



# **The 3rd CUBIST Workshop**

**CUBIST-WS-13**

Simon Andrews, Frithjof Dau (Eds.)



Simon Andrews, Frithjof Dau (Eds.):

## **The 3rd CUBIST Workshop**

**CUBIST-WS-13**



## Preface

This volume contains the papers accepted to the third CUBIST workshop.

CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) is a research project funded by the European Commission under the 7th Framework Programme of ICT, topic 4.3: Intelligent Information Management. The project started in October 2010. CUBIST follows a best-of-breed approach that combines essential features of Semantic Technologies, Business Intelligence and Visual Analytics. CUBIST aims to

- persist federated data in a semantic Data Warehouse; a hybrid approach based on a BI enabled triple store,
- and to provide novel ways of applying visual analytics in which meaningful diagrammatic representations based on Formal Concept Analysis will be used for depicting the data, navigating through the data and for visually querying the data.

As one can see from this description, CUBIST requires expertise from a variety of research fields, which cannot be provided by a single research organization. For this reason, the CUBIST workshop series which addresses the main topics of CUBIST has been set up. The workshops aim to provide a forum for both research and practice for CUBIST-related topics and technologies in order to facilitate interdisciplinary discussions. The first CUBIST workshop was held in conjunction with the 19th International Conference on Conceptual Structures (ICCS 2011), which was held on 25 - 29 July 2011 at the University of Derby, United Kingdom, and attracted submissions from outside the CUBIST consortium. The second workshop was held in conjunction with the 10th International Conference on Formal Concept Analysis (ICFCA 2012) which was held on 6 – 10 May 2012 at the University of Leuven, Belgium. This year's workshop is held in conjunction with the 11th International Conference on Formal Concept Analysis (ICFCA 2013), which took place May 21- May 24 in Dresden, Germany. We are proud that again the workshop received and accepted submissions from outside the CUBIST consortium, which indicates that the project and the workshop are addressing contemporary topics of interest to researchers in the fields. In total we had seven submissions, and six of the submissions were accepted.

We, the chairs, want to express our appreciation to all authors of submitted papers and to the program committee members for their work and valuable comments.

May 2013, Simon Andrews and Frithjof Dau

## **CUBIST-WS-13 Organization**

### **Chairs**

Simon Andrews (Sheffield Hallam University, UK)

Frithjof Dau (SAP AG, Germany)

### **Program Committee**

Axel Schröder (SAP AG, Germany)

Cassio Melo (Centrale Recherche S.A. (CRSA) - Laboratoire MAS, France)

Constantinos Orphanides (Sheffield Hallam University, UK)

Emre Sevinc (Space Applications Services NV, Belgium)

Honour Nwagwu (Sheffield Hallam University, UK)

Katja Pfeiffer (SAP AG, Germany)

Kenneth McLeod (Heriot-Watt University, UK)

Marie-Aude Aufaure (Centrale Recherche S.A. (CRSA) - Laboratoire MAS, France)

Simon Polovina (Sheffield Hallam University, UK)

Yuri Kudryavcev (PMSquare, Australia)

## Table of Contents

JURAJ MACKO. FORMAL CONCEPT ANALYSIS AS A FRAMEWORK FOR BUSINESS INTELLIGENCE TECHNOLOGIES II .....	1
ANDREW TAYLOR, KENNETH MCLEOD AND ALBERT BURGER. SEMANTIC VISUALISATION OF GENE EXPRESSION INFORMATION .....	10
FRITHJOF DAU. AN IMPLEMENTATION FOR FAULT TOLERANCE AND EXPERIMENTAL RESULTS .....	21
RICHARD FALLON AND SIMON POLOVINA. REA ANALYSIS OF SAP HCM; SOME INITIAL FINDINGS .....	31
HONOUR NWAGWU. EVALUATING AND ANALYZING INCONSISTENT RDF DATA IN A SEMANTIC DATASET: EMAGE DATASET .....	44
CONSTANTINOS ORPHANIDES AND GEORGE GEORGIOU. FCAWARE-HOUSE, A PROTOTYPE ONLINE DATA REPOSITORY FOR FCA .....	54





# Formal Concept Analysis as a Framework for Business Intelligence Technologies II

Juraj Macko

Division of Applied computer science  
Dept. Computer Science  
Palacky University, Olomouc  
17. listopadu 12, CZ-77146 Olomouc  
Czech Republic  
email: {juraj.macko}@upol.cz

**Abstract.** Formal concept analysis (FCA) with measures can be seen as a framework for Business Intelligence technologies. In this paper we introduce new ideas about an OLAP cube. We take a focus on a high-dimensional OLAP cube reduction and on a hierarchy of attributes in an OLAP cube.

## 1 Introduction

This paper continues with results proposed in [9] and for more details we will refer on it. The paper is structured as follows: In "Preliminaries" the fundamentals of FCA with measures are described, the formal definition of the *OLAP cube* is shown. "Compressing High-Dimensional OLAP Cube Using FCA With Measures" shows an efficient reduction of an OLAP space using FCA with measures. In "Attribute Hierarchy In OLAP And In FCA With Measures" we discuss different types of hierarchies in OLAP. This paper is supplemented with comprehensive examples. The final part summarizes the results.

## 2 Preliminaries

An input dataset for FCA is a formal context, which is a relation between the set of objects  $X$  and the set of attributes  $Y$ , is denoted by  $\langle X, Y, I \rangle$  where  $I \subseteq X \times Y$ . The concept forming operators  $()^\uparrow$  and  $()^\downarrow$  are defined as  $A^\uparrow = \{y \in Y \mid \text{for each } x \in X : \langle x, y \rangle \in I\}$  and  $B^\downarrow = \{x \in X \mid \text{for each } y \in Y : \langle x, y \rangle \in I\}$ . A formal concept of the formal context  $\langle X, Y, I \rangle$  is denoted by  $\langle A, B \rangle$ , where  $A \subseteq X$  and  $B \subseteq Y$ .  $\langle A, B \rangle$  is a formal concept iff  $A^\uparrow = B$  and  $B^\downarrow = A$ . The set  $A$  is called an extent and the set  $B$  an intent. A set of all formal concepts of  $\langle X, Y, I \rangle$  is denoted by  $\mathcal{B}(X, Y, I)$  and equipped with a partial order  $\leq$  forms a concept lattice of  $\langle X, Y, I \rangle$ .

**Definition 1 (Measure of Object and Attribute [9] ).** A Measure of the object is mapping  $\Phi : X \rightarrow \mathbb{R}^+$  and a Measure of the attribute is mapping  $\Psi : Y \rightarrow \mathbb{R}^+$ .

