smart cities. Therefore the smart response feedback model incorporates a two way communication process in order to record any grievances of the residents and cities to adapt continuously to the needs of the citizens. This information input is used to provide better smart facilities and make the smart cities tailored to the needs of their citizens [15].

### 3.10 Smart ICT Maturity

The smart ICT maturity is an umbrella activity, which calculates the maturity of ICT services in each stage of the proposed evaluation framework. The smart cities need to calculate the maturity of the ICT services in order to safeguard the optimal functioning of the smart cities. The ICT maturity is a concept that has evolved over the years. It is an area that the smart city management considers a vital component of efficient working of the smart cities. The maturity is calculated with the help of algorithms, which cater to the specifications of various elements and variables [15]. The measurement of maturity is done with the help of high-powered computing. When the date is analyzed, seasoned analysis interprets it. Then only a final decision is made and incorporated in the functioning of the smart city [16].

Figure 2 provides the dependency levels as the service integration pyramid in the creation of smart and mature ICT services.

## 4 Case Study and Findings

Most countries are focusing on making their cities smart, while others have successfully developed their smart cities. Copenhagen has smart energy, Kyoto has smart government, Stockholm has smart traffic, Zurich has smart healthcare, and New York has smart crime fighting, are some examples [17]. We have chosen Barcelona, Amsterdam, Stockholm, Hong Kong, and Dubai as our test cases under the following criteria for the benchmarking. Barcelona is considered as one of the best examples of a smart city. Amsterdam is considered because of its long lasting policy of creating sustainability through a process of environmental pollution control, energy reduction leading to behavior of responsible and smarter inhabitants. Hong Kong with its exploding population and structural implementation of smart technology and Stockholm with its participatory democracy and smart city innovation. Dubai is chosen because it is a relatively new city that build a reputation as a vibrant city in the Middle East. Riyadh is chosen as it is a city not so open according to the global standards and has a long journey to be established as a smart city.

**Table 1.** Analysis of Framework for smart cities based on Characteristics, factor, and Services

| Characteristics | Factor | Services |
|---|---|---|
| Smart Infrastructure | • Attractively of Natural Conditions.<br>• Pollution.<br>• Environmental Protection.<br>• Sustainable Resource Management. | • Telecommunication<br>• Decrease in the Amount of Waste Collection and the Number of Trash trucks<br>• Maintenance of Sanitary at Public Area<br>• Decrease in Illumination Electricity |

| | | |
|---|---|---|
| | | • River Protection<br>• Barrier-Free Space<br>• Maintenance of Environment Quality<br>• Energy-Efficient Housing<br>• Efficient Transportation<br>• Power Grids<br>• Cloud Services<br>• Digitizing and Eliminating Paper<br>• Unified Communications (UC).<br>• IT Equipment for Safety<br>• Intelligent Traffic Light<br>• Waste Management System<br>• Smart Lighting. |
| Smart Citizen | • Level of Qualification<br>• Affinity to Lifelong Learning<br>• Social and Ethnic Plurality<br>• Flexibility<br>• Creativity<br>• Cosmopolitanism/Open Mindedness<br>• Participation in Public Life | • E- Learning Portals<br>• E- Services<br>• Digital Forums<br>• Integration of Immigrants<br>• Spending More Time with Family and Community.<br>• Information Lifecycle Management (ILM).<br>• Education Welfare and Unemployment Insurance.<br>• Capability Standard Cosmopolitanism.<br>• Education Levels |
| Smart Governance | • Participation in decision-making<br>• Public and social services<br>• Transparent governance<br>• Political strategies & Perspectives | • Virtual Tours<br>• Electronic Identity Cards (e-IDs).<br>• E- Government Services<br>• Electronic Poll<br>• Digital Audit. |
| Smart Mobility | • Local Accessibility<br>• (Inter-) National Accessibility<br>• Availability of ICT-infrastructure<br>• Sustainable, Innovative and Safe Transport Systems | • E-planning<br>• Transportation Services<br>• E- Learning Portals |
| Smart Economy | • Innovative spirit<br>• Entrepreneurship<br>• Economic Image & Trademarks<br>• Productivity<br>• Flexibility of labor market<br>• International Embedded-ness<br>• Ability to Transform. | • Telecommunication<br>• Citizen Mutual Fund<br>• E- Services<br>• E-Banking<br>• Digital Forums<br>• City Share Trading<br>• Customer Relationship Management (CRM) Software |
| Smart Lifestyle | • Cultural Facilities<br>• Health Conditions<br>• Individual Safety<br>• Housing Quality<br>• Education Facilities<br>• Touristic Attractively<br>• Social Cohesion. | • Virtual Office Space<br>• Mobile Health Clinic<br>• E-Planning<br>• Transportation Services<br>• Educational Services<br>• E-Information Services |

| | | |
|---|---|---|
| Smart Technology Adoption | • Accommodation Quality<br>• Tourism<br>• Academic Quality. | • Virtual Tours<br>• Assisted Living<br>• E-Commerce<br>• E-Parking<br>• E-Travel Planning<br>• E-Public Information Dissimilation |
| Smart Service Integration | • Access<br>• Technology Availability<br>• Citizen Adoptability | • Open Data to Reduce TCO Service-Oriented Architectures (SOA)<br>• High Speed Internet<br>• Free Wi-Fi |
| Smart Response Feedback | • Technology Integration | • Digital Forums<br>• Electronic Poll<br>• Digital Audit |

Overall, Barcelona has implemented all characteristics that are present in any smart city.

**Table 2.** Quantitative Masure of Factor Among Smart Cities

| **City** | **Percentage** | **Smartness Level** |
|---|---|---|
| **Barcelona** | **100%** | **9** |
| **Amsterdam** | **90%** | **8** |
| **Dubai** | **50%** | **5** |
| **Hong Kong** | **70%** | **6.5** |
| **Stockholm** | **80%** | **7** |
| **Riyadh** | **0%** | **0** |

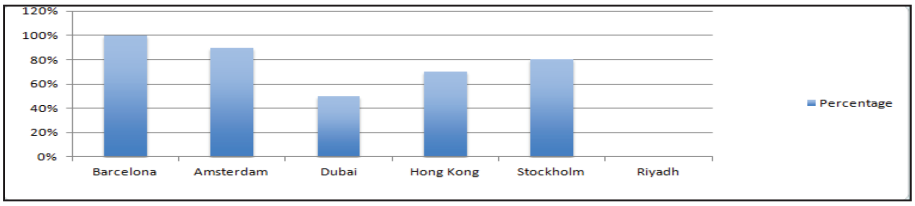



**Fig. 3.** The graphical comparison of smart cities

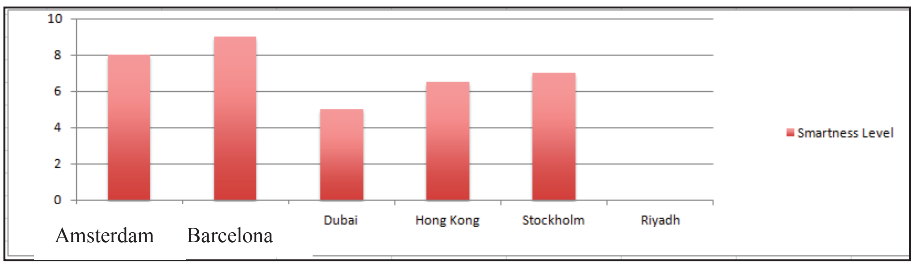



**Fig. 4.** The graphical comparison of smart cities

**Table 3.** Comparison of services among smart cities

| Stockholm | Hong Kong | Dubai | Barcelona | Amsterdam | Services |
|---|---|---|---|---|---|
| X | X | X | X | X | Telecommunication |
| X | X | | X | X | Maintenance of Sanitary at Public Area. |
| | X | X | X | X | Decrease in Illumination Electricity. |
| X | | | X | X | Maintenance of Environment Quality. |
| X | X | X | X | X | Efficient Transportation |
| X | X | X | X | X | Cloud Services. |
| X | | X | X | X | E-Learning Portals |
| X | X | X | X | X | E-Services |
| X | X | X | X | X | Digital Forums |
| X | X | | X | X | Electronic Polls |
| X | X | | X | | Digital Audit |
| | X | | X | | E- Government Services |
| X | | | X | | E-Planning |
| | X | | X | X | Transportation Services |
| X | X | X | X | | E- Learning Portals |
| | X | X | X | X | Online Investment |
| | X | X | X | | City Share Trading |
| X | | X | X | X | Mobile Health Clinic |
| X | X | X | X | | E-Planning |
| X | X | | X | X | Transportation Services |
| | | X | X | | Educational Services |
| X | X | | X | | E-Information Services |
| | X | X | X | | E-Commerce |
| X | | | X | X | E-Parking |
| | | | X | | E-Travel Planning |
| X | X | | X | X | High speed Internet |

## 5 Conclusion

The research has introduced the smart city concept and consolidated different viewpoints in research arena of smart city frameworks into a consolidated framework that maps smart city characteristics to factors representing ICT services. The study discusses the modalities of smart cities and how smart cities enrich the life of its citizens through efficient and optimal management of resources. The research takes a three-layered approach towards developing a framework that can help in evaluating a smartness of the city. The first layer talks about the smart infrastructure, smart citizen and smart governance. The model then progresses to layer two that consists of smart mobility, smart economy and smart lifestyles of the citizens. These are very critical steps in the framework of smart cities. The model then progresses to the final stage or the layer three. Layer three is: smart technology adaptation, smart service integration and smart response feedback model. All these three layers combined result in ICT maturity that eventually qualifies a city into transforming to a smart city according to the smart city benchmarking approach described in this research. The benchmarking is applied to smart cities to validate the process using Barcelona, Amsterdam, Stockholm, Hong Kong, Dubai, and Riyadh. The findings are incorporated to show how the ICT maturity is reached with the help of the three layered model with an analysis of real world examples of cities. Therefore this research has provided a benchmarking approach filling a gap in the evaluation approaches of city's smartness qualification based on ICT service maturity level.

## References


1. Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., & Meijers, E. (2007). *Smart cities Ranking of European medium-sized cities*. Vienna University of Technology.
2. Bowerman, B., Braverman, J., Taylor, J., Todosow, H., & Von Wimmersperg, U. (2000, September). The vision of a smart city. In *2nd International Life Extension Technology Workshop, Paris*.
3. Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of urban technology*, *18*(2), 65-82.
4. Toppeta, D. (2010). *The Smart City Vision: How Innovation and ICT Can Build Smart, "Livable", Sustainable Cities*. The Innovation Knowledge Foundation.
5. Washburn, D., Sindhu, U., Balaouras, S., Dines, R. A., Hayes, N., & Nelson, L. E. (2009). Helping CIOs Understand "Smart City" Initiatives. *Growth*, *17*.
6. Washburn, D., Sindhu, U., Balaouras, S., Dines, R. A., Hayes, N., & Nelson, L. E. (2009). Helping CIOs Understand "Smart City" Initiatives. *Growth*, *17*.
7. Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of urban technology*, *18*(2), 65-82.
8. Hafedh Chourabi, Taewoo Nam,Shawn Walker,J. Ramon Gil-Garcia, (2012). *Understanding Smart Cities: An Integrative Framework.* 45th Hawaii International Conference
9. Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M. & Portugali, Y. (2012). Smart cities of the future. *European Physical Journal-Special Topics*, *214*(1), 481.
10. Ojo, A., Curry, E., & Janowski, T. (2014). Designing Next Generation Smart City Initiatives-Harnessing Findings And Lessons From A Study Of Ten Smart City Programs.
11. Hollands, R. G. (2008). Will the real smart city please stand up? Intelligent, progressive or entrepreneurial?. *City*, *12*(3), 303-320.
12. Plouffe, C. R., Hulland, J. S., & Vandenbosch, M. (2001). Research report: richness versus parsimony in modeling technology adoption decisions—understanding merchant adoption of a smart card-based payment system.*Information systems research*, *12*(2), 208-222.
13. Lee, J. H., Phaal, R., & Lee, S. H. (2013). An integrated service-device-technology roadmap for smart city development. *Technological Forecasting and Social Change*, *80*(2), 286-306.
14. Zygiaris, S. (2013). Smart city reference model: Assisting planners to conceptualize the building of smart city innovation ecosystems. *Journal of the Knowledge Economy*, *4*(2), 217-231.
15. Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., & Morris, R. (2011). Smarter cities and their innovation challenges. *Computer*, *44*(6), 32-39
16. Shepard, M. (2011). Sentient city: Ubiquitous computing, architecture, and the future of urban space. The MIT press.
17. I-Canada, Creating an Intelligent Nation, Community By Community. (2011) www.icanadanetwork.ca/about-i-canada/

# A Realtime Associative Computing System for Interactive Information Exchange in a Multi-database Environment

Fuminori TSUNODA[a,1] and Yasushi Kiyoki[b]
[a] *East Japan Railway Company, Keio univ.*
[b] *Keio univ.*

**Abstract.** A multi-database environment is commonly important for creating new values by integrating heterogeneous data resources. We have designed a realtime association computing system for interactive information exchange among multi-databases. The metadatabase system organized to measure a relationship of interactive data and define a feedback control output to a system. In this paper, we present the applicability of this method to a multidatabase on railway information. We show several experimental results which have been obtained by associative computing for two different databases as a multidatabase environment. By those results, we clarify the effectiveness of the associative computing method in the actual multi-databases.

**Keywords.** Multidatabase, Feedback System, Optimization, Train Control

## 1. Introduction

A lot of sensor data have sent to databases on a wide - area network, the system has evolved rapidly. We have opportunities to obtain significant information from a lot of heterogeneous databases. However, when there is incessant interaction among several databases, it is difficult to design a metadatabase to gain an appropriate output. The associative rule between databases is changing at all times, it is not always true that the metadatabase utilizes multiple databases.

With the rapid progress of global network and database technologies, a large number of legacy databases are connected to the wide-area network. Those databases have been constructed and accessed independently in the wide-area network environment. By implementing an interconnection mechanism among these legacy databases, the values of the legacy databases gain significantly[2]. Particularly, it is effective to introduce the concept of spatial and temporal database computations to a metalevel of multidatabase environments, because this concept realizes the interconnection among heterogeneous databases according to spatial and temporal contexts. The meta-level of a multidatabase system means an abstracted and higher layer of local databases, and it would be constructed independently to the local systems. The interconnection, according to spatial and temporal contexts, means to join databases by computing spatio-temporal semantics and spatio-temporal relationships

[1] Corresponding Author.

which involve contexts defined in specific relationships among local databases. In this paper, we present a system architecture and an implementation method of a multidatabase system which realizes the interconnection by computing spatial and temporal relationships according to spatial and temporal contexts.

In a conventional approach for querying and integrating heterogeneous databases, the properties among data values like equality, synonymity, similarity or topology, which are defined as relationships, must be described statically as couples of pattern descriptions. The pattern descriptions are computed by using the pattern-matching technique with pattern descriptions of another database. This process which we call a relationship conversion realizes computational mechanisms for relationships rather than equality in the conventional approach. Currently, a large number of accessible databases are connected to the global network, therefore, an overhead for generating and updating the static descriptions becomes heavy to convert many kinds of relationships among the large number of databases to the single computational mechanism of equality. There are several research activities for converting the relationships to equality among heterogeneous databases automatically or semi-automatically. For example, an evaluation method of schema similarities using neural networks[9] and an evaluation method for equality among data values of heterogeneous databases using ontology[5] have been studied.

Our research goals are to provide a system framework for computing the relationships between data values of heterogeneous databases and to realize data integration for heterogeneous databases in the railway network environment.

There is not a data integration method with the mechanisms for computing many kinds of relationships comprehensively. It means that the conventional data integration systems which compute specific relationships have been designed and implemented independently. Therefore, they have their own application scopes, operations and data formats independently to others. The framework for integrating these data integration systems to share their functionalities and data representations is essentially needed. The systems in [1][10] compute topological, directional and connective relationships among temporal representations in an one-dimensional Euclidean space. The system in [4] computes topological, directional and connective relationships among spatial representations in a three-dimensional Euclidean space.

To realize a framework for computing spatial and temporal relationships between data values of heterogeneous databases dynamically, we formalize a hybrid computational method that integrates computational systems for spatial and temporal relationships. This method is designed in a meta-level, which corresponds to an abstracted level of local systems, independently to the implementations of legacy databases. Information resources with spatial and temporal representations are increasing in a wide area computer network. Spatial and temporal computation technology can be applied to information retrieval for those resources from spatial and temporal viewpoints.

In this paper, we present a new computational method for spatial and temporal evaluation method between spatial and temporal representations in the railway network environment. We have designed a realtime association computing system for interactive information exchange among multi-databases. The metadatabase possesss some functions to produce an appropriate output.

The rapid urbanization has progressed in the world, the realization of so-called "Smart City" is required. “Smart City” aims at the total optimization of an energy, water, transportation, and so on. Therefore, the data  of each resources are necessary to grasp the situations, the multi-database system performs a valuable role to promote efficiency.

A railroad as a mass rapid transit system is desired much more efficiency. Especially, during the time of a traffic distubance, a railroad operator manages to ensure the transport capacity. Many studies for improving the operation have been made with the view of the railway operators, using the data which the operators can control and grasp directly, such as train delay time or the number of suspended trains. But now it is needed for analysis that the operations' consequence of passenger flow and the measurement data of passenger movement.

The traffic information environment has been changed these days. The sensor data of train cars and the log data of signals on the trucks can be obtained through the network on a real-time basis. But the utilization of these data is limited. In this study, applying multi - database solutions for those databases, we have shown that the railroad operators can improve their operation. We have designed a train control method witch realizes maximum transport capacity during a traffic disturbance.

## 2. Situation modeling

### *2.1. The data relationship*

In this study, we have targeted the correlation data that stored in each database system. Even if the systems are managed and operated independently, the object targeted by the systems, it is often to be viewed the same object from a different angle. Then, Sometimes there are some relationships between the data, hierarchical, containment, causal, and so on.

Figure 1. is shown the image that different system handle partly the same object.

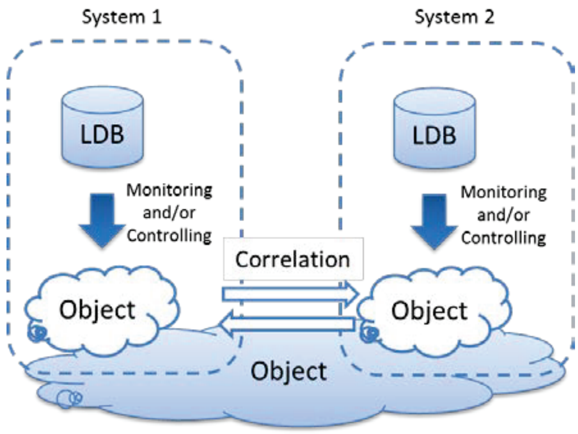



**Figure 1.** An image of some systems handling the same object

We especially have targeted causal interactive relation data. The data of the objects are influenced each other and one of the objects causes the other like a synergy effect. Synergy effect is difficult to understand the situation, but multidatabase method would be one of the key solutions.

The interaction data means that it has a featuring situation which the objects influence each other, the degree of change became more and more stronger. For example, if the data of the pressure gauge and the thermometer has been stored in a separate database, it is impossible to predict the change with only one of the data. Based on a gas state equation, temperature rises the higher the pressure, the pressure rises a higher temperature.

### *2.2. System Structure*

Generally, a simulation is valid for analyzing a complex phenomena. However, When there is a system for controlling the certain phenomena, it often needs the high responsiveness in order to control it speedy. In recent years, although the computer performance has been dramatically improved, it is difficult to apply the feedback based on the simulation results to the system to response in seconds. So far a multi-database and active database system, discussion of response and computer resources was little. However, in order to meet the system needs for handling growing amount of data, a mechanism for the response is important.

When we handle the data with the interaction, it is not enough to manage one database to monitor the threshold. It is necessary to monitor the threshold in each database. A database system is insufficient to produce the control data as a total optimization. The system needs to contact the plurality of database systems and work in conjunction. The system has a system architecture with the following three features.

1. Real-time associative computing

Because the event occurs always sporadically, it is invalid to aggregate the data always. But the data is needed to make a condition rule for the event, the metalevel system obtain the local database data at regular interval. The system start to associate and correspond to solve the problem in cooperation only when necessary, as a result the system suppresses the addition of excessive computer resources. Before the event occurs, the metalevel system has generated the condition rule. Therefore the system runs quickly compared with the system which generate the rule at the each time.

2. Automatically update with the rule condition

Creation role of the control data is updated every certain period of time. The rule is different by the time or day, week, and so on. Based on the analysis of new events occur, The system have a mechanism for setting the data acquisition rules for each of the database systems.

3. Reduction of network capacity

Actually, there are network load problems to aggregate the local databases. In addition, real-time collaboration system is divided into rule generation mechanism with

The proposed system has three steps: Rule Definition and Sharing, Associative Judgement Process and Control Data Generating Process.

#### *2.2.1. Rule Definition and Sharing*

At the first, it is determined the processing rule for the system. The local databases save and transmit the data to metadatabase at the regular intervals. The processing rule is generated at the metalevel system(1). Based on the result, the condition rule of generating control data is updated(2). Next the processing modules' activation rule are updated by the condition rule of metadatabase system. Figure 2. is shown the image of this step.

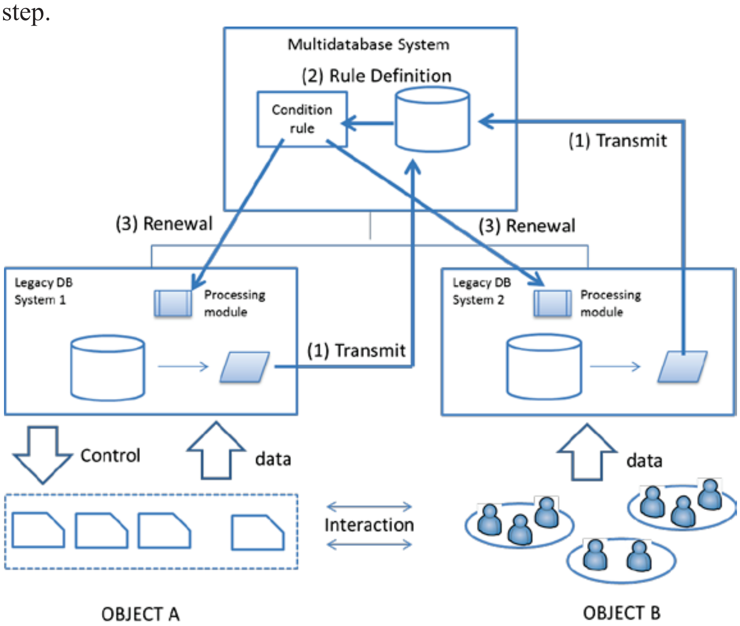



**Figure 2.** Rule Difinition and Sharing

### *2.2.2. Associative Judgement Process*

Next, Associative judgement process are shown at the figure 3.

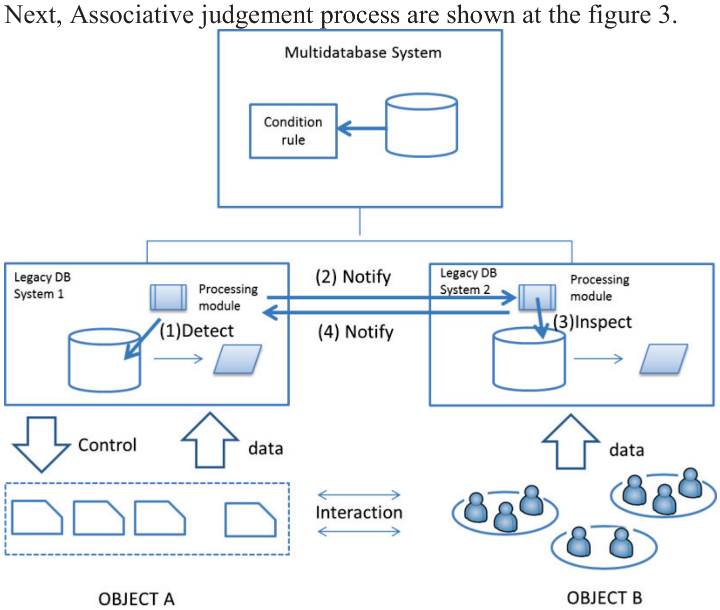



**Figure 3.** Associative Judgement Process

A processing module monitoring the local database. If the module detects the data(1), such as exceed the threshold, the module notifies the related module of other local databases(2). If a module detects the data, it is not possible to determine whether to process, because starting the process is sure to need multiple data for judgement in the case of handling the interactive data. Therefore the module which is notified by another local database's module start to inspect the value stored in the databases(3) and notify the result to the other module(4). By checking among local databases and inspecting own databases, the exchange data has been able to reduce.

### *2.2.3. Control Data Generating Process*

If the processing module of local databases detect the exceed of the threshold in the local databases, it starts to work associatively. Figure 4 shows the process flow. The local databases transmit recent data to the metalevel system(1), the metalevel system generates the temporaly database(2). For generating a control data, the tempolary database is calculated based on the condition rule(3). As a result, the control data is generated and transmitted to the local system which control the objects(4). Lastly, Legacy system control the object(5).

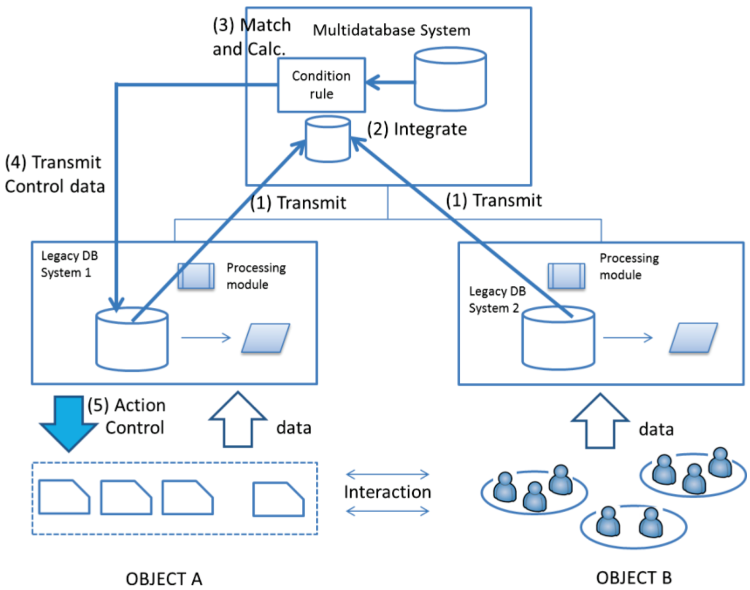



**Figure 4.** Control Data Generating Process

## 3. Data sets

### *3.1. Passenger loading data*

The recent railway train car is eq1uipped with many electronic control devices. In order to check the operation, various sensors are mounted as well. Brake power of the car is controlled according to the load of the car, because it is important to reduce impulse for the front and rear. Therefore, sensors attached to a spring of the carriage for supporting the load of the vehicle body. The data are transmitted to the control system as the value of the load in real time. The value of this load, since it is approximately proportional to the number of people riding, and is a numerical value that can be known roughly the number of passengers in the car. This data structure is as follows.

- Train number (unique)
- Station1 (unique)
- Station2 (unique)
- Planned time
- The number of loading passengers

Figure 5. is shown the sample data that summed a day in each section along a line.

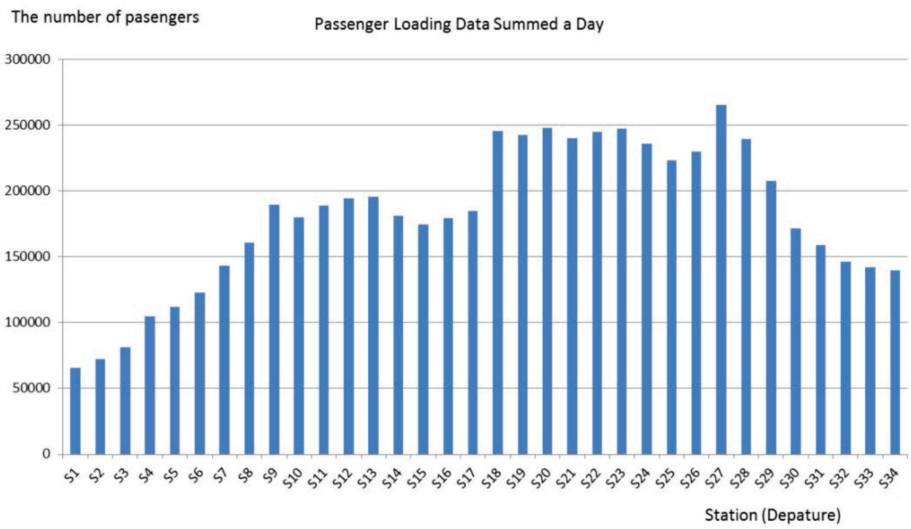



**Figure 5.** Passenger loading data a day

### *3.2. Signaling log data*

Generally, railway divides the line to the track in order to prevent collisions of trains, each track is to be ensured a safe. The front of each track has traffic lights are placed, the train runs by checking it. The traffic signals are controlled at the central control center, passage time of trains is notified to the server of central control center.

Basically, train behaves as an autonomous distributed object, the signal can control the train behavior, such as starting time of stations, for keeping a constant distance.

Data is acquired in this system is the time when the shorted each track, which can be accurately obtained in seconds. Because acceleration from departed the station is no difference by the train, departure time data at the staion is useful to examine the interval as compared to other trains. This data structure is as follows.

- Train number (unique)
- Station (unique)
- Depaturing time of the station
- The number of loading passengers

Figure 6. is shown the sample data that calculated the duration between the trains. An accident occurred in the morning on a day, some durations are extended. The area enclosed by the dots line shows the extended durations due to the accident.

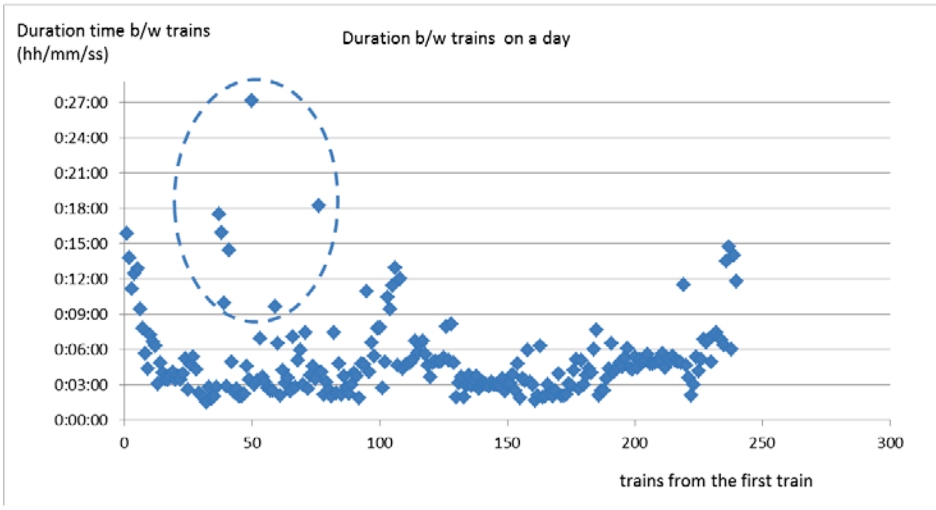



**Figure 6.** Duration between trains on a day

Realizing the smart city, the efficiency of public transport is very necessary. In the field of railways, in recent years, log data of the sensor data and the signal of the trains are able to be obtained through the network in real time. In this study, We propose the architecture of the system with real-time feedback mechanism to realize the train control method to improve the transport capacity. When the accident occurred on the railroad, congestion has happened. After that, the dispatcher makes an effort to recover the operation. Then, there is an interactive causal relation between the congestion and train interval. We propose a system that has a train control method to realize the transportation capacity improvement in railway environment.

In this study, the relationship of the congestion and the train interval control is mainly targeted at the train delay. Previously, the signal data based on the train control system (a train delay data) had been acquired in real time, it is difficult to grasp a congestion situation in real time. However, in recent years, sensors which have equipped on the vehicle and the enhancement of the wayside of network, congestion rate data can now be obtained in real time.

## 4. Experimental Simulation

Figure 7 shows relationship between train interval and loading rate at the section of A sta. and B sta. Basically, the trains run on a 3 minute interval.

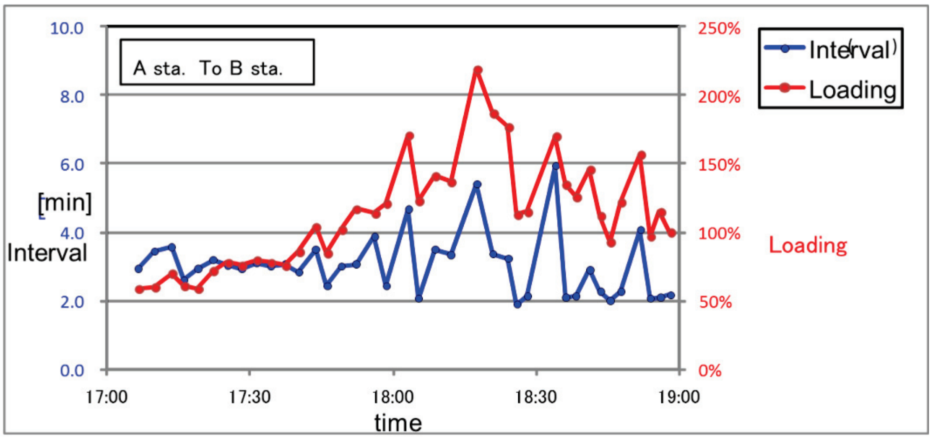



**Figure 7.** The relationship between train interval and loading rate

For this time period, it is considered that the train starts to delay when the interval exceeds 4 minutes and the loading exceeds 150%. Therefore, the figure(4 min. and 150%) is set to the processing module at each local database.
Discussions on above, the controllable object is just a train. Therefore, for minimizing the delay, we have to manage the train on a certain rule.

In this case, the rule at a station was determined like below

A) When trains interval exceed 4 min., check the loading rate.
B) When the loading data exceed 150%, check the interval time
C) The condition meets A) and B), generate control data to stop one train before until the condition A) or B).

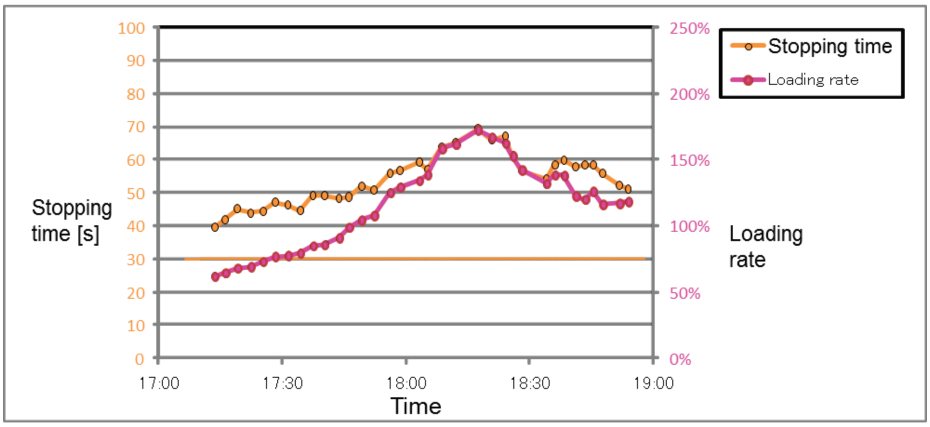



**Figure 8.** The ideal operation for the train stopping time

The rule A) is set to a local system which manages the train control. The rule B) is also set to another local system which monitors the passenger loading of the train. In addition, The rule C) is installed to the condition of metadatabase.

Based on the above condition, the experimental simulation was conducted. Figure 8. is shown the simulation data. The control system manages a stopping time of trains at the station based on the rule. The result shows that the loading data has not an extremely peak, the train runs more smoothly.

## 5. Conclusion

In this paper, we have presented the data integration method for heterogeneous legacy databases by combining spatio-temporal relationships between trains and passengeres in the railway environment. The feature of the system is to provide computational mechanisms of both static and dynamic relationships between trains and passengers, and they are integrated in the meta-level architecture in a multi-database environment.

The metadatabase system organized to measure a relationship of interactive data and define a feedback control output to a system. In this paper, we have presented the applicability of our method to a multidatabase on railway information. We have shown experimental results which have been obtained by associative computing for two different databases as a multidatabase environment.

As the future research, we will design and implement a computational system for data mining in the metalevel system and an active database mechanism which deals with the computations for spatial and temporal relationships among heterogeneous databases in a large scale of railway environment.

## References


[1] J.F. Allen, Maintaining Knowledge about TemporalIntervals, Communications of the ACM, Vol.26,1983, 832-843.

[2] Egenhofer, M.J.: “Spatial Relations: Models, Inferences, and their Future Application”, Proc. Advanced Database Symposium ’96, separate volume, Japan (1996)

[3] Egenhofer, M.J., Rashid, A. and Shariff, B.M.: “Metric Details for Natural-LanguageSpatial-Relations”, ACM Transaction on Information System, Vol.16, No.4, pp.295–321 (1998)

[4] Guttman, A. “R-Trees: A Dynamic Index Structure for Spatial Searching”, Proc. of the 1984 ACM SIGMOD Int’l Conf on Mgmt of Data, pp.45–57 (1984)

[5] Ishibashi, N., Hosokawa, Y. and Kiyoki, Y.: “A Spatial and Temporal Data Integration Method for Heterogeneous Database Environments”, Proc. the IASTED International Conference APPLIED INFORMATICS Symposium 2. Networks, Parallel and Distributed Processing, and Applications, pp.323–330 (2001).

[6] Y. Hosokawa, N. Ishibashi, Y. Yashiro and Y.Kiyoki, A Data Integration Method Realizing Evaluation for Temporal and Spatial Relationships in a Multidatabase Environment, Information Processing Society of Japan Transactions on Databases, Vol.40, No.SIG 8(TOD4), 1999, 95-111.

[7] M.W. Bright, A.R. Hurson and S. Pakzad, A Taxonomy and Current Issues in Multidatabase Systems, Computer, 25(3), 1992, 50-60.

[8] Y. Kiyoki, T. Kitagawa and Y. Hitomi, A fundamental framework for realizing semantic interoperability in a multidatabase environment, Journal of Integrated Computer-Aided Engineering, 2(1), 1995, 3-20.

[9] W.S. Li and C. Clifton, Semantic Integration in Heterogeneous Database Using Neural Networks, Proceedings of the 20th VLDB Conference, 1994, 1-12.

[10] A. Morishima and H. Kitagawa, A Data Modeling and Query Processing Schema for Integration of Structured Document Repositories and Relational Databases, Proceedings of the 5th International Conference on Database Systems for Advanced Applications (DASFAA '97), 1997, 145-154.
[11] N. Yoshida, Y. Kiyoki and T. Kitagawa, An Associative Search Method Based on Symbolic Filtering and Semantic Ordering for Database Systems, Proc. of 7th IFIP 2.6 Working Conference

# MONETARY POLICY TOPIC EXTRACTION BY USING LDA -Termination of Asian Financial Crisis-

Yukari SHIROTA[a,1] Takako HASHIMOTO[b] Tamaki SAKURA[c] and Basabi CHAKRABORTY[d]

[a] *Faculty of Economics, Gakushuin University*
[b] *Faculty of Commerce and Economics, Chiba University of Commerce*
[c] *Japan Center for Economic Research*
[b] *Faculty of Software and Information Science, Iwate Prefectural University*

**Abstract.** In this paper, the minutes of the monetary policy of the Bank of Japan, the central bank of Japan, for the year 1998 has been analyzed for the extraction of the topics concerning Asian financial crisis. The currency crisis started in Thailand in 1997 and spread toward many Asian countries. We analyzed the Monetary Policy Meeting minutes by text mining technologies. Especially we conducted topic extraction from the meeting minutes using Latent Dirichlet Allocation (LDA) model and the time series of the changes of extracted topic ratios are shown. From the analysis results, one topic which seems to be the Asian financial crisis related topic can be found. The topic ratio change curve clearly corresponds to the calm down of economic indices such as currency exchange ratios and market interest rates in the Asian countries.

**Keywords.** Topic extraction, Dirichlet Allocation Model, Monetary policy minutes by Bank of Japan, Asian financial crisis

## 1. Introduction

Minutes of the Monetary Policy Meetings of the Bank of Japan (abbrev. BOJ), the central bank of Japan, are the most important sources of documents regarding the information about governmental monetary policy in Japan. The BOJ's Policy Board decides the basic strategies for the monetary policy at Monetary Policy Meetings. The Policy Board discusses the local and the global economic and financial situations and then decides an appropriate guideline for money market operations[1].

The objective of our research is to develop automatic methodologies for grasping the key points of the financial policy meeting minutes with the use of text mining technologies. Generally the human analysts read the minutes to understand the policymaking process. But the topics in the minutes are complicated and composed of various types, thus sometimes the human reader becomes confused in understanding the

[1] Corresponding Author.

underlying policymaking process. For example, a great variety of words including the "exchange rate fluctuation" and "Europe sovereign issue" appear in the minutes. So an automatic analysis process (software) that can produce several summarized topics from the meeting minutes is of utmost importance.

The analysts of minutes are not only interested in the topic contents (meanings) but also they are interested in the following things: (1) the time of appearance of the topic, (2) the duration of the topic, and (3) the cause factor of the topic which might be domestic factor or an international one.

One of our research goal is to measure policy-related economic uncertainty. Uncertainty in policy minutes becomes high when the future path of government policy is uncertain. As the policy uncertainty level is related to the economic risk, if the level is high, businesses become reluctant to investment. We think that it is significant to clarify the level of the uncertainty of the Japan.