**Definition 2 (Value of Extent and Intent [9] ).** The Value of extent is mapping  $v : A_{\mathcal{B}(X,Y,I)} \rightarrow \mathbb{R}^+$  defined as  $v(A) = \odot_{x \in A} \Phi(x)$ , where  $\odot$  is either the symbol for the sum  $\Sigma$  (the "sum" operation) or the symbol for cardinality  $|A|$  or the arbitrary aggregation function  $\Theta$ .  $A$  is an extent of the formal concept  $\langle A, B \rangle \in \mathcal{B}(X, Y, I)$ . Similarly, the value of the intent is mapping  $w : B_{\mathcal{B}(X,Y,I)} \rightarrow \mathbb{R}^+$  defined as  $w(B) = \odot_{y \in B} \Psi(y)$ , where  $B$  is an intent of the formal concept  $\langle A, B \rangle \in \mathcal{B}(X, Y, I)$

The Database table is a relation  $r$  on the relation scheme  $R = \{A_1, A_2 \dots, A_n\}$  defined as a set of mappings  $\{t_1, t_2 \dots, t_m\}$  from  $R$  to  $\mathcal{D}$  where  $\mathcal{D}$  is a set of all  $D$  - domains of attributes  $A$ ,  $n$  is the number of the columns and  $m$  the number of rows in a database table (see [4]). Domains in  $\mathcal{D}$  are divided into the two groups:  $H_k \in \mathcal{H}$ - dimensions and  $M_s \in \mathcal{M}$  - measures, where  $k \in [1; |\mathcal{H}|]$ ,  $s \in [1; |\mathcal{M}|]$  and  $M_s \subseteq \mathbb{R}^+$ .

**Definition 3 (OLAP Cube space, OLAP Cube [9]).** The space for the OLAP cube is a cartesian product  $C = L^{H_1} \times \dots \times L^{H_k} \times \dots \times L^{H_{|\mathcal{H}|}}$ , where  $L = \{0, 1\}$ . The OLAP cube is a mapping  $\sigma : C \rightarrow \mathbb{R}^+$  and is defined as  $\sigma(h_1, \dots, h_n) = \odot_{i=1}^m t_i(M_s)$  such that  $\{t_i(A_j)\} \supseteq h_j$  for all  $j \in [1; |\mathcal{H}|]$ , where the symbol  $\odot$  stands for the sum operator  $\Sigma$ , the cardinality operator  $||$  or the arbitrary aggregation operator  $\Theta$  and  $|\mathcal{H}|$  is the number of OLAP cube dimensions.

### 3 Compressing High-Dimensional OLAP Cube Using FCA With Measures

In the previous paper [9] we have shown, that FCA with measures can be seen as a generalized OLAP. OLAP uses data which are organized in dimensions. As a direct consequence is, that the scaled attributes (using a nominal scale [1]) from one domain are mutually exclusive. FCA with values enables to analyze the data which are not organized in dimensions, thus those which are independent (see the example with cars and components taken from [9] shown in Table 4). This fact means, that we can work with a relational (binary) data as well. When the attributes with a binary domain are used, usually there is a relatively big amount of such attributes. It implies a high-dimensional OLAP cube. Recall from [9], that the size of an OLAP cube is  $(|H_1| + 1) \times \dots \times (|H_{|\mathcal{H}|}| + 1)$ . The expression "+1" means, that using the domain  $H_1 = \{BMW, SKODA, FIAT\}$  we consider such situation, when no attribute is selected. In a binary case we have two possibilities only (an attribute is selected or not), so a space of such cube will be  $2^{|\mathcal{H}|}$ , where  $|\mathcal{H}|$  is a number of domains (all domains are binary in this case). Hence, the space of the OLAP cube is exponential wrt. number

of attributes. FCA with measures enables to compress such exponential space. Consider  $Y$  as a set of attributes in FCA. Number of formal concepts (which contains intents, closed sets of attributes) is usually significantly lower than a powerset  $2^Y$ , because a real dataset is usually sparse. Using FCA with measures we can replace OLAP cube with a concept lattice with values and we do not lose any information comparing to OLAP. This compression can be used also for the attributes with a many-valued domain. In Figure 1 the example of such compression is shown. From the database table (i), OLAP cube is computed with the space  $(3 + 1) \times (2 + 1) = 12$  cells (ii). In (iii) the formal context (using the nominal scaling) is shown and finally in (iv) the concept lattice is depicted. The concept lattice has only 10 concepts with the values (1 trivial concept is just technical, with no value). Two OLAP cube cells (in (ii) are highlighted using gray color) are missing in the compressed concept lattice. Consider the well known dataset "Mushroom", which contains 23 original attributes (22 + 1 class considered as an attribute) where a cardinality of the domains is between 2 and 12. Using a formula for the OLAP space we get  $7,36 \times 10^{16}$  of cells in the OLAP cube. Comparing to the amount of the formal concepts, which is  $2,39 \times 10^{05}$  we get the space reduced approximately by  $10^{12}$ . In the Table 1 we can see the original OLAP spaces comparing to the reduced ones using five well-known datasets (see the highlighted items with a significant space compression).

TradeMark	Country	Price in 000 EUR
BMW	Germany	30
BMW	France	35
SKODA	Germany	20
SKODA	France	25
FIAT	France	13

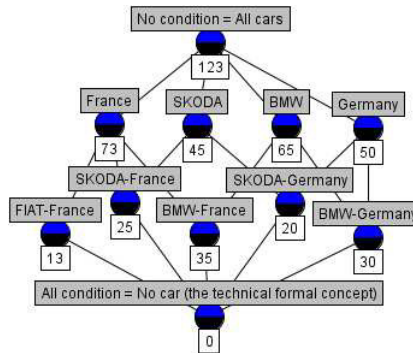
(i) Database

	All countries	France	Germany
All trademarks	123	73	50
FIAT	<b>13</b>	13	<b>0</b>
SKODA	45	25	20
BMW	65	35	30

(ii) OLAP Cube

Car nr.	BMW	SKODA	FIAT	Germany	France	Price in 000 EUR
1	x			x		30
2	x			x	x	35
3		x		x		20
4		x		x	x	25
5			x	x		13

(iii) Formal context with measures



(iv) Concept lattice with extent values

**Fig. 1.** OLAP space compression - example

	Dataset				
	Mushrooms	Adults	Cars	Wine	Tic-tac-toe
Nr. of original attributes (before scaling)	23	14	6	13	29
Nr. of formal attributes (after scaling)	119	124	25	68	29
Nr. of objects	8 124	32 561	1 728	178	958
original OLAP space (nr. of cells in OLAP cube)	7,36E+16	7,64E+10	4,00E+04	5,22E+10	7,86E+05
compressed OLAP space (nr. of formal concepts)	2,39E+05	1,06E+06	1,26E+04	2,54E+04	5,95E+04
<b>compression ratio</b> (compressed / original)	<b>3,25E-12</b>	<b>1,39E-05</b>	3,16E-01	<b>4,86E-07</b>	7,57E-02

**Table 1.** OLAP space compression, source of datasets: <http://fcarepository.com>

Remark 1: Not the whole OLAP cube is presented to a user and also not the whole lattice with values is presented to user. The data are stored using the different way, but the presentation to user can be the same (e.g. using pivot tables, pivot charts or other repost).