Regarding measurements of policy uncertainty, there are researches by Baker et al. in which they attempt to link to the vigour of the economy [4, 5] with the policy uncertainty. Half of their economic-policy uncertainty index comes from a computerized reasoning of relevant references in 10 newspapers. In particular, they identify articles containing 'uncertainty' or 'uncertain', 'economic' or 'economy', and one or more of the following terms: 'congress', 'deficit', 'federal reserve', 'legislation', 'regulation' or 'white house' (including related terms like 'regulatory' or 'the fed') [6].

We have conducted experiments for topic extraction from the blogs and minutes previously just after East-Japan great earthquake in March 2011 [7] which was also a time when the uncertainty level in economy was high. The Asian financial crisis was an ordeal for Asia and then Japan was suffering from nonperforming loans handling. Therefore, Japan's economic uncertainty level was high from 1997 to 1998.

The Monetary Policy Meetings of BOJ are held once or twice in a month, for one or two days. Monetary policy has a significant influence on the daily lives of the public and the Bank needs to clarify to the public the content of its decision and its decision-making processes [1]. Hence, the point of the minutes is soon published on the web by the Bank of Japan in both Japanese and English[2]. 10 years after the date of meeting, the long version of the minutes are published on the web, as well as the short points of the minutes. In this paper, we would like to analyse the Japanese monetary policy at the time when the Asian financial crisis happened.

A currency crisis, a rapid decrease of the currency exchange rate, makes often the financial crisis occur in which the value of financial institutions or assets drops rapidly[3, 4]. The Asian currency crisis started in July, 1997 and got calm down in 1998. The afflicted countries were mainly Thailand, Philippines, Malaysia, Indonesia, and South Korea (Republic of Korea). Then the world economic conditions were as a whole in a slowdown state because the Asian currency crisis led to Asian financial crisis and because Russian economic scene was also bad. The financial crises were associated with a panic or a run on the banks, in which investors sold off assets or withdrew money from savings accounts with the expectation that the value of those assets would drop if they remain at a financial institution[2].

In this paper, we would like to analyse the Monetary Policy Minutes from January 1998 to December 1998. Though we are also interested in analysing policy minutes of

[2] Investopedia dictionary: http://www.investopedia.com/terms/f/financial-crisis.asp

1997, the published minutes by BOJ are from 1998 only. The minute text that we used for the topic extraction is the Japanese version, because we think that the Japanese version can convey more clearly the discussion as the Board meetings are conducted in Japanese language. The topic extraction model used in our work is the Latent Dirichlet Allocation (LDA) model.

From the analysis results, one topic which seems to be the Asian financial crisis related topic has been found clearly. The topic ratio change curve clearly corresponds to the calm down of economic indices such as currency exchange ratios and market interest rates in the Asian countries. In section 2, the topic extraction method using the LDA model has been explained. In section 3, we shall present related economic indices of Asian countries from 1997 to 1998. In section 4, topic extraction results by the LDA model are explained followed by the final section containing conclusion.

## 2. Latent Dirichlet Allocation (LDA) Model

In the section, the LDA model used here for the topic extraction is explained in a simple manner. The LDA is a widely-used multi-topic document model based on Bayesian intefrence method[5]. In the LDA model, each topic is supposed to have a set of related words. One document is supposed to have several topics. The topic distribution for document "i" may be (0.7, 0.3) (in the case of an economic related document) or (0.2, 0.8) (in the case of a disaster related document). To express the possible various distributions, we use Dirichlet distribution by using a hyper parameter $\boldsymbol{\alpha}$. On the same way, we define per-topic word distribution by Dirichlet distribution by using another hyper parameter $\boldsymbol{\beta}$. The used symbols are as follows:

$\boldsymbol{\alpha}$ is the parameter of the Dirichlet prior on the per-document topic distributions,

$\boldsymbol{\beta}$ is the parameter of the Dirichlet prior on the per-topic word distribution,

$\theta_i$ is the topic distribution for document $i$,

$\phi_k$ is the word distribution for topic $k$,

$z_{ij}$ is the topic for the $j$th word in document $i$, and

$w_{ij}$ is the specific word.

The $w_{ij}$ 's are the only observable variables, and the other variables are latent variables. $\boldsymbol{\phi}$ is a Markov matrix of size is $K \times V$ ($V$ is the dimension of the vocabulary, K is the number of topics) , each row of which denotes the word distribution of a topic. The LDA generative process for a corpus $\boldsymbol{D}$ consisting of M documents each of length $N_i$ is as follows :

1. Choose $\theta_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where $i \in \{1,...,M\}$ and $\mathrm{Dir}(\boldsymbol{\alpha})$ is the Dirichlet distribution for parameter $\boldsymbol{\alpha}$
2. Choose $\phi_k \sim \mathrm{Dir}(\boldsymbol{\beta})$, where $k \in \{1,...,K\}$
3. For each of the word positions $i, j$, where $j \in \{1,...,N_i\}$, and $i \in \{1,...,M\}$
    (a) Choose a topic $z_{ij} \sim \mathrm{Multinomial}(\theta_i)$.
    (b) Choose a word $w_{ij} \sim \mathrm{Multinomial}(\phi_{z_{ij}})$.

The Dirichlet-multinomial distriutions are defined in machine learning textbooks[8]. We want to obtain an estimate of $\boldsymbol{Z}$ that gives high probability to the words that appear in the corpus. $z_{ij}$ represents the topic for the $j$th word in document $i$. This problems becomes the maximum a posteriori estimation of $P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. By an integration concerning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, the expression becomes a simple one, $P(\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$. Therefore, we want to obtain Z so that $P(\mathbf{Z}|\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is maximum. The $\boldsymbol{W}$ is given data. The cost of the calculation is too high because the estimation space size is the number of topics ($K$) to the power of the dimension of the vocabulary ($V$), $K^V$. Namely each word has $K$ options independently. So instead of that, a random walk search method by Gibbs sampling is widely used[6]. We used the R packaged offered by “The Comprehensive R Archive Network abbreviated as CRAN titled “lda: Collapsed Gibbs sampling methods for topic models” developed by Jonathan Chang[3].

## 3. Asian Financial Crisis

In this section, the Asian financial crisis using the economic data has been explained.

The start of the crisis was as follows[3]; In Thailand during the early 1990s, the currency, the baht, were pegged at 25 to the dollar. Meanwhile in the US interest rates were very low. Thai companies and individuals took advantage of these circumstances to borrow hefty sums in dollars, feeling confident that they would be able to earn enough baht to pay off their dollar debts by exchanging them at the bargain rate of 25 to the dollar. The result was an explosion of speculative investment.

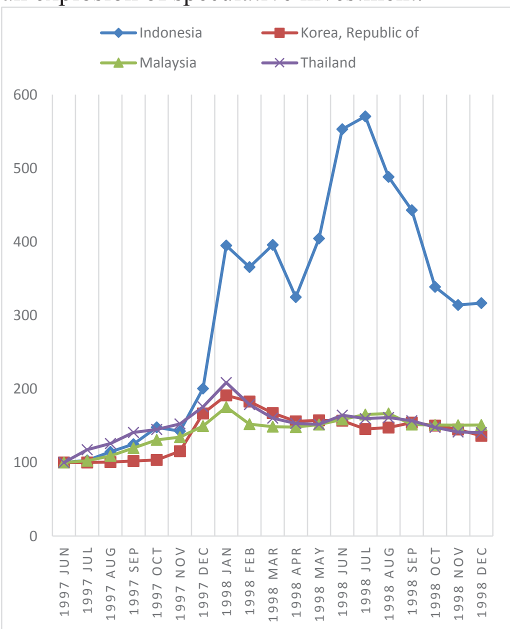



**Figure 1.** Short National currency per U.S. dollar, period average. The base is June 1997. The greater value corresponds to weaker currency against a US dollar. (The data is cited from IMF.)

[3] CRAN web page: http://cran.r-project.org/

The Asian currency crisis had spread to Thailand, Philippines, Malaysia, Indonesia and South Korea (Republic of Korea). Figure 1 shows each country's currency exchange rate against US dollar. The base value is set to be 100 percent in June 1997, just before the crisis. As shown here, the currency crisis in Indonesia is the worst. The worst period in the countries except Indonesia was January, 1998. The worst period in Indonesia was July, 1998. From the data, we can see that Indonesia's recovery was delayed.

During the crisis the country wise interest rates were also rapidly increased and changed (see Figure 2). Concerning the interest rate, the worst increase was also seen in Indonesia and the second worst was seen in South Korea. The base value is set to be 100 percent in June 1997, just before the crisis.

From both the data, we found that the turmoil peak was in January 1998 and then the financial conditions were settled gradually.

In 1998 Japan's economy was in a prolonged slump and in quite a severe situation. Some major Japanese financial institutions had failed. The Japan's financial uncertainty level was so high owing to the nonperforming-loan problem of the financial institutions. In addition, the Asian financial crisis made Asian economic conditions worse, which led the growth of Japanese exports to the Asian countries to fall. Hence, Japan's economy was suffering from domestic and overseas problems at the same time and they anticipated that the recession would continue.

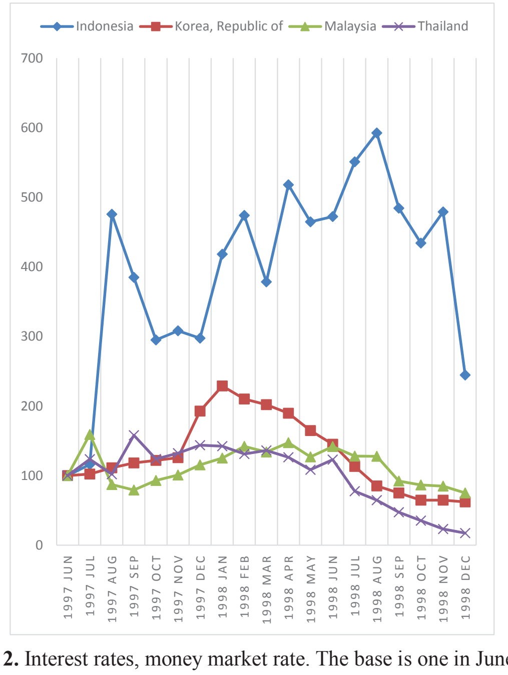



**Figure 2.** Interest rates, money market rate. The base is one in June 1997.
(The data is cited from IMF.)

## 4. Topic Extraction Results

In this section, the topic extraction results by LDA model is presented.
We have conducted the monthly topic extraction by the LDA model. Figure 3 shows the topic ratio time series changes. First, the LDA model generates the word frequencies for each topic for the whole one year text data. Then each month text data is input to the generated total model so we can obtain each month's topic ratio.

In LDA model, we have to decide in advance the number of topics. In this work, we decided the number to be six after some experimentation, as with the six topics, the Asian financial crisis topic clearly appeared. Table1 shows the comparison of frequently appeared word list of each topic. Table 1 gives us which words are concentrated on which topic. Table 2 shows the most frequently appeared words of each topic.

First, we would like to examine the Asian financial crisis topic which is the number 2.

**Topic 2: Asian currency/financial crisis**

The corresponding words are "lowering commodity prices", "bank", "Asia", "exchange rate", "East Asia", "South Korea", "Indonesia", "IMF", "recovery", "China", "Thailand", "trading", "turmoil", "stock (backlog)", and "decreased production". From the words and the appearance frequencies of these words on Topic 2, we think that the topic is related to Asian currency/financial crisis and its impact on Japan. The decline of the topic ratios curve shown in Figure 3 also corresponds to the fact that Asian financial crisis was stabilized in 1998 other than Indonesia (See Figure 1 and 2). As shown in Figure 1, the uncertain level on currency exchange rates was highest in January, 1998. With the decrease of exchange rates, Topic 2 ratio also decreased. This means that in the Board meetings, as the concerns about the Asian financial crisis diminished, the impacts on Japan's economy became smaller and they moved to other topics.

Topic 2 is the only one overseas issue topic and other five topics are all domestic issue topics. Ranking them from the biggest ratio, the topics (excluding topic 2) are as follows: [Topic 5] a lowering of policy interest rates, [Topic 6] weakened corporate activity, [Topic 4] decline in business fixed investment, [Topic 1] injection of taxpayers' money, and [Topic 3] money market operation. The topics are described below in the above order.

**Topic 5: a lowering of policy interest rate**

The corresponding words are "policy interest rate", "(monetary) easing", "deflation", and "nonperforming loans". From these words, we think that the topic is related to a lowering of policy interest rates by BOJ. The minutes on 9th September says that some meeting members indicated a bleak prospect that Japan's economy was on the verge of a deflationary spiral and that there were also concerns over employment conditions and the nonperforming-loan problem of financial institutions. In 1998, Japan's economy was in quite a severe situation and the paramount issue was nonperforming loan problem of financial institution. Then in the board meeting on 9th September, the lowering of policy interest rate was determined.

The followings are a part from the minutes on 9th September, 1998[4]; *At the conclusion of the Board's discussion, most members supported the view that, in the*

[4] https://www.boj.or.jp/en/mopo/mpmsche_minu/minu_1998/g980909.htm/

*implementation of monetary policy for the intermeeting period ahead, the Bank should further ease the stance of monetary policy, encouraging the uncollateralized overnight call rate to decline to around 0.25 percent, while providing more ample liquidity when judged necessary to maintain the stability of the financial markets.*

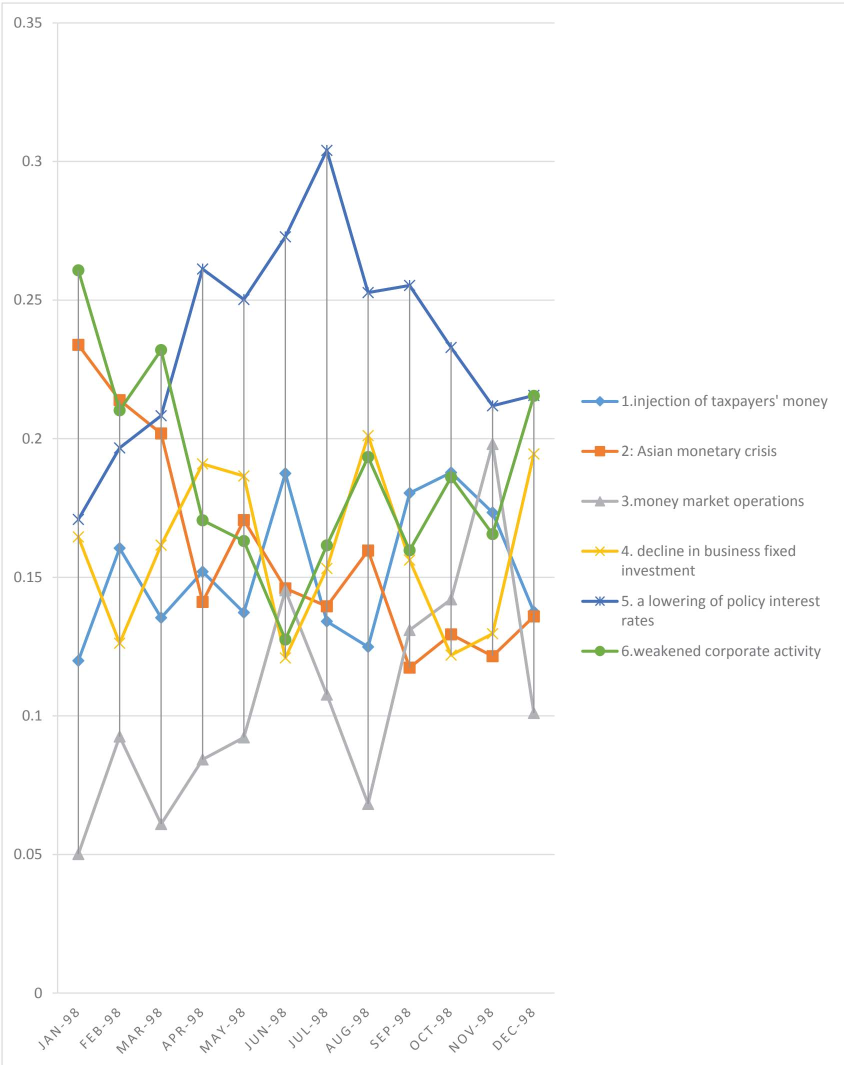



**Figure 3.** The topic ratio changes generated by the LDA model.

This topic represented a high uncertainty and the topic ratios through 1998 was much greater than others as shown in Figure 3. However, the peak of this topic was in July, not in September. We think that this is because economic indicators released on July 28 gave the big impact on Japan's economy which showed a considerable decline in business fixed investment and deterioration in employment and income conditions. The topic title

**Table 1.** The most frequently appeared word list of six topics generated by the LDA model. As the original word is Japanese, they are interpreted to sometimes two words. The color marked words are strongly connected words.

| Topic # | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| market | 371 | 22 | 0 | 0 | 0 | 0 |
| stock prices | 0 | 311 | 3 | 0 | 100 | 0 |
| slump | 0 | 463 | 0 | 0 | 0 | 0 |
| expectation | 0 | 0 | 0 | 0 | 350 | 212 |
| special | 0 | 0 | 0 | 0 | 62 | 70 |
| cutting tax | 0 | 0 | 0 | 0 | 260 | 22 |
| business conditions | 475 | 0 | 0 | 493 | 0 | 0 |
| public | 104 | 0 | 0 | 2 | 0 | 0 |
| fund | 684 | 0 | 0 | 0 | 0 | 336 |
| bank | 0 | 353 | 235 | 0 | 13 | 0 |
| accounting capital | 1 | 0 | 0 | 77 | 358 | 0 |
| uncertainty | 60 | 0 | 17 | 0 | 0 | 16 |
| operation | 0 | 24 | 342 | 0 | 0 | 0 |
| interest rate | 0 | 299 | 0 | 0 | 1309 | 0 |
| risk | 442 | 0 | 0 | 0 | 345 | 0 |
| concern | 114 | 0 | 21 | 252 | 0 | 0 |
| confidence | 0 | 0 | 0 | 0 | 174 | 0 |
| Japan | 0 | 0 | 0 | 0 | 0 | 64 |
| premium | 85 | 0 | 0 | 0 | 0 | 103 |
| loan-out | 392 | 0 | 0 | 0 | 0 | 228 |
| corporate bond | 73 | 0 | 0 | 0 | 0 | 89 |
| Asia | 0 | 460 | 0 | 0 | 0 | 0 |
| exchange rate | 0 | 301 | 0 | 0 | 19 | 32 |
| dollar | 0 | 530 | 0 | 0 | 0 | 0 |
| export | 0 | 0 | 0 | 530 | 0 | 0 |
| bind | 0 | 0 | 0 | 0 | 0 | 55 |
| government bond | 194 | 0 | 0 | 1 | 0 | 0 |
| bubble | 0 | 0 | 0 | 122 | 0 | 0 |
| commodity prices | 0 | 679 | 0 | 0 | 0 | 0 |
| pressure | 0 | 113 | 2 | 0 | 156 | 0 |
| deflation | 0 | 0 | 0 | 0 | 322 | 0 |
| East-Asia | 0 | 70 | 0 | 0 | 0 | 0 |
| South Korea | 0 | 126 | 0 | 0 | 0 | 0 |
| private | 167 | 0 | 0 | 71 | 37 | 0 |
| debt | 0 | 1 | 0 | 0 | 0 | 67 |
| foreign money | 0 | 0 | 0 | 0 | 0 | 161 |
| Indonesia | 0 | 94 | 0 | 0 | 0 | 0 |
| IMF | 0 | 65 | 1 | 0 | 0 | 0 |
| recovery | 0 | 178 | 0 | 0 | 338 | 0 |
| fund management | 20 | 1 | 0 | 0 | 0 | 107 |
| German | 0 | 86 | 0 | 0 | 0 | 0 |
| oil | 0 | 0 | 0 | 0 | 0 | 63 |
| decrease | 0 | 0 | 0 | 0 | 0 | 392 |
| central bank | 0 | 16 | 57 | 0 | 1 | 0 |

| Topic # | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| China | 0 | 109 | 0 | 0 | 0 | 0 |
| prices | 0 | 0 | 4 | 372 | 0 | 0 |
| consumption | 0 | 53 | 0 | 0 | 545 | 318 |
| individual | 0 | 0 | 0 | 172 | 105 | 0 |
| labor | 0 | 0 | 0 | 184 | 0 | 7 |
| employment | 0 | 14 | 23 | 539 | 0 | 0 |
| increase | 0 | 0 | 0 | 0 | 2 | 304 |
| sentiment | 0 | 0 | 0 | 2 | 186 | 0 |
| production | 0 | 0 | 0 | 498 | 0 | 0 |
| facilities | 0 | 0 | 0 | 0 | 0 | 656 |
| investment | 0 | 0 | 0 | 1005 | 0 | 0 |
| order received | 0 | 21 | 0 | 43 | 5 | 4 |
| inflation | 0 | 8 | 3 | 0 | 73 | 0 |
| Thailand | 2 | 56 | 0 | 0 | 0 | 0 |
| balance of payments | 0 | 0 | 0 | 93 | 0 | 0 |
| import | 0 | 7 | 0 | 4 | 0 | 248 |
| maintenance | 0 | 0 | 0 | 0 | 437 | 0 |
| trading | 0 | 52 | 0 | 0 | 1 | 32 |
| building construction | 0 | 0 | 0 | 54 | 0 | 0 |
| automobile | 0 | 0 | 0 | 147 | 0 | 0 |
| turmoil | 0 | 76 | 4 | 0 | 6 | 0 |
| domestic demand | 0 | 0 | 0 | 3 | 0 | 83 |
| quantitative | 0 | 0 | 0 | 1 | 136 | 0 |
| GDP | 0 | 0 | 0 | 200 | 0 | 0 |
| manufacturing | 0 | 0 | 0 | 247 | 0 | 0 |
| machinery | 1 | 0 | 2 | 92 | 0 | 0 |
| deterioration | 185 | 0 | 0 | 0 | 0 | 360 |
| easing | 0 | 0 | 0 | 0 | 414 | 0 |
| loan-out | 0 | 0 | 0 | 0 | 0 | 71 |
| wage | 0 | 31 | 0 | 73 | 0 | 0 |
| stock (backlog) | 0 | 378 | 0 | 0 | 0 | 0 |
| bankruptcy | 0 | 0 | 0 | 0 | 57 | 60 |
| decreased production | 0 | 75 | 0 | 0 | 0 | 0 |
| wholsale | 0 | 14 | 2 | 73 | 2 | 13 |
| tax | 0 | 0 | 0 | 0 | 82 | 0 |
| ratio of exchange | 30 | 0 | 0 | 0 | 0 | 36 |
| money supply | 62 | 0 | 0 | 0 | 0 | 94 |
| liquidity | 380 | 0 | 0 | 0 | 0 | 0 |
| deposit | 0 | 0 | 28 | 0 | 172 | 0 |
| asset | 0 | 0 | 0 | 0 | 229 | 0 |
| commercial banks | 0 | 0 | 0 | 0 | 0 | 67 |
| nonperforming loans | 0 | 0 | 0 | 0 | 205 | 0 |
| rating | 0 | 0 | 1 | 0 | 0 | 56 |
| restructuring | 0 | 0 | 0 | 127 | 0 | 0 |

we made implies both the final decision which was the lowing of policy interest rate, and the decision making process discussions.

**Table 2.** The word list of six topics generated by the LDA model. The coloring is just a mark of a concept group or just a mark.

| # | 1 | # | 2 | # | 3 | # | 4 | # | 5 | # | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| capital | 684 | commodity prices | 679 | market | 875 | investment | 1005 | interest rate | 1309 | corporate | 1322 |
| policy | 672 | dollar | 530 | operation | 476 | problem | 904 | policy | 912 | facilities | 656 |
| economic condition | 475 | lift | 492 | problem | 424 | minus | 696 | measurement | 891 | deteriorate | 602 |
| risk | 442 | tumble | 463 | determination | 417 | market | 625 | effect | 853 | sentiment | 392 |
| lending | 392 | Asia | 460 | rate | 367 | employment | 539 | consumption | 545 | decrease | 392 |
| creditworth | 387 | stock | 378 | operation | 342 | export | 530 | opinipon | 443 | worse | 360 |
| liquidity | 380 | domestic | 375 | supply | 330 | production | 498 | maintenance | 437 | capital | 336 |
| market | 371 | stock prices | 311 | CP | 240 | economic conditions | 493 | easing | 414 | consumption | 318 |
| government | 358 | currency exchange | 301 | bank | 235 | USA | 442 | institution | 385 | small firm | 305 |
| stability | 211 | interest rate | 299 | over | 160 | corporate | 397 | finance | 372 | increase | 304 |
| JGB | 194 | currency exchange | 271 | night | 149 | prices | 372 | capital | 358 | import | 248 |
| BOJ | 187 | recovery | 178 | external | 94 | concern | 252 | Japan | 357 | lending | 228 |
| worsen | 185 | South Korea | 126 | guided | 82 | manufacturing | 247 | expectation | 350 | expectation | 212 |
| private | 167 | China | 109 | bankruptcy | 70 | oberseas | 228 | risk | 345 | income | 206 |
| policy interest rate | 163 | Indonesia | 94 | execution | 69 | commodity | 227 | recovery | 338 | overseas | 195 |
| lending | 151 | caution | 93 | submit | 69 | GDP | 200 | deflation | 322 | indicator | 191 |
| additional | 144 | stability | 91 | soundness | 68 | housing | 196 | household finance | 304 | foreign currency | 161 |
| measurement | 141 | peak | 89 | lemding | 65 | labor | 184 | cutting tax | 260 | turmoil | 161 |
| lowering | 120 | German | 86 | discussion | 60 | unemployment | 173 | nonperforming loans | 205 | spread | 144 |

**Topic 6: weakened corporate activity**

The second biggest domestic topic is Topic 6. The related words are “corporate”, “facilities”, “deteriorate”, and “small firm”. In addition, there are words related to government bonds such as “Japan premium” and “(yield) spread” which can be interpreted as bad effects related words on Japanese corporate activity. Considering these, we think the topic is titled as “weakened corporate activity”. The topic ratio curve is U-shaped with the bottom in July. The bottom corresponds to the peak of Topic 1 which is injection of taxpayers’ money in June. We think that the government policy implementation made effects on the weakened corporate activities.

**Topic 4: decline in business fixed investment**

From the word frequencies in Table 1 and 2, we think that this topic is related to Japan's economic conditions. The minutes in August says as follows: *In the Board's discussion of the current economic situation, many members expressed the view that economic conditions in general had continued to deteriorate.*

We found that this Topic 4 time series changes was an inverse movement against the time series changes of Topic 1 which concerns the BOJ policies.

**Topic 1: injection of taxpayers' money**

From the word list, this topic seems to be related to BOJ policies to secure the stability of the finance markets. Especially, from Table 2, we found the word "taxpayers' money" on this topic. The co-occurrence word of that is in general "injection". Then, although there is no word "injection" in this topic word list, we make the title "injection of taxpayers' money" on this topic.

**Topic 3: money market operation**

This topic is considered to be related to money market operation, because the most frequent words are "(money) market operation". We think Topic 3 is a part of Topic 5. Topic 5 handles various financial policies such as monetary easing and handling of nonperforming bank loans. On the other hand, Topic 3 focuses just on the market operation which encouraged the uncollateralized overnight call rate to decline around 0.25 percent. In addition, the discussions before reaching the decision and the voting results are included in Topic 3.

From the topic extraction results, we made groups of topics as follows:

(A) Overseas problem: economic conditions: [Asian monetary crisis]

(B) Domestic problem: BOJ policy: [a lowering of policy interest rates],[injection of taxpayers' money], [money market operation]

(C) Domestic problem: economic conditions: [decline in business fixed investment], [weakened corporate activity]

## 5. Conclusion

In this work, we have conducted topic extraction using LDA model on the BOJ Monetary Policy Meeting minutes in 1998. In the period, Japan's economy was in very severe conditions and suffered from some financial institutions nonperforming loan problems. From the resultant time series changes of topic ratios, we found the Asian financial crisis topic. The topic movement corresponds to the currency exchange rate movement and both were decreasing curve from the January 1998. In addition, from the word list generated by LDA model, the topic has strong relationship to Asian financial crisis.

Other five topics are considered to be domestic related topics. Among them, there are three BOJ policy topics; they are [a lowering of policy interest rates], [injection of taxpayers' money], and [money market operation]. Other two topics are considered to be related to domestic economic conditions; they are [decline in business fixed investment] and [weakened corporate activity].

We think that these results can make us understand instantly the minute contents. For example, BOJ's biggest concern in 1998 was possibility of deflation and Board meeting members made the decision of a lowering of policy interest rate, as shown by

Topic 1. As far as the topic ratios of this result are concerned, the highest uncertainty was found in July 1998 through 1998. The peak of Topic 5 seems to show the most uncertain period in Japan's economy. On the other hand, as the curve of Topic 2 ratio shows, the uncertain level by the Asian financial crisis was calm down in 1998.

We would like to continue analyzing the BOJ Monetary Policy Meeting minutes by topic extraction methods in future.

## Acknowledgement

We thank Hirohiko Okumura, an emeritus professor of Gakushuin University for thoughtful comments that improved this paper. We thank Prof Tetsuji Kuboyama for his wide range of knowledge about machine learning that helps our research. This research was partly supported by funds from Gakushuin University Computing Centre as its special research project in 2015 and by JSPS KAKENHI Grant Number 15K00314 of which representative is Prof Takako Hashimoto.

## References

[1] BOJ. (10th November, 2014). *Outline of Monetary Policy*. Available: https://www.boj.or.jp/en/mopo/outline/index.htm/

[2] BOJ. (2014). *Minutes of the Monetary Policy Meetings 2014*. Available: https://www.boj.or.jp/mopo/mpmsche_minu/minu_2014/index.htm/ (https://www.boj.or.jp/en/mopo/mpmsche_minu/minu_2014/index.htm/)

[3] P. Stalker, *The No-Nonsense Guide to Global Finance* New Internationalist Pubns Inc 2009.

[4] T. Economist. (1997). *Mahathir, Soros and the currency markets*. Available: http://www.economist.com/node/101043

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research,* vol. 3, pp. 993-1022, 2003.

[6] T. L. Griffiths and M. Steyvers, ""Finding scientific topics," *Proceedings of the National Academy of Sciences,* vol. 101 (Suppl. 1), pp. 5228–5235, 2004.

[7] Hashimoto, T., T. Kuboyama, and Y. Shirota, Topic Detection about the Great Japan East Earthquake based on Emerging Modularity. Proc. of 22nd European Japanese Conference on Information Modelling and Knowledge Bases,Prague, Czech Republic , June 4-9 2012.

[8] Bishop, C, Pattern Recognition and Machine Learning (Information Science and Statistics), 2nd printing, Springer, 2010.

# Finding an Optimal Configuration of the Feed-forward Neural Network

Radoslav ŠTRBA[1], Jakub ŠTOLFA[1] , Svatopluk ŠTOLFA[1]
[1]*Department of Computer Science*
*VŠB – Technical University of Ostrava,*
*Faculty of Electrical Engineering and Computer Science*
*708 33, Ostrava – Poruba, Czech Republic*

**Abstract.** In this paper we present an algorithm for finding an optimal configuration of the artificial neural network that is used for the classification in our use case based effort estimation tool. This approach is based on feed-forward artificial neural network and is trained using the back-propagation training algorithm. Our goal is to find the optimal number of hidden neurons and the optimal number of training iterations to be able to reach maximal accuracy of neural network during the estimations. We demonstrate the usage of the proposed algorithm and its result on the estimation example that contains training and testing datasets of UseCases obtained from real software project development.

**Keywords.** Neural Networks, Feed-forward, Back-propagation, Softmax, Classification, UseCases

## Introduction

Our intention is to develop a decision supporting tool that will help project managers automatically classify UseCase scenarios that are made during the first phase of software development. Based on that classification, the estimations of selected parameters are made to support the manager's decision about the project in next phases [13, 14, 15]. One of the algorithms that are used in our tool for the classification is a feed-forward neural network algorithm. Implemented feed-forward neural network is trained using the back-propagation training algorithm. Our goal is to find the optimal number of hidden neurons and optimal number of training iterations to reach neural network with maximal accuracy for our purpose.

The literature review section provides general information about neural networks and also overview of applications of neural network to solve classification problem is provided. Next section describes reasons for architecture selection of the feed-forward neural network. Experiment section includes information about training and testing process. The last section shows measured results of classification based on proposed algorithm. Furthermore, these results are compared with results of the classification based on manual configuration of the neural network.

## 1. Literature review

### *1.1. Neural networks*

The neural network produces appropriate reactions to the outer world. A reaction to impulses is created in the core of the central nervous system. The most basic element of neural network is the neuron. Every neuron is connected to a large number of other neurons and together they create a neural network. When a neuron is excited a signal is transmitted to target neurons. [2], [5], [6]

It is important to understand which problems call for a neural network based approach. Neural networks are particularly useful for solving problems that cannot be expressed as a series of steps, such as recognizing patterns, classification, series prediction, and data mining. [1]

Training a neural network is the process of finding a set of weight and bias values, so that for a given set of inputs, the outputs produced by the neural network are very close to some target values. You need some historical data for training of the neural network. Training of the neural network searches for a set of weights and biases that most accurately predicts value from input data. Once you have these weight and bias values, you could apply them to an upcoming dataset to predict the results. [4], [7], [8]

### *1.2. Classification using neural network*

The classification algorithms such as Neural Network and Support Vector Machine (SVM) are currently used in various datasets and showing a good classification result as experimented by Young et al [16]. There have been, however, little studies in applying back propagation techniques for email classifications. The main disadvantage of back propagation techniques is that they require considerable time for parameter selection and training. On the other hand, previous research has shown that back-propagation in neural networks can achieve very accurate results, that are sometimes more accurate than those of the symbolic classifiers. [12]

We did a lot of previous work in the field of classification of normalized UseCases using SOM, SVM Fuzzy Rules. The average precision is not satisfactory in some cases. We are still trying to find the best solution that will help to estimate an effort of the software projects. More information about our SOM, SVM and Fuzzy Rules approaches is described in papers [13], [14], [15].

In a classification problem, vectors are assumed to be members of finite number of classes or categories. It is not easy to give the general guidelines for construction a suitable neural network solution of a classification task. At the beginning the data must be processed and input vectors with missing component values must be either discarded or missing values must be substituted with their probable values. The choice of the proper network architecture is crucial task. The numbers of neurons in the input layer and in the output layer are given by the task to be solved. However, the choice of the proper number of hidden layers and the choice of the proper numbers of neurons in each of them is a difficult problem. More neurons means better approximation ability, however, as we saw before, networks with great approximation ability are susceptible to the over-fitting [9].

To choose correctly the output functions of neurons in the output layer is the trickiest point concerning the network architecture. A good choice can be made if the theoretical results, explained in the preceding paragraphs, are taken into account. [9].

## 2. Design of feed-forward neural network to solve classification task

Most development of neural networks today is based upon manual design and configuration. A person who has knowledge about the specific application area specifies a network architecture, configuration and activation dynamics. This state of affairs is perhaps not surprising, given that the general space of possible neural networks is so large and complex that automatically searching for an optimal network architecture may in general be computationally intractable or at least impractical for complex applications [10], [11].

### *2.1. Layers of neural network*

Our feed-forward network begins with input layer. The input layer is connected to a hidden layer. Hidden layer is connected directly to the output layer. There is one hidden layer. The output layer of the neural network is what actually presents a pattern to the external environment. The number of input and output neurons is directly related to the intended use of neural network. In this case the neural network is used to solve a classification problem. Items are classified into two separate groups. We have output neuron for each group, so at all we have two output neurons. You can see designed neural network in Table 1.

**Table 1.** Overview of neural network architecture

| Layer | Initial number of neurons | Activation function |
|---|---|---|
| Input | 9 | Hyperbolic Tangent (min = -20, max = 20) |
| Hidden | 12 | Hyperbolic Tangent (min = -20, max = 20) |
| Output | 2 | Softmax Function |

### *2.2. Activation functions of hidden layer and output layer*

The hyperbolic tangent activation function has been chosen for hidden layer and it returns both positive and negative values. When graphed, the hyperbolic tangent function looks very similar to the log-sigmoid function. The important difference is that *tanh* function returns a value between -1 and +1 instead of between 0 and 1. The algebraic expression of hyperbolic tangent activation function is bellow (1).

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} = \frac{1 - e^{-2x}}{1 + e^{-2x}} \quad (1)$$

The most usual kind of classification problem is the problem with two mutually excluded classes. A network with only one output neuron can be used. In this case, we try to use output neurons with softmax output functions. It will facilitate computation of weight updates as they can be calculated according to the classical back-propagation scheme. [9]

We've decided to choose the softmax output activation function and two binary neurons, one for each category categories. The output layer consists of two neurons.

There are two groups that input items are assigned into. These groups are mapped on values of output parameter called extended work on Use-Case, which can take value 0 (false) or value 1 (true). For output layer, we've decided to use the softmax activation function (2).

The softmax activation function is popular as activation function for neural networks. It converts an arbitrary real-valued vector into a multinomial probability vector. It's used in classification problems. This is version of winner-take-all nonlinearity, in which maximum output is transformed to 1.0. [17]

$$h(z) = \frac{e^{z}}{\sum_{i=1}^{n} e^{z_i}} \tag{2}$$

*2.3. Mean Squared Error*

For our classification network we used Mean squared error (MSE) metrics, because the output targets are discrete. The definition for the MSE is given in the following equation (3), where n is the number of patterns in the validation set, m is the number of components in the output vector, o is the output of a single neuron *j*, *t* is the target for the single neuron j, and each input pattern is denoted by vector *i.*

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(o_{ij}-t_{ij})^2}{n} \tag{3}$$

## 3. An experiment based on automatic configuration of the neural network

*3.1. Training and testing data*

The whole datasets consists of 1041 UseCases from years 2008-2013. These items are divided into 6 datasets. Each dataset includes last 10 testing items. The number of training items in dataset for year 2008 is 385, for year 2009 is 569, for year 2010 is 634, for year 2011 is 738, for year 2012 is 953 and for year 2013 is 1041 items (*Acc1 – Acc6*). We use three sets of parameters: descriptive, structural and really evaluated parameters. [13]

Descriptive parameters are evaluated from the description of the use case, there were explained in detail in paper [13]. We use following descriptive parameters:
*Overall difficulty* is difficulty of words, rows and paragraphs. *RFC* is project ID

Structural parameters are evaluated by the structural or relational property of the use case on given project. Set of structural parameters was defined as a result from the interview with several senior project managers. We use following structural parameters and values: *NSC* (New = 3 / Standard = 1 / Change = 2) is functionality of the system.
*Concerned activities* (1 / 2 / 3) means how many business processes are touched by the implementation of this use case. *Use-Case type* can obtain value 1 (summary use case) or value 0 (user or sub function). *Remote work* can obtain value 1 (work can be done remote) or value 0 (work must be done onsite). *Implementation remote* (Yes = 1 / No = 0) means if the implementation of Use-Case is done onsite. *Testing level* (Easy = 0 / Normal = 1 / Complex = 2) means, how difficult will be the testing of implemented functionality.

Evaluated parameters are evaluated backward after the project is ended. We use following really evaluated parameter: *Extra work* (1 - additional work turned up, 0 – without additional work) means if there was some additional work then we have expected.

For more information about our data extraction and UseCase transformation approach we refer reader to paper [13]. Experiments based on same datasets, but using different (SVM, SOM and Fuzzy Rules) approach is described in papers [14], [15].

### *3.2. Data preprocessing*

You can see the part of example dataset rows after cleaning process in Table 2. As a part of the cleaning process we removed columns that contains same values for all rows. Parameter *XWork* (extra work) was divided into 2-values binary vector. It is required for the classification using softmax output activation function. Now, the matrix includes 9 columns plus 2 extra work columns (*XWorkYes* and *XWorkNo*). Furthermore, parameters: *UseCase type*, *Work remote* and *Implementation remote* were excluded from the training and testing process because the have equal values for each row.

Parameters of training and testing matrix after data preprocessing are: *N1.) Dataset id, N2.) Dif. Rows, N3.) Dif. Paragraph, N4.) Dif. Words, N5.) Overall dif., N6.) RFC, N7.) N/S/C, N8.) Concerned activities, N9.) Testing level, N10.) XWorkYes, N11.) XWorkNo.*

**Table 2.** Example data

| N1 | N2 | N3 | N4 | N5 | N6 | N7 | N8 | N9 | N10 | N11 |
|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 3 | 3 | 2 | 3 | 1 | 2 | 1 | 1 | 0 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | 0 |

### *3.3. Training and testing*

Initial configuration of the feed-forward neural network includes number of iterations. The algorithm for finding an optimal configuration starts with number of iterations computed using formula (4) and value $n = 1$. Every training cycle the number of iterations grows. Because of architecture of the neural network we've decided to use function with quadratic growth.

$$2\mathrm{n} * (2\mathrm{n})^2 \tag{4}$$

### *3.4. Number of neurons in hidden layer*

Number of hidden neurons is computed as a sum of input and output neurons minus expression $(n - 1)$. Variable $n$ is incremented every training cycle with linear time complexity. The goal is decrement number of neurons after each training cycle.

$$(inNum + outNum) - (n - 1) \tag{5}$$

*3.5. Proposed finding algorithm*

We propose an algorithm for finding optimal configuration of the neural network. The algorithm starts with initial values. These values are number of hidden neurons and number of iterations. After each training cycle the number of training iterations is incremented using formula (4). The parameter: number of hidden neurons is decremented using formula (5). Finding process continues until the accuracy of the neural network is 100% or if the maximal number of training iterations is reached. The maximal possible number of training cycles is 10000 and maximal number of hidden neurons is 12.

## 4. Results

We've used the neural network to classify the dataset of normalized UseCases. These UseCases are divided into two categories (extended work, not extended work). The classification is a process that is closely related to pattern recognition. A neural network trained for classification is designed to take input samples and classify them into groups. In this case, UseCases are classified into two different groups.