Remark 2: Not the whole OLAP cube is materialized in a real application. However a compression ratio is calculated from the whole OLAP cube comparing to compressed one (e.g. see [13]).

In [13] there were presented another solution how to compress OLAP cube, thus using a Dwarf Cube. The authors of [13] claim, that a Petabyte 25-dimensional cube was shrunk this way to a 2.3GB Dwarf Cube. For a detailed description of Dwarf cube we refer to [13]. Here we only compare our approach using a toy example from [13]. In Table 2 data for OLAP are shown and in Table 3 a comparison of tuples is shown for two OLAP representations (Dwarf and FCA with measures). In Table 3 we can see, that Dwarf Cube contains more tuples than

Store	Customer	Product	Price
S1	C2	P2	70
S1	C3	P1	40
S2	C1	P1	90
S2	C1	P2	50

**Table 2.** Data for OLAP

concept lattice. But this is only a toy example. Our hypothesis is, that FCA with measures contains a minimal possible amount of all non-redundant tuples for the OLAP cube compression. A formal proof as well as an experimental study will be part of our future research.

	Tuple in the Dwarf Cube	Tuple derived from an intent
1	$\langle S1, C2, P2 \rangle$	$\langle S1, C2, P2 \rangle$
2	$\langle S1, C2, ALL \rangle$	
3	$\langle S1, C3, P1 \rangle$	$\langle S1, C3, P1 \rangle$
4	$\langle S1, C3, ALL \rangle$	
5	$\langle S1, ALL, P1 \rangle$	
6	$\langle S1, ALL, P2 \rangle$	
7	$\langle S1, ALL, ALL \rangle$	$\langle S1, ALL, ALL \rangle$
8	$\langle S2, C1, P1 \rangle$	$\langle S2, C1, P1 \rangle$
9	$\langle S2, C1, P2 \rangle$	$\langle S2, C1, P2 \rangle$
10	$\langle S2, C1, ALL \rangle$	$\langle S2, C1, ALL \rangle$
11	$\langle ALL, ALL, P1 \rangle$	$\langle ALL, ALL, P1 \rangle$
12	$\langle ALL, ALL, P2 \rangle$	$\langle ALL, ALL, P2 \rangle$
13	$\langle ALL, ALL, ALL \rangle$	$\langle ALL, ALL, ALL \rangle$

**Table 3.** Dwarf Cube vs. FCA with measures

## 4 Attribute Hierarchy In OLAP And In FCA With Measures

In the paper [9] we claim, that FCA with measures is a generalization of OLAP cube. This claim however excludes the case, when a hierarchy of attributes in OLAP is defined. Attributes in a dimension can be split into smaller parts, e.g. in the dimension *Date* we can consider the hierarchy *Year* > *Month*. In FCA with measures we can consider the dimension "Date" and it can be nominally scaled into attributes *Year* and *Month* as well. Consider the following Figure 2. When the original table (i) is scaled (ii) and formal concepts are computed, we get the concepts with the intent  $\{Jan\}$  and  $\{Feb\}$ . Such intents can generally be used e.g. for analyzing the seasonality, however using the hierarchy *Year* > *Month* such intent is not interesting (in this case it is a total amount of all cars sold in January regardless of the year). All other formal concepts are reasonable (i.e. total amount in one year or total amount in one month of the particular year). There are two possibilities how to deal with such problem. The first approach

Obj.	Date
1	Jan, 2011
2	Feb, 2011
3	Jan, 2012
4	Feb, 2012

(i)  
Original data

Obj.	2011	2012	Jan	Feb
1	×		×	
2	×			×
3		×	×	
4		×		×

(ii)  
Nominal scaling

Obj.	2011	2012	2011 Jan	2011 Feb	2012 Jan	2012 Feb
1	×		×			
2	×			×		
3		×			×	
4		×				×

(iii)  
Hierarchical scaling

**Fig. 2.** Hierarchy of attributes

is just to scale the original data from (i) using a hierarchy (iii). Such scaling directly enables to avoid undesired formal concepts with intents such as  $\{Jan\}$  and  $\{Feb\}$  (Note: The undesired formal concept with the intent *all attributes* technically remains just to form a lattice).

Another option is to use AD formulas proposed in [11, 12]. An *AD formula* over a set  $Y$  of attributes is an expression  $A \sqsubseteq B$ , where  $A, B \subseteq Y$ .  $A \sqsubseteq B$  is true in  $K \subseteq Y$  if whenever  $A \cap K \neq \emptyset$ , then  $B \cap K \neq \emptyset$ . For a given set  $T$  of AD formulas over  $Y$  and a formal context  $\langle X, Y, I \rangle$  we get the concept lattice constrained by  $T$ , which is denoted by  $\mathcal{B}_T(X, Y, I)$ . Such lattice consists of formal concepts of  $\langle X, Y, I \rangle$  in which all AD formulas from  $T$  are true. For more details we refer to [11, 12]. In our example we can use AD formula  $\{Jan, Feb\} \sqsubseteq \{2011, 2012\}$ , which means: whenever we have a month in an intent of a formal concept (here *Jan* or *Feb*), we need also to have a year in intent (here 2011 and 2012). In other words, a year is hierarchically higher than a month. Constraining the original concept lattice by AD formula, undesired formal concepts will be avoided. A formal concept analysis with measures using AD formula can be seen as a generalization of OLAP with hierarchies.