*4.1. Results of experiment using automatic configuration*

Results obtained from classification of UseCase datasets are showed in Table 3. Each *training cycle* parameter is complete neural network training process. Parameters *Acc1-Acc6* means testing accuracy value of the neural network. This accuracy is computed using six different training datasets (*Acc1 – Acc6*). The number of training items in each next dataset has increased, because training and testing items from previous dataset are added. Last 10 items of each dataset we are always using as testing items. The bold printed value is the highest value of accuracy of neural network trained using provided dataset.

For example, let us discuss the maximum precision neural network using second training dataset achieved in the first *training cycle.* This accuracy is achieved with 2744 *training iterations* and six *hidden neurons* and it is showed in column with name *Acc2* and value is 100 in Table 3.

**Table 3.** Results

| Training cycle | Iterations | Hidden neurons | Acc1 | Acc2 | Acc3 | Acc4 | Acc5 | Acc6 |
|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 12 | **50** | 60 | 80 | 60 | **90** | 60 |
| 2 | 64 | 11 | 50 | 60 | 80 | 60 | 80 | 60 |
| 3 | 216 | 10 | 50 | 70 | 80 | 80 | 80 | 60 |
| 4 | 512 | 9 | 30 | 80 | 70 | 90 | 60 | **90** |
| 5 | 1000 | 8 | 20 | 80 | **100** | **100** | 60 | 90 |
| 6 | 1728 | 7 | 20 | 80 | 100 | 100 | 70 | 90 |
| 7 | 2744 | 6 | 20 | **100** | 90 | 100 | 70 | 90 |
| 8 | 4096 | 5 | 20 | 100 | 90 | 100 | 70 | 90 |
| 9 | 5832 | 4 | 20 | 100 | 90 | 100 | 70 | 90 |
| 10 | 8000 | 3 | 20 | 100 | 90 | 100 | 70 | 90 |

*4.2. Compare best results reached by manual configuration*

The difference between accuracy of *manually* configured neural network and *automatically* configured neural network is shown in Table 4. For both configurations, the neural network is tested for six cases (*Acc1 – Acc6*). It means that for training and testing we used six datasets of UseCases.

The first row shows highest value of accuracy of *manually* configured neural network. The manually configured neural network includes seven hidden neurons. This optimal number of hidden neurons is the same for all datasets.

Second row shows highest value of accuracy of *automatically* configured neural network. In other words, second row shows the best results (bold printed values) from result Table 3.

**Table 4.** Best results of classification using manual and automatic configuration

| Configuration | Acc1 | Acc2 | Acc3 | Acc4 | Acc5 | Acc6 |
|---|---|---|---|---|---|---|
| Manual | 30 | 100 | 100 | 100 | 80 | 90 |
| Automatic | 50 | 100 | 100 | 100 | 90 | 90 |

## 5. Conclusion and future work

This paper presents and discusses the usage of automatically set configuration of feed-forward neural network. The network is used then for the classification of the use cases in our decision supporting tool for software project estimation. The experiment shows that the automatically set configuration network provides results that are same or better in some cases then the manually set network results. These results seem to be promising for our purposes so far. Anyway, more experiments have to be done to demonstrate the general applicability even for our purposes.

Future work will be focused on research in field of automated design of Feed-Forward neural network for classification and the application of our algorithm to the other classification problems as well.

## References


[1] Heaton Jeff. Introduction to neural networks, ISBN: 1-60439-008-5

[2] H. T. Siegelmann and E. D. Sontag. Turing computability with neural networks. Applied Mathematics Letters 4(6), pages 77–80, 1991.

[3] A. Frank and A. Asuncion. Uci machine learning repository, 2010

[4] D. Fedorcak and I. Vondrak. Self-organization of labeled data using growing neural gas. Submitted for Neural Network World, 2009.

[5] De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of timeseries forecasting. International Journal of Forecasting, 22,443–473.

[6] Ghiassi .M , Nangoy. S(2009). A dynamic artificial neuralnetwork model for forecasting nonlinear processes. Computers& Industrial Engineering .76(5), 302–316.

[7] J. S. Henneke, S. T. Rachev, F. J. Fabozzi, and M. Nikolove, “MCMC-based estimation of Markov Switching ARMA-GARCH models,” Applied Economics, vol. 43, no. 3, pp. 259–271, 2011.
[8] A. Azadeh, M. Saberi, and M. Anvari, “An integrated artificial neural network algorithm for performance assessment and optimization of decision making units,” Expert Systems with Applications, vol. 37, no. 8, pp. 5688–5697, 2010.
[9] Vesely, A. Classification and prediction with neural networks. In Berka, P., Rauch, J., Zighed D. A., editors. Data Mining and Medical Knowledge Management: Cases and Applications (76-107) (2009).
[10] Mehrotra K., Mohan CK., Ranka S.: Elements of Artificial Neural Networks. MIT, Cambridge, MA, 1997.
[11] Jung J., Reggia J. A.: The Automated Design of Artificial Neural Networks Using Evolutionary Computation, Studies in Computational Intelligence (SCI) 92, 19–41 (2008)
[12] Ayodele T., Zhou S., Khusainov R.: Email Classification Using Back Propagation Technique, International Journal of Intelligent Computing Research, 2010.
[13] Štolfa, J., Štolfa, S., Kobĕrský, O., Kopka, M., Kožuszník, J. & Snášel, V. 2012, "Methodology for estimating working time effort of the software project", CEUR Workshop Proceedings, pp. 25.
[14] Štolfa, J., Kobĕrský, O., Krömer, P., Štolfa, S., Kopka, M. & Snášel, V. 2013, "Comparison of fuzzy rules and SVM approach to the value estimation of the use case parameters", Proceedings of the 2013 Joint IFSA World Congress and NAFIPS Annual Meeting, IFSA/NAFIPS 2013, pp. 789.
[15] Stolfa S., Stolfa S., Kromer P., Kobersky O., Kopka M. and Snasel V.: Fuzzy Rules and SVM Approach to the Estimation of Use Case Parameters, Innovations in Bio-inspired Computing and Applications Advances in Intelligent Systems and Computing Volume 237, 2014, pp 37-47
[16] Youn, S. a. (2006). A Comparative Study for Email Classification. *JOURNAL OF SOFTWARE , 2* (3), 1-13.

# Comparison of Support Vector Machines and Neural Network Models on Real Data

**Boštjan Brumen[1], Ivan Rozman[1], Aleš Černezel[1]**

[1] University of Maribor, Faculty of Electrical Engineering and Computer science, Smetanova 17, Si-2000 Maribor, Slovenia

e-mail: [1](bostjan.brumen, i.rozman, ales.cernezel)@uni-mb.si

***Abstract.***

Background: During the discussion at the previous EJC conference held in Kiel with one of the participants an interesting question arose: which classification algorithm should be used as the first choice when modelling the data and building a classification tree, Support Vector Machines (SVM) or Neural Networks (NN)?

Objective: The objective of this paper is to find out whether there is any significant difference between the performance of the Support Vector Machines and of the Neural Network classification algorithms.

Methods: A Weka-based multilayer perceptron (MLP) neural network and a Support Vector Machines classification algorithms were applied to a set of datasets (n=121) from publicly available repositories (UCI) in step wise k-fold cross-validation and an error rate was measured in each step. First, four different functions, i.e. power, linear, logarithmic, exponential, were fit to the measured error rates. Where the fit was statistically significant (n=54) for all functions for both algorithms, we measured the average mean squared error rate for each function and its rank. The Wilcoxon's signed rank test was used to test whether the differences between ranks are significant.

Results: In a total of 54 datasets, the SVM algorithm using the exponential function was better than NN (*P=0,023*). Average mean squared error across all datasets was at 0,000251 for SVM's exponential function statistically significantly different from NN's MSE at 0,000313.

Conclusion: In the area of human cognitive performance the exponential function was found to be the best fit for a description of an individual learner. The NN and SVM algorithms are not different from each other in the capacity of capturing the interrelations in the data. However, SVM does outperform NN in case exponential function is used to describe the performance.

## 1. Introduction

During the discussion at the previous EJC conference held in Kiel with one of the participants an interesting question arose: which classification algorithm should be used as the first choice when modelling the data and building a classification tree, Support Vector Machines or Neural Networks. Namely, artificial learners have not received a high volume of research in terms of a description of their behavior [1]. Ideally, a description of a learning problem would be a functional dependency between the data, the learning algorithm's internal specifics and its performance (e.g. error). This way we could analytically determine the output (error rate) based on the input (data, selected learner). To answer the question "Which one is better?", one would only compare two analytical models for each of the sets (learner, data).

Unfortunately, such models have not yet been devised. Standard numerical (and other statistical) methods become unstable when using large data sets [2]. Different theoretical approaches provide estimates for the size of the confidence interval on the training error under various settings of the problem of learning from examples. Vapnik-Chervonenkis theory [3] is the most comprehensive description of learning from examples. However, it has some limits (e.g. oracle is never wrong) that make it difficult for real-life implementations, as described in detail [4]. Some results for a specific learner and for specific type of data can be found in the literature, e.g. [5], but no general analytical solution is available.

Without an accepted theoretical model at hand, we can empirically measure a learner's performance on as many tasks (data sets) as possible. However, not much research was conducted on the description of performance of neural networks or support vector machines on a large scale comparison using several different data sets. Complementarily, there was some work on classification trees which can be predicted by a power law [6]. Several authors have either confirmed this or have been building on their results [7, 8]. But, a more recent research conducted by Singh has produced some evidence against the power law [9].

The main contribution of this paper is the answer to the question: "Are any differences related to the choice of the algorithm, i.e. is any of the chosen algorithms performing better?

## 2. Method

We have chosen a multilayer artificial neural network classifier, which is freely available from the Waikato Environment for Knowledge Analysis (WEKA) project toolkit [10, 11] version 3.6.8, with standard built-in settings and initial values.

The computer used was equipped with Windows-7 (x64) operating system, an Intel i5-650 processor and 8 GB of DDR3 RAM. For statistical analyses we used IBM SPSS version 21.

### 2.1. Data collection

We used publicly available datasets from University of California at Irvine (UCI) Machine Learning Repository [12]. We selected the datasets where the problem task is classification; the number of records in a dataset was larger than 200 and the number of instances exceeded the number of attributes (i.e. the task was classification, not feature selection).

The UCI repository contains datasets in ".data" and ".names" format while Weka's native format is ARFF. Therefore we used files available from various sources, such as TunedIT [13], Håkan Kjellerstrand' weka page [14, 15] and Kevin Chai's page [16]. We gathered 121 datasets.

We used only the original or larger datasets where several ones were available and ignored any separate training or test set, or any associated cost model.

### 2.2. Data pre-processing

We followed the following steps for obtaining the error rate curve (i.e. learning curve) [17]:

1. Data items in a data set are randomly shuffled
2. First, $n_{i=1}$=50 items are chosen
3. Build decision trees using k-fold cross-validation on sample size of $n_i$ [18, 19]; $k$ was set to 10 [8, 18-20];
4. Measure the error rate for each tree in 10-fold run and average the result over 10 runs
5. Store the pair ($n_i$=sample size, $e_i$=error)
6. The number of items in a data set is increased by 10; $n_{i+1}$:=$n_i$+10
7. Repeat steps 3-6 until all data items in a dataset are used.

### 2.3. Fitting a curve model to the measured data

The next step in our research was to fit a model to the error rate curves. We used four different functions, as in Equations 1-4:

| | | |
|---|---|---|
| power (POW): | $f(x) = p_1 + p_2 x^{p_3}$ | Eq. 1 |
| linear (LIN): | $f(x) = p_1 + p_2 x$ | Eq. 2 |
| logarithm (LOG): | $f(x) = p_1 + p_2 \log x$ | Eq. 3 |
| exponential (EXP): | $f(x) = p_1 + p_2 e^{p_3 x}$ | Eq. 4 |

The functions do not have the same number of parameters ($p_i$). They all include the constant $p_1$ and coefficient $p_2$, in addition to potent $p_3$ for the power and the exponential function. Based on the specifics of the problem and the speed of convergence we limited the parameters to the following intervals:

- $p_1$ to interval [0, 1] (error rate cannot be less than 0 and more than 1)

- $p_2$ to interval [0, 100] for power function and to [-100, 0] for the others, and
- $p_3$ to interval [-100, 0] (error rate is decreasing hence $p_3$needs to be negative)

We used the open-source GNU Octave software [21] and the built-in Levenberg-Marquardt's algorithm [22, 23], also known as the damped least-squares (DLS) method, for fitting the function parameters to the data.

The inputs to the algorithm were vector $x$ (sample sizes *n*), vector $y$ (error rates *e*), initial values of parameters $p_i$ ([0,01; 1; -0,1] for POW, [0,1; -0,001] for LIN, [0,1; -0,01] for LOG and [0,01; 0,1; -0,01] for EXP), function to be fit to vectors $x, y$ (power, linear, logarithm, or exponential), partial derivatives of functions with respect to parameters $p_i$, and limits of parameters $p_i$ (as described above).

The algorithm's output were vector of functional values of fitted function for input $x$, vector of parameters $p_i$, where minimum mean squared error was obtained, and a flag whether the convergence was reached or not.

## 3. Results

For each dataset and for each of the chosen algorithms we tested the claim that the underlying data can be modeled by the probability density functions POW, LIN, LOG and EXP, respectively. We used the Pearson's chi-squared test ($\chi^2$), also known as the chi-squared goodness-of-fit test or chi-squared test for independence, where the null hypothesis was $H_0$: $r_\mu = 0$ or there is no correlation between the population and the model [24], at $\alpha$=0,05. Table 1 lists the results: the values in bold are *P* values indicating that the null hypothesis is rejected, the number of degrees of freedom (df), and the coefficient of determination $R^2$ (indication how well a regression line fits a set of data). N/A indicates that the model was not calculated because the Levenberg-Marquardt's algorithm suggested a constant model and hence $\chi^2$ cannot be computed.

**Table 1.** Datasets and respective P-values and $r^2$ values for goodness of fit of a model to data for each of the two algorithms

| Dataset | df | POW p SVM | POW $r^2$ SVM | POW p NB | POW $r^2$ NB | LIN p SVM | LIN $r^2$ SVM | LIN p NB | LIN $r^2$ NB | LOG p SVM | LOG $r^2$ SVM | LOG p NB | LOG $r^2$ NB | EXP p SVM | EXP $r^2$ SVM | EXP p NB | EXP $r^2$ NB |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ada_agnostic | 449 | **0,000** | 0,78 | **0,000** | 0,74 | **0,000** | 0,28 | **0,000** | 0,31 | **0,000** | 0,62 | **0,000** | 0,63 | **0,000** | 0,82 | **0,000** | 0,80 |
| ada_prior | 449 | **0,000** | 0,47 | 1,000 | 0,00 | **0,000** | 0,24 | 1,000 | 0,00 | **0,000** | 0,42 | 1,000 | 0,00 | **0,000** | 0,52 | **0,000** | 0,22 |
| analcatdata_authorship | 77 | **0,000** | 0,03 | **0,000** | 0,85 | **0,000** | 0,20 | **0,000** | 0,47 | **0,000** | 0,28 | **0,000** | 0,74 | **0,000** | 0,34 | **0,000** | 0,92 |
| analcatdata_braziltourism | 34 | **0,000** | 0,88 | **0,000** | 0,85 | **0,000** | 0,90 | **0,000** | 0,81 | **0,000** | 0,91 | **0,000** | 0,87 | **0,000** | 0,94 | **0,000** | 0,89 |
| analcatdata_broadwaymult | 21 | 0,808 | 0,00 | **0,008** | 0,27 | 0,603 | 0,01 | **0,017** | 0,22 | 0,807 | 0,00 | **0,008** | 0,27 | 0,607 | 0,01 | **0,006** | 0,29 |
| analcatdata_dmft | 72 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| analcatdata_halloffame | 126 | **0,000** | 0,75 | 1,000 | 0,00 | **0,000** | 0,54 | 1,000 | 0,00 | **0,000** | 0,70 | 1,000 | 0,00 | **0,000** | 0,69 | 1,000 | 0,00 |
| analcatdata_marketing | 29 | 0,413 | 0,02 | **0,000** | 0,35 | 0,104 | 0,08 | **0,000** | 0,34 | 0,406 | 0,02 | **0,000** | 0,35 | 0,107 | 0,08 | **0,000** | 0,35 |
| analcatdata_reviewer | 30 | 0,900 | 0,00 | **0,000** | 0,33 | 0,369 | 0,03 | **0,000** | 0,42 | 0,899 | 0,00 | **0,000** | 0,34 | 0,378 | 0,02 | **0,000** | 0,42 |
| anneal | 82 | **0,000** | 0,62 | 1,000 | 0,00 | **0,000** | 0,55 | 0,721 | 0,00 | **0,000** | 0,62 | 1,000 | 0,00 | **0,000** | 0,59 | 0,152 | 0,02 |
| anneal.ORIG | 82 | **0,000** | 0,88 | **0,000** | 0,32 | **0,000** | 0,54 | **0,022** | 0,06 | **0,000** | 0,78 | **0,000** | 0,21 | **0,000** | 0,88 | **0,000** | 0,39 |
| audiology | 15 | **0,000** | 0,81 | **0,000** | 0,81 | **0,000** | 0,81 | **0,000** | 0,88 | **0,000** | 0,82 | **0,000** | 0,84 | **0,000** | 0,83 | **0,000** | 0,87 |
| australian | 61 | **0,000** | 0,65 | **0,000** | 0,38 | **0,000** | 0,66 | **0,013** | 0,09 | **0,000** | 0,65 | **0,002** | 0,15 | **0,000** | 0,66 | **0,000** | 0,46 |
| autos | 13 | **0,000** | 0,88 | **0,003** | 0,45 | **0,000** | 0,71 | **0,000** | 0,61 | **0,000** | 0,81 | **0,002** | 0,47 | **0,000** | 0,87 | **0,000** | 0,58 |
| badges_plain | 22 | **0,000** | 0,93 | **0,000** | 0,93 | **0,000** | 0,73 | **0,000** | 0,73 | **0,000** | 0,87 | **0,000** | 0,87 | **0,000** | 0,92 | **0,000** | 0,92 |
| balance-scale | 55 | **0,000** | 0,69 | **0,000** | 0,80 | **0,011** | 0,11 | **0,000** | 0,52 | **0,000** | 0,32 | **0,000** | 0,73 | **0,000** | 0,74 | **0,000** | 0,84 |
| baseball-hitter | 25 | **0,000** | 0,92 | **0,000** | 0,47 | **0,000** | 0,67 | **0,000** | 0,63 | **0,000** | 0,82 | **0,000** | 0,34 | **0,000** | 0,89 | **0,000** | 0,62 |
| baseball-pitcher | 13 | **0,000** | 0,97 | 0,119 | 0,15 | **0,000** | 0,87 | 0,641 | 0,01 | **0,000** | 0,95 | 0,428 | 0,04 | **0,000** | 0,96 | **0,034** | 0,27 |
| BC | 21 | 0,511 | 0,02 | 0,827 | 0,00 | 0,571 | 0,01 | 0,277 | 0,05 | 0,513 | 0,02 | 0,822 | 0,00 | 0,569 | 0,01 | 0,289 | 0,05 |
| Billionaires92 | 16 | 1,000 | 0,00 | 0,832 | 0,00 | 1,000 | 0,00 | 0,446 | 0,03 | N/A | 0,00 | 0,830 | 0,00 | 1,000 | 0,00 | 0,450 | 0,03 |
| biomed | 13 | **0,000** | 0,66 | 1,000 | 0,00 | **0,001** | 0,52 | 1,000 | 0,00 | **0,000** | 0,61 | 1,000 | 0,00 | **0,000** | 0,66 | 1,000 | 0,00 |
| breast-cancer | 21 | **0,000** | 0,80 | **0,000** | 0,78 | **0,000** | 0,70 | **0,000** | 0,60 | **0,000** | 0,78 | **0,000** | 0,73 | **0,000** | 0,78 | **0,000** | 0,79 |
| breast-w | 62 | 1,000 | -0,27 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 0,959 | 0,00 | 1,000 | 0,00 | 0,087 | 0,04 | 1,000 | 0,00 |
| car | 165 | **0,000** | 0,91 | **0,000** | 0,88 | **0,000** | 0,49 | **0,000** | 0,73 | **0,000** | 0,80 | **0,000** | 0,90 | **0,000** | 0,92 | **0,000** | 0,93 |
| cars_with_names | 33 | **0,000** | 0,83 | **0,000** | 0,60 | **0,000** | 0,68 | **0,000** | 0,68 | **0,000** | 0,80 | **0,000** | 0,62 | **0,000** | 0,82 | **0,000** | 0,68 |
| CH | 312 | **0,000** | 0,86 | **0,000** | 0,86 | **0,000** | 0,21 | **0,000** | 0,26 | **0,000** | 0,44 | **0,000** | 0,60 | **0,000** | 0,86 | **0,000** | 0,92 |
| cmc | 140 | **0,000** | 0,71 | **0,000** | 0,59 | **0,000** | 0,21 | **0,000** | 0,17 | **0,000** | 0,36 | **0,000** | 0,31 | **0,000** | 0,76 | **0,000** | 0,66 |
| colic | 29 | **0,000** | 0,65 | **0,000** | 0,65 | **0,000** | 0,60 | **0,000** | 0,46 | **0,000** | 0,66 | **0,000** | 0,60 | **0,000** | 0,69 | **0,000** | 0,68 |
| colic.ORIG | 29 | **0,000** | 0,59 | 1,000 | 0,00 | **0,000** | 0,60 | 1,000 | 0,00 | **0,000** | 0,60 | 1,000 | 0,00 | **0,000** | 0,61 | 1,000 | 0,00 |
| cps_85_wages | 46 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| credit-a | 61 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| credit-g | 92 | **0,000** | 0,79 | **0,000** | 0,48 | **0,000** | 0,68 | **0,000** | 0,38 | **0,000** | 0,78 | **0,000** | 0,44 | **0,000** | 0,78 | **0,000** | 0,40 |
| credit | 41 | **0,000** | 0,35 | **0,000** | 0,86 | **0,001** | 0,24 | **0,000** | 0,79 | **0,000** | 0,33 | **0,000** | 0,87 | **0,000** | 0,38 | **0,000** | 0,88 |
| csb_ch12 | 153 | **0,000** | -0,01 | 1,000 | 0,00 | **0,000** | 0,09 | 1,000 | 0,00 | **0,000** | 0,16 | 1,000 | 0,00 | **0,000** | 0,55 | 1,000 | 0,00 |
| csb_ch9 | 316 | **0,000** | 0,48 | **0,000** | 0,56 | **0,000** | 0,43 | **0,000** | 0,26 | **0,000** | 0,48 | **0,000** | 0,40 | **0,000** | 0,49 | **0,000** | 0,57 |
| cylinder-bands | 46 | **0,000** | 0,77 | **0,000** | 0,78 | **0,000** | 0,59 | **0,000** | 0,55 | **0,000** | 0,72 | **0,000** | 0,69 | **0,000** | 0,74 | **0,000** | 0,73 |
| db3-bf | 39 | **0,000** | 0,73 | **0,000** | 0,75 | **0,000** | 0,52 | **0,000** | 0,73 | **0,000** | 0,62 | **0,000** | 0,76 | **0,000** | 0,69 | **0,000** | 0,78 |
| dermatology | 29 | **0,000** | 0,35 | **0,000** | 0,81 | **0,018** | 0,17 | **0,000** | 0,59 | **0,006** | 0,22 | **0,000** | 0,73 | **0,000** | 0,52 | **0,000** | 0,80 |
| diabetes | 69 | **0,000** | 0,35 | **0,000** | 0,42 | **0,000** | 0,31 | **0,000** | 0,30 | **0,000** | 0,35 | **0,000** | 0,42 | **0,000** | 0,41 | **0,000** | 0,49 |
| ecoli | 26 | **0,001** | 0,32 | **0,000** | 0,66 | **0,000** | 0,41 | **0,000** | 0,47 | **0,001** | 0,33 | **0,000** | 0,59 | **0,000** | 0,40 | **0,000** | 0,68 |
| eucalyptus | 66 | **0,000** | 0,83 | **0,000** | 0,54 | **0,000** | 0,76 | **0,000** | 0,48 | **0,000** | 0,83 | **0,000** | 0,54 | **0,000** | 0,85 | **0,000** | 0,53 |
| eye_movements | 1086 | **0,000** | 0,69 | **0,000** | 0,22 | **0,000** | 0,59 | **0,000** | 0,11 | **0,000** | 0,69 | **0,000** | 0,16 | **0,000** | 0,61 | **0,000** | 0,14 |

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| genresTrain | 1242 | **0,000** | 0,97 | 1,000 | 0,00 | **0,000** | 0,51 | **0,001** | 0,01 | **0,000** | 0,87 | **0,000** | 0,02 | **0,000** | 0,86 | **0,000** | 0,55 |
| gina_agnostic | 339 | **0,000** | 0,81 | **0,000** | 0,69 | **0,000** | 0,69 | **0,000** | 0,47 | **0,000** | 0,81 | **0,000** | 0,66 | **0,000** | 0,79 | **0,000** | 0,67 |
| gina_prior | 339 | **0,000** | 0,50 | 0,834 | 0,00 | **0,000** | 0,64 | 0,073 | 0,01 | **0,000** | 0,55 | 0,831 | 0,00 | **0,000** | 0,69 | 1,000 | 0,00 |
| gina_prior2 | 339 | **0,000** | -0,48 | **0,000** | 0,70 | **0,000** | 0,11 | **0,000** | 0,07 | **0,000** | 0,34 | **0,000** | 0,20 | **0,000** | 0,91 | **0,000** | 0,74 |
| GL | 14 | **0,012** | 0,33 | N/A | 0,00 | **0,003** | 0,43 | 1,000 | 0,00 | **0,011** | 0,34 | N/A | 0,00 | **0,004** | 0,42 | 1,000 | 0,00 |
| glass | 14 | 0,271 | 0,08 | 1,000 | 0,00 | 0,228 | 0,09 | N/A | 0,00 | 0,269 | 0,08 | N/A | 0,00 | 0,227 | 0,09 | 1,000 | 0,00 |
| haberman | 23 | **0,000** | 0,60 | N/A | 0,00 | **0,000** | 0,64 | 1,000 | 0,00 | **0,000** | 0,62 | 1,000 | 0,00 | **0,000** | 0,66 | 1,000 | 0,00 |
| HD | 23 | **0,000** | 0,46 | 0,580 | 0,01 | **0,000** | 0,54 | 0,583 | 0,01 | **0,000** | 0,49 | 0,580 | 0,01 | **0,000** | 0,54 | 0,582 | 0,01 |
| heart-c | 23 | **0,000** | 0,50 | 1,000 | 0,00 | **0,001** | 0,38 | 1,000 | 0,00 | **0,000** | 0,47 | 1,000 | 0,00 | **0,000** | 0,52 | 1,000 | 0,00 |
| heart-h | 22 | **0,029** | 0,18 | 0,242 | 0,06 | **0,006** | 0,28 | 0,078 | 0,12 | **0,025** | 0,19 | 0,229 | 0,06 | **0,007** | 0,27 | 0,082 | 0,12 |
| heart-statlog | 19 | **0,000** | 0,56 | **0,000** | 0,69 | **0,036** | 0,19 | **0,010** | 0,27 | **0,004** | 0,34 | **0,001** | 0,44 | **0,000** | 0,62 | **0,000** | 0,74 |
| HO | 29 | **0,000** | 0,74 | 1,000 | 0,00 | **0,000** | 0,61 | N/A | 0,00 | **0,000** | 0,70 | 1,000 | 0,00 | **0,000** | 0,69 | 1,000 | 0,00 |
| HY | 309 | **0,000** | 0,24 | **0,000** | -0,12 | **0,000** | 0,53 | **0,000** | 0,10 | **0,000** | 0,26 | **0,000** | 0,25 | **0,000** | 0,52 | **0,000** | 0,81 |
| hypothyroid | 370 | 1,000 | 0,00 | 0,992 | 0,00 | 1,000 | 0,00 | 0,996 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | **0,006** | 0,02 | **0,000** | 0,50 |
| ionosphere | 28 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| irish | 42 | **0,000** | 0,27 | **0,000** | 0,77 | 0,623 | 0,01 | **0,000** | 0,86 | 0,262 | 0,03 | **0,000** | 0,85 | **0,000** | 0,53 | **0,000** | 0,89 |
| jEdit_4.0_4.2 | 20 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | N/A | 0,00 |
| jEdit_4.2_4.3 | 29 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| jm1 | 1081 | 1,000 | 0,00 | **0,000** | 0,48 | 1,000 | 0,00 | **0,000** | 0,12 | 1,000 | 0,00 | **0,000** | 0,35 | **0,000** | 0,08 | **0,000** | 0,66 |
| kc1 | 203 | 1,000 | 0,00 | **0,045** | 0,02 | N/A | 0,00 | 0,088 | 0,01 | 1,000 | 0,00 | **0,043** | 0,02 | 1,000 | 0,00 | **0,001** | 0,06 |
| kc2 | 45 | 1,000 | 0,00 | **0,001** | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | -0,01 |
| kc3 | 38 | 1,000 | 0,00 | 0,098 | 0,07 | 1,000 | 0,00 | 0,137 | 0,05 | 1,000 | 0,00 | 0,098 | 0,07 | 1,000 | 0,00 | 0,072 | 0,08 |
| kdd_ipums_la_97-small | 694 | **0,000** | 0,65 | **0,000** | 0,47 | **0,033** | 0,01 | 1,000 | 0,00 | **0,000** | 0,18 | **0,001** | 0,02 | **0,000** | 0,82 | 1,000 | 0,00 |
| kdd_ipums_la_98-small | 741 | 1,000 | 0,00 | **0,000** | 0,26 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | **0,000** | 0,09 | **0,000** | 0,32 |
| kdd_ipums_la_99-small | 877 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 0,216 | 0,00 |
| kdd_synthetic_control | 52 | **0,000** | 0,91 | **0,000** | 0,70 | **0,000** | 0,68 | **0,000** | 0,53 | **0,000** | 0,87 | **0,000** | 0,67 | **0,000** | 0,93 | **0,000** | 0,69 |
| kr-vs-kp | 312 | **0,000** | 0,86 | **0,000** | 0,79 | **0,000** | 0,36 | **0,000** | 0,19 | **0,000** | 0,71 | **0,000** | 0,44 | **0,000** | 0,92 | **0,000** | 0,78 |
| kropt | 2798 | **0,000** | 0,94 | **0,000** | 0,90 | **0,000** | 0,36 | **0,000** | 0,41 | **0,000** | 0,79 | **0,000** | 0,83 | **0,000** | 0,94 | **0,000** | 0,95 |
| landsat | 636 | **0,000** | 0,26 | 1,000 | 0,00 | **0,000** | 0,40 | 1,000 | 0,00 | **0,000** | 0,27 | 1,000 | 0,00 | **0,000** | 0,40 | **0,004** | 0,01 |
| letter | 1992 | **0,000** | 0,99 | **0,000** | 0,94 | **0,000** | 0,49 | **0,000** | 0,09 | **0,000** | 0,89 | **0,000** | 0,39 | **0,000** | 0,92 | **0,000** | 0,82 |
| liver-disorders | 27 | 0,215 | 0,05 | 1,000 | 0,00 | 0,313 | 0,04 | N/A | 0,00 | 0,217 | 0,05 | 1,000 | 0,00 | 0,150 | 0,07 | 1,000 | 0,00 |
| mc1 | 939 | 1,000 | 0,00 | 1,000 | 0,00 | **0,000** | 0,11 | 0,999 | 0,00 | 1,000 | 0,00 | 0,998 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| mfeat-factors | 192 | **0,000** | 0,93 | **0,000** | 0,93 | **0,000** | 0,40 | **0,000** | 0,28 | **0,000** | 0,71 | **0,000** | 0,53 | **0,000** | 0,88 | **0,000** | 0,91 |
| mfeat-fourier | 192 | **0,000** | 0,79 | **0,000** | 0,70 | **0,000** | 0,62 | **0,000** | 0,18 | **0,000** | 0,77 | **0,000** | 0,31 | **0,000** | 0,71 | **0,000** | 0,74 |
| mfeat-karhunen | 192 | **0,000** | 0,95 | 0,995 | -0,81 | **0,000** | 0,38 | **0,000** | 0,18 | **0,000** | 0,68 | **0,000** | 0,45 | **0,000** | 0,91 | **0,000** | 0,97 |
| mfeat-morphological | 192 | **0,000** | 0,27 | 0,187 | 0,01 | **0,002** | 0,05 | 0,235 | 0,01 | **0,000** | 0,14 | 0,190 | 0,01 | **0,000** | 0,34 | **0,000** | 0,08 |
| mfeat-pixel | 192 | **0,000** | 0,95 | **0,000** | 0,96 | **0,000** | 0,47 | **0,000** | 0,28 | **0,000** | 0,78 | **0,000** | 0,60 | **0,000** | 0,90 | **0,000** | 0,98 |
| mfeat-zernike | 192 | **0,000** | 0,77 | **0,000** | 0,78 | **0,000** | 0,33 | **0,000** | 0,12 | **0,000** | 0,60 | **0,000** | 0,32 | **0,000** | 0,73 | **0,000** | 0,84 |
| monks-problems-1_test | 36 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | -68,53 | 1,000 | 0,00 |
| monks-problems-2_test | 36 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| monks-problems-3_test | 36 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 0,429 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| mozilla4 | 1547 | **0,000** | 0,59 | 1,000 | 0,00 | **0,000** | 0,15 | 1,000 | 0,00 | **0,000** | 0,45 | 1,000 | 0,00 | **0,000** | 0,73 | 1,000 | 0,00 |
| MU | 805 | **0,000** | 0,81 | **0,000** | 0,94 | **0,000** | 0,20 | **0,000** | 0,74 | **0,000** | 0,52 | **0,000** | 0,95 | **0,000** | 0,79 | **0,000** | 0,92 |
| mushroom | 805 | **0,000** | 0,71 | **0,000** | 0,94 | **0,000** | 0,13 | **0,000** | 0,64 | **0,000** | 0,38 | **0,000** | 0,92 | **0,000** | 0,75 | **0,000** | 0,90 |
| mw1 | 33 | **0,026** | 0,04 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 0,898 | 0,00 | 1,000 | 0,00 | **0,000** | 0,31 | 1,000 | 0,00 |
| nursery | 1288 | **0,000** | 0,68 | **0,000** | 0,51 | **0,000** | 0,22 | **0,000** | 0,21 | **0,000** | 0,53 | **0,000** | 0,44 | **0,000** | 0,75 | **0,000** | 0,54 |
| optdigits | 554 | **0,000** | 0,91 | **0,000** | 0,98 | **0,000** | 0,41 | **0,000** | 0,21 | **0,000** | 0,76 | **0,000** | 0,52 | **0,000** | 0,89 | **0,000** | 0,93 |
| page-blocks | 540 | **0,000** | 0,72 | N/A | -0,77 | **0,000** | 0,36 | 1,000 | 0,00 | **0,000** | 0,63 | **0,005** | 0,01 | **0,000** | 0,67 | 0,985 | 0,00 |
| pc1 | 103 | **0,000** | 0,73 | **0,000** | 0,94 | **0,000** | 0,47 | **0,000** | 0,44 | **0,000** | 0,70 | **0,000** | 0,74 | **0,000** | 0,81 | **0,000** | 0,92 |
| pc3 | 149 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | **0,000** | 0,00 | 1,000 | 0,00 | 0,994 | 0,00 | 1,000 | 0,00 | 0,934 | 0,00 |
| pc4 | 138 | **0,000** | 0,11 | **0,000** | 0,09 | **0,000** | 0,14 | **0,000** | 0,25 | **0,000** | 0,11 | **0,000** | 0,10 | **0,000** | 0,14 | **0,000** | 0,24 |
| pendigits | 1092 | **0,000** | 0,97 | **0,000** | 0,88 | **0,000** | 0,33 | **0,000** | 0,22 | **0,000** | 0,73 | **0,000** | 0,57 | **0,000** | 0,89 | **0,000** | 0,79 |
| primary-tumor | 26 | **0,022** | 0,17 | **0,000** | 0,40 | 0,254 | 0,05 | **0,006** | 0,24 | 0,185 | 0,06 | **0,003** | 0,28 | 0,255 | 0,05 | **0,000** | 0,65 |
| prnn_fglass | 14 | **0,000** | 0,68 | 1,000 | 0,00 | **0,000** | 0,69 | N/A | 0,00 | **0,000** | 0,69 | N/A | 0,00 | **0,000** | 0,71 | 1,000 | 0,00 |
| prnn_synth | 17 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| rmftsa_propores | 21 | **0,016** | 0,23 | **0,001** | 0,38 | **0,003** | 0,32 | **0,000** | 0,49 | **0,014** | 0,24 | **0,001** | 0,39 | **0,003** | 0,32 | **0,000** | 0,48 |
| schizo | 26 | N/A | 0,00 | 0,907 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | N/A | 0,00 | 1,000 | 0,00 |
| scopes-bf | 55 | **0,000** | 0,87 | 0,916 | 0,00 | **0,000** | 0,57 | 0,944 | 0,00 | **0,000** | 0,77 | 0,962 | 0,00 | **0,000** | 0,87 | **0,042** | 0,07 |
| SE | 309 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 0,950 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| segment | 223 | **0,000** | 0,57 | 1,000 | 0,00 | **0,000** | 0,20 | 1,000 | 0,00 | **0,000** | 0,40 | 0,947 | 0,00 | **0,000** | 0,59 | 1,000 | 0,00 |
| sick | 370 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 0,943 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 | 1,000 | 0,00 |
| sonar | 13 | 0,091 | 0,18 | **0,045** | 0,24 | 0,712 | 0,01 | 0,364 | 0,06 | 0,395 | 0,05 | 0,164 | 0,12 | **0,032** | 0,27 | **0,028** | 0,28 |
| soybean | 61 | **0,000** | 0,91 | **0,000** | 0,95 | **0,000** | 0,37 | **0,000** | 0,71 | **0,000** | 0,62 | **0,000** | 0,90 | **0,000** | 0,92 | **0,000** | 0,92 |
| spambase | 453 | **0,000** | 0,87 | **0,000** | 0,51 | **0,000** | 0,51 | **0,000** | 0,47 | **0,000** | 0,81 | **0,000** | 0,52 | **0,000** | 0,82 | **0,000** | 0,52 |
| splice | 311 | **0,000** | 0,88 | **0,000** | 0,96 | **0,000** | 0,67 | **0,000** | 0,47 | **0,000** | 0,89 | **0,000** | 0,81 | **0,000** | 0,91 | **0,000** | 0,90 |
| sylva_agnostic | 1432 | **0,000** | 0,81 | 0,996 | 0,00 | **0,000** | 0,18 | 0,999 | 0,00 | **0,000** | 0,55 | 1,000 | 0,00 | **0,000** | 0,84 | **0,000** | 0,46 |
| sylva_prior | 1432 | **0,000** | 0,57 | 1,000 | 0,00 | **0,000** | 0,06 | 1,000 | 0,00 | **0,000** | 0,26 | 1,000 | 0,00 | **0,000** | 0,73 | **0,000** | 0,20 |
| tic-tac-toe | 88 | **0,000** | 0,79 | 0,130 | 0,03 | **0,000** | 0,17 | **0,002** | 0,10 | **0,000** | 0,39 | 0,123 | 0,03 | **0,000** | 0,86 | **0,002** | 0,10 |
| ticdata_categ | 575 | **0,000** | 0,15 | 1,000 | 0,00 | **0,000** | 0,49 | 1,000 | 0,00 | **0,000** | 0,18 | 1,000 | 0,00 | 1,000 | 0,00 | 0,998 | 0,00 |
| titanic | 213 | **0,000** | 0,64 | **0,000** | 0,76 | **0,000** | 0,13 | **0,000** | 0,16 | **0,000** | 0,39 | **0,000** | 0,46 | **0,000** | 0,78 | **0,000** | 0,87 |
| train | 492 | **0,000** | 0,83 | **0,000** | 0,78 | **0,000** | 0,35 | **0,000** | 0,40 | **0,000** | 0,69 | **0,000** | 0,71 | **0,000** | 0,88 | **0,000** | 0,82 |
| usp05 | 13 | **0,000** | 0,80 | **0,000** | 0,84 | **0,000** | 0,66 | **0,000** | 0,82 | **0,000** | 0,77 | **0,000** | 0,84 | **0,000** | 0,82 | **0,000** | 0,85 |
| V1 | 36 | **0,000** | 0,47 | 1,000 | 0,00 | 0,070 | 0,08 | 1,000 | 0,00 | **0,006** | 0,18 | N/A | 0,00 | **0,000** | 0,61 | 0,206 | 0,04 |
| vehicle | 77 | **0,000** | 0,93 | 1,000 | 0,00 | **0,000** | 0,60 | 1,000 | 0,00 | **0,000** | 0,84 | 1,000 | 0,00 | **0,000** | 0,94 | 1,000 | 0,00 |
| visualizing_fly | 75 | **0,000** | 0,75 | 1,000 | 0,00 | **0,000** | 0,77 | 1,000 | 0,00 | **0,000** | 0,76 | 1,000 | 0,00 | **0,000** | 0,80 | 1,000 | 0,00 |
| VO | 36 | N/A | -0,06 | 1,000 | 0,00 | **0,002** | 0,23 | 0,775 | 0,00 | **0,000** | 0,36 | 1,000 | 0,00 | **0,000** | 0,58 | N/A | 0,00 |
| vote | 36 | **0,003** | 0,20 | 1,000 | 0,00 | **0,000** | 0,30 | 1,000 | 0,00 | **0,002** | 0,23 | 1,000 | 0,00 | **0,000** | 0,29 | 1,000 | 0,00 |
| vowel | 91 | **0,000** | 0,92 | **0,000** | 0,63 | **0,000** | 0,92 | **0,000** | 0,56 | **0,000** | 0,75 | **0,000** | 0,63 | **0,000** | 0,97 | **0,000** | 0,59 |
| waveform-5000 | 492 | **0,000** | 0,76 | **0,000** | 0,35 | **0,000** | 0,27 | **0,000** | 0,20 | **0,000** | 0,57 | **0,000** | 0,28 | **0,000** | 0,79 | **0,000** | 0,31 |

Please note that $R^2$ value is not a square of R. $R^2$ can yield negative values where the predictions that are being compared to the corresponding outcomes have not been derived from a model-fitting procedure using those data, and when fitting non-linear functions to data [25].