In [14] there were presented some types of a hierarchy used in OLAP cube, but not all types of a hierarchy can be defined using AD formula. In this paper a preliminary results are presented (all examples of hierarchies are taken from [14]) :

1. simple hierarchies (represented by a tree)
  - (a) symmetric hierarchy:
$$\{Department A\} \sqsupseteq \{Category 1\}, \{Department A\} \sqsupseteq \{Category 2\},$$

$$\{Category 1\} \sqsupseteq \{Product 1\}, \{Category 1\} \sqsupseteq \{Product 2\}, \{Category 2\} \sqsupseteq$$

$$\{Product 3\}, \{Category 2\} \sqsupseteq \{Product 4\}$$
  - (b) asymmetric hierarchy:
$$\{bank X\} \sqsupseteq \{branch 1\}, \{bank X\} \sqsupseteq \{branch 2\}, \{bank X\} \sqsupseteq$$

$$\{branch 3\}, \{branch 1\} \sqsupseteq \{agency 11\}, \{branch 1\} \sqsupseteq \{agency 12\},$$

$$\{branch 3\} \sqsupseteq \{agency 31\}, \{branch 3\} \sqsupseteq \{agency 32\}, \{agency 11\} \sqsupseteq$$

$$\{ATM 111\}, \{agency 11\} \sqsupseteq \{ATM 112\}$$
  - (c) generalized hierarchy:
$$\{area A\} \sqsupseteq \{branch 1\}, \{area A\} \sqsupseteq \{branch 2\}, \{branch 1\} \sqsupseteq \{class 1\},$$

$$\{class 1\} \sqsupseteq \{profession A\}, \{class 1\} \sqsupseteq \{profession B\}, \{profession B\} \sqsupseteq$$

$$\{customer X\}, \{profession B\} \sqsupseteq \{customer Y\}, \{branch 1\} \sqsupseteq \{sector 1\},$$

$$\{sector 1\} \sqsupseteq \{type A\}, \{sector 1\} \sqsupseteq \{type B\}, \{type B\} \sqsupseteq \{customer Z\},$$

$$\{type B\} \sqsupseteq \{customer K\}$$
2. non-strict hierarchy:
$$\{division A\} \sqsupseteq \{Section 1, Section 2, Section 3\}, \{Section 1, Section 2, Section 3\} \sqsupseteq$$

$$\{employee X\}$$

This approach we can use also on for attributes which are not organized in dimensions by telling which group of independent attributes is more important than other group. In the example with cars (see the Tables 4 and 5) there are attributes Air Conditioning (*AC*), Airbag (*AB*), Antilock Braking System (*ABS*), Tempomat (*TMP*), Extra Guarantee (*EG*) and Automatic Transmission (*AT*). We can say, that  $\{AB, ABS\}$  are more important (because of security) than  $\{AC, TMP, EG, AT\}$  (which are used just for a higher comfort). AD formula in this case is  $\{AB, ABS\} \sqsubseteq \{AC, TMP, EG, AT\}$ , which means, that we

will care about values of formal concepts (e.g. the *Total Price*) only for such cars, which possess at least one of the security attributes *AB* or *ABS* (in the Table 5 labeled by \*).

	1. <i>AC</i>	2. <i>AB</i>	3. <i>ABS</i>	4. <i>TMP</i>	5. <i>EG</i>	6. <i>AT</i>	$\Phi(X)$ = Price in EUR
Car1	x	x					16 000
Car2		x	x	x			12 000
Car3		x	x	x	x		14 000
Car4	x			x	x		16 000
Car5	x				x		12 000
Car6	x	x	x				12 000
Car7		x	x	x			12 000
Car8			x				14 000
Car9							16 000
Car10		x					12 000
Car11		x		x			12 000
Car12	x	x	x	x	x	x	14 000
Car13		x		x			16 000
Car14	x	x		x	x	x	16 000
Car15			x		x		14 000
Car16	x	x					12 000
Car17	x	x					12 000
Car18		x	x	x			16 000
Car19			x				16 000
Car20	x	x	x	x			14 000
$\Psi(Y)$ = Price in EUR	1 000	500	800	600	250	100	

**Table 4.** The formal context of the cars, the additional components, the price of the car and the price of the component [9]

## 5 Conclusion

FCA with measures as a new area is just on the beginning. Based on our preliminary research it appears, that FCA with measures can significantly reduce a space of OLAP cube. FCA with measures can also be used as a generalization of OLAP even different hierarchy of attributes is included. In the future research we will focus on the detailed experimental research, where we will compare other reducing techniques of an OLAP cube space with our approach.

## References

1. Ganter B., Wille R.: *Formal Concept Analysis. Mathematical Foundations*. Springer, Berlin, 1999.

	Extent Cars	Intent Components	Extent value Total Price	secure cars
1	$X$ - all cars	$\emptyset$	278 000	
2	{2, 3, 4, 7, 11, 12, 13, 14, 18}	{ $TMP$ }	128 000	
3	{3, 4, 5, 12, 14, 15, 20}	{ $EG$ }	100 000	
4	{1, 4, 5, 6, 12, 14, 16, 17, 20}	{ $AC$ }	124 000	
5	{1, 2, 3, 6, 7, 10, 11, 12, 13, 14, 16, 17, 18, 20}	{ $AB$ }	190 000	*
6	{2, 3, 6, 7, 8, 12, 15, 18, 19, 20}	{ $ABS$ }	138 000	*
7	{3, 4, 12, 14}	{ $EG, TMP$ }	60 000	
8	{4, 5, 12, 14, 20}	{ $EG, AC$ }	72 000	
9	{2, 3, 7, 11, 12, 13, 14, 18}	{ $AB, TMP$ }	112 000	*
10	{3, 12, 14, 20}	{ $AB, EG$ }	58 000	*
11	{3, 12, 15, 20}	{ $ABS, EG$ }	56 000	*
12	{1, 6, 12, 14, 16, 17, 20}	{ $AC, AB$ }	96 000	*
13	{2, 3, 6, 7, 12, 18, 20}	{ $AB, ABS$ }	94 000	*
14	{4, 12, 14}	{ $AC, TMP, EG$ }	46 000	
15	{3, 12, 14}	{ $TMP, EG, AB$ }	44 000	*
16	{12, 14, 20}	{ $EG, AB, AC$ }	44 000	*
17	{2, 3, 7, 12, 18}	{ $AB, ABS, TMP$ }	68 000	*
18	{3, 12, 20}	{ $ABS, EG, AB$ }	42 000	*
19	{6, 12, 20}	{ $ABS, AC, AB$ }	40 000	*
20	{12, 14}	{ $AB, EG, AC, TMP, AT$ }	30 000	*
21	{3, 12}	{ $AB, EG, TMP, ABS$ }	28 000	*
22	{12, 20}	{ $AB, AC, EG, ABS$ }	28 000	*
23	{12}	$Y$ - all components	14 000	*

**Table 5.** The formal concepts with the extent value [9]

2. Codd E.F., Codd S.B., and Salley C.T.: Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate *Codd & Date* (1993)
3. Wang Z., Klir G.: *Generalized measure theory*, Springer, New York, 2009
4. Maier D.: *The theory of relational databases*, Computer Science Press, Rockville, 1983
5. Kuznetsov S. D., Kudryavtsev A.: A mathematical model of the OLAP cubes, Programming and Computer Software, 2009, Vol. 35, No. 5, pp. 257–265. Pleiades Publishing, Ltd., 2009.
6. Calvo T., Kolesárová A., Komorníková M., Mesiar R. *Aggregation operators: Properties, classes and construction methods* Aggregation Operators: New Trend and Applications, p. 3-106 , Eds: Calvo T., Mayor G., Mesiar R., Physica Verlag, (Heidelberg 2002)
7. Belohlavek R., Vychodil V.: *Background Knowledge in Formal Concept Analysis: Constraints via Closure Operators*. ACM SAC 2010, 1113–1114.
8. Belohlavek R., Vychodil V.: *Formal concept analysis with constraints by closure operators*. In: H. Scharfe, P. Hitzler, and P. Ohrstrom (Eds.): Proc. ICCS 2006, Lecture Notes in Artificial Intelligence 4068, pp. 131-143, Springer-Verlag, Berlin Heidelberg, 2006.
9. Macko J.: *Formal Concept Analysis as a Framework for Business Intelligence Technologies*. In: F. Domenach, D.I. Ignatov, and J. Poelmans (Eds.): ICFCA 2012, LNAI 7278, Springer, Heidelberg, 2012, pp. 195-210.
10. Kanovsky J., Macko J.: *ConSeQueL - SQL Preprocessor Using Formal Concept Analysis with Measures* CUBIST 2012 workshop



11. Belohlavek R., Sklenář V.: Formal concept analysis constrained by attribute-dependency formulas. In: B. Ganter and R. Godin (Eds.): ICFCA 2005, *Lect. Notes Comp. Sci.* **3403**, pp. 176–191, Springer-Verlag, Berlin/Heidelberg, 2005.
12. Belohlavek R., Vychodil V.: Formal concept analysis with background knowledge: attribute priorities. *IEEE Trans. Systems, Man, and Cybernetics, Part C* Volume 39 Issue 4, July 2009, pp. 399–409. DOI: 10.1109/TSMCC.2008.2012168.
13. Sismanis Y., Deligiannakis A., Roussopoulos N., Kotidis Y.: Dwarf: Shrinking the petacube *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, p.464-475
14. Malinowski E., Zimányi E. : Hierarchies in a multidimensional model: From conceptual modeling to logical representation. *Data & Knowledge Engineering* 59.2 (2006): 348-377

# Semantic visualisation of gene expression information

Andy Taylor<sup>1</sup>, Kenneth McLeod<sup>1</sup>, and Albert Burger<sup>1,2</sup>

<sup>1</sup> Heriot-Watt University, Edinburgh, EH14 4AS, UK

ajt17 | kcm1 | a.g.burger@hw.ac.uk

WWW home page: <http://www.macs.hw.ac.uk/bisel>

<sup>2</sup> MRC Human Genetics Unit, Edinburgh, EH4 2XU, UK

**Abstract.** The Edinburgh Mouse Atlas of Gene Expression (EMAGE) publishes the results of gene expression experiments on the mouse embryo. Whilst this resource uses visual mechanisms to display the result of a single experiment, it currently provides no technique for the visual navigation of data. Ideally, a semantic visual navigation mechanism would exist. Our work focuses on trying to understand the requirements for such a mechanism. To this end, a prototype solution (based on sunburst visualisations) is being built. This paper presents the prototype and reports on the initial feedback from users.

**Keywords:** semantic visualisation, gene expression information, big data, usability

## 1 Introduction

Due to high throughput experiments and information technology, there are now many big data resources available for biologists. It is increasingly clear that users require assistance when navigating and searching this information. One way of providing support is through appropriate visualisation. Visualisation can be used to help users explore data or help users interpret information.

While such problems are widespread, this work focuses exclusively on a single use case. The Edinburgh Mouse Atlas of Gene Expression (EMAGE) [11] publishes online gene expression information for the developmental mouse. EMAGE provides a variety of tools to help users search the data; however, it does not enable users to visually navigate the data. This work aims to address this gap.

Although there are many different technologies and techniques that might be applicable (see Section 3), this work focuses on the application of sunburst visualisations. The goal is to develop a prototype application that will inform the creation of a real world solution. As such, the objective is not to create a powerful application that perfectly meets the needs of EMAGE's user community, but instead to understand the requirements of that group and learn how we may satisfy those needs in future.

This paper reports on the use case and motivation for this work, outlines the prototype currently under development and relates the initial feedback from potential end users.

Section 2 describes the use case in which this work is set, and reviews the existing visualisations that are employed within the use case. Section 3 considers related work before Section 4 describes how we chose the visualisation to focus on. Sunburst visualisations are discussed in Section 5. Subsequently, Section 6 reviews the customisation of Sunburst visualisations for use within the current use case. Section 7 features a discussion and conclusions are presented in Section 8.

## 2 EMAGE - a mouse atlas of gene expression

This paper focuses on a biological resource as its use case. That resource, EMAGE [11], publishes online gene expression information for the developmental mouse.

A gene is a unit of instructions that provides directions for one essential task, i.e., the creation of a protein. Gene expression information describes whether or not a gene is expressed (active) in a location. Such information allows biologists to discover relationships between genes, in particular when genes are active in the same location.