For Neural Network algorithm it can be observed that out of 121 datasets, (only) 74 are such that all the models can be used to describe the data, whereas this number increases to 84 in the case of the SVM algorithm.

A natural question arises: is the SVM algorithm with 84 cases consistently better in capturing the underlying data interrelations than the NN with 74 cases and hence better, or is it due to a chance alone? Namely, there are cases where NN performed better than SVM (see e.g. the "analcatdata broadwaymult" and "dmft" dataset). For each of the algorithms we counted how many models were successfully fit to the data in each dataset.

We performed the Wilcoxon's signed rank test to check for the median differences in the number of wins between each algorithm. More formally, we tested the following null hypothesis: $H_0$:$\mu_{SVM}=\mu_{NN.}$ With *P=0,174* (α=0,05) we can not reject the null hypothesis, meaning the SVM is not consistently better in capability of capturing the interrelations in data than the NN algorithm. Thus, the research question is partially answered: neither SVM nor NN is better in capturing the data. But how well the models fit to the data? Are the models produced by the SVM any better than those produced by NN?

To answer the second part of the question we eliminated those datasets where the models do not describe the classifier's performance adequately from our further study because they cannot be compared. These datasets are such that the selected algorithm cannot be used, i.e. the algorithm is inappropriate for the problem domain. In these cases other algorithms and/or approaches need to be used [26]. We ended up with 54 datasets.

From the vector of fitted function's values ($f$) and from the vector $y$ we calculated the mean squared error (MSE) of $j^{th}$ dataset (DS), using Equation 5:

$$MSE_{DS_j} = \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n} \qquad \text{Eq. 5}$$

where $n$ is the number of input points, i.e. the size of a vector, for each individual data set $DS_j$. MSE describes how well the observed points fit to the modelled function.

The average MSEs for NN and SVM were calculated for each of the functions (data presented in Table 2). None of the eight calculated MSE value sets (two models four functions each) is normally distributed (*P<0,000* for all). For this reason we performed the non-parametric Wilcoxon tests. Here, ranks (instead of values) are compared; the model with lower average MSE gets assigned rank 1, the other rank 2. In other words, we calculated an average MSE for a given dataset and a model's function. For example, for the dataset "ada-agnostic", the average MSEs for the power function were 0,000118 and 0,000259 for SVM and NN models, respectively. Here, SMV's MSE was assigned rank 1 and NN's MSE rank 2.

We compared if models and the corresponding function's MSEs are statistically different from each other across all datasets.

Table 2: MSEs and their ranks for POW, LIN, LOG, EXP functions and models (SVM, NN)

| Dataset | POW avg (MSE) SVM | RANK | POW avg (MSE) NN | RANK | LIN avg (MSE) SVM | RANK | LIN avg (MSE) NN | RANK | LOG avg (MSE) SVM | RANK | LOG avg (MSE) NN | RANK | EXP avg (MSE) SVM | RANK | EXP avg (MSE) NN | RANK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ada_agnostic | 0,000118 | 1 | 0,000259 | 0 | 0,000232 | 1 | 0,000278 | 0 | 0,000142 | 1 | 0,000274 | 0 | 0,000114 | 1 | 0,000249 | 0 |
| ada_prior | 0,000189 | 1 | 0,000311 | 0 | 0,000213 | 1 | 0,000387 | 0 | 0,000193 | 1 | 0,000364 | 0 | 0,000171 | 1 | 0,000289 | 0 |
| anneal | 0,000115 | 1 | 0,000179 | 0 | 0,000107 | 1 | 0,000391 | 0 | 0,000100 | 1 | 0,000220 | 0 | 0,000094 | 1 | 0,000147 | 0 |
| australian | 0,000141 | 1 | 0,000858 | 0 | 0,000191 | 1 | 0,000786 | 0 | 0,000170 | 1 | 0,000846 | 0 | 0,000142 | 1 | 0,000780 | 0 |
| autos | 0,000472 | 1 | 0,001616 | 0 | 0,001074 | 1 | 0,001993 | 0 | 0,001163 | 1 | 0,002065 | 0 | 0,000405 | 1 | 0,001400 | 0 |
| badges_plain | 0,000066 | 1 | 0,000108 | 0 | 0,000142 | 0 | 0,000097 | 1 | 0,000089 | 1 | 0,000106 | 0 | 0,000064 | 1 | 0,000094 | 0 |
| balance-scale | 0,000071 | 1 | 0,000261 | 0 | 0,000153 | 1 | 0,000387 | 0 | 0,000103 | 1 | 0,000296 | 0 | 0,000054 | 1 | 0,000255 | 0 |
| baseball-hitter | 0,001045 | 0 | 0,000883 | 1 | 0,002430 | 0 | 0,001102 | 1 | 0,002246 | 0 | 0,002040 | 1 | 0,001986 | 0 | 0,000411 | 1 |
| baseball-pitcher | 0,000139 | 1 | 0,001733 | 0 | 0,000623 | 1 | 0,001624 | 0 | 0,000529 | 1 | 0,001572 | 0 | 0,000162 | 1 | 0,001521 | 0 |
| biomed | 0,000064 | 1 | 0,000518 | 0 | 0,000074 | 1 | 0,000456 | 0 | 0,000067 | 1 | 0,000480 | 0 | 0,000062 | 1 | 0,000449 | 0 |
| breast-cancer | 0,000658 | 0 | 0,000358 | 1 | 0,001044 | 0 | 0,000346 | 1 | 0,000920 | 0 | 0,000356 | 1 | 0,000643 | 0 | 0,000345 | 1 |
| cars_with_names | 0,000519 | 0 | 0,000302 | 1 | 0,000722 | 0 | 0,000262 | 1 | 0,000583 | 0 | 0,000278 | 1 | 0,000433 | 0 | 0,000238 | 1 |
| CH | 0,000187 | 0 | 0,000162 | 1 | 0,000335 | 1 | 0,000474 | 0 | 0,000222 | 0 | 0,000210 | 1 | 0,000076 | 1 | 0,000115 | 0 |
| credit | 0,000204 | 1 | 0,000692 | 0 | 0,000181 | 1 | 0,000732 | 0 | 0,000183 | 1 | 0,000686 | 0 | 0,000178 | 1 | 0,000661 | 0 |
| csb_ch12 | 0,000142 | 1 | 0,000174 | 0 | 0,000117 | 1 | 0,000177 | 0 | 0,000105 | 1 | 0,000171 | 0 | 0,000054 | 1 | 0,000164 | 0 |
| db3-bf | 0,000999 | 0 | 0,000628 | 1 | 0,001938 | 0 | 0,000592 | 1 | 0,002001 | 0 | 0,000620 | 1 | 0,000657 | 0 | 0,000566 | 1 |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| diabetes | 0,000346 | 1 | 0,000419 | 0 | 0,000308 | 1 | 0,000424 | 0 | 0,000342 | 1 | 0,000421 | 0 | 0,000305 | 1 | 0,000407 | 0 |
| ecoli | 0,000261 | 1 | 0,000491 | 0 | 0,000253 | 1 | 0,000481 | 0 | 0,000260 | 1 | 0,000487 | 0 | 0,000253 | 1 | 0,000471 | 0 |
| eucalyptus | 0,000614 | 0 | 0,000502 | 1 | 0,000617 | 1 | 0,001112 | 0 | 0,000601 | 1 | 0,000767 | 0 | 0,000535 | 1 | 0,000542 | 0 |
| eye_movements | 0,000144 | 1 | 0,000326 | 0 | 0,000191 | 1 | 0,000319 | 0 | 0,000155 | 1 | 0,000322 | 0 | 0,000145 | 1 | 0,000300 | 0 |
| genresTrain | 0,000083 | 1 | 0,000150 | 0 | 0,000659 | 1 | 0,001083 | 0 | 0,000219 | 1 | 0,000295 | 0 | 0,000189 | 1 | 0,000254 | 0 |
| gina_agnostic | 0,000254 | 1 | 0,000428 | 0 | 0,000293 | 1 | 0,000612 | 0 | 0,000250 | 1 | 0,000473 | 0 | 0,000232 | 1 | 0,000350 | 0 |
| gina_prior2 | 0,000769 | 0 | 0,000273 | 1 | 0,001140 | 1 | 0,001149 | 0 | 0,000784 | 0 | 0,000576 | 1 | 0,000111 | 1 | 0,000344 | 0 |
| heart-h | 0,000328 | 1 | 0,000450 | 0 | 0,000317 | 1 | 0,000448 | 0 | 0,000328 | 1 | 0,000450 | 0 | 0,000318 | 1 | 0,000442 | 0 |
| HO | 0,000444 | 0 | 0,000349 | 1 | 0,000660 | 0 | 0,000520 | 1 | 0,000561 | 0 | 0,000360 | 1 | 0,000360 | 0 | 0,000341 | 1 |
| kr-vs-kp | 0,000418 | 0 | 0,000247 | 1 | 0,000781 | 0 | 0,000645 | 1 | 0,000456 | 0 | 0,000243 | 1 | 0,000305 | 0 | 0,000127 | 1 |
| kropt | 0,000125 | 1 | 0,000846 | 0 | 0,000781 | 1 | 0,001512 | 0 | 0,000237 | 1 | 0,001867 | 0 | 0,000117 | 1 | 0,000253 | 0 |
| landsat | 0,000151 | 1 | 0,000166 | 0 | 0,000145 | 1 | 0,000160 | 0 | 0,000151 | 1 | 0,000159 | 0 | 0,000867 | 0 | 0,000356 | 1 |
| letter | 0,000073 | 1 | 0,000101 | 0 | 0,002165 | 0 | 0,001217 | 1 | 0,000420 | 1 | 0,000520 | 0 | 0,000299 | 0 | 0,000284 | 1 |
| mfeat-factors | 0,000207 | 0 | 0,000131 | 1 | 0,000753 | 0 | 0,000392 | 1 | 0,000395 | 0 | 0,000212 | 1 | 0,000179 | 0 | 0,000127 | 1 |
| mfeat-fourier | 0,000691 | 0 | 0,000251 | 1 | 0,001043 | 0 | 0,000500 | 1 | 0,000777 | 0 | 0,000274 | 1 | 0,000665 | 0 | 0,000261 | 1 |
| mfeat-karhunen | 0,000156 | 1 | 0,000301 | 0 | 0,001156 | 0 | 0,000925 | 1 | 0,000580 | 0 | 0,000507 | 1 | 0,000183 | 1 | 0,000189 | 0 |
| mfeat-pixel | 0,000168 | 0 | 0,000081 | 1 | 0,000502 | 1 | 0,000571 | 0 | 0,000245 | 0 | 0,000226 | 1 | 0,000135 | 0 | 0,000107 | 1 |
| mozilla4 | 0,000143 | 0 | 0,000111 | 1 | 0,000205 | 0 | 0,000156 | 1 | 0,000156 | 0 | 0,000126 | 1 | 0,000103 | 0 | 0,000091 | 1 |
| MU | 0,000067 | 0 | 0,000018 | 1 | 0,000027 | 1 | 0,000028 | 0 | 0,000018 | 1 | 0,000019 | 0 | 0,000012 | 1 | 0,000016 | 0 |
| mushroom | 0,000099 | 0 | 0,000051 | 1 | 0,000039 | 1 | 0,000043 | 0 | 0,000030 | 0 | 0,000029 | 1 | 0,000021 | 0 | 0,000018 | 1 |
| nursery | 0,000092 | 1 | 0,000145 | 0 | 0,000131 | 1 | 0,000253 | 0 | 0,000098 | 0 | 0,000083 | 1 | 0,000081 | 0 | 0,000056 | 1 |
| optdigits | 0,000088 | 1 | 0,000106 | 0 | 0,000225 | 0 | 0,000197 | 1 | 0,000113 | 0 | 0,000099 | 1 | 0,000071 | 1 | 0,000077 | 0 |
| page-blocks | 0,000045 | 0 | 0,000042 | 1 | 0,000068 | 1 | 0,000106 | 0 | 0,000050 | 1 | 0,000059 | 0 | 0,000041 | 0 | 0,000039 | 1 |
| pc4 | 0,000105 | 1 | 0,000188 | 0 | 0,000102 | 1 | 0,000224 | 0 | 0,000104 | 1 | 0,000193 | 0 | 0,000123 | 1 | 0,000171 | 0 |
| pendigits | 0,000061 | 0 | 0,000050 | 1 | 0,000519 | 0 | 0,000387 | 1 | 0,000199 | 0 | 0,000158 | 1 | 0,000105 | 1 | 0,000108 | 0 |
| scopes-bf | 0,000113 | 1 | 0,000179 | 0 | 0,000215 | 0 | 0,000165 | 1 | 0,000139 | 1 | 0,000142 | 0 | 0,000111 | 1 | 0,000117 | 0 |
| segment | 0,000195 | 0 | 0,000096 | 1 | 0,000217 | 0 | 0,000205 | 1 | 0,000174 | 0 | 0,000126 | 1 | 0,000131 | 0 | 0,000094 | 1 |
| soybean | 0,000163 | 1 | 0,000411 | 0 | 0,000780 | 0 | 0,000478 | 1 | 0,000448 | 0 | 0,000381 | 1 | 0,000141 | 1 | 0,000344 | 0 |
| spambase | 0,000057 | 1 | 0,000236 | 0 | 0,000128 | 1 | 0,000526 | 0 | 0,000071 | 1 | 0,000360 | 0 | 0,000067 | 1 | 0,000235 | 0 |
| sylva_agnostic | 0,000028 | 0 | 0,000026 | 1 | 0,000029 | 1 | 0,000033 | 0 | 0,000016 | 1 | 0,000020 | 0 | 0,000012 | 1 | 0,000013 | 0 |
| sylva_prior | 0,000089 | 0 | 0,000059 | 1 | 0,000053 | 0 | 0,000036 | 1 | 0,000042 | 0 | 0,000026 | 1 | 0,000026 | 0 | 0,000017 | 1 |
| titanic | 0,000239 | 0 | 0,000217 | 1 | 0,000382 | 0 | 0,000233 | 1 | 0,000282 | 0 | 0,000218 | 1 | 0,000176 | 1 | 0,000202 | 0 |
| train | 0,000230 | 1 | 0,000325 | 0 | 0,000464 | 0 | 0,000334 | 1 | 0,000268 | 1 | 0,000322 | 0 | 0,000203 | 1 | 0,000305 | 0 |
| usp05 | 0,000095 | 1 | 0,000285 | 0 | 0,000106 | 1 | 0,000285 | 0 | 0,000093 | 1 | 0,000284 | 0 | 0,000087 | 1 | 0,000269 | 0 |
| vehicle | 0,000596 | 1 | 0,000990 | 0 | 0,001395 | 0 | 0,001331 | 1 | 0,000767 | 1 | 0,001008 | 0 | 0,000493 | 1 | 0,001036 | 0 |
| VO | 0,000305 | 0 | 0,000114 | 1 | 0,000183 | 0 | 0,000142 | 1 | 0,000159 | 0 | 0,000124 | 1 | 0,000127 | 0 | 0,000099 | 1 |
| vowel | 0,001209 | 1 | 0,001244 | 0 | 0,001114 | 1 | 0,001591 | 0 | 0,002663 | 0 | 0,001623 | 1 | 0,000779 | 0 | 0,000584 | 1 |
| waveform-5000 | 0,000141 | 1 | 0,000180 | 0 | 0,000263 | 1 | 0,000281 | 0 | 0,000173 | 1 | 0,000212 | 0 | 0,000133 | 1 | 0,000166 | 0 |
| SUM | 0,000269 | 32 | 0,000364 | 22 | 0,000518 | 32 | 0,000541 | 22 | 0,000401 | 31 | 0,000451 | 23 | 0,000251 | 35 | 0,000313 | 19 |

As can be observed, the NN's EXP had rank-sum of 19 and an average MSE of 0,000313; SVM's EXP had rank-sum of 35 and the average MSE of 0,000251. Please note that the rank is an ordinal value and hence calculating its mean value is inappropriate [24, p. 472].

Finally, the main research question was tested: which model was better? To rephrase, was SVM with the rank-sum of 35 and average MSE of 0,000251 significantly better than NN with rank-sum of 19 and average MSE of 0,000313? The same type of question was asked for the other types of functions.

Formally, to test the significance of difference in MSEs we used Wilcoxon's related samples test. The null hypotheses, the mean of differences between model's $M_i\ (MSE)$ and $M_j\ (MSE)$ for a given function equals 0, were as follows:

- $H1_0$:$\mu_{MSE/powerSVM}$=$\mu_{MSE/powerNN}$;
- $H2_0$:$\mu_{MSE/linearSVM}$=$\mu_{MSE/linearNN}$;
- $H3_0$:$\mu_{MSE/logaritmicSVM}$ =$\mu_{MSE/logaritmicNN}$; and
- $H4_0$:$\mu_{MSE/exponentialSVM}$ =$\mu_{MSE/exponentialNN}$.

Table 4 lists the results of Wilcoxon's signed rank test analysis for all four comparisons, with values in bold indicating significance at $\alpha$=0,05 level.

Table 3: Wilcoxon signed rank test for different function models

| Pair # | Pair | Sig. (2-tailed) |
|---|---|---|
| Pair 1 | POW (rank SVM) – POW (rank NN) | 0,067 |
| Pair 2 | LIN (rank SVM) – LIN (rank NN) | 0,295 |

| Pair # | Pair | Sig. (2-tailed) |
|---|---|---|
| Pair 3 | LOG (rank SVM) – LOG (rank NN) | 0,204 |
| Pair 4 | EXP (rank SVM) – EXP (rank NN) | **0,023** |

The results show that exponential function's average rank for the SVM algorithm is significantly different at $\alpha$=0,05 threshold from average rank of NN algorithm ($P$=0,023). Thus, all the above mentioned hypotheses $H1_0$ to $H3_0$ need to be rejected with the exception of $H4_0$.

This means that SVM is performing better than NN when exponential function is used for the modelling. As shown in previous works, exponential function is the most appropriate for the modeling of algorithm's performance.

## 4. Conclusion

In this paper we conducted analyses of error rate curves produced by a selected multilayer perceptron neural network classifier and a support vector machines classifier. The results show that, in average, we cannot reject the hypothesis that both algorithms are equally good in capturing the internal relationships in data (P=0,174). However, each of the algorithms can be superior in limited specific cases.

Since both algorithms are performing equally well, the further investigation went into the goodness-of-fit measure. The question was, does one algorithm model the data more precisely than the other one? The answer is that when using POW, LIN, and LOG functions, there are no differences (*P=0,067*, *P=0,295* and *P=0,204*, respectively). However, when using the EXP function, SVM algorithm is statistically significantly outperforming the NN (*P=0,023*).

The contribution of the work is important in many perspectives: firstly, the first choice when modeling data using classification is the SVM algorithm using exponential function. Secondly, based on this finging, early in the learning phase one can fit the model's parameters and estimate the final error rate. In case the estimated final performance is lower than the one required, one can modify the learner's parameters early in the process. Thirdly, the results of our experiment show that some datasets exist where modelling of the artificial learner's performance is not successful due to the inability of a learner to properly capture the data interrelations, and that some algorithms are more adequate for a specific task than the others. This too could be detected early in the learning process.

## Acknowledgements

This work was partially supported by the Slovenian Research Agency under grant numbers 1000-11-310138 and P2-0057 (B).

## References

1. Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques. INFORMATICA (Ljubljana) 2007; 31(3): 249-268.
2. Dzemyda G and Sakalauskas L. Large-Scale Data Analysis Using Heuristic Methods. Informatica (Lithuan.) 2011; 22(1): 1-10.
3. Vapnik VN. Estimation of Dependences Based on Empirical Data. NY: Springer-Verlag 1982.
4. Brumen B, Jurič MB, Welzer T, Rozman I, Jaakkola H and Papadopoulos A. Assessment of classification models with small amounts of data. Informatica (Lithuan.) 2007; 18(3): 343-362.
5. Dučinskas K and Stabingiene L. Expected Bayes Error Rate in Supervised Classification of Spatial Gaussian Data. Informatica (Lithuan.) 2011; 22(3): 371-381.
6. Frey LJ and Fisher DH. Modeling decision tree performance with the power law. Seventh International Workshop on Artificial Intelligence and Statistics; 1999. Ft. Lauderdale, Fl, USA: Morgan Kaufmann.
7. Last M. Predicting and Optimizing Classifier Utility with the Power Law. 7th IEEE International Conference on Data Mining. ICDM Workshops 2007.; 2007. Omaha, Nebraska, USA: IEEE; DOI: 10.1109/icdmw.2007.31.

8. Provost F, Jensen D and Oates T. Efficient progressive sampling. Fifth International Conference on Knowledge Discovery and Data Mining; 1999. San Diego: ACM.
9. Singh S. Modeling Performance of Different Classification Methods: Deviation from the Power Law. Project Report. Nashville, Tennessee, USA: Vanderbilt University, Department of Computer Science; 2005.
10. Witten IH and Frank E. Data Mining: Practical machine learning tools and techniques. 2nd ed.San Francisco, CA, USA: Morgan Kaufmann; 2005. ISBN: 0120884070
11. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P and Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009; 11(1): 10-18.
12. Asuncion A and Newman D. UCI Machine Learning Repository. 2010; Available from: http://archive.ics.uci.edu/ml/datasets.html. (Archived by WebCite® at http://www.webcitation.org/6C2hgsRrX).
13. TunedIT. TunedIT research repository. 2012; Available from: http://tunedit.org/search?q=arff&qt=Repository. Accessed: 2012-12-12. (Archived by WebCite® at http://www.webcitation.org/6CqplN6Xr).
14. Kjellerstrand H. My Weka page. 2012; Available from: http://www.hakank.org/weka/. Accessed: 2012-12-12. (Archived by WebCite® at http://www.webcitation.org/6Cqq5pQtZ).
15. Kjellerstrand H. My Weka page/DASL. 2012; Available from: http://www.hakank.org/weka/DASL/. Accessed: 2012-12-12. (Archived by WebCite® at http://www.webcitation.org/6CqqCwPmy).
16. Chai K. Kevin Chai Datasets. 2012; Available from: http://kevinchai.net/datasets. Accessed: 2012-12-12. (Archived by WebCite® at http://www.webcitation.org/6CqqWlQEp).
17. Brumen B, Hölbl M, Harej Pulko K, Welzer T, Heričko M, Jurič MB and Jaakkola H. Learning Process Termination Criteria. Informatica (Lithuan.) 2012; 23(4): 521-536.
18. Cohen PR. Empirical methods for artificial intelligence. Cambridge, MA, USA: MIT press; 1995. ISBN: 9780262032254
19. Weiss SM and Kulikowski CA. Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo, CA, USA: Morgan Kaufmann; 1991. ISBN: 978-1558600652
20. McLachlan GJ, Do K-A and Ambroise C. Analyzing microarray gene expression data. Hoboken, NJ, USA: Wiley; 2004. ISBN: 0471226165
21. Eaton JW. GNU Octave. 2012; Available from: http://www.gnu.org/software/octave/. Accessed: 2012-12-12. (Archived by WebCite® at http://www.webcitation.org/6CqyEvDKU).
22. Marquardt DW. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of the Society for Industrial and Applied Mathematics 1963; 11(2): 431-441. DOI: 10.2307/2098941.
23. Levenberg K. A Method for the Solution of Certain Non-Linear Problems in Least Squares. Quarterly of Applied Mathematics 1944; 2: 164–168.
24. Argyrous G. Statistics for research: With a guide to SPSS, 3rd ed. Thousand Oaks, CA, USA: SAGE Publications Ltd.; 2011. ISBN: 1849205957
25. Colin Cameron A and Windmeijer FAG. An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics 1997; 77(2): 329-342. DOI: http://dx.doi.org/10.1016/S0304-4076(96)01818-0.
26. Medvedev V, Dzemyda G, Kurasova O and Marcinkevicijus V. Efficient Data Projection for Visual Analysis of Large Data Sets Using Neural Networks. Informatica (Lithuan.) 2011; 22(4): 507-520.

# Real-time Sensing, Processing and Actuating Functions of 5D World Map System: A Collaborative Knowledge Sharing System for Environmental Analysis

Shiori SASAKI [a, b] and Yasushi KIYOKI [a]
[a]*Keio University, Japan*, [b] *Chulalongkorn University, Thailand*
{sashiori, kiyoki}@sfc.keio.ac.jp

**Abstract.** This paper presents real-time sensing, processing and actuating functions of a collaborative knowledge sharing system called 5D World Map System, and the applications in the field of multidisciplinary environmental research and education. The first objective is to integrate the analysis of sensing data into a knowledge sharing system with multimedia, based on the framework of Sensing-Processing-Actuation of Cyber-Physical Systems. The proposed real-time sensing, processing and actuating functions enable multiple remote-users to acquire real-time sensing data from multiple sites around the world, perform analytical visualizations of the acquired sensing data by the selected calculation methods on the system to discover the incidental phenomena, and provide the analysed results to related users' terminal equipment automatically. The second objective of the research is to realize a new multidimensional data analysis and knowledge sharing system for a collaborative environment by applying the concept of "differential computing" to the analysis of sensing data. Especially, in the processing function, the time-series difference of the value of each sensor, the differences between the values of multiple sensors in a selected area, and the time-series differences between the values of multiple sensors, are calculated, to detect an environmental incident and estimate the possibility of occurrence of the same kind of incident in the neighboring sites within the same area. By using 5D World Map System integrated with these functions, the users are able to perform a global analysis on the environmental sensing data along with the related multimedia data on a single view of time-series maps, based on the spatiotemporal and semantic correlation calculations.

**Keywords.** global environment, information system, sensor data, differential computing, spatiotemporal database, multimedia, semantic associative search, data mining, mobile computing, ubiquitous computing,

## 1. Introduction

This paper presents real-time sensing, processing and actuation functions of a collaborative knowledge sharing system called 5D World Map System, and its applications in the field of multidisciplinary environmental research and education.

In the field of environmental engineering, environmental measurement with sensing data is seemed to be the base of the researches, and several analytical methods and environmental modelling have been studied strenuously. On the other hand, in the field of information and communication technology, several ideas and designs of multimedia sharing systems have been proposed, and sophisticated techniques to share the knowledge have been implemented in various applications.

On reflection, we have been constructed a collaborative knowledge sharing system called 5D World Map System [4][5], which provides various functionalities to share and visualize various types of multimedia data. Though this system is effective for sharing the information and knowledge for environmental analysis, the analytical functions for real-time sensing data had not been realized yet.

In this research, we propose that a combination of the multimedia sharing and visualizing functions and the real-time sensing data analysis functions makes environmental analysis much richer and deeper, which contributes to activities of collaborative knowledge creation.

The first objective is to integrate the analysis of sensing data into a knowledge sharing system with multimedia, based on the framework of Sensing-Processing-Actuation of Cyber-Physical Systems (CPS) [1][2][3].

In the research field of CPS, it is proposed and promoted that “a scientific and engineering CPS discipline should advance the conceptualization and realization of future societal-scale systems characterized by: (a) deep integration and pervasiveness of real-time processing, sensing, and actuation across logical and physical heterogeneous domains; and (b) systematic analysis of the interactions between engineering structures, information processing, humans and the physical world”. [3]

To integrate the real-time analysis of sensing data and the multimedia sharing system based on the framework, we propose a set of real-time sensing, processing and actuating functions on 5D World Map System, which enables multiple remote-users to acquire real-time sensing data from multiple sites around the world, perform analytical visualizations of the acquired sensing data by the selected calculation methods on the system to discover the incidental phenomena, and receive the analysed results by their own terminal equipment automatically.

The second objective of the research is to realize a new multidimensional data analysis and knowledge sharing system for collaborative environmental research by applying the concept of “differential computing” [4] to the analysis of sensing data.

By using differential computing, “important factors that change natural environment are highlighted,” and “the highlighted factors are visualized by using our Multi-dimensional World Map, which makes it possible to view the nature environment changes in the view of history, geographic, etc.” [4].

Specifically, in the proposed processing function, the time-series difference of the value of each sensor, the differences between the value of multiple sensors in a selected area, and the time-series differences from the differences between the values of multiple sensors, are calculated, to detect an environmental incident and estimate the possibility of occurrence of the same kind of incident in the neighboring sites within the same area.

By using the 5D World Map System integrated with these functions, the users are able to perform a global analysis on the environmental sensing data along with the related multimedia data on a single view of time-series maps, based on the spatiotemporal and semantic correlation calculations.

In this paper, we introduce the overview of the 5D World Map System that have been proposed in [4][5] in Section 2, the basic method of the proposed real-time sensing, processing and actuation functions in Section 3, the implementation method of the proposed functions in Section 4, and the use cases in Section 5 to examine the feasibility of the proposed functions in the real environmental analysis.

## 2. Overview of the 5D World Map System

5D World Map System is a knowledge creation and sharing system which enables semantic, temporal and spatial analysis of multimedia, integrates and visualizes the analyzed results as a 5-dimentional dynamic historical atlas (5D World Map Set) [4][5]. The main feature of this system is to create various context-dependent patterns of environmental/historical/cultural stories according to a user's viewpoints dynamically. This system generates multiple views of semantic and temporal-spatial relationships among multimedia of the multidisciplinary fields such as global environmental issues. A semantic associative search method [6] is applied to this system for realizing the concept that "semantics" of words, features of multimedia, and events according to the "context". Semantically-evaluated and analyzed multimedia data is also mapped dynamically onto a time-series multi-geographical space. This system provides high visibility of semantic correlations between multimedia in time series variation with geographic information [4][5][7].

By applying to the analysis of environmental issues, 5D World Map System can be utilized as a knowledge sharing system to analyze the commonality and difference among regional environments in a global view, and a collaborative knowledge creation system for remote users such as researchers and students of the same environmental issues and phenomena, such as wild fire, ecosystem, climate change, etc., around the world (**Figure 1**).

The applications and the usage of this system are described by several significant study scenarios: (a) Cross-topic environmental studies using multimedia retrieval, (b) Time-series based environmental-change observation, and (c) Cross-space environmental-issue overview [8]-[15].

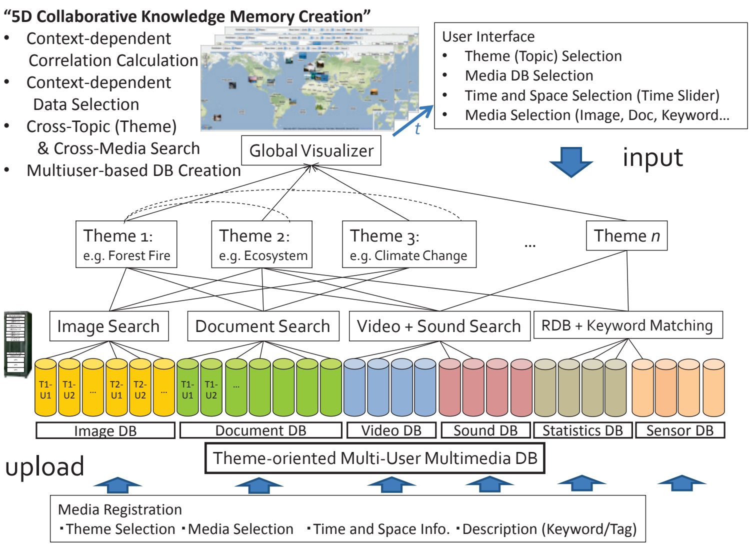



**Figure 1** Overview of 5D World Map System [4][5]

The latest 5D World Map System consists of five main functions: (1) Cross-topic multimedia search by semantic similarity calculation, (2) Multimedia database overview by spatiotemporal information, (3) Media data uploader for multi users, (4) Differential computing for spatiotemporal data, and (5) Historical-geographical information visualization.

### *2.1. Multimedia Search by Semantic Similarity Calculation*

Metadata of multimedia data such as word frequency of documents or color frequency of images are vectorized and mapped onto the semantic or color space for calculating the similarity between a query and the multimedia. By this function, the correlations between (a) multimedia and/or keywords input by a user as a query and (b) multimedia data mapped onto the search spaces are calculated.

### *2.2. Multimedia Database Overview by Spatiotemporal Information*

Various multimedia data (Text, Image, Sound, Movie), within a selected category or across the categories, are visualized onto a set of time-series world maps. By this function, the comparative analysis among various countries, regions, cities from the aspect of differences and similarities is realized, and the time-series changes can be observed .

### *2.3. Multimedia Data Uploader*

Users can upload the multimedia data in various formats (.txt, .pdf, .csv, .jpg, .png, .gif, .mpeg, .kml) on various topics of environmental issues from natural science to social science (e.g. “energy efficiency”, “history of the international agreements on climate change”, etc.) and share them from remote client. By this function, a collaborative database creation and a real-time analysis on the global environment are realized. Users can upload, share, download, edit any contents and re-upload their edited multimedia. GUI support users to input spatiotemporal information (date and place) of media data.

### *2.4. Differential Computing and Visualization for spatiotemporal data*

This function visualizes numerical and statistical data with spatiotemporal information such as population, energy consumption/production, GDP, CO2 emission, ratio of forest coverage, oil dependency rate, etc.. By this function, users can visualize the collected numerical values onto the time-series multi-geographical spaces as a set of colored polygon data or variable-sized markers. In addition, this function enables to visualize the differences extracted from a set of time-series geographical images such as satellite images, aerial photos, and meteorological information images as colors. The differences are calculated from the correspondence relationship between the positions on a set of time-series images and geographical coordinates (latitude and longitude).

### *2.5. Historical-geographical Information Visualization*

This function maps and visualizes by overlaying the geographical information data in the format of Keyhole Markup Language (KML) such as polygon data of landform, ocean area, forest area, iced area, desert area etc., location data of city, road, river etc., statistical data of population, GDP, house index and historical data such as national borders etc. The function creates the geographical information database automatically based on a KML file uploaded from a user with the attributes of date, user name, category etc. Multiple geographical data in KML format can be visualized on the same map at once.

Although these existing functions are effective and useful for the knowledge sharing with multimedia data to analyze the commonality and difference among regional environments in a global view, we realize the sensing, processing and actuating functions for sensing data analysis in addition to the existing functions, and embed these functions into 5D World Map System. The combination of sensing data analysis and multimedia search and mapping enables more effective and deep environmental analysis.

## 3. Sensing, Processing and Actuation Functions of the 5D World Map System

In this section, we describe the following four functions: (1) Real-time Tracking for sensor data, (2) Real-time Alerting for sensor data, (3) Real-time Differential Computing for sensor data, and (4) Analytical Visualization for sensor data. (1) is realized as a sensing function, (2) is realized as an actuation function, (3) and (4) are realized as processing functions. The structure of the system and the relations between these functions and the existing functions are shown in Figure 2.

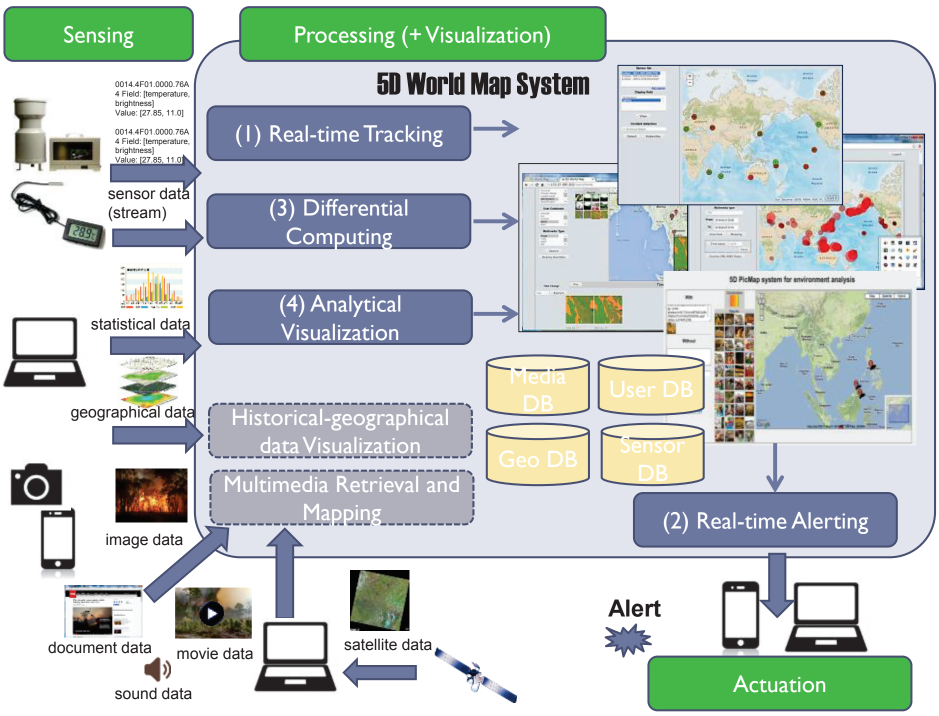



**Figure 2** Overview of the sensing, processing and actuating functions of 5D World Map System

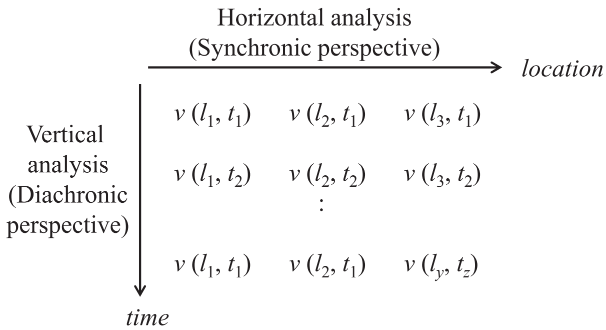



**Figure 3** Basic concepts of horizontal analysis and vertical analysis, and the basic data structure of sensor database for the differential computing for sensor data in 5D World Map System

Figure 3 presents the basic concepts of horizontal analysis and vertical analysis. The horizontal analysis indicates that it calculates the differences among the values of sensor data in multiple locations at a time, and the vertical analysis indicates that it calculates the time-series differences among the values of sensor data in a same location. Figure 3 also presents the basic data structure of sensor database for the differential computing for sensor data in 5D World Map System

*3.1. Real-time Tracking for Sensing Data*

This function is realized as a sensing function in the Sensing-Processing-Actuation framework.

To receive the various formats of values sent from various types of sensing device (e.g. thermometer, water-level indicator, colorimeter, etc.) by every second, identifying the device and the data type is essential. In addition, even from one single sensing device, various types of sensing data (e.g. temperature, humidity, light, etc.) may be sent to the system. Thus, each sensing device is needed to be given a universally unique identifier (UUID). UUID is an identifier standard used in software construction, which enable distributed systems to uniquely identify information without significant central coordination. Also, each sensing device should be given a session token. A session token is a piece of data that is used in network communications (often over HTTP) to identify a session, a series of related message exchanges. Each value $v_i$ sent by each session should include the information of user ID and device name also.

Where each value of a sensor value set $V := \{v_1, v_2, \ldots, v_n\}$ has the attributes of user ID $U := \{u_1, u_2, \ldots, u_m\}$, device name $N := \{n_1, n_2, \ldots, n_l\}$, data type $D := \{d_1, d_2, \ldots, d_o\}$, universally unique identifier $UUID := \{uuid_1, uuid_2, \ldots, uuid_p\}$, token $TK := \{tk_1, tk_2, \ldots, \mathrm{k}t_z\}$, $T := \{time\text{-}stamp\}$, location $L := \{latitude, longitude\}$, the data structure of the sensor database for real-time tracking $DB_{track}$ is expressed by using a collection of sets $S_\lambda := \{U, N, D, UUID, TK, T, L\}$ as follows:

$$DB_{track} = \prod_{\lambda \in \Lambda} S_\lambda$$

When a sequence of attributes of value $sq_i$ = ($v_i$, $u_i$, $n_i$, $d_i$, $uuid_i$, $tk_i$, *time-stamp*, *latitude*, *longitude*) is sent from a unique sensing device, the system receives and stores it into the sensor database for real-time tracking $DB_{track}$ in an order according to the time stamp. A sequence of a value

### *3.2. Real-time Alerting with Sensing Data*

This function is realized as a actuating function in the Sensing-Processing-Actuation framework.

When a user $u_i$ registers a value of threshold $th_\varepsilon$ (ε > 0) for a specific data type $d_i$ of a specific device $n_i$, the system calculates the difference between the threshold $th_\varepsilon$ and every value in every second, and sends the sequence of value and the difference between the $th_\varepsilon$ and the value with a message as an alert to the user. Where a sensor value set $V := \{v_1, v_2, \ldots, v_n\}$, a value of threshold $th_\varepsilon$ (ε > 0), and comparison operators ю $:= (<, >, \leq, \geq, =)$ are given, the alerting function $f_{alert}$ is described as:

$$f_{alert} : V \rightarrow (v_i, |(v_i - th_\epsilon)| \mid v_i \text{ ю } th_\varepsilon)$$

### *3.3. Real-time Differential Computing for Sensing Data*

This function is realized as a processing function in the Sensing-Processing-Actuation framework.

To detect any environmental incident in an area by analyzing the characteristics of sensor data in the area, we design four types of differential operations as a trial.

(1) Differential Operation 1 (*Op*1): Calculation of difference between the current value of each sensor and the threshold $th_\varepsilon$ for incident detection using comparison operators ю $:= (<, >, \leq, \geq, =)$
(2) Differential Operation 2 (*Op*2): Calculation of difference between maximum and minimum values of the multiple sensors in an area
(3) Differential Operation 3 (*Op*3): Calculation of difference between the value (t-1) and the value (t) of each sensor
(4) Differential Operation 4 (*Op*4): Calculation of time-series differential ratio between the value (t-1) and the value (t) of the differences between the values of multiple sensors

Where each value of a sensor value set $V := \{v_1, v_2, \ldots, v_n\}$ and a sequence of attributes of value $sq_i$ = ($v_i$, $u_i$, $n_i$, $d_i$, $uuid_i$, $tk_i$, *time-stamp*, *latitude*, *longitude*) are given, *Op*1, *Op*2, *Op*3 and *Op*4 are described as follows.

$$Op1 : \left(Sq_i(t),\ Sq_j(t)\right) \rightarrow (v_i(t) \mid v_i \text{ ю } th_\varepsilon)$$

where ю $:= (<, >, \leq, \geq, =)$

$$Op2 : \left(Sq_i(t),\ Sq_j(t)\right) \rightarrow max \mid v_i(t) - v_j(t) \mid$$

$$Op3 : \left( (Sq_i(t),\ uuid_i)\ , (Sq_j(t+1), uuid_i) \right) \rightarrow$$
$$|\ (v_i(t),\ uuid_i)\ -\ (v_j(t+1), uuid_i)\ |$$

$$Op4 : \left( (Sq_i(t),\ uuid_i)\ , (Sq_k(t+1), uuid_j), (Sq_j(t+1),\ uuid_i)\ , (Sq_l(t+1), uuid_j) \right)$$
$$\rightarrow max|\ \left( (v_i(t),\ uuid_i)\ -\ (v_j(t+1), uuid_i) \right) - \left( (v_k(t),\ uuid_j)\ -\ (v_l(t+1), uuid_j) \right) |$$

The typical usage of each operator is described as follows. *Op*1 is used for detecting an environmental incident or abnormality in a selected area or all the areas in the world. For example, the dangerous water-level of a river region is detected, if the selected data type is "water-level". *Op*2 is used for calculating the range of values of a specific data type in a selected area or all the areas in the world. For example, how much the amount of rain fall varies in a county is able to be detected, if the selected data type is "rain-fall amount". *Op*3 is used for detecting the moment when the value of a place changes drastically. For example, the moment when a fire occurs in a specific region is able to be detected, if the selected data type is "temperature". *Op*4 is used for detecting the place where an incident or abnormal phenomena happens in a specific area, and estimating the possibility of occurrence of the similar incident in the neighboring places in the next moment. For example, the possibility of occurrence of water pollution in a downstream area of a river is able to be estimated by using "water-quality" data type, if the difference between an upstream area and a midstream area increases suddenly.

### *3.4. Analytical Visualization for Sensing Data*

This function is also realized as a processing function in the Sensing-Processing-Actuation framework.

The sensor database created by the process of Section 3.1 is converted in the different format of table for the visualization on the 5D World Map. In this process, UUID, Token and device name are deleted because these are not needed in the visualization.

Where each value of sensor data $\{v_1, v_2, ..., v_n\}$ has the attributes of user ID $U := \{u_1, u_2, \ldots, u_m\}$, category name $C := \{c_1, c_2, \ldots, c_g\}$, keyword $K := \{k_1, k_2, \ldots, k_h\}$, value ID $VID := \{vid_1, vid_2, \ldots, vid_k\}$, $T := \{time\ stamp\}$, location $L := \{latitude, longitude\}$, the data structure of the sensor database for analytical visualization $DB_{vis}$ is expressed by using a collection of sets $P_\lambda := \{U, C, K, VID, T, L\}$ as follows:

$$DB_{viz} = \prod_{\lambda \in \Lambda} P_\lambda$$

Where all the sensor value set $V := \{v_1, v_2, ..., v_n\} \in DB_{vis}$ and a set of characters for selection CH := $\{ch_1, ch_2, \ldots, ch_q\}$ are given, a sequence of attributes of value for visualization $s_{ui} = (v_i, u_i, k_i, vid_i, time\text{-}stamp, latitude, longitude)$ is selected by the value selector $Vs$.

$$V_S : CH, V \rightarrow \{S_{ui} \mid k_i \supseteq CH\}$$

From the selected sequences of value $s_{ui}$, a set ($v_i$, $vid_{i,}$ *time-stamp*, *latitude*, *longitude*) is projected by the value projector $Pv$, and mapped onto a time-series world map sets as sizable colored markers.

$$P_v : S_{ui}, V \rightarrow \{ v_i, vid_{i,}\ time\text{-}stamp, latitude, longitude\}$$

## 4. Implementation

The 5D World Map System with sensing, processing and actuating functions enables users to access the sensing data obtained from various sensors of various platforms that have been installed in remote research institutions of collaborative partners around the world, analyze and visualize the sensing data according to their own viewpoint (by setting threshold or selecting the methods of differential calculation), and receive an alert of the sensed abnormal value. Furthermore, by registering their own sensors used in their researches to this system, the users are able to automatically calculate, analyze and visualize the data on the map, and share the analyzed results with other users around the world.

### *4.1. Real-time Tracking, Alerting and Differential Computing for Sensing Data*

In the implementation, a prototype system to store various kinds of real-time sensing data (temperature, humidity, brightness, rainfall, water quality, etc.) from various sensing devices in the database is created. As shown in Figure 4, the system receives and stores these kinds of data, performs the spatiotemporal difference calculation, and maps the analyzed results onto 5D World Map Set. The objective of the implementation is to construct a database system for real-time sharing and visualization of stream-type sensing data, which is sent from various sensing devices and accessible to the remote users.

For the prototype implementation, the sensor database is created by using open-source NoSQL database MongoDB [16]. The other implementation environment is described as below.

- OS: Mac OS X, Windows Server 2008 R2 Enterprise
- Software: ArcGIS Server 10.2 later, MongoDB, Netbeans IDE 6.8 later, CakePHP Framework, Apache HTTP Server, GlassFish Server
- Language: HTML, Java, Python, PHP, CakePHP, JavaScript

As a client sensing device, we selected Oracle Sun SPOT JAVA Development Kit [17] for monitoring temperature and lighting, and developed a program to send the sensing data by using Sun SPOT Manager [18] and Java SE Development Kit 8 [19]. The API to be provided by the system is implemented in JSON format by JavaScript [20]. The data processing from client devices to the server is described in Figure 3.

The sensing data is sent from Sun SPOT sensor board via radio wave to the processor board (base station) connected to the host computer. The host computer receives the sensing data via a program to send it to the 5D World Map server through the internet network as a stream data. The program to send the sensing data to the server should include the sensor information provided as JSON format in API, consisting of user ID, token, time, latitude, longitude, UUID, device name and data type.

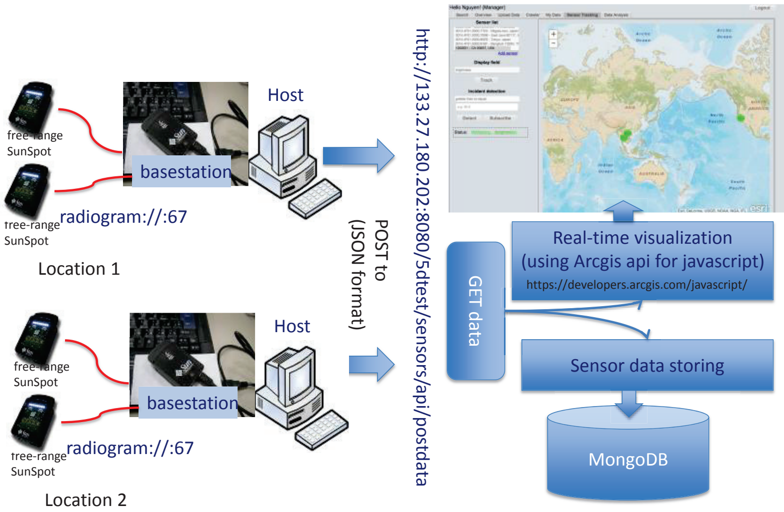
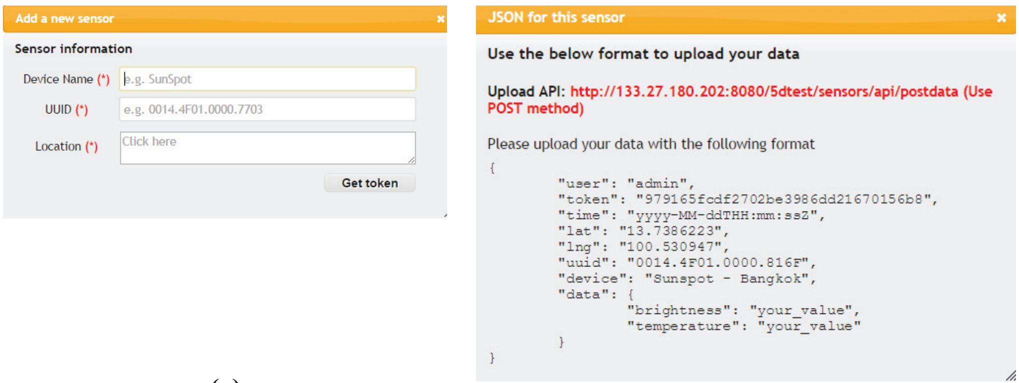



**Figure 4** Data processing from a client sensing devices to the 5D World Map System server




**(a)** **(b)**

**Figure 5 (a)** Interface for sensing device registration in 5D World Map System, and **(b)** API to send sensor data from sensing devices to the server, provided by 5D World Map System (When a user register device name, UUID and address to (a), the JSON format to send the sensing data to the server including user ID, token, time, latitude, longitude, UUID, device name, data type is provided to the unique user by the system as shown in (b).

As shown in **Figure 5**, after a user logs in the 5D World Map System, the user can register his/her own sensing devices with the information of device name, UUID and address as client devices of the system by using the interface shown in **Figure 5(a)**. When the user completes the registration, the JSON format to send the sensing data to the server including user ID, token, time, latitude, longitude, UUID, device name, data type is provided by the system as shown in **Figure 5(b)**. To send the sensing data from their own devices, the sensor information provided as JSON format in API, consisting of user ID, token, time, latitude, longitude, UUID, device name and data type is needed to be embed to the program.

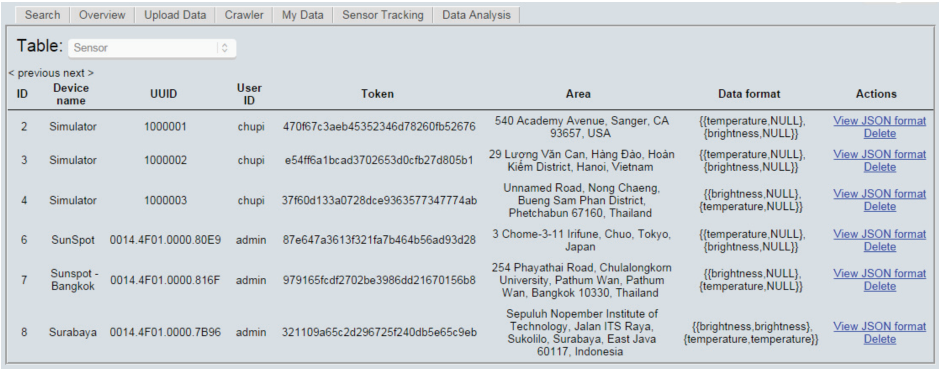

Search | Overview | Upload Data | Crawler | My Data | Sensor Tracking | Data Analysis

Table: Sensor

< previous next >

| ID | Device name | UUID | User ID | Token | Area | Data format | Actions |
|---|---|---|---|---|---|---|---|
| 2 | Simulator | 1000001 | chupi | 470f67c3aeb45352346d78260fb52676 | 540 Academy Avenue, Sanger, CA 93657, USA | {{temperature,NULL}, {brightness,NULL}} | View JSON format Delete |
| 3 | Simulator | 1000002 | chupi | e54ff6a1bcad3702653d0cfb27d805b1 | 29 Lượng Văn Can, Hàng Đào, Hoàn Kiếm District, Hanoi, Vietnam | {{temperature,NULL}, {brightness,NULL}} | View JSON format Delete |
| 4 | Simulator | 1000003 | chupi | 37f60d133a0728dce9363577347774ab | Unnamed Road, Nong Chaeng, Bueng Sam Phan District, Phetchabun 67160, Thailand | {{brightness,NULL}, {temperature,NULL}} | View JSON format Delete |
| 6 | SunSpot | 0014.4F01.0000.80E9 | admin | 87e647a3613f321fa7b464b56ad93d28 | 3 Chome-3-11 Irifune, Chuo, Tokyo, Japan | {{temperature,NULL}, {brightness,NULL}} | View JSON format Delete |
| 7 | Sunspot - Bangkok | 0014.4F01.0000.816F | admin | 979165fcdf2702be3986dd21670156b8 | 254 Phayathai Road, Chulalongkorn University, Pathum Wan, Pathum Wan, Bangkok 10330, Thailand | {{brightness,NULL}, {temperature,NULL}} | View JSON format Delete |
| 8 | Surabaya | 0014.4F01.0000.7B96 | admin | 321109a65c2d296725f240db5e65c9eb | Sepuluh Nopember Institute of Technology, Jalan ITS Raya, Sukolilo, Surabaya, East Java 60117, Indonesia | {{brightness,brightness}, {temperature,temperature}} | View JSON format Delete |

**Figure 6** Examples of sensor information registered in the sensor database in 5D World Map System (The information of each sensor device is stored with attributes of sensor ID, device name, UUID, Token, Area (address), data type .)

**Figure 6** shows the examples of sensor information registered in the sensor database in 5D World Map System (The information of each sensor device is stored with attributes of sensor ID, device name, UUID, Token, Area (address), data type .)

**Figure 7** shows the visualization results of sensor data onto 5D World Map set. The location of each sensor is appeared in the world map, and the value of each sensor is represented as the circled marker. If a user specifies sensor and data type for displaying on the map from the lists of registered sensors and data types, the selected sensor data will appear on the map. A circle and measured value will be displayed at sensor position. If the user inputs a threshold value in the "incident detection" and selects a comparison operator, the differences between the threshold and each sensor value are calculated and the corresponding marker(s) will blink.

**Figure 8(a)** shows the registration interface for subscription of alert, and **Figure 8(b)** shows an example of alert message with a sequence of sensing data information. When the users observe the visualization results of sensor data in the world map, and think the results deserve to watch constantly, they are able to register their e-mail addresses to the system so that they can receive the calculation result based on the minimum interval they set. By specifying condition, you can receive alert by e-mail when sensor detect abnormal value.

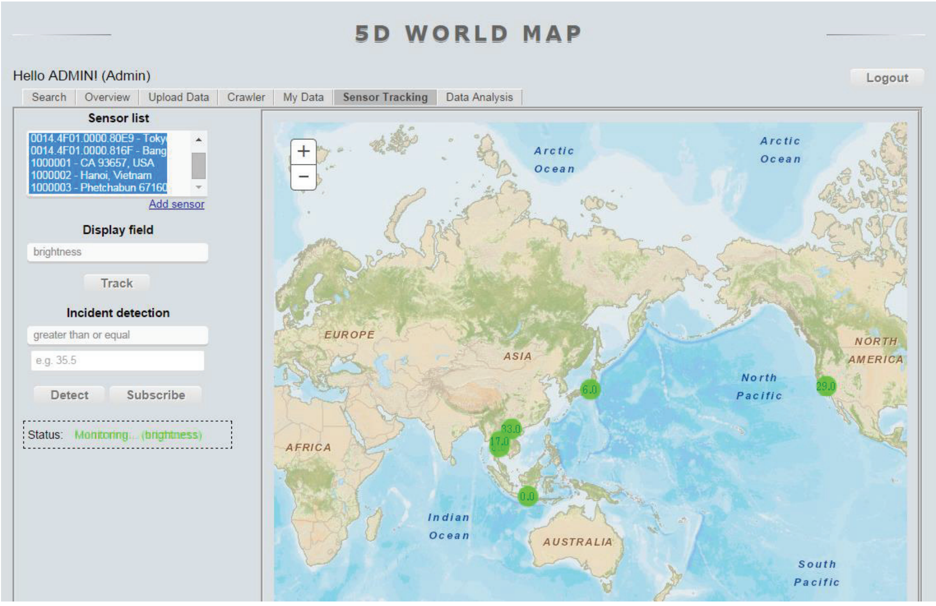
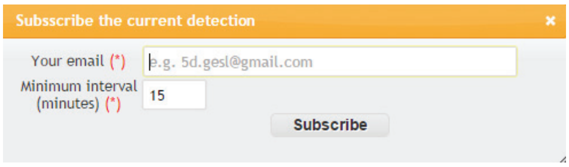



**Figure 7** Visualization results of sensor data onto 5D World Map set (The value of each sensor (to measure brightness, in this case) is represented as the circled marker. If a user inputs a threshold value in the "incident detection" and selects a comparison operator, the differences between the threshold and each sensor value are calculated and the corresponding marker(s) will blink.)




**(a)**


**(b)**

**Figure 8** **(a)** Registration interface for subscription of alert, and **(b)** an example of alert message with a sequence of sensing data information

### *4.2. Analytical Visualization for Sensing Data*

The past sensor data stored in the sensor database can be also analyzed and visualized subsequently on the world map using the analytical visualization function of 5D World Map System. The past sensor data is stored as CSV files.

When a user selects the category (topic) and/or the user, and the time period for comparison, each data are represented as a circled, resizable and colored marker. The size of marker from large to small and the color of marker from vivid to dark represent the relative value of each data in the set of values. In addition, when the user clicks the "time lapse" button and selects the granularity of the time by day/month/year, the change of geographical distribution of sensor data values with the lapse of time is shown as an animation.

**Figure 9** shows an example of selected data table for visualization. This shows the color of marker ranges from vivid to dark, and it represents the relative value of each data in the set of values - the depth of earthquakes, which occurred around the world during the period from Aug. 23th to Aug. 29th, 2014 [21]. **Figure 10** shows the visualization results of the data shown in Figure 9. When the user clicks the "time lapse" button and selects the granularity of the time by day, the time-series change of geographical distribution is shown as an animation.

View Data

Search:

| Filename | Id | Value | Date | Location | Lat,Lon | |
|---|---|---|---|---|---|---|
| arcgis-csv.csv | 181 | 0 | 2014-08-23 19:39:41 | 69km SE of Cordova, Alaska | 60.0657,-144.95 | ☑ |
| arcgis-csv.csv | 45 | 0 | 2014-08-27 11:04:19 | 67km ESE of Lakeview, Oregon | 41.8905,-119.6292 | ☑ |
| arcgis-csv.csv | 162 | 0 | 2014-08-24 10:34:55 | 67km ESE of Lakeview, Oregon | 41.8997,-119.6338 | ☑ |
| arcgis-csv.csv | 71 | 0.1 | 2014-08-26 16:24:53 | 83km ENE of Cantwell, Alaska | 63.5465,-147.3212 | ☑ |
| arcgis-csv.csv | 13 | 0.1 | 2014-08-28 18:19:20 | 97km NNW of Larsen Bay, Alaska | 58.2797,-154.8442 | ☑ |
| arcgis-csv.csv | 205 | 1 | 2014-08-23 02:50:13 | 176km NNE of Mayo, Canada | 64.9658,-134.083 | ☑ |
| arcgis-csv.csv | 81 | 10.34 | 2014-08-26 10:19:04 | 128km W of Hofn, Iceland | 64.4491,-17.8233 | ☑ |
| arcgis-csv.csv | 125 | 10.5 | 2014-08-25 04:24:44 | 6km SW of Napa, California | 38.2588,-122.3433 | ☑ |
| arcgis-csv.csv | 213 | 10.7 | 2014-08-22 23:13:42 | 67km NE of Kotzebue, Alaska | 67.3003,-161.4311 | ☑ |
| arcgis-csv.csv | 52 | 11.1 | 2014-08-27 06:35:52 | 3km W of American Canyon, California | 38.1803,-122.3035 | ☑ |
| arcgis-csv.csv | 126 | 11.3 | 2014-08-25 04:20:44 | 6km NW of American Canyon, California | 38.2202,-122.3128 | ☑ |
| arcgis-csv.csv | 232 | 11.93 | 2014-08-22 13:30:35 | 53km ESE of Maneadero, Mexico | 31.6186667,-116.0101667 | ☑ |
| arcgis-csv.csv | 145 | 169.5 | 2014-08-24 18:05:58 | 74km WNW of Sand Point, Alaska | 55.6589,-161.5392 | ☑ |
| arcgis-csv.csv | 123 | 198.88 | 2014-08-25 05:15:50 | 132km N of Lae, Papua New Guinea | -5.5236,147.0757 | ☑ |
| arcgis-csv.csv | 69 | 203.56 | 2014-08-26 16:48:10 | 151km E of Ndoi Island, Fiji | -20.6319,-177.2443 | ☑ |
| arcgis-csv.csv | 47 | 225.61 | 2014-08-27 07:59:34 | 113km ESE of San Pedro de Atacama, Chile | -23.219,-67.1626 | ☑ |
| arcgis-csv.csv | 82 | 440.41 | 2014-08-26 09:27:21 | Izu Islands, Japan region | 30.4076,138.2271 | ☑ |
| arcgis-csv.csv | 129 | 594.64 | 2014-08-25 03:20:27 | 300km SW of Ndoi Island, Fiji | -22.8235,179.5558 | ☑ |

Export

**Figure 9** Example of selected data table for visualization (The data of the depth of earthquakes around the world during the period from Aug. 23th to Aug. 29th, 2014 is shown. Data Source: USGS [21])

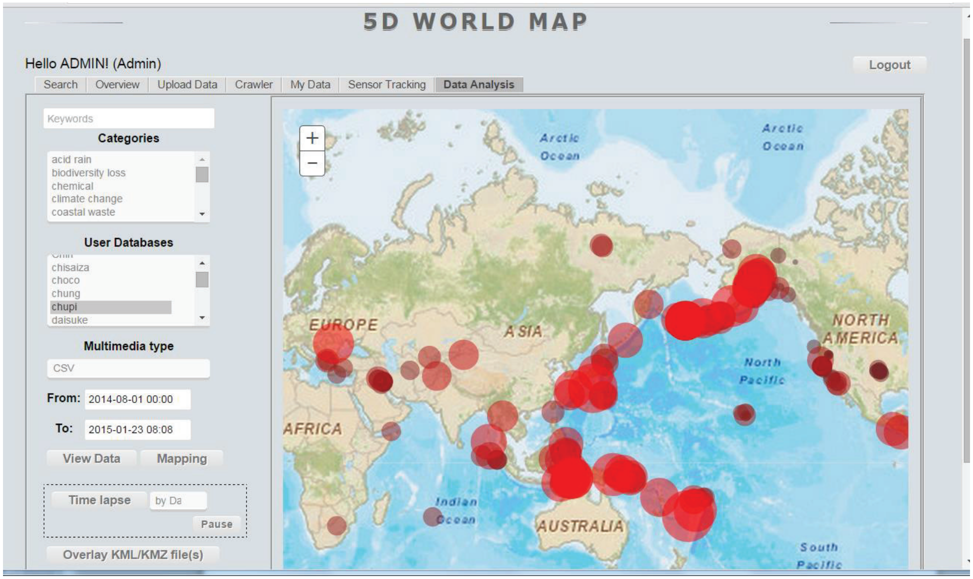


**Figure 10** Visualization of sensing data on the time-series world map set (The data of the depth of earthquakes around the world during the period from Aug. 23th to Aug. 29th, 2014 are shown. Data Source: USGS [21])

## 5. Experiments and Use Cases

### *5.1. Operation Experiments on Real-time Tracking and Differential Computing for Sensing Data*

To examine the functionality of Sensing-Processing-Actuation for sensing data of 5D World Map system, we performed several operation experiments using an implemented prototype system.

First, regarding real-time tracking for sensing data, **Figure 11** shows an example of sequence of each value of sensing data, sent from two sensors. This figure shows that the value of light of Sensor 1 changed when we illuminated a light to the Sensor 1, and the value of temperature of Sensor 1 had gone up when we set it on a heater, while the value of temperature of Sensor 2 had gone down when we set it in a fridge.

**Table 1** shows the log of real sensing data sent from sensing devices in three places – Tokyo (Japan), Bangkok (Thailand) and Surabaya (Indonesia). Based on this log data, we created a virtual data set for testing the operations described in Section 3. The created virtual data is shown in **Table 2**.

Because the differential calculation of temperature in remote countries seems not to be meaningful, we assume that this virtual data shows some kinds of values in emergency (e.g. water level of a river) in several locations in a specific area, for example, an upstream area [A], an midstream area [B] and an downstream area [C] of a river.

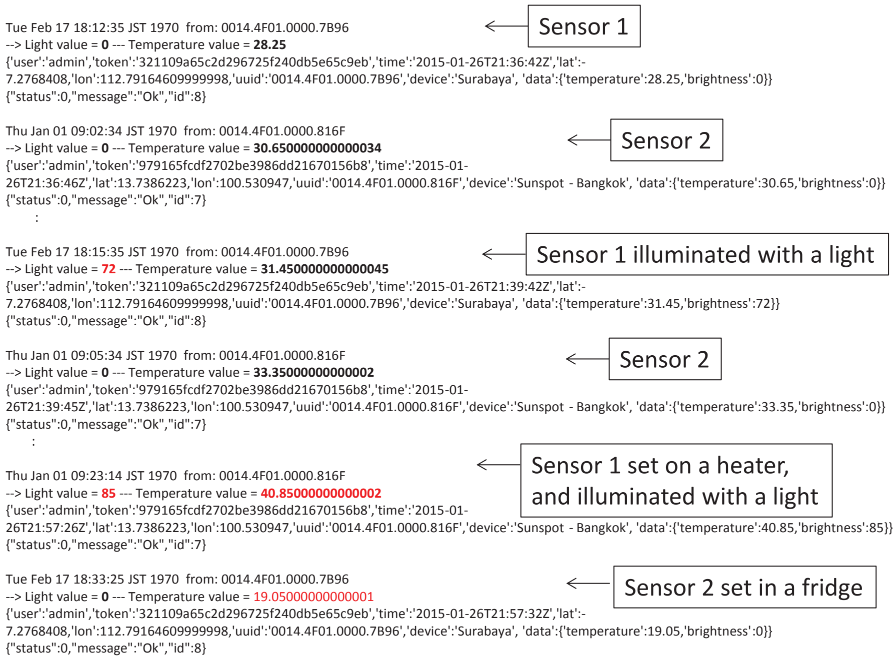



**Figure 11** Examples of sequence of sensing data sent from two sensing devices to the server by every three second (The timing of value-change detection of brightness, and the timing of value-change detection of temperature are indicated.)

**Table 1** Examples of real sensing data sent from sensing devices in multiple places

| Time | UUID | Location name | value |
|---|---|---|---|
| 2015-01-16T04:45:28.000Z | 0014.4F01.0000.80E9 | SunSpot - Tokyo | { ″temperature″ : 24.35, ″brightness″ : 53 } |
| 2015-01-16T04:45:29.000Z | 0014.4F01.0000.816F | Sunspot - Bangkok | { ″temperature″ : 29.05, ″brightness″ : 1 } |
| 2015-01-16T04:45:36.000Z | 0014.4F01.0000.7B96 | Sunspot - Surabaya | { ″temperature″ : 24.75, ″brightness″ : 0 } |
| 2015-01-16T04:45:38.000Z | 0014.4F01.0000.80E9 | SunSpot - Tokyo | { ″temperature″ : 24.15, ″brightness″ : 57 } |
| 2015-01-16T04:45:39.000Z | 0014.4F01.0000.816F | Sunspot - Bangkok | { ″temperature″ : 29.05, ″brightness″ : 1 } |
| 2015-01-16T04:45:46.000Z | 0014.4F01.0000.7B96 | Sunspot - Surabaya | { ″temperature″ : 24.55, ″brightness″ : 0 } |
| 2015-01-16T04:45:48.000Z | 0014.4F01.0000.80E9 | SunSpot - Tokyo | { ″temperature″ : 24.15, ″brightness″ : 55 } |
| 2015-01-16T04:45:49.000Z | 0014.4F01.0000.816F | Sunspot - Bangkok | { ″temperature″ : 29.05, ″brightness″ : 0 } |
| 2015-01-16T04:45:56.000Z | 0014.4F01.0000.7B96 | Sunspot - Surabaya | { ″temperature″ : 24.75, ″brightness″ : 0 } |
| 2015-01-16T04:45:58.000Z | 0014.4F01.0000.80E9 | SunSpot - Tokyo | { ″temperature″ : 24.15, ″brightness″ : 13 } |
| 2015-01-16T04:45:59.000Z | 0014.4F01.0000.816F | Sunspot - Bangkok | { ″temperature″ : 29.05, ″brightness″ : 0 } |
| 2015-01-16T04:46:06.000Z | 0014.4F01.0000.7B96 | Sunspot - Surabaya | { ″temperature″ : 30.45, ″brightness″ : 0 } |
| 2015-01-16T04:46:08.000Z | 0014.4F01.0000.80E9 | SunSpot - Tokyo | { ″temperature″ : 23.75, ″brightness″ : 15 } |
| 2015-01-16T04:46:09.000Z | 0014.4F01.0000.816F | Sunspot - Bangkok | { ″temperature″ : 29.25, ″brightness″ : 0 } |
| 2015-01-16T04:46:16.000Z | 0014.4F01.0000.7B96 | Sunspot - Surabaya | { ″temperature″ : 30.65, ″brightness″ : 0 } |
| 2015-01-16T04:46:18.000Z | 0014.4F01.0000.80E9 | SunSpot - Tokyo | { ″temperature″ : 23.75, ″brightness″ : 15 } |
| 2015-01-16T04:46:19.000Z | 0014.4F01.0000.816F | Sunspot - Bangkok | { ″temperature″ : 29.25, ″brightness″ : 1 } |
| 2015-01-16T04:46:26.000Z | 0014.4F01.0000.7B96 | Sunspot - Surabaya | { ″temperature″ : 30.45, ″brightness″ : 0 } |

**Table 2** shows an example of calculation result of difference by Differential Operation 1 (Op1). The dangerous level of value in place B is detected, based on the threshold = 30, while **Table 3** shows an example of calculation result of difference by Differential Operation 2 (Op2). In this calculation by Op2, when and how much the value changes drastically in each location are detected. **Table 4** shows an example of calculation result of difference by Differential Operation 3 (Op3). In this calculation,

when, between where and how much the value varies in this area are detected. Finally, **Table 5** shows an example of calculation result of difference by Differential Operation 4 (Op4). In this calculation, the time when the difference between place A and place B increases suddenly is detected. From this result, the possibility of occurrence of the similar phenomena between place B and place C in the next timing might be estimated.

**Table 2** Example of calculation result of difference by Differential Operation 1 (Op1) (The dangerous level of temperature in place B is detected, based on the threshold = 30.)

| | UUID | 0014.4F01.0000.80E9 | 0014.4F01.0000.7B96 | 0014.4F01.0000.816F |
|---|---|---|---|---|
| | Location name | SunSpot - Tokyo [A] | Sunspot - Surabaya [B] | Sunspot - Bangkok [C] |
| Time | *t*1 | 24.35 | 24.75 | 29.05 |
| | *t*2 | 24.15 | 24.55 | 29.05 |
| | *t*3 | 24.15 | 24.75 | 29.05 |
| | *t*4 | 24.15 | 30.45 | 29.05 |
| | *t*5 | 23.75 | 30.65 | 29.25 |
| | *t*6 | 23.75 | 30.45 | 29.25 |

**Table 3** Example of calculation result of difference by Differential Operation 2 (Op2): (When and how much the temperature changes drastically are detected.)

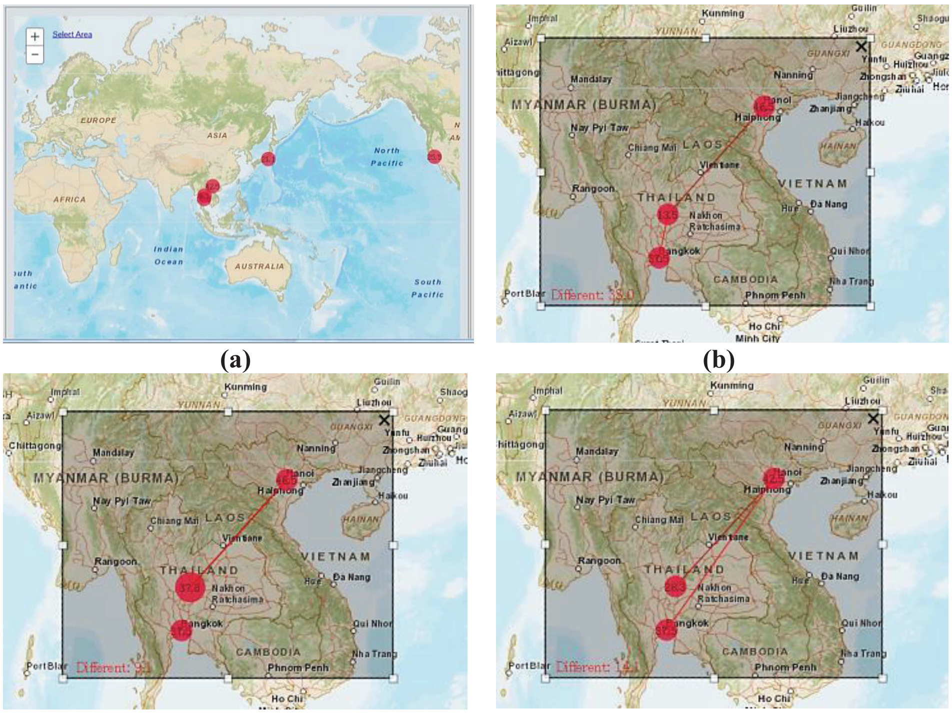
| | | Location | | |
|---|---|---|---|---|
| | | [A] | [B] | [C] |
| Difference | $\|v(t1)-v(t0)\|$ | 0 | 0 | 0 |
| | $\|v(t2)-v(t1)\|$ | 0.2 | 0.2 | 0.2 |
| | $\|v(t3)-v(t2)\|$ | 0.0 | 0.2 | 0.0 |
| | $\|v(t4)-v(t3)\|$ | 0.0 | 2.9 | 0.0 |
| | $\|v(t5)-v(t4)\|$ | 0.4 | 0.0 | 0.6 |
| | $\|v(t6)-v(t5)\|$ | 0.0 | 0.2 | 0.0 |

**Table 4** Example of calculation result of difference by Differential Operation 3 (Op3): (When, between where and how much the temperature varies in this area are detected.)

| | | Difference | | |
|---|---|---|---|---|
| | | $\|A-B\|$ | $\|B-C\|$ | $\|C-A\|$ |
| Time | *t*1 | 0.4 | 4.3 | 4.7 |
| | *t*2 | 0.4 | 4.5 | 4.9 |
| | *t*3 | 0.6 | 4.3 | 4.9 |
| | *t*4 | 6.3 | 1.4 | 4.9 |
| | *t*5 | 6.9 | 1.4 | 5.5 |
| | *t*6 | 6.7 | 1.2 | 5.5 |

**Table 5** Example of calculation result of difference by Differential Operation 4 (Op4): (The time when the difference between place A and place B increases suddenly is detected. The possibility of occurrence of the similar phenomena between place B and place C in the next timing can be estimated.)

| | | Difference | | |
|---|---|---|---|---|
| | | $\|A-B\|$ | $\|B-C\|$ | $\|C-A\|$ |
| Difference | $\|v(t1)-v(t0)\|$ | 0 | 0 | 0 |
| | $\|v(t2)-v(t1)\|$ | 0.0 | 0.2 | 0.2 |
| | $\|v(t3)-v(t2)\|$ | 0.2 | 0.2 | 0.0 |
| | $\|v(t4)-v(t3)\|$ | 5.7 | 2.9 | 0.0 |
| | $\|v(t5)-v(t4)\|$ | 0.6 | 0.0 | 0.6 |
| | $\|v(t6)-v(t5)\|$ | 0.2 | 0.2 | 0.0 |



(a) (b)

(c) (d)

**Figure 12** Example of visualization results of the difference calculation on the 5D World Map: (a) Result of real-time tracking, (b) selecting an area and difference calculation (c) and (d)

**Figure 12** shows examples of visualization results of the difference calculation on the 5D World Map. Figure 12(a) shows the result of real-time tracking, and Figure 12(b) shows the interactive interface to select an area and the difference calculation results by *Op2*. The difference between minimum value and maximum value is indicated in the lower left of the rectangle of selected area. Figure 12(c) shows the difference calculation results by *Op3* and *Op4*. The combination of two places where time-series difference happened drastically is shown by a line which connects the two places, and a place where the time-series difference happened incidentally is represented in a larger circled marker.

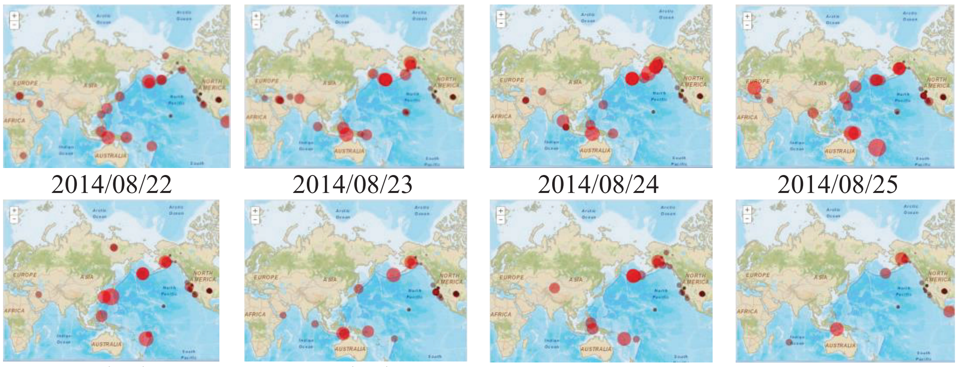
*5.2. Use cases of Analytical Visualization for Sensing Data*

*5.2.1 Earthquake analysis*

This case shows the analysis of the depth of earthquakes, which occurred around the world during the period from Aug. 23th to Aug. 29$^{th}$, 2014 [21]. The data table has been shown in Figure 9.

**Figure 13** shows the visualization results of the time-series change of geographical distribution of the data. From the results, we can observe intuitively that there is a point where deep earthquakes had happened through the whole period (eg.

Alaska), and there is a emergent timing that deep earthquakes happened in Fiji (2014/08/25) and consequently in Japan (2014/08/26)



2014/08/22 2014/08/23 2014/08/24 2014/08/25

2014/08/26 2014/08/27 2014/08/28 2014/08/29

**Figure 13** Time-series change of geographical distribution of the depth of earthquakes, which occurred around the world during the period from Aug. 23th to Aug. 29$^{th}$, 2014

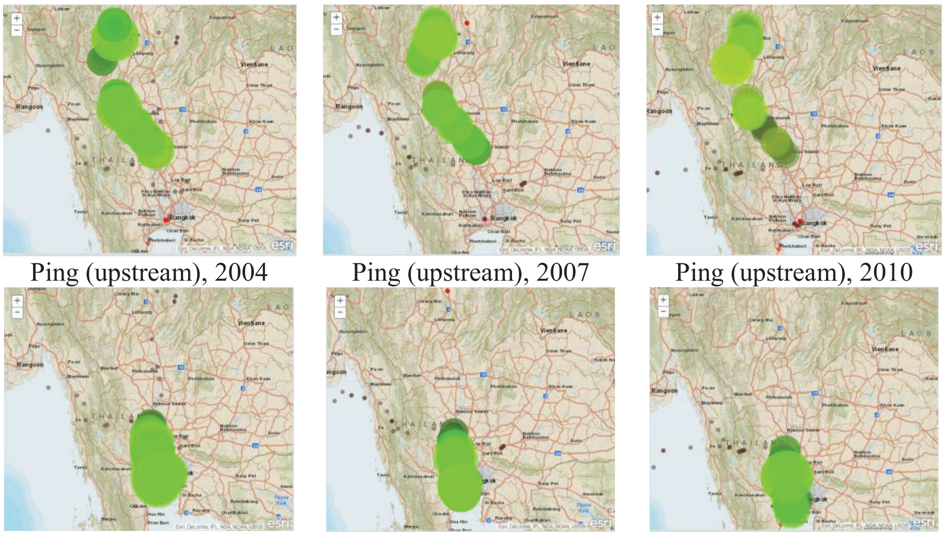
*5.2.2 Water Quality Analysis*

This case shows the analysis of water quality data in Thailand from 2004-2014 [22]. The study areas of water quality in river of Thailand are Ping River in the northern part, Chao Phraya River in the central part and Thajen River in the south-western part of central, which have several sampling points.

**Figure 14** shows the visualization results of three important indexes of water quality: Dissolve Oxygen (DO), Biological Oxygen Demand (BOD) and Total coliform bacteria, in 2004, 2007 and 2010. **Figure 14** shows the results when the index of Total coliform bacteria is focused.

**(a)** **(b)** **(c)**

**Figure 14** Visualization results of three important indexes of water quality in Thailand: **(a)** Dissolve Oxygen (DO), **(b)** Biological Oxygen Demand (BOD) and **(c)** Total coliform bacteria from 2004 - 2010



Ping (upstream), 2004 | Ping (upstream), 2007 | Ping (upstream), 2010

Thajern (downstream), 2004 | Thajern (downstream), 2007 | Thajern (downstream), 2010

**Figure 15** Visualization results of the index of Total coliform bacteria in Ping River (upstream) and Thajern River (downstream) in 2004, 2007 and 2010

From **Figure 16**, we can observe intuitively that the water pollution in a downstream area (Thajen River) is worse than an upstream are (Ping River). The cause can be presumed in a water cycle, which has an accumulation process from initial station, and in human activity beside the rivers. From the time line, we can observe that the water quality of Ping River (upstream) becomes better according to the lapse of time, while that of Thajern River (downstream) is not improved much, because of the increase of the development of domestic and industry in that area.

## 6. Conclusion

In this paper, we have presented real-time sensing, processing and actuating functions of a collaborative knowledge sharing system called 5D World Map System, and several applications in the field of multidisciplinary environmental research and education.