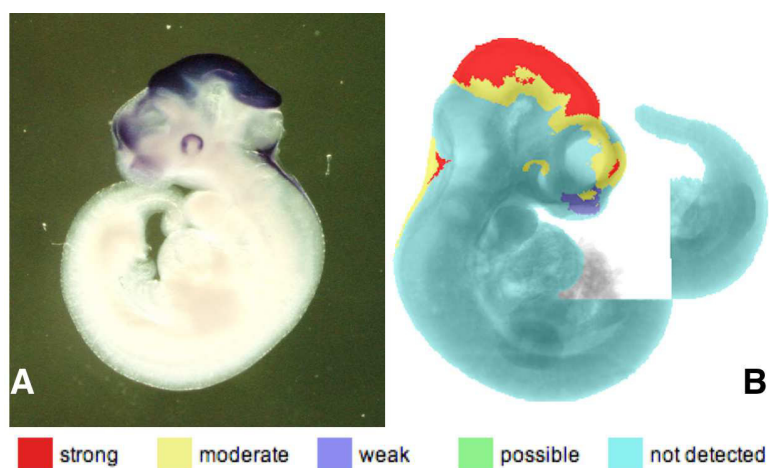
The gene expression information is obtained by experimenting on a mouse embryo. Each embryo corresponds to a point in time of the *developmental mouse*: the mouse from conception until birth. The time window is split into 26 distinct periods called Theiler Stages (TS). Each stage has its own anatomy, and corresponding anatomy called EMAP [10]. Moreover, there are a number of 3D models representing different stages of the developmental mouse.

The result of an in-situ hybridization (ISH) experiment is documented as an image displaying an area of a mouse (from a particular TS) in which some subsections of the mouse are highly coloured, as depicted in Figure 1(A). Areas of colour indicate that the gene is expressed in that location. Furthermore, the image provides some indication of the level (strength) of expression: the more intense the colour, the stronger the expression. Results are analysed manually under a microscope. A human expert determines in which tissues the gene is expressed, and at what level of expression. Strength information is described using natural language terms such as strong, moderate, weak or present. For example, the gene *bmp4* is strongly expressed in the future brain from TS15. These statements are so-called *textual annotations*. Textual annotations represent the structured version of a subset of the original unstructured data (e.g., Figure 1(A)).

Textual annotations capture whatever information the researcher wishes to present. They may be incomplete (if the researcher is only interested in the heart, (s)he will not create textual annotations for the brain) or documented at a high granularity (the textual annotation will report the gene being expressed in the heart rather than the sub-component in which it is actually found).

In an attempt to provide a more complete and precise set of results, experimental images (e.g. Figure 1(A)) can be mapped onto 3D models of the mouse creating the *spatial annotation* depicted in Figure 1(B). These spatial annota-

tions are normally generated by EMAGE, whilst the textual annotations are produced by the researchers who performed the experiment.



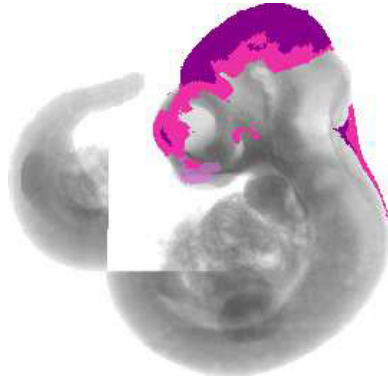
**Fig. 1.** A sample image of (A) an experimental result, and (B) associated spatial annotation, from the EMAGE database.

This work uses the textual annotations.

## 2.1 Existing EMAGE visualisations

The current range of visualisations employed within EMAGE concentrate on displaying gene expression information within the context of its location within the mouse. For example, Figure 1 (B) clearly shows where the gene *Oxt2* is expressed within TS17 (this image is taken from experiment EMAGE:1411). An alternative representation of the same information can be seen in Figure 2. In this depiction, the intensity of the colour indicates the level of expression; the greater the intensity the higher the level of expression, e.g., purple = strong and pink = moderate.

Both Figures 1 and 2 use the idea of displaying the location of gene expression on a standardised model of the mouse, with the level of expression indicated through the use of colour. Moreover, Figures 1 and 2 present the result of a single experiment (EMAGE:1411) as a spatial annotation. Accordingly, they display all the gene expression information that was obtained from the experiment. In contrast, the textual annotations report all the gene expression information that the researcher wanted to report. Currently, there is no mechanism to visualise the results of only the textual annotations. Nor is there a mechanism to represent the results (textual or spatial annotations) of multiple experiments. Instead, a user must read multiple lists of textual annotations or look at multiple spatial



**Fig. 2.** An alternative representation of Figure 1 (b) in which the intensity of the colour is used to represent the level of expression, i.e., purple = strong, pink = moderate and mauve = weak.

annotation images (e.g. Figure 1(B)). Within this work, we seek to provide a visualisation tool, that enables the above tasks for the user.

EMAGE does not provide any visualisation to summarise the expression information across time or between multiple genes. For instance, it is not possible to see how the expression information for *Oxt2* compares to that of the gene *Bmp4*. Nor is it possible to see the way in which the location of *Bmp4* changes over time as the mouse develops.

### 3 Related work

Increasingly data-sets within the life sciences are approaching sizes which are not manageable by humans and as such usable visualisations are vital in helping human researchers navigate this data [6].

Accordingly, the life sciences are a very active area for visualisation. For example Heat Maps have been used to demonstrate Anisotropic Flocking Behaviour [1], Hive Plots [8] have been used to visualise gene co-expression and a variety of Partition Graphs [7] have been used to visualise ancestry.

Phylogenetic trees, e.g. [2], are the visualisations traditionally used to represent differences between species, and then to analyse those differences statistically.

Circos [5] was designed for the visualisation of genomic data, in particular the relationships between different cancer genomes. The CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) project uses Circos style diagrams as one mechanism for the representation of association rules. Cytoscape [9] takes this a step further, and provides a mechanism to visualise networks. This can be used to represent any biological network, for example protein interaction networks [4].

Cytoscape uses Force-directed Graphs to visualise networks. If the network can be simplified into a tree structure, Sunburst visualisations [13] or Icicle visualisations can be used instead. An icicle [15] is a sunburst transformed from polar to cartesian co-ordinates.

## 4 EMAGE focus group

Resources do not allow for the creation of an application with a wide array of visualisation techniques. Accordingly, it was decided to focus on a single visualisation; however, which one?

This question was answered by the EMAGE focus group: a small number of EMAGE staff who were assembled to guide this work. In the first meeting the focus group was presented with a range of different visualisation techniques and asked which they preferred. Their choice would be the mechanism featured in the application.

The favourite option was a sunburst visualisation (e.g. Figure 3; the reasons for this are simple. Firstly, the anatomy is a tree structure therefore a visualisation technique designed to display tree structures is highly appropriate. Secondly, unlike force-directed graphs (e.g., Cytoscape) there are no edges within the sunburst. This means that there are no crossed edges and no question of how to best layout the nodes. Similar results have been reported elsewhere, e.g., [13].

## 5 Sunburst visualisations

Essentially a sunburst (e.g., see Figure 3) takes information organised within a tree structure, and displays the tree structure in a radial layout. Assuming the information is organised as a tree, as opposed to a graph, no organisational data is lost.

The size and position of the blocks within the sunburst are used to indicate the structure, and organisation, of the data. Data attribute values are presented by colouring the nodes.

The centre of a sunburst diagram is the root node of the tree, with children of the root node being the first layer of blocks in the sunburst. Children sit directly around in the next layer of the sunburst, and so on, until the leaf nodes are reached at the edge of the diagram.

It is possible to zoom into a node by double clicking it. This causes the parent node, of the clicked node, to become the central node of the updated sunburst, and thus gives more prominence to the node of interest by making it larger. To move up a level, the user should double click on the central node. In this manner, a user may navigate up and down the internal tree structure.

## 6 Sunbursts for gene expression

EMAGE uses the EMAP anatomy (and corresponding ontology) to describe the anatomical space of the mouse embryo. The full anatomy is a DAG (Directed



**Fig. 3.** A generic sunburst diagram: each block represents a node within a tree. The order and position of the blocks recreates the structure of the tree.

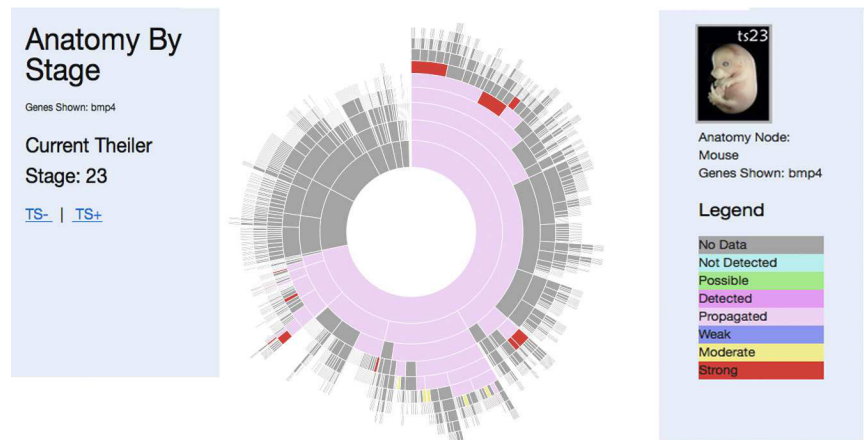
Acyclic Graph); however, for computational and presentational reasons a simplified tree representation exists too. It is the EMAP tree that features within our sunburst diagrams. Because the mouse anatomy changes greatly over time, there is one sunburst for every Theiler Stage (Figure 4 shows the sunburst for TS23 with the expression profile of *Bmp4*).

Each node in the diagram represents a tissue in the mouse apart from the root node, which is the mouse itself. The node's colour is used to present the level of expression for that node, the colour scheme chosen mimics the original EMAGE colour scheme (see Figure 1).

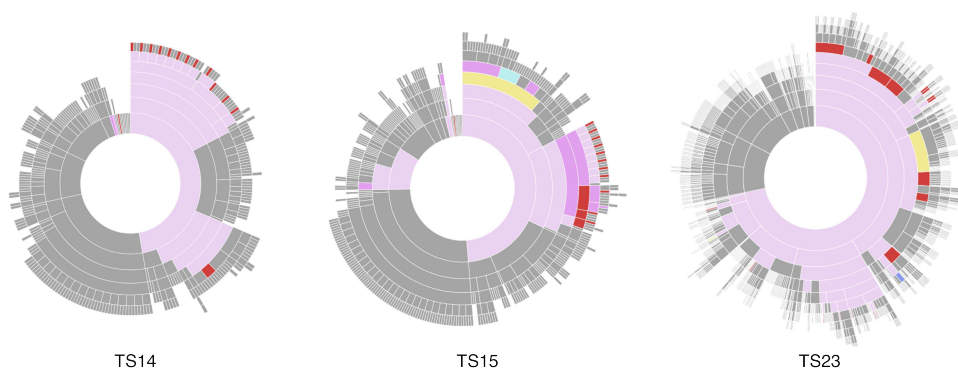
When the user moves the mouse over a node, the box in the top right corner is updated to show the name of the tissue that node represents. Additionally, if the node contains gene expression information that is displayed.

The top left corner (Figure 4) provides a navigation box, that allows the user to move from stage to stage. In this way the user can watch the expression profile change over time. Alternatively, the same effect can be achieved by showing multiple sunbursts side by side: Figure 5 contains sunbursts for *Ssh* in stages TS14, 15 and 23.

Moreover, it is possible to show the expression profile for multiple genes in the same sunburst providing a visual way to determine which locations the genes have in common. Figure 6 shows the expression profile for over 50 genes. The genes are listed in the box in the top left corner. The coloured nodes (in the sunburst) indicate where at least one of the genes is expressed. If the mouse is moved over one of the coloured nodes the box in the top right corner is updated to show the tissue name and the list of genes expressed there (with associated



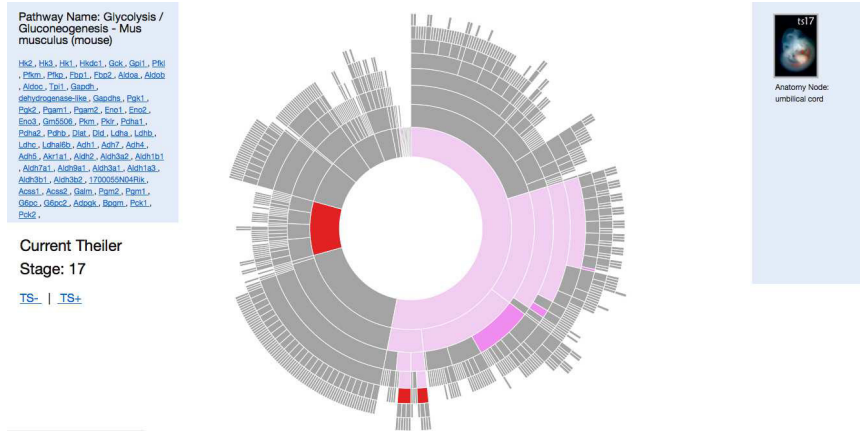
**Fig. 4.** Expression profile of *Bmp4* in TS23. Top left box allows navigation through other stages. Box in top right provides details of whichever tissue (node) the mouse is hovering over.



**Fig. 5.** Depicting the expression profile of the gene *Ssh* across time.



strengths). If multiple genes are expressed, at different strengths, in the same node the highest level of expression is used to colour the node.