The proposed real-time sensing, processing and actuating functions enable multiple remote-users to acquire real-time sensing data from multiple sites around the world, perform analytical visualizations of the acquired sensing data by the selected calculation methods on the system to discover the incidental phenomena, and receive the analysed results by their own terminal equipment automatically.

By using 5D World Map System with multimedia integrated with these functions, the users are able to perform a global analysis on the environmental sensing data along with the related multimedia data on a single view of time-series maps, based on the spatiotemporal and semantic correlation calculations. The combination of sensing data analysis and multimedia search and mapping enables more effective and deep environmental analysis.

**Acknowledgment**
This research is partially supported by Rachadapisek Sompote Fund for Postdoctoral Fellowship, Chulalongkorn University. Also, I would like to express my gratitude for Ms. Diep Thi-Ngoc Nguen on her support in the implementation of the system.

## References

[1] Edward A. Lee, "Cyber Physical Systems: Design Challenges," Object Oriented Real-Time Distributed Computing (ISORC), 11th IEEE International Symposium on 5-7 May, 2008, pp. 363-369.
[2] Edward A. Lee, "Cyber-physical systems - are computing foundations adequate?," Position Paper for NSF Workshop On Cyber-Physical Systems: Research Motivation, Techniques and Roadmap, October 16-17, 2006.
[3] Ragunathan (Raj) Rajkumar, Insup Lee, Lui Sha, John Stankovic, "Cyber-Physical Systems: The Next Computing Revolution", DAC '10 Proceedings of the 47th Design Automation Conference, ACM New York, NY, USA, 2010, Pages 731-736, 2010.
[4] Kiyoki, Yasushi, and Xing Chen. "Contextual and Differential Computing for the Multi-Dimensional World Map with Context-Specific Spatial-Temporal and Semantic Axes." Information Modelling and Knowledge Bases XXV 260 (2014): 82.
[5] Sasaki, S., Takahashi, Y, Kiyoki, Y., "The 4D World Map System with Semantic and Spatiotemporal Analyzers, " Information Modelling and Knowledge Bases, Vol.XXI, IOS Press, pp. 1 - 18, 2010.
[6] Kiyoki, Y., Kitagawa, T., A semantic associative search method for knowledge acquisition, Information Modelling and Knowledge Bases VI, H. Kangassalo et al eds. IOS Press 1995 pp. 121-130.
[7] Kiyoki, Y., Sasaki, S., Nguyen N. T., Nguyen, T. N. D., "Cross-cultural Multimedia Computing with Impression-based Semantic Spaces", Conceptual Modelling and Its Theoretical Foundations, Lecture Notes in Computer Science, Springer, pp.316-328, March 2012.
[8] Nguyen, D. T. N., Sasaki, S. and Kiyoki,Y., "5D World PicMap: Imagination-based Image Search System with Spatiotemporal Analyzers," Proceedings of The IADIS e-Society 2011 (ES 2011), Avila, Spain, pp.272 - 278, March 9th - 13th, 2011.
[9] Mori, M., Sasaki, S. and Kiyoki, Y., "A Bird-World Image-Map System with Spatiotemporal and Color Analysis Functions," 10th International Conference on IEEE ICT and Knowledge Engineering (ICT-KE 2012), Siam University, Bangkok, Thailand, pp.104 - 111, 21 Nov - 23 Nov 2012.
[10] Itabashi, Y., Sasaki, S., and Kiyoki, Y., "Distinctive-Color Analytical Visualization for Cross-Cultural Image Computing with 5D World Map," Knowledge Creation & Intelligent Computing 2013, South Bali, Indonesia, pp.37-44, March 2013.
[11] Fujioka, H., Sasaki, S., and Kiyoki, Y., "An Analytical Visualization System of Historical, Geographical and International Trading Information for 5D World Map," Knowledge Creation & Intelligent Computing 2013, South Bali, Indonesia, pp.51-58, March 2013.
[12] Komori, Y., Sasaki, S., and Kiyoki, Y., "Descriptive Bibliography Search System with Historical Geographic 5D World Map Visualization," Knowledge Creation & Intelligent Computing 2013, South Bali, Indonesia, pp.132-139, March 2013.
[13] Amane KOBAYASHI, Shiori SASAKI, Yasushi KIYOKI, An Analytical Visualization of Global Political Effects with 5D World Map System: Cases of Anti-governmental Movement to Regional Upheaval in the Middle East, The Third Indonesian-Japanese Conference on Knowledge Creation and Intelligent Computing (KCIC) 2014, Malang, Indonesia, accepted 8 pages, March 25-26, 2014.
[14] Wahjoe Tjatur SESULIHATIEN, Shiori SASAKI, Yasushi KIYOKI, Building Public E-Health with 5D World Map System (EH5D) : Awareness Scenario of Dengue Spreading in Surabaya, The Third Indonesian-Japanese Conference on Knowledge Creation and Intelligent Computing (KCIC) 2014, Malang, Indonesia, accepted 8 pages, March 25-26, 2014.
[15] Kang, S., Kobayashi, A., Tomobe, K., Hashim, I. H. M., Sasaki, S., Morita, K., "Effects of Renewable Energy Policies in Asia-Pacific: Application of a 5D World Map System," 2014 Biennial Conference on Sustainable Business, Energy and Development in Asia (COSA2014), Hiroshima, Japan, March 17-19, 2014.
[16] http://www.mongodb.org/
[17] http://www.sunspotworld.com/products/
[18] http://www.sunspotworld.com/SPOTManager/
[19] http://www.oracle.com/technetwork/java/javase/overview/index.html
[20] http://www.json.org/
[21] Data source: USGS, http://earthquake.usgs.gov/earthquakes/feed/v1.0/
[22] Data source: the Metropolitan Waterworks Authority and Pollution Control Department, Ministry of Natural Resources and Environment in Thailand, http://www.pcd.go.th

# SPARQL-based framework for semantically-based event processing

Ana SASA BASTINOS[a,1], Dejan LAVBIC[b]
[a]*fluidOperations fluid Operations AG, Walldorf, Germany*
[b]*Faculty of Computer and Information Science, University of Ljubljana, Slovenia*

**Abstract.** Event-driven architecture and complex event processing have become important topics in achieving business reactivity and proactivity. Even though many technologies to support this have been developed, there is still a need for providing support for complex events based on different levels of required expressivity. On the other hand, semantic technologies and semantic-based systems have emerged and are becoming more and more used, whereas there is no recognized solution for event processing in information systems based on semantic technologies. In this paper, we address these issues. We present a framework for ontology-based support for complex events, which allows for semantically enriched event definitions and automatic recognition. In order to automatically recognize complex events, an appropriate event definition basis has to be defined. The paper represents a continuation of our previous work by enhancing this basis and developing a SPARQL-based framework to provide a richer mechanism for event definitions and more efficient event detection. As a proof-of-concept, we provide an implementation of this framework using the semantic integration platform - Information Workbench. The implementation is available as an app that runs on top of the Information Workbench platform.

**Keywords.** Event processing, complex event, ontology, SPARQL, RDF, OWL.

## 1. Introduction

With increasing demands for agility and reactivity of business processes, scientific circles and leading information technology companies have paid a lot of attention to EDA (event-driven architecture) and CEP (complex event processing). EDA and CEP enable event-driven systems - systems in which actions result from business events [1]. Even though many technologies to support this have been developed [2], there is still no recognized approach available that would provide semantically-enriched support for complex event processing based purely on standard semantic technologies. Existing EDA and CEP approaches do not take into consideration different expressivity requirements that are needed in definition of a large number of diverse complex events. Because of this, detection of complex events that require semantically expressive descriptions is not automated and experts are required to monitor business operation in order to determine if a complex event has occurred. Such an approach can decrease reactiveness and proactiveness of an organization [3][4].

We addressed some of these issues in our previous work presented in [3], where we discussed ontology-based framework for complex events, and later in [4] by

[1] Corresponding Author.

enhancing that basis and applying the aspectual model from the field of linguistics to our event ontology. In this paper, we take an alternative approach to defining the event ontology and defining event conditions. We do not use the Semantic Web Rule Language (SWRL) in order to construct rules that define when an event occurs, as we did in our previous work. Instead, we use SPARQL-based approach, which makes our framework more widely applicable to a wide range of available SPARQL-based tools and systems. Furthermore, this approach is not limited only to OWL-based systems, but can be used with any RDF-based system.

## 2. Background and related work

### *2.1. Events, event-driven architecture and complex event processing*

The business operations of today's enterprises are heavily influenced by numerous of internal and external business events. With the Event Driven Architecture and particularly the Complex Event Processing (CEP), many research activities have been dedicated to technologies required for identifying complex correlations in these large amounts of event data right after its appearance [2].

An event is anything that happens, or is contemplated as happening [5]. An entity is considered event-driven when it acts in response to an event. EDA is a paradigm that describes an approach to information systems development with a focus on developing an architecture that has the ability to detect events and react intelligently to them [6]. EDA represents a complement to the service-oriented architecture (SOA), which has become one of the most recognized paradigms in information systems development in the recent years [7]. By enhancing the paradigm of SOA, enterprises can gain improve their ability for business transformation by implementing event-driven architectures that automatically detect and react to significant business events [6][8]. An important part of every EDA that enables and predetermines to what level a system is able to detect and respond to complex events is CEP. CEP refers to computing that performs operations on complex events, including reading, creating, transforming, abstracting, or discarding them [5].

The event-related terminology that we use in this paper, complies to the Event Processing Glossary – Version 2.0 [5]. We use the following terms from the glossary:

- Event (event object): An object that represents, encodes, or records an event, generally for the purpose of computer processing.
- Event type: A class of event objects.
- Simple event: An event that is not viewed as summarizing, representing, or denoting a set of other events.
- Complex event: An event that summarizes, represents, or denotes a set of other events.
- Derived event: An event that is generated as a result of applying a method or process to one or more other events.

### *2.2. Ontologies*

An ontology represents an explicit specification of a conceptualization [9]. There are several languages available for ontology representation, such as DAML, CGs, OIL, DAML+OIL, OWL (OWL 1, OWL 2). To support definition and processing of events, our work is based on OWL 2 [10]. The reasons for this are a rich and useful group of ontology features for defining and describing the domain, a high level of support, and its XML foundations, which make it appropriate to be used in conjunction with other Web technologies. The main concepts of OWL are Classes, Properties, and Individuals [10]. This means that besides an ontology that provides a conceptual representation, an OWL document can also comprise instance level descriptions of an enterprise. The main types of OWL properties are object properties and data properties. An object property relates an instance to another instance, while a data property relates an instance to a literal.

Resource Description Framework (RDF) is a directed, labeled graph data format for representing information in the Web. RDF data model has a form of subject–predicate–object expressions. RDF expressions are known as triples in the RDF terminology. A common way to persist RDF data in its native representation is in a triple store.

OWL 2 ontologies can be used along with information written in RDF, and OWL 2 ontologies themselves are primarily exchanged as RDF documents [10]. In order to define the event types, we use the SPARQL 1.1 query language [11]. SPARQL (SPARQL Protocol and RDF Query Language) is a set of specifications that provide languages and protocols to query and manipulate graph content on the Web or in an RDF store.

### *2.3. EDA and CEP approaches based on semantic technologies*

The use of ontologies in the field of EDA and CEP has been identified as a suitable approach for the semantic definition of events by many researchers. Cheng et al. [12] have proposed a framework for context-aware processing of business rules in event-driven architectures. They have developed a context ontology in order to resolve a problem of inconsistent dictionaries at knowledge sharing and merging of rules. A key difference between their approach and ours is that they focus on achieving semantic interoperability, whereas our goal is to use ontology to define and detect complex events. Sen and Ma [13] have proposed an approach for combining reactive rules with ontologies. They have used ontologies to capture the context in which certain behaviour is appropriate, together with techniques for finding similarities with the primary objective to enable detection of similar complex event patterns. Their research is related to our approach with the key difference that they do not talk about how complex event types can be defined by the ontology, and how they can be used for detection in the context of EDA. Moser et al. [14] have proposed semantic correlation of events by using ontologies. Their work is based on the observation that the event correlation is necessary for CEP to relate events obtained from various sources in order to detect patterns and situations of the business context. Paschke [15][16] has proposed a language for semantic design of CEP patterns and a Web infrastructure for distributed rule-based event processing multi-agent eco-systems. Adaikkalavan and Chakravarthy [17] propose an interval-based event specification language developed by expressing simple and composite events that are part of active rules. By contrast with these two approaches, our approach builds on standard semantic languages, does not define a specific event-processing language, and allows using available query engines and tools that support these standards. Several authors propose domain-specific definition and detection of events by using ontologies, such as [18],[19],[20], and [21]. On the other hand, our work is generic and not domain specific. Our study of related work has shown that there is no generic complex-event processing approach available to be applied to information systems based on semantic technologies based purely on standard semantic languages and protocols.

## 3. Overview of the SPARQL-based framework for event processing

In order to enable CE definitions, detection and triggering based on ontologies and the OWL language, we have developed an event-driven framework based on the publish-subscribe model. It is illustrated in Fig. 1.

The core part of the framework is the RDF database containing the ontologies and instance data. The ontology is composed of two parts: the base event ontology, and the domain ontology. The base event ontology is a generic part of the framework that can be reused in different domains and environments. We have developed it with the purpose of providing the common basis for event types and events. It provides classes and relationships for event instances (events) and event types. More details on the event ontology can be found in section 4.

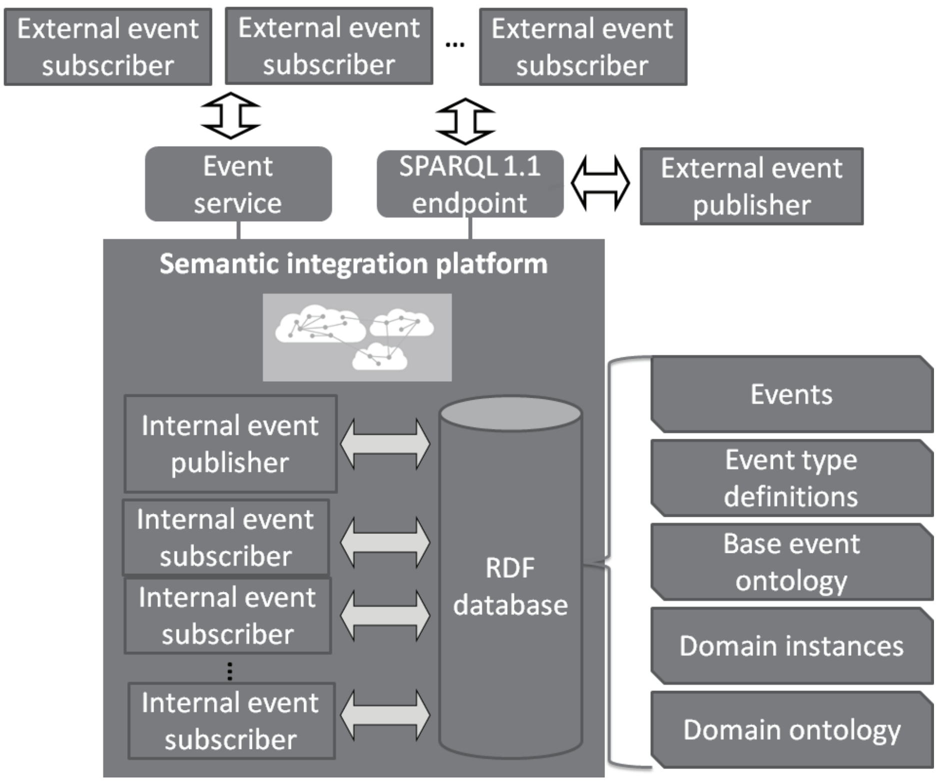



**Figure 1.** Overview of the SPARQL-based framework for event processing.

Domain ontology is the environment specific ontology that is used in event types to specify domain-specific conditions for event triggering. Any OWL-based or RDF-based domain ontology can be used in this place.

Both base event and domain ontology are persisted in the RDF database. The RDF database also contains information about event types, event instances, and domain instances.

Producing the events and informing other components about the event that interests them is based on the publish-subscribe model. The event publisher detects when an event occurs based on the event type definition. If an event is detected, it publishes it into the RDF database as a new event together with its semantic properties. The event publisher can be a component internal to the semantic integration system (internal event publisher), or an external component (external event publisher). An external event publisher uses the SPARQL 1.1. endpoint in order to get the information from the RDF database and to publish events. In order to publish events, the external event publisher uses the CONSTRUCT SPARQL 1.1. query types.

Event subscribers can subscribe to events by specifying the event type(s) they are interested in. They can find the events invoking a parametrized predefined SPARQL query on the database, where the parameter is the event type they are interested in. Similarly, as in case of the publisher, a subscriber can be a component that is either internal or external to the system. External subscribers can use the provided SPARQL 1.1 endpoint or a customized endpoint (event service) to retrieve events based on the event type.

The publish-subscribe model used in our approach can be compared to the JMS publish-subscribe model, where the publisher publishes messages to a topic. Instead of publishing messages to a topic, in our case a new event is created of a given event type. Based on the event type, the subscriber can always find the events of the type they are interested in. In this comparison, an event type plays the role similar to the role of a topic in the JMS publish-subscribe model.

## 4. Base event ontology and event type definitions

Base event ontology defines classes and relationships that provide the basis for defining, detecting and publishing of events (Fig. 2). It is implemented using OWL 2 DL ontology language.

Event ontology distinguishes between the *Event* class and the *Event Type* class. Members of the *Event Type* class represent event definitions which contain information about when an event of the type occurs. Every event type carries information that uniquely defines when an event of this type occurs. On the other hand, members of the Event class are event instances with concrete information about an event that occurred. An instance of the Event class is related to its event type via *Has event type* functional object property.

An event type defines an event expression via the *Has event condition expression* property. Event expressions are the key part of our support for CEP. They represent the semantical expression that uniquely defines when an event occurs. They should be written in an agreed upon RDF-based or OWL-DL syntax. In our study, we use two types of SPARQL 1.1 query forms, i.e. the SELECT query form, and the CONSTRUCT query form. A SELECT SPARQL query returns all, or a subset of, the variables bound in a query pattern match. For the event publisher, its result indicates whether an event of the given type has occurred. If the result is an empty set, then the event did not occur. A CONSTRUCT QUERY form returns a graph, which the provider adds into the triple store, provided that the statements resulting from the query do not yet exist. The SELECT SPARQL query is used to define simple and underived complex events, while the CONSTRUCT query form is used to represent derived events.

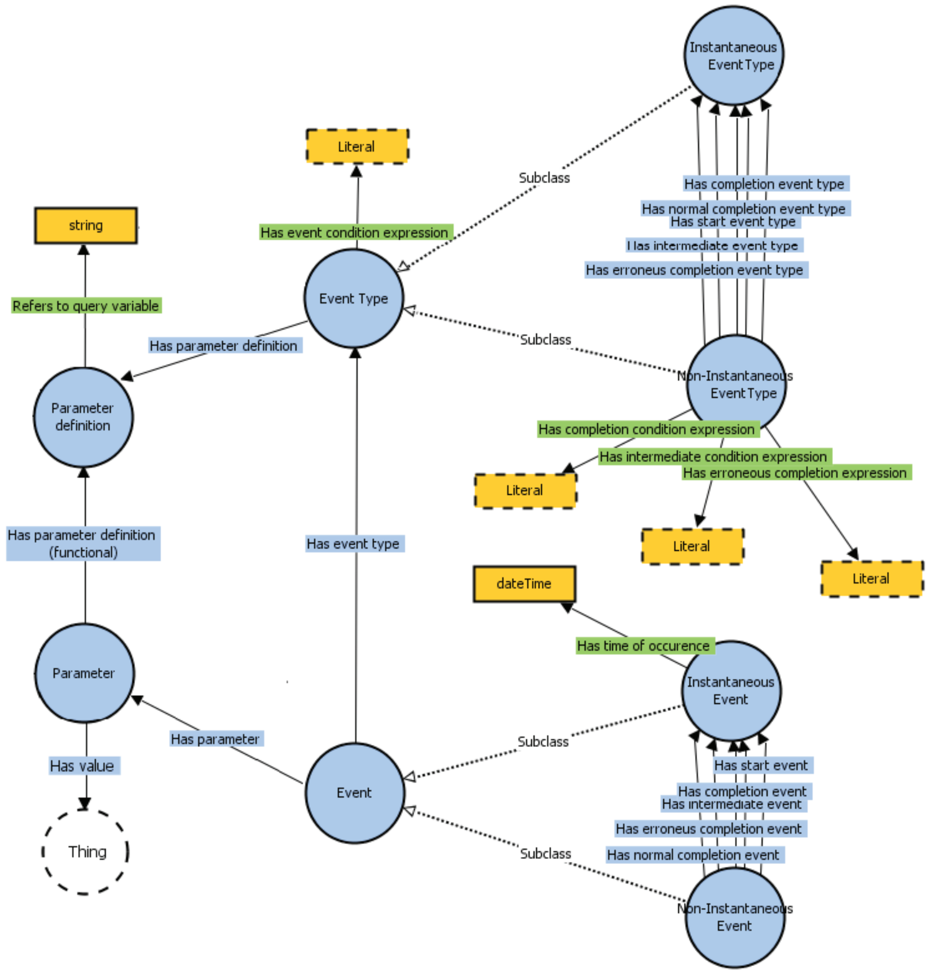



**Figure 2.** Base event ontology overview.

The condition defining when an event occurs may be true for a longer period of time, even though it actually refers to only one event. For example, the condition for an event type “Out of credit” could be that the credit balance is 0. However, this event should only be triggered once during the time when the credit balance is 0. For this reason, we may need to uniquely identify an occurrence of an event in a way that would prevent incorrect publishing of more than one event instance. For this reason, an additional check that verifies the uniqueness of the event must be included in the conditional event expression for such cases. In the given example, the unique identifier of the event would be the time when the user runs out of credit and the unique user name. Therefore, the event condition expression must also include the check if the event with such a unique identifier already exists. If it does not, a new event will not be published.

*4.1. Instantaneous events vs. non-instantaneous events*

In order to support different events, we distinguish between two basic types of events:

- instantaneous events, which are characterized by a point in time when they occur, and
- non-instantaneous events, which have a duration.

To reflect this in the ontology, the *Event type* class has two subclasses: *Instantaneous Event Type* and *Non-Instantaneous Event Type*. Correspondingly, Event class has two corresponding subclasses *Instantaneous Event* and *Non-Instantaneous Event.*

In comparison with an instantaneous event, a non-instantaneous event has a duration. About non-instantaneous events, one can ask a question in a continuous tense, such as “is it happening?”. This is not relevant for instantaneous events.

The conceptualization of the non-instantaneous event type, allows us to treat it in a different way that instantaneous events, which is particularly important for the event subscribers. An event subscriber may be interested when a non-instantaneous event starts, when it ends, whether it ends with one or another event type, and may also be interested in certain intermediate events that can occur between the start and the end if this event. In order to support different types of end events, the object property *Has completion event type* has two subproperties: *Has normal completion event type*, *Has erroneous completion event type*. Furthermore, one can define intermediate events using the property *Has intermediate event type*.

One can observe that the non-instantaneous event is actually a specific form of a complex event, since it is defined by related occurrence of one or more instantaneous events. An alternative approach would be to represent the non-instantaneous events by two or more separate events using the instantaneous event concept, for example by defining a separate event type designating its start, and by defining its dependent intermediate and end event types. However, with such approach the subscriber would have to be aware of all the possible subevents and subscribe to all of them. In our approach, we include the non-instantaneous event as its own concept in order to simplify its definition and implementation, especially compared to managing several separate event types that really represent one complex event. It is a very common type of an event that is widely applicable. Because of the conceptual model that supports non-instantaneous event types, implementation of the mechanism that allows sending the corresponding information to the subscribers is very straightforward.

*4.2. Parameter definitions and parameters*

Besides the event expression, an event type has one or more parameter definitions. A parameter definition defines which semantic properties an event has. When an event occurs, it can carry some information. An event instance is related to this information via its *Has parameter* object property. Every parameter points to exactly one parameter definition. As our approach is based on semantic technologies, the parameters are also part of the semantic world. Each parameter has a value that is either an IRI of an OWL class or a data type. An event type can have zero, one or more parameter definitions. If it has more than one, this means that the event can be related to different instances of different OWL classes or different data values with their own meaning. Event publisher retrieves the information about the event parameters from the queries using the property *Refers to query variable*, which maps the parameter definition to the query variable. In this way, the publisher has a uniquely identifiable value that it can assign to each event parameter.

## 5. Example

In order to illustrate the concepts presented in the previous section, this section provides example event type definitions and parameter definitions. Example domain of user account credit balance is chosen, because it is easy to understand while it provides a complex-enough basis to show the different concepts introduced in this paper.

A very common event type in the example domain is the *Credit change event type*. Its OWL definition written in the turtle format is given in Table 1. In this and the following examples, the namespace prefixes *action* and *credit* stand for namespaces from the domain ontology, whereas the namespace prefixes *event* and *eventIns* refer to the event ontology namespace and event instance namespace, respectively.

**Table 1.** Credit change event type

```
eventIns:CreditChangeEventType rdf:type event:EventType, event:InstantaneousEventType ;
                        rdfs:label "Credit change event type" ;
                        event:hasEventConditionExpression """

SELECT distinct ?newCreditBalance ?oldCreditBalance ?time1 ?timeOfLastChange ?user WHERE {
  ?user a <http://www.fluidops.com/User>.
?record a action:Record .
?record credit:refersToUser ?user.
?record action:startDate ?timeOfLastChange .
?record action:createdBy action:CreditBalanceChange.
?record credit:creditBalance ?newCreditBalance .

?record1 a action:Record .
?record1 credit:refersToUser ?user.
?record1 action:startDate ?time1 .
?record1 action:createdBy action:CreditBalanceChange.
?record1 credit:creditBalance ?oldCreditBalance .

FILTER NOT EXISTS {
               ?record2 a action:Record .
               ?record2 credit:refersToUser ?user.
               ?record2 action:startDate ?time2 .
               ?record2 action:createdBy action:CreditBalanceChange.
               FILTER (?timeOfLastChange > ?time2 && ?time1 < ?time2 ).
}

FILTER (?oldCreditBalance != ?newCreditBalance && ?timeOfLastChange > ?time1) .

FILTER NOT EXISTS {
               ?event a event:Event .
               ?event event:hasEventType eventIns:CreditChangeEventType.
               ?event event:hasParameter ?param .
               ?param event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
               ?param event:hasValue ?timeOfLastChange.
               ?event event:hasParameter ?userParam .
```

```
            ?userParam event:hasParameterDefinition eventIns:userParameterDefinition.
            ?userParam event:hasValue ?user.
}}
                    """ ;
                    event:hasParameterDefinition eventIns:newCreditBalanceParameterDefinition ,
                                        eventIns:oldCreditBalanceParameterDefinition ,
                                        eventIns:timeOfLastChangeParameterDefinition ,
                                        eventIns:userParameterDefinition .
```

The event expression of the *Credit change event type* defines that the event of this type occurs, when two credit records exist with a different credit balance, both of the action *CreditBalanceChange*, and there is no other record of this action that was recorded between their action times. To ensure uniqueness of the event, the second "FILTER NOT EXISTS" block of the query is used to determine that the same user does not already have a *Credit change event type* with the same time stamp.

One can observe that the Credit change event type points to four parameter definitions, which indicates that each event of this type will have exactly four parameters. An example of the *New credit balance parameter definition* is shown in Table 2. The relationship between the variable *newCreditBalance* in the SELECT SPARQL query of the event type definition and the parameter definition is given with the property *Refers to query variable* in the parameter definition.

**Table 2.** New credit balance parameter definition

```
eventIns:newCreditBalanceParameterDefinition rdf:type event:ParameterDefinition ;
               rdfs:label "New credit balance parameter definition";
               event:refersToQueryVariable "newCreditBalance"^^xsd:string .
```

Table 3 shows an example of a derived event type definition. *Out of credit event type* can be derived from the *Credit change event type*. The corresponding event condition expression of this event type is a CONSTRUCT SPARQL query, which searches for events of type *Credit change event type* with the new credit balance of 0. If there is such a credit change event type, the event publisher will add a statement that this event is also of the *Out of credit event type*, unless such a statement already exists.

**Table 3.** Out of credit event type

```
eventIns:ZeroCreditEventType rdf:type event:InstantaneousEventType, event:EventType ;
                    rdfs:label "Out of credit event" ;
                    event:hasEventConditionExpression """
CONSTRUCT  {
            ?event event:hasEventType eventIns:ZeroCreditEventType .
} WHERE              {
            ?event a event:Event .
            ?event event:hasEventType eventIns:CreditChangeEventType .
            ?event event:hasParameter ?param .
            ?param event:hasParameterDefinition eventIns:newCreditBalanceParameterDefinition.
            ?param event:hasValue ?newCreditBalance .
  FILTER(?newCreditBalance = 0).
}
"""^^xsd:string .
```

Table 4 represents an example of a non-instantaneous event type, i.e. the *Subscription renewal event type*. An event of this type starts when an *Order credit request event* occurs. An *Order credit request event* is a derived event of the *Out of credit event*, which happens if the *Out of credit event* occurs when the user has automatic subscription renewal turned on (Table 5). The Subscription renewal event type completes when a *Positive credit change event* occurs after the start event and the event completion condition returns a result. The completion condition expression also relates the start and the completion events.

**Table 4.** Subscription renewal event type

```
eventIns:SubscriptonRenewalEventType rdf:type event:NonInstantaneousEventType;
                          rdfs:label "Subscription renewal event type" ;
                          event:hasCompletionEventType eventIns:PositiveCreditChangeEventType ;
                          event:hasStartEventType eventIns:OrderCreditRequestEventType ;
                          event:hasCompletionConditionExpression
""""
SELECT distinct ?nonInstEvent ?startEvent ?completionEvent WHERE {
  ?startEvent a event:Event .
  ?startEvent event:hasEventType eventIns:OrderCreditRequestEventType .
  ?startEvent event:hasParameter ?userParam1 .
  ?userParam1 event:hasParameterDefinition eventIns:userParameterDefinition.
  ?userParam1 event:hasValue ?user.
  ?startEvent event:hasParameter ?lastChange1 .
  ?lastChange1 event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
  ?lastChange1 event:hasValue ?time1.

  ?nonInstEvent event:hasStartEvent ?startEvent .
  ?nonInstEvent event:hasEventType eventIns:SubscriptonRenewalEventType .

  ?completionEvent a event:Event .
  ?completionEvent event:hasEventType eventIns:PositiveCreditChangeEventType .
  ?completionEvent event:hasParameter ?userParam2 .
  ?userParam2 event:hasParameterDefinition eventIns:userParameterDefinition.
  ?userParam2 event:hasValue ?user.
  ?completionEvent event:hasParameter ?lastChange2 .
  ?lastChange2 event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
  ?lastChange2 event:hasValue ?time2.

  FILTER(?time2 > ?time1) .

  FILTER NOT EXISTS {
            ?startEvent2 a event:Event .
            ?startEvent2 event:hasEventType eventIns:OrderCreditRequestEventType .
   ?startEvent2 event:hasParameter ?userParamS2 .
   ?userParamS2 event:hasParameterDefinition eventIns:userParameterDefinition.
   ?userParamS2 event:hasValue ?user.
   ?startEvent2 event:hasParameter ?lastChangeS2 .
   ?lastChangeS2 event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
   ?lastChangeS2 event:hasValue ?timeS2.
            FILTER(?timeS2 < ?time2 && ?timeS2 > ?time1) .
  }

  FILTER NOT EXISTS {
            ?event3 a event:Event .
            ?event3 event:hasEventType eventIns:PositiveCreditChangeEventType.
            ?event3 event:hasParameter ?userParam3 .
            ?userParam3 event:hasParameterDefinition event:userParameterDefinition.
            ?userParam3 event:hasValue ?user.
            ?event3 event:hasParameter ?lastChange3 .
            ?lastChange3 event:hasParameterDefinition event:timeOfLastChangeParameterDefinition.
            ?lastChange3 event:hasValue ?time3.
            FILTER(?time3 < ?time2 && ?time3 > ?time1) .
            }

   FILTER NOT EXISTS {
                          ?nonInstEvent           event:hasCompletionEvent ?completionEventX .
   }
```

```
}
"""^^xsd:string .
```

**Table 5.** Order credit request event type

```
eventIns:OrderCreditRequestEventType rdf:type event:InstantaneousEventType, event:EventType;
                        rdfs:label "Order credit request event type" ;
                        event:hasEventConditionExpression
"""
CONSTRUCT   {
?event event:hasEventType eventIns:OrderCreditRequestEventType .
} WHERE {
?event a event:Event .
?event event:hasEventType eventIns:ZeroCreditEventType .
?event event:hasParameter ?param .
?param event:hasParameterDefinition eventIns:userParameterDefinition.
?param event:hasValue ?user.
?user credit:hasAutomaticSubscriptionRenewal ?renewal.
FILTER(?renewal = xsd:boolean('true')).
}
"""^^xsd:string .
```

Using the proposed approach, a number of different event types of various complexities can be defined, while basing their event expressions either on the domain information, either on the event information, or both. A complex event from the example domain is an event that is published if the user runs out of credit three times within the last five days. If that happens, we may want to offer the user a subscription or an option of automatic subscription renewal. The event expression for such an event is given in Table 6.

**Table 6.** Repeated Zero Credit Event Type

```
eventIns:RepeatedZeroCreditEventType rdf:type event:EventType, event:InstantaneousEventType ;

                        rdfs:label "Repeated zero credit event type" ;

                        event:hasEventConditionExpression """
CONSTRUCT   {
            ?e3 event:hasEventType eventIns:RepeatedZeroCreditEventType .
} WHERE {
            ?user a <http://www.fluidops.com/User>.
            ?e1 a event:Event.
            ?e1 event:hasEventType eventIns:ZeroCreditEventType.
            ?e2 a event:Event.
            ?e2 event:hasEventType eventIns:ZeroCreditEventType.
            ?e3 a event:Event.
            ?e3 event:hasEventType eventIns:ZeroCreditEventType.
            FILTER(?e1 != ?e2 && ?e2 != ?e3).
            ?e1 event:hasParameter ?userParam1 .
            ?userParam1 event:hasParameterDefinition eventIns:userParameterDefinition.
            ?userParam1 event:hasValue ?user.

            ?e2 event:hasParameter ?userParam2 .
            ?userParam2 event:hasParameterDefinition eventIns:userParameterDefinition.
            ?userParam2 event:hasValue ?user.
            ?e3 event:hasParameter ?userParam3 .
            ?userParam3 event:hasParameterDefinition eventIns:userParameterDefinition.
            ?userParam3 event:hasValue ?user.
            ?e1 event:hasParameter ?timeParam1 .
```

```
        ?timeParam1 event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
        ?timeParam1 event:hasValue ?time1 .
        ?e2 event:hasParameter ?timeParam2 .
        ?timeParam2 event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
        ?timeParam2 event:hasValue ?time2 .
        ?e3 event:hasParameter ?timeParam3 .
        ?timeParam3 event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
        ?timeParam3 event:hasValue ?time3 .
        BIND (<http://www.fluidops.com/service/timestamp>() AS ?now) .
        BIND (<http://www.fluidops.com/service/timestamp>(?time1) AS ?time1TS) .
        FILTER(?now - ?time1TS < 432000000).
        BIND (<http://www.fluidops.com/service/timestamp>(?time2) AS ?time2TS) .
        FILTER(?now - ?time2TS < 432000000).
        BIND (<http://www.fluidops.com/service/timestamp>(?time3) AS ?time3TS) .
        FILTER(?now - ?time3TS < 432000000).
        FILTER(?time3 > ?time2 && ?time2 > ?time1) .
        FILTER NOT EXISTS {
        ?event a event:Event .
        ?event event:hasEventType eventIns:RepeatedZeroCreditEventType.
        ?event event:hasParameter ?param .
        ?param event:hasParameterDefinition eventIns:timeOfLastChangeParameterDefinition.
        ?param event:hasValue ?time3.
        ?event event:hasParameter ?userParam .
        ?userParam event:hasParameterDefinition eventIns:userParameterDefinition.
        ?userParam event:hasValue ?user.
        }
}
"""^^xsd:string .
```

## 6. Proof of concept

As a proof-of-concept, we provide an implementation of our SPARQL-based framework for event processing using the Information Workbench semantic integration platform.

The Information Workbench (Fig. 3) is a Web-based open platform for Linked Data and Big Data solutions in the enterprise developed by fluid Operations [22]. It exists as a community edition that is freely available, as well as a commercial edition. Our approach is based on the community edition. Information Workbench uses data providers that collect data from internal and external sources, convert it into the semantic Resource Description Framework (RDF) format, and interlink it. The integrated data set can be stored in a central repository or managed virtual layer over federated data sources. The provider architecture is extensible and supports the on-demand integration of additional data. Information Workbench is based on semantic technology standards, such as RDF, OWL and SPARQL. It offers a SDK for building apps to support your individual scenarios and requirements. Adapting and extending the functionality can be achieved using the open API interface, semantic data integration through data providers, rules, workflows and an extensible pool of predefined widgets for creating dynamic visualizations and building reports.

Information Workbench integrates the OpenRDF Sesame triple store and, among other things, it provides a SPARQL 1.1 endpoint. For these reasons, our proof-of-concept implementation was developed using the Information Workbench and is available as an app that runs on top of the Information Workbench platform. It can be found and downloaded in the fluidOps AppCenter [23].

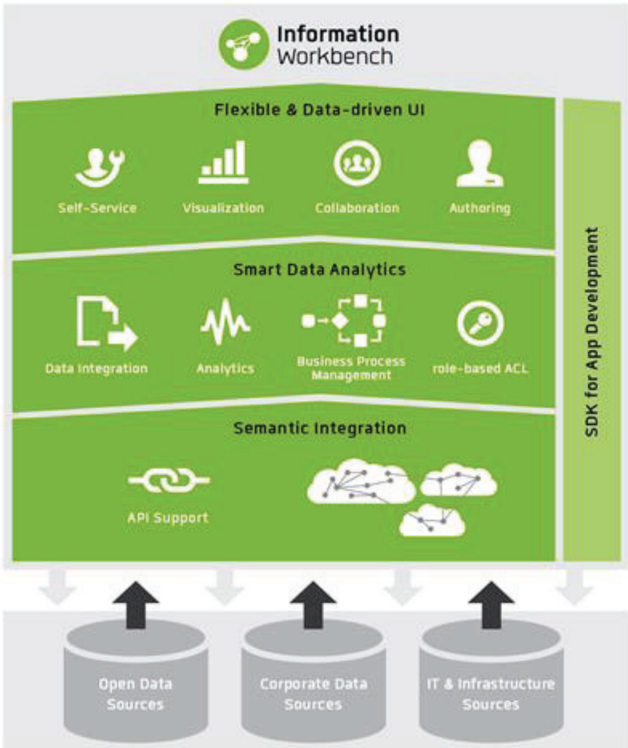



**Figure 3. Information Workbench**

In our app, the base event ontology, event type definitions and events reside in the OpenRDF Sesame triple store. Event publisher is internal and is implemented as a data provider which reacts on event type expressions in order to publish new events which it integrates with the event data into the triple store. In order to facilitate creation of internal event subscribers, a generic event subscriber component was developed. Specific event subscribers can extend this component and define the event type(s) they want to subscribe to. These can be chosen from the library of existing event types in the triple store. External event subscribers can use the SPARQL 1.1 endpoint service to retrieve information about the events. In order to define new event types, a special user interface is provided which allows the user to define different aspects of an event type, including the name, description, event expression and parameter definitions.

## 7. Discussion and conclusion

In this paper, we have presented a novel approach to defining and detecting complex events, which provides highly expressive complex event definitions using RDF and OWL ontologies and the SPARQL query language. It is a generic framework that can be applied to different application domains. A proof-of-concept implementation was developed and is available online.

Compared to other, non-semantic based event processing approaches, our approach offers several advantages. Firstly, it is based on W3C standards, especially RDF, OWL and SPARQL, which means that not special event-processing language or formalism, needs to be introduced. Consequently, no special tools that would understand and process specific formalisms need to be developed, as existing available open-source and commercial systems and tools that support these standards can be used. For example, in our proof-of-concept implementation the community edition of the fluidOps Information Workbench together with the Sesame open source Java framework for processing RDF data were used.

Secondly, it supports definition of complex event types with semantic expressivity requirements of various complexities. Besides, using the OWL- and SPARQL- based semantic approach to event processing has several intrinsic advantages. Our framework integrates event and domain information, thus enabling to perform more complete and accurate inferences about complex event occurrences. As the event expressions can use the domain ontology as well as the event ontology, an OWL-DL reasoner can be used in order to detect events based on the information inferred from the integrated asserted models.

Furthermore, with our mechanism of defining the derived events, one can implicitly define events with a taxonomical structure. An event type can have several specializations, for example each with specific additional conditions. This allows for greater flexibility and allows for working with different levels of detail. Based on the information that is available, more or less specific event can be published. In this way, different subscribers can subscribe to different levels of event abstractions.

## References


[1] D. Bugaite and O. Vasilecas, "Events Propagation from the Business System Level," in *Information Systems Development*, C. Barry, M. Lang, W. Wojtkowski, K. Conboy, and G. Wojtkowski, Eds. Springer US, 2009, pp. 1105–1116.

[2] Julian Krumeich, Benjamin Weis, Dirk Werth, and Peter Loos, "Event-Driven Business Process Management: where are we now?," *Bus. Process Manag. J.*, vol. 20, no. 4, pp. 615–633, Jul. 2014.

[3] A. Sasa Bastinos and O. Vasilecas, "Ontology-based support for complex events," *Elektron. Ir Elektrotechnika*, vol. 7, pp. 83–88, 2011.

[4] A. Sasa Bastinos and M. Krisper, "Context ontology for Event-Driven Information Systems," in *The Seventh International Conference on Systems*, Saint Gilles, Reunion Island, 2012.

[5] D. Luckham, W. R. Schulte, J. Adkins, P. Bizzaro, H.-A. Jacobsen, A. Mavashev, B. M. Michelson, P. Niblett, and D. Tucker, "Event Processing Glossary – Version 2.0." Event Processing Technical Society, Jul-2011.

[6] H. Taylor, A. Yochem, L. Phillips, and F. Martinez, *Event-Driven Architecture: How SOA Enables the Real-Time Enterprise*, 1 edition. Upper Saddle River, NJ: Addison-Wesley Professional, 2009.

[7] A. Sasa, M. B. Juric, and M. Krisper, "Service-Oriented Framework for Human Task Support and Automation," *IEEE Trans. Ind. Inform.*, vol. 4, no. 4, pp. 292–302, Nov. 2008.

[8] K. Chandy and W. R. Schulte, *Event Processing: Designing IT Systems for Agile Companies*, 1 edition. New York: McGraw-Hill Osborne Media, 2009.

[9] T. R. Gruber, “Toward Principles for the Design of Ontologies Used for Knowledge Sharing,” *Int. J. Hum.-Comput. Stud.*, vol. 43, no. 5–6, pp. 907–928, Nov. 1995.

[10] “OWL 2 Web Ontology Language Document Overview (Second Edition), W3C Recommendation.” W3C, 11-Dec-2012.

[11] “SPARQL 1.1 Overview, W3C Recommendation.” W3C, 21-Mar-2013.

[12] S. -y. Cheng, W. -r. Jih, and J. Y. -j. Hsu, “Context-aware Policy Matching in Event-Driven Architecture,” in *Proceeding of the AAAI 2005 Workshop: Contexts and Ontologies: Theory, Practice and Applications*, Pittsburgh, Pennsylvania, USA, 2005, pp. 140–141.

[13] S. Sen and J. Ma, “Contextualised Event-driven Prediction with Ontology-based Similarity,” in *Proceedings of the 2009 AAAI Spring Symposium, Intelligent Event Processing*, 2009, pp. 73–79.

[14] T. Moser, H. Roth, S. Rozsnyai, R. Mordinyi, and S. Biffl, “Semantic Event Correlation Using Ontologies,” in *On the Move to Meaningful Internet Systems: OTM 2009*, R. Meersman, T. Dillon, and P. Herrero, Eds. Springer Berlin Heidelberg, 2009, pp. 1087–1094.

[15] A. Paschke, “Reaction RuleML 1.0 for Rules, Events and Actions in Semantic Complex Event Processing,” in *Rules on the Web. From Theory to Applications*, A. Bikakis, P. Fodor, and D. Roman, Eds. Springer International Publishing, 2014, pp. 1–21.

[16] A. Paschke and H. Boley, “Rule responder: rule-based agents for the semantic-pragmatic web,” *Int. J. Artif. Intell. Tools*, vol. 20, no. 06, pp. 1043–1081, Dec. 2011.

[17] R. Adaikkalavan and S. Chakravarthy, “Event Specification and Processing for Advanced Applications: Generalization and Formalization,” in *Database and Expert Systems Applications*, R. Wagner, N. Revell, and G. Pernul, Eds. Springer Berlin Heidelberg, 2007, pp. 369–379.

[18] A. Briassouli, S. Dasiopoulou, and I. Kompatsiaris, “Ontology-Based Trajectory Analysis for Semantic Event Detection,” in *International Conference on Semantic Computing, 2007. ICSC 2007*, 2007, pp. 735–742.

[19] A. R. J. François, R. Nevatia, J. Hobbs, R. C. Bolles, and J. R. Smith, “VERL: an ontology framework for representing and annotating video events,” *IEEE Multimed.*, vol. 12, no. 4, pp. 76–86, Oct. 2005.

[20] T. Metzke, A. Rogge-Solti, A. Baumgrass, J. Mendling, and M. Weske, “Enabling Semantic Complex Event Processing in the Domain of Logistics,” in *Service-Oriented Computing – ICSOC 2013 Workshops*, A. R. Lomuscio, S. Nepal, F. Patrizi, B. Benatallah, and I. Brandić, Eds. Springer International Publishing, 2013, pp. 419–431.

[21] H. Paulheim, “Efficient Semantic Event Processing: Lessons Learned in User Interface Integration,” in *The Semantic Web: Research and Applications*, L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Eds. Springer Berlin Heidelberg, 2010, pp. 60–74.

[22] fluid Operations, “Information Workbench,” *Information Workbench*. [Online]. Available: http://www.fluidops.com/en/portfolio/information_workbench/.

[23] fluid Operations, “fluidOps AppCenter.” [Online]. Available: http://appcenter.fluidops.net/.

# A Conceptual Model of Fishery in Resource-Event-Agent Framework

Sohei ITO [a,1], Kunimasa AOKI [a] and Kazuaki KAJITORI [a]

[a] *Department of Fisheries Distribution and Management, National Fisheries University, Japan*

**Abstract.** In this paper we present a conceptual model of fishery based on Resource-Event-Agent (REA) framework. We identify the concepts and relationships among them existing in fishery. The knowledge structure of fishery is formally presented so that the conceptual model be a formal ontology. For this, we introduce a formal semantic structure and a logical language. The meaning of the fishery-specific concepts and relationships is described as logical axioms. Through this formal approach, we discuss the adequateness and compatibility of REA framework to model fishery.

**Keywords.** fisheries management, enterprise modelling, REA framework, formal ontology, temporal logic

## 1. Introduction

Fishery is a sector involving the capture of fishes or marine products. Fishery is the core of fishing industry which spans food processing, aquaculture, the refrigeration industry, transportation, distribution and retailing. In fishery, fishermen make their livelihood by fishing. Hence fishery can be seen as an economic activity.

In Japan, many fishing entities have management difficulties while a portion of fishermen earn a lot of money. Due to this problem, many fishing entities cannot find successors and the number of entities are rapidly decreasing [14]. To prevent such decline of fishery, we need to identify the reasons of this crisis in fishery management and find a measure to improve the management structures of fishery. For this purpose, we need an conceptual model which articulates the structures of fishery as a starting point.

The structure of fishery seems simple. Fishermen capture or harvest fishes (in wider sense - the word ‘fish’ includes aquatic animals like mollusks and crustaceans) and sell them to make money. However, we must consider many other entities involving in this activity such as fishing vessels, fishing gear, buyers, and so on. Hence we are interested in what ‘personae’ appear and what relationships exist among them in fishery. In other words, we are interested in the conceptual model of fishery. In this paper, we identify the concepts and relationships among them in fishery. Not only we identify the fishery-specific concepts and relationships, but we are to describe the meaning of them in order

[1] Corresponding Author: Sohei Ito, National Fisheries University, 2-7-1 Nagata-Honmachi, Shimonoseki, Japan; E-mail: ito@fish-u.ac.jp.

to specify the knowledge structure (conceptualisation) of fishery. Therefore our purpose is to construct an *ontology* of fishery.

As we mentioned, fishery is in essence an economic activity. Thus it is natural that the conceptual model of fishery is described in the framework of business process ontology such as AIAI Enterprise Ontology [15], Toronto Virtual Enterprise Ontology (TOVE) [2], $e^3$-*value*$^{TM}$ [6] and Resource-Event-Agent (REA) Enterprise Ontology [4,5]. However, these ontologies focus on the secondary and the tertiary sectors and it is not clear whether fishery which belongs to the primary sector can be modelled.

In this paper, we describe the conceptual model of fishery based on Resource-Event-Agent (REA) framework. The reason we use REA framework is that REA provides generic notions in economic activities and focuses on three core concepts - resource, event and agent - which are also central in fishery. Moreover, to discuss the meaning of fishery-specific concepts and relationships, REA framework has an advantage that it has the formal model and ontology [11]. It enhances the precise understanding of the knowledge structure of fishery. Another reason to choose REA framework is that REA is suitable for designing accounting information systems or ERP systems for enterprise management. Once we establish the conceptual model of fishery based on REA framework, it can serve as the ground of such practical systems.

To the best of our knowledge, this is the first attempt to apply REA framework to modelling the primary industry. Therefore this work is a key test of the adequateness and expressivity of REA framework for modelling the primary industry.

The rest of this paper is organised as follows. Section 2 reviews REA framework and introduces REA concepts and relationships. Section 3 presents a conceptual model of fishery based on REA framework. Fishery-specific concepts and structure of models are identified and the constraints on them are informally presented as axioms. In section 4 we formalise the REA-based structure of fishery and axioms. The semantic structure of REA and the logical language for describing formal ontology of REA is introduced. The ontology is presented as a set of axioms. Section 5 discusses the adequateness and expressivity of REA framework for modelling fishery. We also reason about the problem we found in modelling fishery and how we solve it. The final section offers conclusions and some future directions.

## 2. REA Framework

The original REA paper by McCarthy [13] focused on a data structure as a solid foundation of enterprise accounting system. REA framework was later extended to an enterprise ontology which provides an application neutral data model to represent organisational structures and business processes of enterprises [4,5] with the potential to be implemented as ERP systems [16,17,18].

The REA framework and ontology have evolved in the last decade and several versions exist in current literature. Now REA framework is a two-layer model consisting of operational level and policy level [5]. The operational level describes the accountability infrastructure including the economic activities that take place in a company, the resources that are acquired and consumed, and the agents who participate in economic activities. Plainly speaking, operational level models ‘what is’ and ’what has happened’. On the other hand, the policy level describes ‘what should, could, or must be’ in the enterprise. Policy level specifications define constraints and guidelines under which an enterprise operates. In this paper, we are only interested in the operational level models of REA since what we are trying is to model actual structures or phenomena in fishery.

The explanation about REA framework in the following is based on Hruby's reference model [10] because it has a logical formalisation [11] upon which we will reason about the compatibility of REA framework for modelling fishery.

The concepts in REA framework are *economic resources*, *economic events* and *economic agents*.

Economic resources are something that have utility and are under control of an enterprise. Economic resources are not limited to tangible things. Examples of economic resources are products, services, tools, money, et cetera.

Economic agents are individual or organisation that participates in the economic events related to the enterprise. Economic agents are not limited to the member of the enterprise. For example, a customer of the enterprise will be economic agents if we model a deal between the enterprise and the customer. Economic agents have a power or right to use or dispose of economic resources. Examples of economic agents are customers, vendors, employees and enterprises.

Economic events are phenomena which cause changes in economic resources. Economic events may occur instantaneously. Economic events are divided into increment ones and decrement ones depending on how the value of economic resources changes through economic events. Examples of economic events are sales of goods, rentals, use of services and so on.

We have several relationships amongst the concepts in REA framework to model structures of economic activities. The relationships as associations between resources and events are *outflow*, *inflow*, *consume*, *use* and *produce*. For the relationships as associations between agents and events, we have *provide* and *receive*.

REA captures two types of processes - *exchange* and *conversion* processes - to model economic activities of enterprises. *Exchange* is a process in which an enterprise transfers some rights on a resource to other economic agents, and receives some rights on other resources from other economic agents in return. The following rules apply for any application model representing the REA exchange process [3,10].

- Every *increment economic event* must be related by *exchange* duality to a *decrement economic event*, and vice versa.
- Every *increment economic event* must be related by *inflow* relationship to an *economic resource*.
- Every *decrement economic event* must be related by *outflow* relationship to an economic resource.
- Every *economic event* must be related by a *provide* relationship to an *economic agent*, and by a *receive* relationship to an *economic agent*. At runtime, these two agents must represent entities with different economic interests.

*Inflow* and *outflow* relationship between an *economic resource* and *increment economic event* indicate that the enterprise increases and decreases the resource as a result of the event, respectively. Increase and decrease of a resource mean increase and decrease of value of the resource, respectively. *Provide* relationship designates the *economic agent* who loses the rights to the *economic resource* in the *economic event*. *Receive* relationship designates the *economic agent* who receives the rights to the *economic resource* in

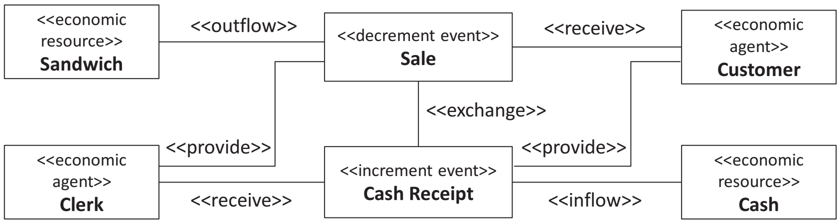



**Figure 1.** An example of REA model of exchange.

the *economic event*. Here we can consider several rights on resources such as ownership rights, usage rights, copy rights and so on. An important fact is that they contribute to the value of the resource for the enterprise. In this work, we take the simplest standpoint for the *inflow* and *outflow* of rights on resources, that is, they are increase or decrease of *quantity* of resources[2], because it is easy to understand and be kept track of. However, we still leave the semantics of *provide* and *receive* relationships as losing or receiving the rights on resources, since *economic agents* who participate in *economic events* in exchange process should have some 'rights' on resources which are exchanged in the process. For example, a shop clerk should have a right to handle goods or cashes in the shop. After the sale, his right on goods decreases (since some goods are sold out). Instead his right on cash (in cashier) increases (since he can utilise more cash).

An example of exchange process is presented in Fig. 1. Here the *economic agent* **Clerk** provides the *economic resource* **Sandwich** to the *decrement event* **Sale**. The *decrement event* **Sale** is the decrement aspect of this deal. The increment aspect of this deal is modelled as the *increment event* **Cash Receipt** which is associated with *exchange* duality relationship to the *decrement event* **Sale**. Through the *economic event* **Cash Receipt**, the *economic agent* **Clerk** receives the *economic resource* **Cash**. The role of the *economic agent* **Customer** is the opposite in this process. **Customer** receives the **Sandwich** in **Sale** and provides **Cash** in **Cash Receipt**.

*Conversion* is a process in which an enterprise uses or consumes resources in order to produce new resources or change their properties. The following rules apply for the REA conversion process.

- Each *increment economic event* must be related by *conversion* duality relationship to a *decrement economic event* and vice versa.
- Each *increment economic event* must be related by *produce* relationship to an economic resource.
- Each *decrement economic event* must be related either by *use* or by *consume* relationship to an *economic resource*.
- Each *economic event* must be related by both *provide* and *receive* relationship to an *economic agent*.

*Produce* relationship between an *economic resource* and an *economic event* indicates that the *economic resource* is produced by the *economic event*. It includes creation of a new product or a change of some product feature (e.g. quality) such as repair, transport

[2]There are economic activities which cannot be (naturally) modelled in this simple view. In this work, however, it does not cause troubles.

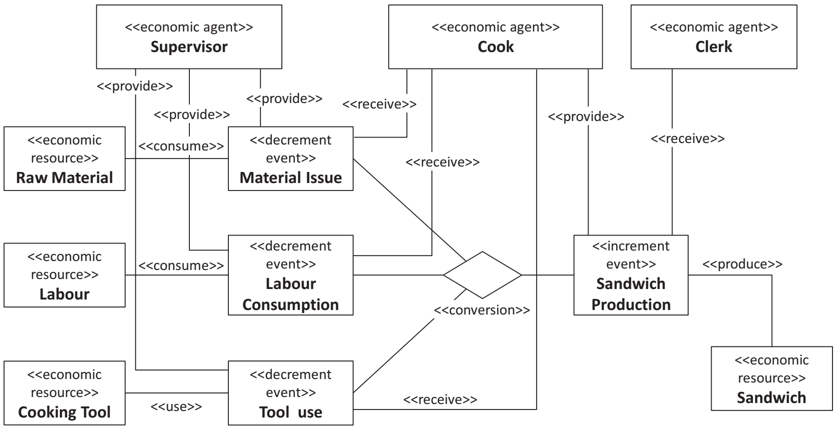



**Figure 2.** An example of REA model of conversion.

and maintenance. *Consume* relationship between an *economic resource* and an *economic event* indicates that the *economic resource* is completely used up in the *economic event* and is no longer available after the event. *Use* relationship is similar to *consume* relationship except that used resources are still available after the event. *Provide* and *receive* relationships between an *economic agent* and an *economic event* designate who lose and receive control over the *economic resources* as a result of the *economic events*. Note that the enterprise does not always control its resources. For example, if the enterprise uses a maintenance service for some resources, the control over the resources should be transferred to the service provider.

In Fig. 2, we present an example REA model of conversion process. It models how the economic resource **Sandwich** is produced by consuming the economic resources **Raw Material** and **Labour**, and by using the economic resource **Oven**. The economic agent **Supervisor** provides them and the economic agent **Cook** receives them to produce **Sandwich**. After the production of **Sandwich**, the economic agent **Clerk** receives it. In this process, three decrement events **Material Issue**, **Labour Consumption** and **Oven use** are associated to the increment event **Sandwich Production** with conversion duality.

## 3. A Conceptual Model of Fishery in REA

Fishery is an economic activity in which fishermen catch fishes or marine products and sell them to earn money. Therefore, in principle, fishery can be divided into two parts – *fishing* and *fish sale*. The purpose of this section is to model these two parts.

The actual fishery spans the various type of fishing and fish sale. For example, fishing can be categorised by target fish species such as tuna, cod, sardine, etc. or by fishing methods such as gillnet fishing, trawl fishing, purse seining, longline fishing, et cetera. Fish sale also has a lot of types. Most typical in Japan is sale at a market in a fishing port. Fishermen bring their products to the market, then wholesale traders, marine product processing industries and retail companies come and buy them. Other types of fish sale are cooperative shipping and cooperative marketing, direct marketing (i.e. fishermen directly sell to consumers) and so on. Cooperative shipping and cooperative marketing are conducted by local Fisheries cooperatives. A local cooperative gathers products of cooperative members (fishermen) and ship them to other markets or mass retailers, or consign them to other distributors.

Categorisation and types of fishery may depend on countries or areas. Any type of fishery will have a common structure as well as its specific structure. In this study we focus on a typical small-scale entity in which a fisherman is the manager, called *fishing household* who engages in coastal fishing. Nevertheless, the core concepts and structures will be common in other types of fishery.

### *3.1. Concepts in Fishery*

We articulate the concepts in fishery in terms of Resource-Event-Agent framework. The bold-faced terms in the following are what we identified as specific concepts in fishery. Note that they represent 'types' of entities and not individuals. We precisely define the meaning of the concepts in formal framework later (section 4).

#### *3.1.1. Resource in Fishery*

Economic resources in fishery consist of things which is necessary to operate fishing, and products which are targets of fishing, i.e. marine products.

We designate the economic resource type which is an abstract collection of targets of fishing as **Fish**. This concept is a collective term comprehending not only 'fish' (in biological meaning) but also other aquatic animals such as Cephalopoda (e.g. octopuses and squids), crustaceans (e.g. shrimps and crabs), shellfishes, et cetera. Fishermen's objective in fishing is to increase the resources of type **Fish** and exchange it to increase his resource of type **Capital**. Thus **Capital** is also an economic resource type appearing in fishery.

There are several things necessary to operate fishing. Such things are also regarded as economic resources in fishery. **Fishing vessel** is a generic term of ships which engage in fishery. It is an essential production tool in fishery. The structure, equipment, nautical instruments, marine products processing implements, storage equipment, etc. are designed according to the size of the fishing vessel, target fish species and fishing gear.

Fishing gear is the equipment used with the aim to operate fishing and has a direct effect in fishing. Examples of fishing gear are fishing nets, pole and lines, fishing weirs and so on. Each type of fishing gear is further classified into several kinds; for example, fishing nets are classified into gillnets, trawl nets, lift nets, purse seines, set nets and so on. There is auxiliary fishing gear in fishing which is auxiliary equipment to support fishing with the aim to increase fishing efficiency such as fishing lights and fishfinders. It also includes equipments to support operation of fishing gear such as net haulers, line haulers, trawl winches, et cetera. We treat collectively both fishing gear and auxiliary fishing gear and categorise them as **Fishing gear**.

**Fuel** is a vital component in fishery since fishing vessels and fishing equipment are operated on it.

There are many supplies needed in fishery such as fish boxes to place fishes, ice to keep fishes fresh, fishing bait to attract fishes (in some kind of fishing), et cetera. We represent the type of all such economic resources as **Supplies**.

### *3.1.2. Event in Fishery*

Economic events in fishery mainly consist of fishing and sale. In REA framework, every economic event should be either increment ones or decrement ones. Moreover, each increment economic event must be related to a decrement economic event, and vice versa. This means we must capture every economic process in two aspects – increment and decrement. This principle applies to the case of fishery.

First we consider fishing. Fishing is an activity to catch fishes. Its purpose is to increase the resource **Fish**. Therefore **Fishing** can be considered as a type of increment economic events.

What is the decrement part of **Fishing**? To operate fishing, fishermen use or consume a lot of things. They use **Fishing vessel** to sail on the sea. To sail fishing vessels, they consume **Fuel**. They use **Fishing gear** to catch fishes. They use or consume many **Supplies**. For example, they use fish boxes to place fishes and consume ice to keep fishes fresh. Hence, we can identify these decrement event types: **Sailing**, **Fishing gear use** and **Supplies use**. These are dual event types of increment event type **Fishing**.

Next we consider fish sale. The purpose of fish sale is to sell fishes and obtain money to increase fishermen's **Capital**. The decrement part of this activity decreases **Fish**. The increment part of this activity increases **Capital**. So we designate the type of decrement events as **Fish sale** and the type of increment events as **Capital increase**.

As we mentioned in section 2, there are two types of duality – exchange duality and conversion duality – in REA framework. Obviously, the duality between **Fish sale** and **Capital increase** is exchange duality. What about the type of duality between **Fishing** and its corresponding decrement events **Sailing**, **Fishing gear use** and **Supplies use**? Fishing activities can be regarded as a kind of production process of **Fish**. Of course, fishing is not 'production' in an usual sense like production of industrial goods or products. It is 'catching'. Nevertheless, fishing can be seen as the change of the property of fishes from non-available state to available state (in view of fisherman). The notion of conversion processes in REA captures change of properties of economic resources. Thus this activity is regarded as conversion process in REA.

We can consider other activities such as fuel procurement, maintenance of fishing vessels and fishing gear, supplies procurement, transportation of fishes and so on. To concentrate on the core structure of fishery, we omit these activities.

### *3.1.3. Agent in Fishery*

Economic agents in fishery consist of **Fisherman** and **Buyer**. **Fisherman** is a type of economic agents who engage in fishing and sell fishes. **Buyer** is a type of economic agents who buy fishes from **Fisherman**. It may include wholesale traders, retail companies, customers and so on.

Since we focused on fishing household, we do not consider other employees in fishing. The model of larger fishery such as fishery companies will be interesting research, but is not the purpose of this study and we leave it for future research.

## *3.2. An Example Model of Fishery*

Before we introduce domain-specific axioms of fishery, we show an example model of fishery in a diagrammatic way. Through construction of an example model, we can observe a domain-specific structure or axioms of fishery.

An example fishery model is a skipjack pole-and-line fishing. In this fishing, a fisherman finds a fishing point (by using a fishfinder or observing sea surface), scatters bait to attract skipjacks then cast a fishing lure into the sea with a pole-and-line. If a fish take the lure, the fisherman lands it onto the boat. Landed fishes are stored with ice to be kept fresh.

This fishing part is modelled in REA framework as the conversion process depicted in Fig. 3. In this figure, the notation of entities (with a colon and an underline) is similar as UML object diagrams. For example, an economic resource **KOYO-MARU : Fishing vessel** means that **KOYO-MARU**[3] is an instance of resource type **Fishing vessel**. For readability, we omit underlines (and colons if the instance name is omitted) in the text.

The central activity of fishing process is divided into four economic events. Increment event **Fishing** produces the economic resource **Skipjack : Fish**. The dual events of **Fishing** are **Sailing**, **Fishing gear use** and **Supplies use**. In decrement event **Sailing**, fishing vessel **KOYO-MARU** is used and **Fuel** is consumed. In decrement event **Fishing gear use**, the resource **pole-and-line** from the resource type **Fishing gear** is used. In decrement event **Supplies use**, two resources of the resource type **Supplies** are related to it. A resource **ice** is consumed (to keep fishes fresh) and **bait** is consumed (to attract skipjacks).

In this fishing process, we have only one agent **John** of agent type **Fisherman**. To determine how he participates in the four economic events in this process needs deeper analysis. In REA framework, each economic event needs both the provider and recipient. Since this fishing process is a conversion process, the semantics of *provide* and *receive* relationships should be understood as losing or receiving control over economic resources. In the increment event **Fishing**, he must participate as recipient since he obtains control over the resource **Skipjack** since he will come to be able to sell (or utilise for other purpose, e.g. eat) it.

The problem is how he participates in the three decrement events **Sailing**, **Fishing gear use** and **Supplies use**. Since he is the owner of the resources used or consumed in these events, he has control over these resources. To allow someone to use or consume these resources, he must be the provider. Nevertheless, the one who uses and consumes them is himself. Hence he must also be the recipient. How do we model such 'self-transfer' of control? This problem is also pointed out by Ito and Vymětal [11]. Unfortunately, the simple solution that the fisherman **John** is simultaneously related to e.g. **Sailing** with both *provide* and *receive* relationship causes problem in defining formal semantics of REA framework [11]. Therefore we introduce a new relationship *prepare* in REA framework to represent such self-transfer of control. Note that if we consider a model of company in which the manager or supervisor is different from fishermen who are better to be called employees, this self-transfer problem does not arise. Then the manager or supervisor participate in these events as the provider and fishermen will be recipients.

The fish sale part of this skipjack fishery is depicted in Fig. 4. The fish sale is modelled as an exchange process in REA. This exchange process models the exchange of fisherman's resource **Skipjack** and **Cash**. The decrement part is represented as the decrement event **Fish sale** and the increment part is modelled as **Cash receipt**. Thus **Fish sale** causes outflow of the resource **Skipjack** and **Cash receipt** causes inflow of the resource **Cash**. The provider of **Fish sale** is clearly **John**, the fisherman and the recipient is some

[3] In Japan, names of fishing vessels usually have the suffix '-MARU'.

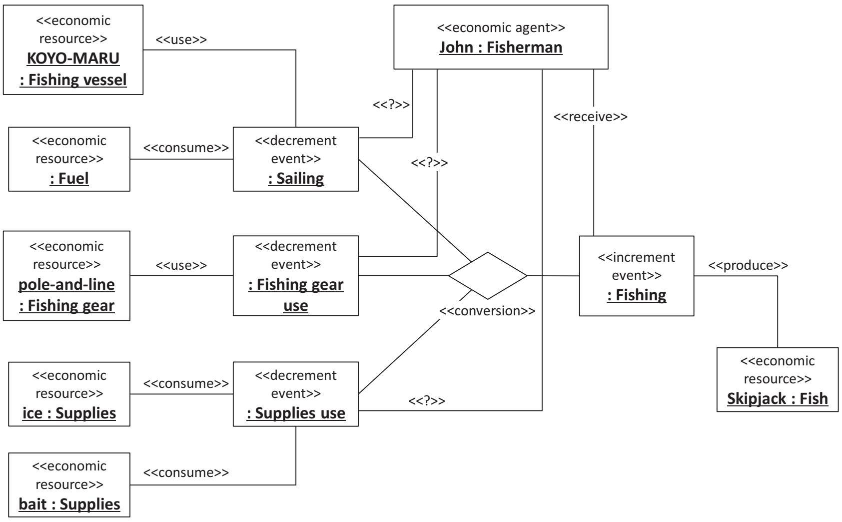



**Figure 3.** An example of REA model of fishing process.

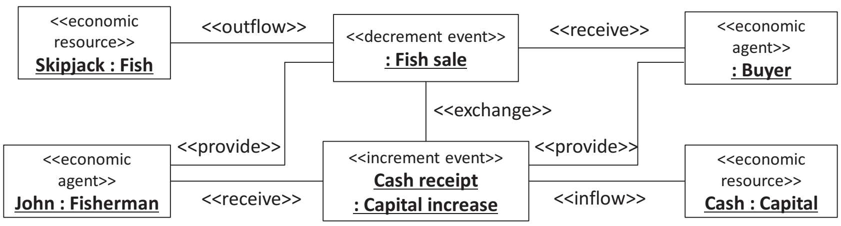



**Figure 4.** An example of REA model of fishery (sale part).

**Buyer**. On the contrary, the provider of **Cash receipt** is the **Buyer** and the recipient is **John**.

Note that there is another type of exchange such as credit sale in real trade. In that case, his **Capital** in this model will be **Accounts payable** and the increment event will be **Accounts payable increase** but the structure of the model does not change.

The entire model of this fishery is the combination of the fishing process and the fish sale process. It is depicted as in Fig.5.

### *3.3. Axioms of Fishery*

The example fishery model discussed in the previous section gives insights about the axioms about the relationships among entity types specific to fishery. In addition, we have implicit axioms specific to fishery domain which have not been clarified yet.

The axioms about relationships among entity types in fishery are informally stated as follows (‘F’ means ‘Fishery’). First we show axioms in fishing processes.

**F1** Economic events of type **Fishing** are *increment events*.

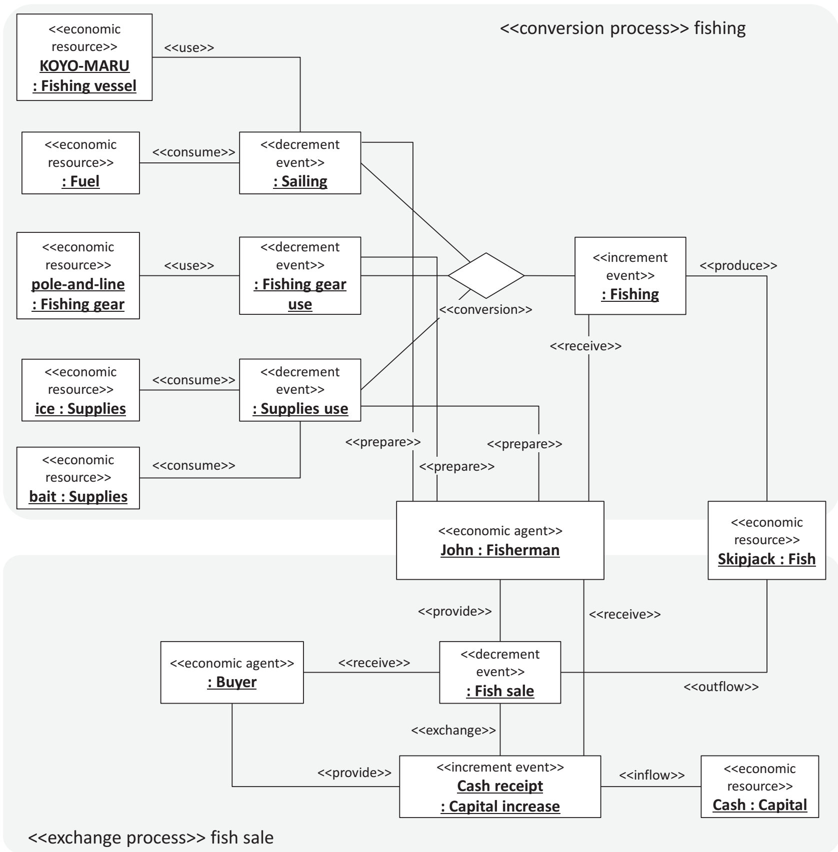



**Figure 5.** An example of REA model of fishery.

**F2** Economic events of type **Sailing**, **Fishing gear use** and **Supplies use** are *decrement events*.

**F3** Economic events of type **Fishing** have *conversion duality* to economic events of type **Sailing**, **Fishing gear use** and **Supplies use**.

**F4** Economic resources of type **Fish** are related to economic events of type **Fishing** by *produce* relationship.

**F5** Economic events of type **Sailing** are related to economic resources of type **Fishing vessel** with *use* relationship, and to economic resources of type **Fuel** with *consume* relationship.

**F6** Economic events of type **Fishing gear use** are related to economic resources of type **Fishing gear** with *use* relationship.

**F7** Economic events of type **Supplies use** are related to economic resources of type **Supplies** with *consume* or *use* relationship.

**F8** Economic agents of type **Fisherman** are related to economic events of type **Fishing** with *receive* relationship, and to economic events of type **Sailing**, **Fishing gear use** and **Supplies use** with *prepare* relationship.

Axiom F1 is a natural postulate since **Fishing** is a kind of producing activity. Axiom F2 says that these event types capture the decrement side of fish production activity. These events use or consume economic resources to operate fishing. Axiom F3 says that the fish production process is conversion process in REA and is modelled as **Fishing** versus **Sailing**, **Fishing gear use** and **Supplies use**. Axiom F4 states the trivial fact that economic events of type **Fishing** is an activity of producing resources of type **Fish**. Axiom F5 states that for sailing, we need two types of resources **Fishing vessel** and **Fuel**[4]. F6 says that economic events of type **Fishing gear use** use economic resource of type **Fishing gear**. In the example model, a pole-and-line was used but it may be nets or some other fishing gear for other type of fishing. F7 says that economic events of type **Supplies use** use or consume economic resources of type **Supplies**. In the example model, resources **ice** and **bait** are consumed. Some resources are not consumed but used so that they are able to be used as resources later. Fish boxes are example of such non-consumed supplies. F8 stipulates how an economic agent of type **Fisherman** participate in the economic events in fishing processes. **Fisherman** participates as a recipient in **Fishing**. As we observed in the example model, **Fisherman** is both the manager and worker in the process. Therefore he *prepares* economic resources needed in fishing for himself. Note that this *prepare* relationship is introduced to describe self-transferring of rights or control over resources. Thus *prepare* relationship may not be fishery-specific relationship but will be appear in other conversion processes which include self-transfer.

In fish sale processes, we have the following axioms.

**F9** Economic events of type **Fish sale** are *decrement events*.

**F10** Economic events of type **Capital increase** are *increment events*.

**F11** Economic events of type **Fish sale** have *exchange duality* to economic events of type **Capital increase**.

**F12** Economic resources of type **Fish** are related to economic events of type **Fish sale** with *outflow* relationship.

**F13** Economic resources of type **Capital** are related to economic events of type **Capital increase**.

**F14** Economic agents of type **Fisherman** are related to economic event of type **Fish sale** with *provide* relationship and to economic event of type **Capital increase** with *receive* relationship.

**F15** Economic agents of type **Buyer** are related to economic event of type **Fish sale** with *receive* relationship and to economic event of type **Capital increase** with *provide* relationship.

F9-11 state that fish sale processes are modelled as exchange processes in REA framework whose decrement part is an economic event of type **Fish sale** and increment part is an economic event of type **Capital increase**. F12 says that economic events of **Fish sale** cause outflow of economic resources of type **Fish**. F13 says that economic events of type **Capital increase** cause inflow of economic resources of type **Capital**. Hence F13 and F14 state that fish sale processes are processes in which a fisherman exchanges his resource of **Fish** for cash or something to increase his **Capital**. F14 says a fisherman is the provider in an event of type **Fish sale** and the recipient of an event of

[4]Sailing by non-powered vessels may not consume fuel, but we ignore such case in this study.

type **Capital increase**. As a result he loses **Fish** to increase **Capital**. F15 says a buyer obtains **Fish** and loses something which contributes increase of fisherman's **Capital**.

Now we discuss implicit axioms which did not appear in the diagrammatical representation of fishery models.

Conversion processes in REA framework assume that the products of the processes necessarily increase. It is not generally true in fishing. It may be valid for some fishing method such as trawl fishing, however, it depends on the condition of sea or season. Even if fishermen have a bit of catch, the caught fishes are sometimes unsalable due to size or fish species. Moreover, pole-and-line or longline fishing targeting large and expensive fishes like tunas, skipjacks, dolphin fishes etc. sometimes result in no catch. Uncertainty of a catch is a common feature of fishery. Therefore we have the following axiom.

**F16** Economic events of type **Fishing** do not necessarily increase economic resources of type **Fish**.

When we see fishes as commercial products, they have a unique feature compared to other kinds of products – *freshness*. Freshness is the main factor which affects the value of fishes. Fishermen cannot leave fishes on a market for the next day[5]. Thus at the level of fishermen all fishes should be sold out (to the wholesale trader or else). As a result, after fish sale fisherman's right on the resources of type **Fish** vanishes and is no longer available for sale. This is uniqueness of fishes as commercial products and we have the following axiom:

**F17** After economic events of type **Fish sale** occurred, the economic resources of type **Fish** which is related by the events are no longer available.

Note that we do not require that all fishes are completely sold out at a single transaction. After a number of transactions, all fishes must be sold out.

Finally, the following axiom states about time scale of fishery processes.

**F18** Fishes are only sold on the day they are caught.

This reflects the fact that fishermen go fishing very early morning and after fishing they bring fishes to a market in a fishing port and sell them on the very day in coastal fishing.

## 4. Formal Ontology of REA-based Fishery Model

In this section we construct formal ontology of REA-based fishery model. Since we use REA framework as a base, the formal ontology is obtained by modifying and extending the formal ontology of REA [11].

First we review the step of constructing formal ontology. According to Gruber [7], 'an ontology is an explicit specification of a conceptualisation'. The interpretation of this definition is various, but we adopt the (slightly modified) version of Guarino's interpretation [8,9]. We interpret 'conceptualisation' as intended semantic structures of a certain knowledge domain, and 'explicit specification' as logical statements which characterise

[5] If a fisherman has refrigeration facilities, he can store frozen fishes for a long time, but it needs another concepts of resource type **Frozen Fish**. We do not consider this case.

the intended semantic structures. In this interpretation we need (i) semantic structures which are suitable to formally represent the knowledge structure of the domain of discourse, (ii) a logical language whose semantics is defined on the semantic structures, and (iii) logical statements (i.e. axioms) which characterise the intended semantic structures of the domain. The logical language includes vocabulary to describe the concepts in the domain, such as *Fish* and *Fisherman*. Therefore, the axioms restrict the correct interpretation of the vocabulary, i.e. what entities are considered as *Fish* or *Fisherman*, and what are the relationship between them.

Since in our study the knowledge structure of fishery is presented in REA framework, we can reuse these components which were already introduced in the formal ontology of REA at operational level [11]. In the following sections, we describe the above three steps to construct formal ontology one-by-one.

### *4.1. Semantic Structure*

Since an economic process generally evolves over time, the semantic structure should be defined over some time structure. Here we choose a linear time structure since REA processes at operational level represent 'what has happened' and we just have a single trace of such process.

In our semantic structure, we have the domain of objects which are semantic representation of entities of REA models (e.g. fishes, fishermen, fishing vessels, and so on). Since the entities appearing in REA framework span several sorts of objects, the object domain is many-sorted. Here we have the set of sorts $\Sigma = \{R, E, A, Q, RC, Da\}$. Sort $R$ represents the sort of *resources*, $E$ *events*, $A$ *agents*, $Q$ *quantity* of resources, $RC$ *rights or control* of agents on resources, and $Da$ *date* of event occurrence.

Then we define the semantic structure as follows:

**Definition 1** (Operational frame)**.** An *operational frame* is a quadruple $\langle D, \rho, \mathbf{P}, \mathbf{F} \rangle$ where $D = \bigcup_{\sigma \in \Sigma} D_\sigma$ is an object domain, $\rho = \rho[0]\rho[1]\ldots$ is a (possibly infinite) sequence of states, $\mathbf{P}$ is a set of intensional relations and $\mathbf{F}$ is a set of intensional functions. Intuitively, $D_R$, $D_E$ and $D_A$ are domains for resources, events and agents, respectively. $D_Q$, $D_{RC}$ and $D_{Da}$ are domains for attributes quantity, rights or control and date, respectively. We assume that $D_Q$ and $D_{Da}$ are well-ordered by a relation $\leq_Q$ and $\leq_{Da}$ respectively, and $D_{RC}$ is partially ordered by $\leq_{RC}$. Naturally we assume we have equalities $=_Q$, $=_{RC}$ and $=_{Da}$ on $D_Q$, $D_{RC}$ and $D_{Da}$ respectively.

The notion of *intensional* relations and functions are introduced to describe situations that the *extensions* of relations and functions differ at different states. For example, a relation $p$ holds between objects $a$ and $b$ at some state, but may not be for another state. Formally, an intensional relation of sort $\sigma_0 \times \cdots \times \sigma_n$ is a total function from $\rho \to \mathfrak{P}(D_{\sigma_0} \times \cdots \times D_{\sigma_n})$, where $\mathfrak{P}(D)$ means the power set of $D$. An intensional function of sort $\sigma_0 \times \cdots \times \sigma_n \to \sigma_{n+1}$ is a total function from $\rho \to (D_{\sigma_0} \times \cdots \times D_{\sigma_n} \to D_{\sigma_{n+1}})$. Therefore, for each intensional relation $p$ and function $f$, $p(\rho[i])$ and $f(\rho[i])$ are a normal relation and function respectively (representing extensions of $p$ and $f$ at state $\rho[i]$).

We write $|\rho|$ for the length of $\rho$. If $\rho$ is infinite, $|\rho| = \infty$. $\rho$ represents how the state of affairs of a process change over time. The time we are referring to is not real but logical. In other words, $\rho$ captures causal relationship of a process. A state of affair of a process is represented by relations and functions. Relations show e.g. who participates in an event and functions show e.g. what is the quantity of a resource.

We have just introduced a formal structure of REA processes of fishery. Next we introduce a logical language so that we syntactically characterise REA processes of fishery.

### *4.2. Logical Language $L^F_{REA}$*

The purpose of our logical language is to characterise REA processes of fishery (formally represented as operational frames) by describing how we interpret the concepts and relationships appearing in REA processes of fishery. For example, how do we interpret the concept **Fish**? What is the difference of **Fish** and **Fuel**? To answer these questions formally, we introduce the logical language $L^F_{REA}$ (F means 'fishery').

The logical language $L^F_{REA}$ consists of vocabulary to describe concepts and relationships of REA processes of fishery, syntax which stipulates how we write formulae and semantics which defines how we interpret formulae on the operational frames.

Since the operational frame of REA processes of fishery is many-sorted, $L^F_{REA}$ is also many-sorted and have the same set of sorts $\Sigma = \{R, E, A, Q, RC\}$. The vocabulary of $L^F_{REA}$ consists of following predicate symbols, relation symbols, function symbols, constant symbols and variable symbols:

1. The predicate symbols of sort $R$: *Fish*, *Fishing_vessel*, *Fuel*, *Fishing_gear*, *Supplies*, *Capital*.
2. The predicate symbols of sort $E$: *IncEv*, *DecEv*, *occur*, *Fishing*, *Sailing*, *Fishing_gear_use*, *Supplies_use*, *Fish_sale*, *Capital_increase*.
3. The predicate symbols of sort $A$: *Fisherman*, *Buyer*.
4. The predicate symbols of sort $E \times R$: *produce*, *consume*, *use*, *inflow*, *outflow*.
5. The predicate symbols of sort $A \times E$: *provide*, *receive*, *prepare*.
6. The predicate symbols of sort $E \times E$: *conversion*, *exchange*.
7. The predicate symbols of sort $Q \times Q$: $\leq_Q$, $=_Q$.
8. The predicate symbols of sort $RC \times RC$: $\leq_{RC}$, $=_{RC}$.
9. The predicate symbols of sort $Da \times Da$: $\leq_{Da}$, $=_{Da}$.
10. The function symbol of sort $R \to Q$: *quantity*.
11. The function symbol of sort $R \times A \to RC$: *rc*.
12. The function symbol of sort $E \to Da$: *date*.
13. The sets of constant symbols $\{Con_\sigma\}_{\sigma\in\Sigma}$. We assume a special symbol $0 \in Con_Q$.
14. The sets of variable symbols $\{Var_\sigma\}_{\sigma\in\Sigma}$.

The predicate symbols correspond to concept names and relationship names. Sort of symbols partially characterises the concepts. For example, *Fish* is a unary predicate symbol of sort $R$. This means an element of *Fish* is a resource. In other words, the concept *Fish* is a collection of economic resources. The properties of *Fish* (e.g. what *Fish* is and what relationship it has to other entities in fishery) is characterised by logical statements to restrict the correct interpretation of the predicate *Fish*.

Note that we use the same symbols $\leq_Q$, $\leq_{RC}$, $=_Q$, $=_{RC}$, $\leq_{Da}$ and $=_{Da}$ as the order relations on the domains $D_Q$ and $D_{RC}$. The formers are syntactic objects and the latter are semantics objects. As we will see later, these symbols are interpreted as the corresponding order relations.

Now we define the set of terms of $L^F_{REA}$.

**Definition 2** (Term)**.** Let $\sigma \in \Sigma$ be a sort. The set $Term_\sigma$ of *terms* of $L^F_{REA}$ is inductively defined as follows:

1. $Con_\sigma \cup Var_\sigma \subseteq Term_\sigma$.
2. If $t_i \in Term_{\sigma_i}$ for $1 \leq i \leq n$ and $f$ is a function symbol of sort $\sigma_1 \times \cdots \times \sigma_n \to \sigma$ then $f(t_1,\ldots,t_n) \in Term_\sigma$.

Next we define atoms of $L^F_{REA}$.

**Definition 3** (Atom)**.** If $t_i \in Term_{\sigma_i}$ for $1 \leq i \leq n$ and $p$ is a relation symbol of sort $\sigma_1 \times \cdots \times \sigma_n$ then $p(t_1,\ldots,t_n)$ is an atom.

Now we define formulae of $L^F_{REA}$. Since $L^F_{REA}$ is intended to be interpreted on operational frames of definition 1, which are linear-time structures, our logic is based on linear temporal logic [1,12].

**Definition 4** (Formula)**.** The *formulae* of $L^F_{REA}$ is inductively defined as follows:

1. Atoms are formulae.
2. If $\varphi$ and $\psi$ are formulae then $\neg\varphi$, $\varphi \wedge \psi$, $\Box\varphi$, $\overline{\Box}\varphi$, $\circ\varphi$ and $\overline{\circ}\varphi$ are formulae.
3. If $\varphi$ is a formula and $x \in Var_\sigma$ then $\forall_\sigma x\varphi$ is a formula, where $\sigma \in \Sigma$.

As we can see, we use the temporal operators $\Box$, $\overline{\Box}$, $\circ$ and $\overline{\circ}$ as well as the propositional operators $\neg$ and $\wedge$, and quantifiers of many-sorted logic $\forall_\sigma$.

Intuitively, $\Box\varphi$ means '$\varphi$ is always true from the current state', $\circ\varphi$ means '$\varphi$ is true in the next state'. The overlined operators $\overline{\Box}$ and $\overline{\circ}$ mean the 'past' direction of $\Box$ and $\circ$. $\neg\varphi$ means '$\varphi$ is not true' and $\varphi \wedge \psi$ means 'both $\varphi$ and $\psi$ are true'. $\forall_\sigma x\varphi$ means '$\varphi$ is true for arbitrary object $x$ of sort $\sigma$'.

Now we formally define the semantics of $L^F_{REA}$ with respect to operational frames.

**Definition 5** (interpretation)**.** Let $\mathfrak{O} = \langle D, \rho, \mathbf{P}, \mathbf{F} \rangle$ be an operational frame. An *interpretation* $I$ of symbols with respect to $\mathfrak{O}$ is a function satisfying the following:

1. $I(c) \in D_\sigma$ if $c \in Con_\sigma \cup Var_\sigma$ for $\sigma \in \Sigma$.
2. $I(p) \in \mathbf{P}$ if $p$ is a predicate symbol where the sort of $p$ and $I(p)$ are the same.
3. $I(f) \in \mathbf{F}$ if $f$ is a function symbol where the sort of $f$ and $I(f)$ are the same.
4. $I(\sim) = \sim$ where $\sim \in \{\leq_Q, =_Q, \leq_{RC}, =_{RC}\}$.

Clause 4 seems a bit confusing. Recall operational frames are equipped with the *relations* $\leq_Q$, $=_Q$, $\leq_{RC}$, and $=_{RC}$ on the domains $D_Q$ and $D_{RC}$. We also introduced *relation symbols* $\leq_Q$, $=_Q$, $\leq_{RC}$, and $=_{RC}$ on the domains $D_Q$ and $D_{RC}$. Thus, clause 4 says we interpret e.g. the *relation symbol* $\leq_Q$ as the *relation* $\leq_Q$ defined on the domain $D_Q$ of the operational frame $\mathfrak{O}$.

An *operational structure* is a pair $\langle \mathfrak{O}, I \rangle$. Let $\mathfrak{M} = \langle \langle D, \rho, \mathbf{P}, \mathbf{F} \rangle, I \rangle$. Due to intensional functions, the interpretation of terms vary according to states.

**Definition 6** (interpretation of terms)**.** The interpretation of terms $(\mathfrak{M}, i)$, where $i$ is a natural number, is defined as follows:

1. $(\mathfrak{M}, i)(c) = I(c)$ if $c \in Con_\sigma \cup Var_\sigma$ for all $\sigma \in \Sigma$.
2. $(\mathfrak{M}, i)(f(t_1,\ldots,t_n)) = I(f)(\rho[i])((\mathfrak{M}, i)(t_1),\ldots,(\mathfrak{M}, i)(t_n))$.

Then we define how we interpret formulae of $L^F_{REA}$ by defining the satisfaction relation $\models$. We write $\mathfrak{M}, i \models \varphi$ to mean the formula $\varphi$ is true at the $i$-th state $\rho[i]$ of the structure $\mathfrak{M}$.

**Definition 7** (satisfaction relation)**.** The satisfaction relation $\models$ is defined as follows:

1. $\mathfrak{M}, i \models p(t_1, \ldots, t_n)$ iff $\langle (\mathfrak{M}, i)(t_1), \ldots (\mathfrak{M}, i)(t_n) \rangle \in I(p)(\rho[i])$.
2. $\mathfrak{M}, i \models \neg\varphi$ iff $\mathfrak{M}, i \not\models \varphi$.
3. $\mathfrak{M}, i \models \varphi \wedge \psi$ iff $\mathfrak{M}, i \models \varphi$ and $\mathfrak{M}, i \models \psi$.
4. $\mathfrak{M}, i \models \forall_\sigma x \varphi$ iff $\mathfrak{M}[x \mapsto d], i \models \varphi$ for all $d \in D_\sigma$.
5. $\mathfrak{M}, i \models \Box\varphi$ iff $\mathfrak{M}, j \models \varphi$ for all $j \geq i$.
6. $\mathfrak{M}, i \models \overline{\Box}\varphi$ iff $\mathfrak{M}, j \models \varphi$ for all $j \leq i$.
7. $\mathfrak{M}, i \models \circ\varphi$ iff $\mathfrak{M}, i+1 \models \varphi$ and $i > |\rho|$.
8. $\mathfrak{M}, i \models \overline{\circ}\varphi$ iff $\mathfrak{M}, i-1 \models \varphi$ and $i > 0$.

$\mathfrak{M}[x \mapsto d]$ represents the structure $\langle \mathfrak{O}, I[x \mapsto d] \rangle$ where the interpretation $I[x \mapsto d]$ is the same as $I$ except it maps $x$ to $d$.

We use the following abbreviations; $\varphi \vee \psi = \neg(\neg\varphi \wedge \neg\psi)$, $\varphi \to \psi = \neg\varphi \vee \psi$, $\varphi \leftrightarrow \psi = (\varphi \to \psi) \wedge (\psi \to \varphi)$, $\Diamond\varphi = \neg\Box\neg\varphi$, $\overline{\Diamond}\varphi = \neg\overline{\Box}\neg\varphi$, $\odot\varphi = \neg\circ\neg\varphi$, $\overline{\odot}\varphi = \neg\overline{\circ}\neg\varphi$. Intuitively, $\varphi \vee \psi$ means '$\varphi$ or $\psi$ is true', $\varphi \to \psi$ means '$\varphi$ implies $\psi$', $\varphi \leftrightarrow \psi$ means '$\varphi$ and $\psi$ are equivalent', $\Diamond\varphi$ means '$\varphi$ is eventually true', $\odot\varphi$ means $\varphi$ is true at the next state or there is no next state.

We assume $\wedge$ and $\vee$ binds more strongly than $\to$, and $\leftrightarrow$ and unary operators binds more strongly than binary operators.

Now we introduce some notations.

**Definition 8.** If we have $\mathfrak{M}, i \models \varphi$ for all $i < |\rho|$, we write $\mathfrak{M} \models \varphi$. The formula $\varphi$ is *valid* if $\mathfrak{M} \models \varphi$ for any operational structure $\mathfrak{M}$ and we just write $\models \varphi$. Let $\Gamma$ be a set of formulae. We write $\mathfrak{M} \models \Gamma$ if $\mathfrak{M} \models \psi$ for any $\psi \in \Gamma$. We write $\Gamma \models \varphi$ if $\mathfrak{M} \models \Gamma \Rightarrow \mathfrak{M} \models \varphi$ for any $\mathfrak{M}$.

### *4.3. Axioms*

In this section we formally state axioms of REA processes of fishery. Since our fishery model is based on REA framework, the axioms of fishery consist of REA axioms and fishery-specific axioms. REA axioms are already formalised in [11]. However, since our interpretation of concepts and relationships is slightly different from those of genuine REA, some REA axioms are modified. In addition, we added the new relationship *prepare*, thus we add an axiom for this new relationship.

First we formalise the axioms F1 to F8 in fishing processes which are informally stated in section 3.3. These axioms stipulate how the concepts in fishery are interrelated.

**F1** $\forall_E e(\mathit{Fishing}(e) \to \mathit{IncEv}(e))$
**F2** $\forall_E e(\mathit{Sailing}(e) \vee \mathit{Fishing\_gear\_use}(e) \vee \mathit{Supplies\_use}(e) \to \mathit{DecEv}(e))$
**F3** $\forall_E e_1(\mathit{Fishing}(e_1) \to \exists_E e_2 \exists_E e_3 \exists_E e_4(\mathit{Sailing}(e_2) \wedge \mathit{Fishing\_gear\_use}(e_3) \wedge \mathit{Supplies\_use}(e_4) \wedge \mathit{conversion}(e_2, e_1) \wedge \mathit{conversion}(e_3, e_1) \wedge \mathit{conversion}(e_4, e_1)))$
**F4** $\forall_E e(\mathit{Fishing}(e) \wedge \mathit{occur}(e) \to \exists_R r(\mathit{Fish}(r) \wedge \mathit{produce}(e, r)))$
**F5** $\forall_E e(\mathit{Sailing}(e) \wedge \mathit{occur}(e) \to \exists_R r_1 \exists_R r_2(\mathit{Fishing\_vessel}(r_1) \wedge \mathit{use}(e, r_1) \wedge \mathit{Fuel}(r_2) \wedge \mathit{consume}(e, r_2)))$

**F6** $\forall_E e(Fishing\_gear\_use(e) \wedge occur(e) \rightarrow \exists_R(Fishing\_gear(r) \wedge use(e,r)))$
**F7** $\forall_E e(Supplies\_use(e) \wedge occur(e) \rightarrow \exists_R(Supplies(r) \wedge use(e,r)))$
**F8-1** $\forall_E e(Fishing(e) \wedge occur(e) \rightarrow \exists a(Fisherman(a) \wedge provide(a,e)))$
**F8-2** $\forall_E e((Sailing(e) \vee Fishing\_gear\_use(e) \vee Supplies\_use(e)) \wedge occur(e) \rightarrow \exists a(Fisherman(a) \wedge prepare(a,e)))$

Axiom F8 is stated in two clauses F8-1 and F8-2. Predicate $occur(e)$ means event $e_1$ occurred at that state. The formal REA ontology discriminates existence of events and occurrence of events [11]. Since REA model is understood as the record of 'what has happened', the events must occur at some time-point. This is captured by *occur* predicate. When an event occurred, resource production, use and consume or rights/control exchange happen. Therefore such relations holds at the very state.

Next we formalise the axioms F9 to F15 in fish sale processes. These axioms describe straightforwardly the relationship among concepts stated informally in section 3.3.

**F9** $\forall_E e(Fish\_sale(e) \rightarrow DecEv(e))$
**F10** $\forall_E e(Capital\_increase(e) \rightarrow IncEv(e))$
**F11** $\forall_E e_1(Fish\_sale(e_1) \rightarrow \exists e_2(Capital\_increase(e_2) \wedge exchange(e_1,e_2)))$
**F12** $\forall_E e(Fish\_sale(e) \wedge occur(e) \rightarrow \exists_R r(Fish(r) \wedge outflow(e,r)))$
**F13** $\forall_E e(Capital\_increase(e) \wedge occur(e) \rightarrow \exists_R r(Capital(r) \wedge inflow(e,r)))$
**F14-1** $\forall_E e(Fish\_sale(e) \wedge occur(e) \rightarrow \exists_A a(Fisherman(a) \wedge provide(a,e)))$
**F14-2** $\forall_E e(Capital\_increase(e) \wedge occur(e) \rightarrow \exists_A a(Fisherman(a) \wedge receive(a,e)))$
**F15-1** $\forall_E e(Fish\_sale(e) \wedge occur(e) \rightarrow \exists_A a(Buyer(a) \wedge receive(a,e)))$
**F15-2** $\forall_E e(Capital\_increase(e) \wedge occur(e) \rightarrow \exists_A a(Buyer(a) \wedge provide(a,e)))$

Formalisation of F16 and F17 is not trivial. We first consider F16 which says economic events of type **Fishing** do not necessarily increase economic resources of type **Fish**. Since events of type **Fishing** are related to resources of type **Fish** by *produce* relationship (from F4), this axiom restricts how we interpret *produce* relationship.

In the formal REA ontology [11], the axiom about the meaning of *produce* relationship is stated as (RA10-1 in [11]):

$$\forall_E e \forall_R r \forall_Q q(produce(e,r) \wedge quantity(r) =_Q q \rightarrow \circ quantity(r) >_Q q).$$

This axiom says that if *produce* relationship holds between event $e$ and resource $r$ and the quantity of $r$ is $q$ at the current state, then the quantity of $r$ becomes larger than $q$ at the next state. In other words, *produce* relationship ensures the increase of the resource. In fishery processes, this is not true since sometimes fishermen do not catch fishes which is stated in axiom F16. Thus axiom F16 can be obtained by slightly modifying the above axiom:

**F16** $\forall_E e \forall_R r \forall_Q q(produce(e,r) \wedge quantity(r) =_Q q \rightarrow \circ quantity(r) \geq_Q q)$

In this axiom the quantity of produced resources is allowed to be the same after production.

Axiom F17 says that after economic events of type **Fish sale** occurred, the economic resources of type **Fish** which is related by the events are no longer available. 'Not available' is somewhat ambiguous. The intention of this axiom is that fishermen must sell out their fishes in fish sale processes. This can be naturally modelled as the quantity of fishes becomes zero. Hence axiom F17 is formalised as follows:

**F17** $\forall_E e \forall_R r(\mathit{Fish\_sale}(e) \wedge \mathit{Fish}(r) \wedge \mathit{outflow}(e,r) \wedge \mathit{occur}(e) \rightarrow \Diamond \mathit{quantity}(r) = 0)$

Axiom F18 states that fishes are sold on the day they are caught. This axiom can be formalised using *date* function on events.

**F18** $\forall_R r \forall_E e \forall_{Da} d(\mathit{Fishing}(r) \wedge \mathit{produce}(e,r) \wedge \mathit{date}(e) =_{Da} d \rightarrow \forall_E e'(\mathit{Fish\_sale}(e') \wedge \mathit{occur}(e') \wedge \mathit{outflow}(e',r) \rightarrow \mathit{date}(e') =_{Da} d))$

By this axiom and F17, we ensure that all fishes are sold out in one day.

Now we introduce axioms for REA extension. First, we introduce an axiom for *prepare* relationship. As we noted in section 3.2, *prepare* relationship represents self-transfer of his control over a resource. Because self-transfer of control cause neither gain nor loss of his control, we can formalise the meaning of *prepare* as follows ('RE' means 'REA Extension'):

**RE1** $\forall_E e \forall_R r \forall_A a \forall_{RC} c(ER(e,r) \wedge \mathit{prepare}(a,e) \wedge rc(a,r) =_{RC} c \rightarrow \circ rc(a,r) =_{RC} c)$

where $ER(e,r) = \mathit{inflow}(e,r) \vee \mathit{outflow}(e,r) \vee \mathit{produce}(e,r) \vee \mathit{consume}(e,r) \vee \mathit{use}(e,r)$.

Since we admit the *prepare* relationship between event and agents in addition to *provide* and *receive*, we modify the REA axiom which says there must be a provider and a recipient for each economic event in conversion processes. The modified REA axiom is as follows:

**RE2** $\forall_E e_1 \forall_E e_2(\mathit{conversion}(e_1,e_2) \wedge \mathit{occur}(e_1) \rightarrow \exists_R r(\mathit{consume}(e_1,r) \vee \mathit{use}(e_1,r)) \wedge ((\exists_A a\, \mathit{provide}(a,e_1) \wedge \exists_A b\, \mathit{receive}(b,e_1)) \oplus \exists_A a\, \mathit{prepare}(a,e_1)))$

where $\oplus$ is the 'exclusive-or' operator.

This formal expression of the axiom needs some explanation. For notational convention, when we write $\mathit{conversion}(e_1,e_2)$ the event $e_1$ is a decrement event and $e_2$ is an increment event in the conversion process (RA2-2 in [11]). In the decrement part of conversion processes we must have either a provider and a recipient, or one agent who prepares resources used or consumed in the event. Note that *prepare* relationship is not allowed for increment events.

We need to introduce axioms for *date* function which is intended to represent the date of event occurrence. We assume that before occurrence the date of an event is undefined and after occurrence the date of an event is determined and does not change afterwards.

**RE3** $\forall_E e(\mathit{occur}(e) \rightarrow \overline{\circ}\overline{\Box}(\neg \exists_{Da} d(\mathit{date}(e) =_{Da} d)) \wedge \exists_{Da} d \Box(\mathit{date}(e) =_{Da} d))$

The rest of REA axioms also apply for fishery processes.

## 5. Discussion

All the axioms introduced in the previous section are the explicit specification of intended structures of REA processes of fishery. That is, the axioms are the formal ontology of REA processes of fishery. We denote the set of axioms of REA processes of fishery as $\Gamma^F_{REA}$. Then, the set of REA processes of fishery corresponds to the set of operational structures which satisfies $\Gamma^F_{REA}$, that is, $\{\mathfrak{M} \mid \mathfrak{M} \models \Gamma^F_{REA}\}$. This formal characterisation of fishery processes serves as a specification of an accounting information system of fishery. What classes we need, what are relationships among them, and what are correct traces of fishing processes and fish sale processes are specified in the formal ontology.

Since we not only add axioms but also modified the axioms of genuine REA (by F16, RE1 and RE2), there is no inclusion between the set of REA processes of fishery and the set of REA processes. Thus we can conclude that REA framework is *not adequate* as is to model fishery. However, we showed through this study that REA framework is still useful to model fishery. The concepts and relationships of REA are naturally used to model the almost all aspects of fishery. This result implies that accounting information system or ERP system of fishery based on REA framework are feasible and how we adapt it to fishery domain is elucidated by our formal ontology of REA-based fishery processes.

Our formal ontology can be used to reason about theorems of fishery processes. Theorems are deduced by the axioms and inference rules of $L^F_{REA}$ which are the same as $L^O_{REA}$ [11]. The following theorem is easily derived from F1 and REA axioms about *provide* relationship.

**Theorem.** $\forall_E e(Fishing(e) \wedge occur(e) \rightarrow \exists_A a\, provide(a,e))$.

This theorem says that there is a provider for each **Fishing** event. Note that we have a recipient of each **Fishing** event by axiom F8, i.e **Fisherman**. However we do not have an axiom about providers of **Fishing** events. Then the question arises: who are providers of **Fishing**, in other words, who loses rights over **Fish**? From practical point of view, one solution to this problem is to leave it undesignated. This is an example of *modelling compromise* and is widely accepted in practical applications [10]. From ontological point of view, however, we must have some provider as the theorem says. One plausible solution is to introduce agent type **Ocean**. An ocean area (e.g. East Chine sea, Pacific Ocean, and so on), which is an instance of **Ocean**, is a provider of **Fishing**. Although ocean is not an entity who insists who has control or ownership on the resource **Fish**, it is a reasonable choice to keep track of the area of production of fishes for commercial purposes or sustainable fish resource management. Of course, to consider ocean as an agent is a broad interpretation of the concept of 'agent' in REA. Nevertheless, such extension is harmless for the purpose of this study and we can adopt it.

## 6. Conclusion

In this paper, we constructed a conceptual model of fishery using REA framework. The concepts appearing in fishery are identified and classified using REA concepts. We formalised the conceptual model of fishery and articulated the semantic structure of fishery processes using a logical language. Since we have an formal explicit specification of the domain-specific structure of fishery, we constructed a formal ontology of fishery.

This study is a first attempt to model the primary industry in REA framework. Through this approach, the usefulness of REA framework is confirmed, though we need slight extensions and modifications.

Our research begins with philosophical motivation, but it opens the path to constructing accounting information systems or ERP systems for fishery management. Thus working for this practical direction is one of the future researches. Other direction is to extend our conceptual model to other type of fishery such as fishery company, or mother-ship type fishery which consists of big ship (called mother ship) and several solo operating catcher boats. These fisheries have different economic structures compared to fishing households which we modelled in this paper. To elucidate the difference will be an interesting research which sheds light on how to improve the management structures. An important further research is to simulate (like [17,18]) and compare our fishery models to identify the reason of management difficulties in fishery and find measures to overcome them. Our work is a first step for this important direction.

## References


[1] E. A. Emerson. Temporal and modal logic. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science, Volume B: Formal Models and Sematics (B)*, pages 995–1072. MIT Press, 1990.

[2] M. S. Fox and M. Gruninger. Enterprise modeling. *AI Magazine*, 19(3):109–121, 1998.

[3] G. L. Geerts and W. E. McCarthy. The ontological foundations of REA enterprise information systems. In *Annual Meeting of the American Accounting Association*, 2000.

[4] G. L. Geerts and W. E. McCarthy. An ontological analysis of the economic primitives of the extended-REA enterprise information architecture. *International Journal of Accounting Information Systems*, 3(1):1–16, 2002.

[5] G. L. Geerts and W. E. McCarthy. Policy-level specification in REA enterprise information systems. *Journal of Information Systems*, 20(2):37–63, 2006.

[6] J. Gordijn and H. Akkermans. Value based requirements engineering: Exploring innovative e-commerce idea. *Requirements Engineering Journal*, 8(2):114–134, 2003.

[7] T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, December 1995.

[8] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. *Int. J. Hum.-Comput. Stud.*, 43(5-6):625–640, December 1995.

[9] N. Guarino. Understanding, building and using ontologies. *Int. J. Hum.-Comput. Stud.*, 46(2-3):293–310, March 1997.

[10] P. Hruby. *Model-Driven Design Using Business Paterns*. Springer-Verlag, 2006.

[11] S. Ito and D. Vymětal. The formal REA model at the operational level. *Applied Ontology*, 8(4):275–300, 2013.

[12] O. Lichtenstein, A. Pnueli, and L. Zuck. The glory of the past. In *Logics of Programs*, volume 193 of *Lecture Notes in Computer Science*, pages 196–218. Springer Berlin Heidelberg, 1985.

[13] W. E. McCarthy. The REA accounting model: A generalized framework for accounting systems in a shared data environment. *The Accounting Review*, 57(3):554–78, 1982.

[14] Ministry of Agriculture, Forestry and Fisheries. Fisheries employment trend survey (in Japanese). `http://www.maff.go.jp/j/tokei/kouhyou/gyogyou_doukou/index.html`.

[15] M. Uschold, M. King, S. Moralee, and Y. Zorgios. The enterprise ontology. *The Knowledge Engineering Review*, 13(1):31–89, 1998.

[16] P. Vandenbossche and H. Wortmann. Why accounting data models from research are not incorporated in ERP systems. In *2nd International REA Technology Workshop*, 2006.

[17] D. Vymětal and C. V. Scheller. MAREA: Multi-agent REA-based business process simulation framework. In *International Scientific Conference ICT for Competitiveness*, pages 301–310, 2012.

[18] D. Vymětal and R. Šperka. MAREA - from an agent simulation application to the social network analysis. In *18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2014, Gdynia, Poland, 15-17 September 2014*, pages 1416–1425, 2014.

# Mutual Resource Exchanging Model in Mobile Computing and its Application to Hybrid Learning

Naofumi YOSHIDA
*Faculty of Global Media Studies, Komazawa University*

**Abstract.** The only way for efficient use of limited resources is exchanging them each other in mobile environment. In this paper, an application for hybrid learning of wave pattern detection is presented. The detection accuracy and speed will be improved by exchanging learning results mutually among integrated detection methods. By this model, flexible and elastic usability can be implemented on mobile devices in mobile computing environment. And hybrid learning of wave pattern detection enables us to elastic and intelligent learning among devices.

**Keywords.** Mutual Resource Exchanging, Hybrid, Learning, Wave Pattern Detection

## 1. Introduction

The resource exchanging is the key function because we have the limitation of the capacity for every single device in mobile computing environment. If there is no chance to generate resources, the only way for efficient use of limited resources is exchanging them each other inside a closed world, in mobile environment. Based on that background, a mutual resource exchanging model[1][3] in mobile computing environment are proposed. In this paper, an application for hybrid learning of wave pattern detection is presented. The detection accuracy and speed will be improved by exchanging learning results mutually among integrated detection methods. By this model, flexible and elastic usability can be implemented on mobile devices in mobile computing environment. And hybrid learning of wave pattern detection enables us to elastic and intelligent learning among devices. Wave pattern detection is fundamental functionality in digital devices because it is a basis of Analog-Digital and Digital-Analog converting.

Figure 1 shows the overview of a mutual resource exchanging model[1][3] in mobile computing. In the conceptual layer, all target devices are summarized as a group of devices inside a certain closed world in mobile computing environment. All resources are connected to the group of devices. In the physical layer, all devices are recognized and resources are owned by each device. If one device requires much resource than the amount of its own, resources owned by other devices are exchanged to the target device. Though these exchanging is performed for every kinds of resources, in the conceptual layer there is no exchanging and we have just one group of target devices.

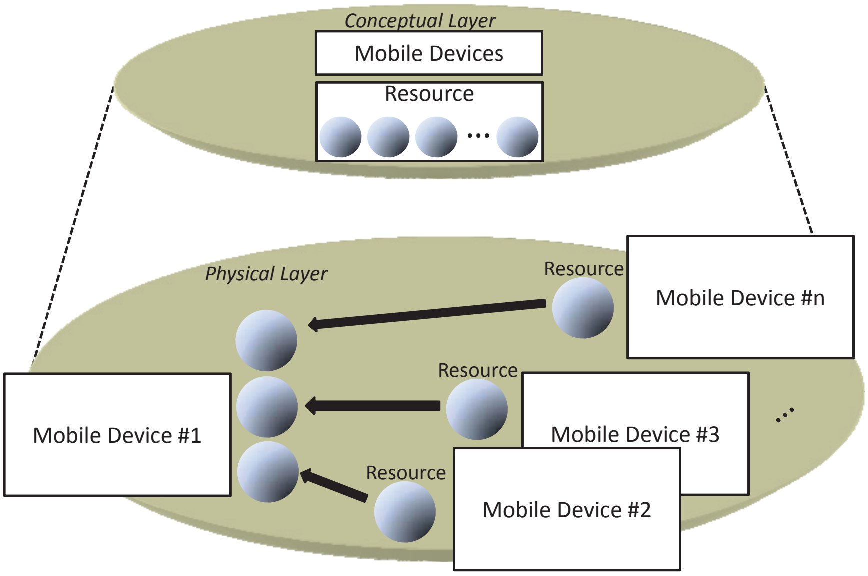



**Figure 1.** Overview of a Mutual Resource Exchanging Model in Mobile Computing [1][3]

## 2. Wave Pattern Detection and Hybrid Learning

Wave pattern detection is key and fundamental functionality in digital devices. Figure 2 shows a typical wave pattern detection among mobile devices using several detector. This method is for the integrated environment of detection methods by using several detectors that have their own features.

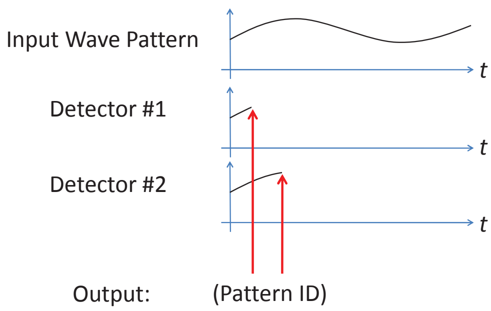



**Figure 2.** Wave Detector among Mobile Devices using Several Detector

By using several detector, accuracy (precision) and speed is more effective than cases using a single detector. Learning mechanisms among several detector improves the effectiveness of wave detection. Figure 3 shows a typical case of wave detection using three detectors environment. There are various kinds of detector, having features about accuracy and speed. Hybrid learning is effective for integrating various kinds detector, shown in Figure 4. In this case Detector #1 is fast but not so accurate.

Detector #2 is fast enough. Detector #3 is not fast but accurate. If we integrate them, we can detect wave patterns fast and accurate. Moreover, if we learn the integrated detection results for Detector #1, we are able to use learning results only by Detector #1. After the learning (Figure 4), we can detect wave patterns fast and accurate only by Detector #1.

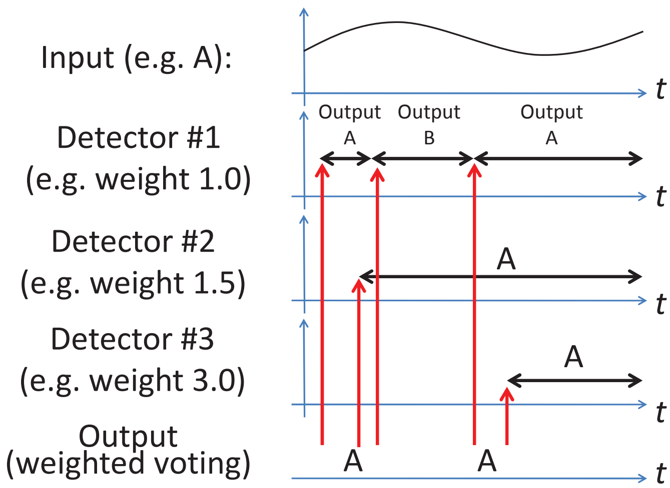



**Figure 3.** Hybrid Learning of Wave Detection using Several Detector (before)

## 3. Application of Wave Detection and Hybrid Learning for Energy Saving

We have many research activities on energy saving[11]. Especially, there are effective visualization methods for electricity consumptions. There are two critical problems. When we must decrease energy consumption of a certain whole building, it is difficult to decide a certain device and decrease its electricity. The second problem is that it is difficult to know which device can be switched off or decrease its power in advance. It means difficulty for knowing device profiles in advance. If we can visualize electricity consumptions for each device, we need to categorize their patters for stable (e.g. network router), periodic transition (e.g. refrigerator), focusing specific duration of time (e.g. microwave oven, cooking heater), depended on surrounded environment (e.g. air conditioner, heating appliance), and control them manually. A typical application to energy saving for decreasing total energy consumption is executed by following steps[2][5].

- Step-1: Configure watt-meters for target devices, setup remote-controllers for each of them, and define the priority table such as Table 1.
- Step-2: Collect time-series data of devices, and detect them by this method.
- Step-3: Switch off or decrease electricity consumption by remote-controllers.
- Step-4: Capture total electricity consumption and check if requirements are satisfied.
- Step-5: If it is not satisfied, go to Step-2.

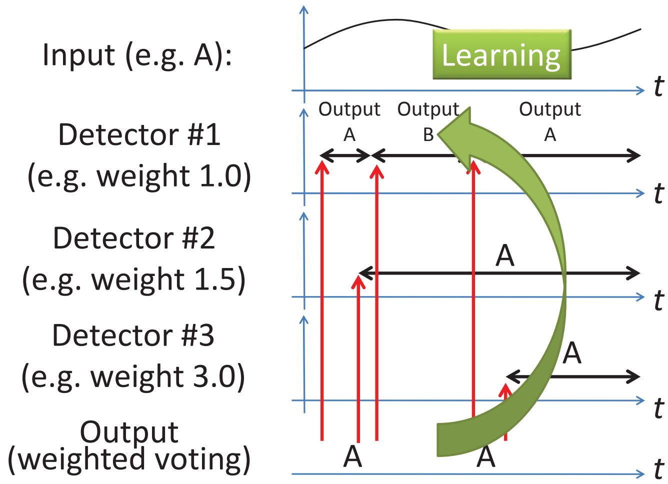



**Figure 4.** Hybrid Learning of Wave Detection using Several Detector (after)

In this application, it is assumed to apply this method to a room or a whole building with many devices. These devices are already used, so it is impossible to replace them to new devices with electricity controllers. It is necessary to keep legacy devices without any hardware replacement. By applying this method, we do not need any hardware replacement. We just attach watt-meters and setup remote-controllers. We have many wave pattern analysis (e.g. [10]), but the controlling electricity is important.

**Table 1.** Example of Priority Table in this Application

| # | Priority | Description |
|---|---|---|
| 1 | High+ | Never switched off (e.g. medical device) |
| 2 | High | Never switched off if power source is available (e.g. communication device) |
| 3 | Programmable | Flexible to decrease electricity (e.g. air conditioning) |
| 4 | Low+ | Can be switched off after notification (e.g. AC/DC adaptor, game device, refrigerator) |
| 5 | Low | Can be switched off (e.g. Lighting during daytime) |

## 4. Applications of Wave Detection and Hybrid Learning

The first of additional applications wave detection and hybrid learning is healthcare. Healthcare is directly important for us. And these days it is easy to implement healthcare application [4] by weight meters, blood pressure meters, electrocardiograms, and motion sensors. It is essential to detect normal and abnormal from various status of human being for healthcare applications. This method is applicable to time-series data analysis such as electrocardiograms and motion sensors. It enables us to detect biophysical status fast and precisely.

The second of applications wave detection and hybrid learning is social media. In the environment of M2M (Machine to Machine) [9], it is required to detect important situation from sensor data more fast and more precisely than machine-to-human environment. This environment is more complicated so it is effective to apply this method, because this method covers wide areas.

As a next step of this field, we can apply this detection method to the combination [6] of P2P (peer-to-peer) and social network. The peer-to-peer device interconnection using social network is effective to analyze personal information because it is robust for privacy protection. Private information is appropriate to treat outside of the cloud computing environment, and peer-to-peer architecture is a key for such circumstances. Actual human-human relationships that are commonly presented in the social network could be introduced to securely communicate confidential information among personal devices.

## 5. Conclusion

In this paper, an application for hybrid learning of wave pattern detection has been presented. By this model, flexible and elastic usability can be implemented on mobile devices in mobile computing environment. The detection accuracy and speed will be improved by exchanging learning results mutually among integrated detection methods. Hybrid learning of wave pattern detection enables us to elastic and intelligent learning among devices. Wave pattern detection is fundamental functionality in digital devices because it is a basis of Analog-Digital and Digital-Analog converting. As future work, algebra formalization can be applied for this model. Also, qualitative and quantitative analysis of this model is important for the evaluation. Furthermore, inclusion of context modelling[7] and cross-cultural affairs[8] are also future work.

## Acknowledge

I would like to thank Prof. Norihiro Ishikawa and Naoki Ishibashi (Komazawa University) for his valuable comments to this paper.

## References

[1] Naofumi Yoshida: Mutual Resource Exchanging Model in Mobile Computing and its Application to Collective Intelligence 3D Movies, Proceedings of the 24th International Conference on Information Modelling and Knowledge Bases EJC 2014, pp.486-493, 2014.

[2] Naofumi Yoshida, Naoki Ishibashi, Norihiro Ishikawa: A Hybrid Device Profile Detection Method and Its Applications, 38th Annual IEEE Computer Software and Applications Conference Workshops (COMPSACW 2014), pp.558 - 560, Vasteras, Sweden, 21-25 July, 2014.

[3] Naofumi Yoshida: A Mutual Resource Exchanging Model in Mobile Computing and its Applications to Universal Battery and Bandwidth Sharing, Proceedings of the 23rd European-Japanese Conference on Information Modelling and Knowledge Bases (EJC2013), pp.282-287, Nara, Japan, June 2013.

[4] Naofumi Yoshida, Daigo Matsubara, Naoki Ishibashi, Nobuo Saito, Norihide Ishikawa, Hikaru Takei, Shoichi Horiguchi: A Health Management Sercice by Cell Phones and its Usability Evaluation, Information Processing Society of Japan Transactions on Consumer Devices & Systems, Vol.2, No.1, pp.28-37, Mar. 2012.

[5] Naofumi Yoshida, Naoki Ishibashi, Masaki Minami, Satoshi Washio, Daigo Matsubara, Nobuo Saito, Norihiro Ishikawa: A Hybrid Device Profile Detection Method and Its Application to Saving Electricity, Multimedia, Distributed, Cooperative, and Mobile Symposium 2012(DICOMO2012), 2012.
[6] Naoki Ishibashi, Naofumi Yoshida, Masaki Minami, Satoshi Washio, Norihiro Ishikawa, and Nobuo Saito: "Machine-Machine Communications Using Relationships in Social Media," The 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet 2012), Vol.4, pp.2688--2691., 2012.
[7] Anneli HEIMBURGER, Yasushi KIYOKI, Tommi KARKKAINEN, Ekaterina GILMAN, Kim KYOUNG-SOOK and Naofumi YOSHIDA: On Context Modelling in Systems and Applications Development, Information Modelling and Knowledge Bases, XXII, 17 pages, May 2011.
[8] Anneli HEIMBURGER, Shiori SASAKI, Naofumi YOSHIDA, Teijo VENALAINEN, Petri, LINNA, Tatjana WELZER, Cross-Cultural Collaborative Systems: Towards Cultural Computing, Information Modelling and Knowledge Bases, Vol. XXI, pp.403-417, 2010.
[9] N. Ishikawa, H. Sumino, T. Kato, J. Hjelm, S. Murakami K. Kitagawa and N. Saito: Mobile Peer-to-Peer Computing for Next Generation Distributed Environments: Advancing Conceptual and Algorithmic Applications (Edited By Boon-Chong Seet): Chapter XVII (Peer-to-Peer Networking Platform and Its Applications for Mobile Phones), Information Science Reference, IGI Global, 2009.
[10] Greveler, Ulrich, Benjamin Justus, and Dennis Loehr. "Multimedia content identification through smart meter power usage profiles." Computers, Privacy and Data Protection, 2012.
[11] Toni Mastelic, Ariel Oleksiak, Holger Claussen, Ivona Brandic, Jean-Marc Pierson, Athanasios V. Vasilakos: Cloud Computing: Survey on Energy Efficiency, Journal ACM Computing Surveys (CSUR), Volume 47 Issue 2, January 2015.

## Impressum




| | |
|---|---|
| **Title:** | Proceedings of the 25th International Conference on Information Modelling and Knowledge Bases (EJC 2015) |
| **Editors:** | Tatjana Welzer, University of Maribor, Slovenia<br>Marko Hölbl, University of Maribor, Slovenia<br>Yasushi Kiyoki, Keio University, Japan<br>Bernhard Thalheim, Christian-Albrechts University at Kiel, Germany<br>Hannu Jaakkola, Tampere University of Technology (Pori), Finland |
| **Technical Editors:** | Lili Nemec Zlatolas, University of Maribor, Slovenia<br>Naofumi Yoshida, Komazawa University, Japan |
| **Page count:** | 324 |
| **Edition:** | 1st |
| **First published:** | 2015 |
| **Publisher:** | University of Maribor, Faculty of Electrical Engineering and Computer Science |
| **Impression:** | 50 USB flash drive units |

ISBN 978-961-248-486-6

9 789612 484866


Not for sale