**Fig. 6.** Depicting the combined expression profiles of over 50 genes at TS17. Top left box lists the genes featured. Top right box shows the tissue the mouse is hovering over, and lists the genes expressed there (with associated level of expression).

## 7 Discussion

The basic functionality currently exists as a live system, which has been shown to the focus group. One aspect that is popular with users is the ability of the sunburst to visualise the results of multiple experiments within a single diagram. Currently, it is not possible to do this within EMAGE. One of the main disadvantages of this approach is that it only visualises textual annotations, which are often less precise/complete than spatial annotations. Moreover, it only presents textual annotations from EMAGE when it could show annotations from complementary resources (e.g., GXD [12]) too.

Whilst initial feedback was broadly positive, testing revealed some flaws with the controls: it is too difficult to change the gene(s) shown in the sunburst. Once corrected, the prototype will be shown to the focus group and their feedback incorporated. When the focus group are happy with the tool a more thorough evaluation, with a wider set of participants, will take place.

As EMAGE is one of the three CUBIST use cases it is worthwhile comparing this prototype with the CUBIST dashboard (i.e., CUBIX [3]), which also uses sunbursts. Within our work, the sunburst is used to visualise the mouse anatomy and colour indicates where a gene is expressed. In contrast, CUBIX uses sunbursts to visualise the results of Formal Concept Analysis (e.g., [14]). Rather than tissues, the nodes of the CUBIX sunburst are “concepts” and thus represent a collection of entities, for example, tissues, genes and/or Theiler Stages.

Clearly, this is early work. There is much to be considered and reviewed. For example, currently our approach to handle time (change in Theiler Stage) is to display multiple sunbursts (one for each stage), it seems impossible to do anything else. Sunburst visualisations are based on a tree structure (in this case the tree represents the anatomy at a single Theiler Stage). Whilst it would be possible to merge multiple trees by adding a new root node, this would likely lead to a diagram too complex for real world use. However, if the requirement for a sunburst is removed, it may be possible to display all the relevant information in a single diagram by switching away from a tree-based representation.

Sunburst visualisations are ideal for presenting tree structures; however, if the structure is a graph they lose information. Whilst the EMAP mouse anatomy is a directed acyclic graph (DAG) it has a simplified tree representation too. Therefore it is easy to represent the mouse anatomy as a tree, and thus sunburst. Yet, in doing so information is lost. There are many different ways of organising the mouse anatomy contained within the DAG; the “correct” way depends entirely on context. In our approach only one organisation is presented. Although this representation is the most commonly used it is not always ideal. The solution may be to offer a series of different trees/sunbursts, one for each different organisation.

Despite all the obvious problems with the sunburst, it has one attribute that is vital for this application: it is easy to understand. There is no point in creating a presentation mechanism that captures all the data and relationships if the EMAGE community find it too complex to use. A balance must be struck between presenting as much information as possible and providing a tool that users are comfortable with. Understanding where this balance lies is one of the key tasks of the current prototype.

## 8 Conclusion

During this paper we have presented a discussion of the ways in which sunburst visualisations can be used to present meaningful depictions of gene expression profiles. These profiles show where a gene is active, and how active that gene is. Sunburst visualisations enable a summary of the profiles to be displayed, and can present changes in the profile over time or an aggregation of multiple gene profiles. The latter is a powerful tool that enables a biologist to quickly determine what genes have in common; something that cannot be achieved with existing visualisation mechanisms in our use case. By allowing the gene and Theiler Stage to be changed, the sunbursts allow a user to visually browse the gene expression information, another feature that is currently missing from the use case.

The visualisations are being developed in partnership with biological experts from the EMAGE database of mouse gene expression. An EMAGE focus group enables us to appropriately target and test our work. Once complete, we aim to undertake a summative evaluation in order to accrue knowledge that may be applied to the development of a real world tool.

## Acknowledgements

This work is part of the CUBIST project (Combining and Uniting Business Intelligence with Semantic Technologies), funded by the European Commission's 7th Framework Programme of ICT under topic 4.3: 'Intelligent Information Management'.

The authors are thankful for the advice and guidance provided by members of the EMAGE team, and for the comments provided by the anonymous reviewers.

## References

1. M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1231–1237, 2008.
2. A. Boc, B. Diallo Alpha, and V. Makarenkov. T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic Acids Research*, 40(W1):W5373–W579, 2012.
3. Melo C., Orphanides C., M<sup>c</sup>Leod K., Aufaure M-A., Andrews S., and Burger A. A conceptual approach to gene expression analysis enhanced by visual analytics. In *Proceedings of the 20th ACM symposium on applied computing*. Coimbra, Portugal, March 2013.
4. Agapito G., Guzzi P.H., and Cannataro M. Visualization of protein interaction networks: problems and solutions. *BMC Bioinformatics*, 14(Suppl 1):S1, 2013.
5. Krzywinski M. I., Schein J.E., Birol I., Connors J., Gascoyne R., Horsman D., Jones S.J., and Marra M.A. Circos: An information aesthetic for comparative genomics. *Genome Research*, 2009.
6. Goecks J., Nekrutenko A., Taylor J., and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.
7. Baroni M., Semple C., and Steel M. A framework for representing reticulate evolution. *Annals of Combinatorics*, 8(4):391–408, 2005.
8. Krzywinski M., Birol I., Jones S., and Marra M. Hive plots - rational approach to visualizing networks. *Briefings in Bioinformatics*, 2011.
9. Cytoscape 2.8: new features for data integration and network visualization. M.E. Smoot and K. Ono and J. Ruscchinski and P.l. wang and T. Ideker. *Bioinformatics*, 27(3):431–432, 2011.
10. Baldock R and Davidson D. *Anatomy ontologies for bioinformatics: principles and practise*, chapter The Edinburgh Mouse Atlas, pages 249–265. Springer Verlag, 2008.
11. Venkataraman S., Stevenson P., Yang Y., Richardson L., Burton N., Perry T.P., Smith P., Baldock R.A., Davidson D.R., and Christiansen J.H. EMAGE - Edinburgh Mouse Atlas of Gene Expression: 2008 update. *Nucleic Acids Research*, 36(1):D860–D865, 2007.
12. J. H. FInger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse gene expression database (gxd): 2011 update. *Nucleic Acids Research*, 30(suppl 1):D835–D841, 2011.

13. J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of human-computer studies*, 53(5):663–694, 2000.
14. S. Andrews and K. McLeod. Gene co-expression in mouse embryo tissues. In *Proceedings of the 1st CUBIST (Combining and Uniting Business Intelligence with Semantic Technologies) workshop*, 2011.
15. Y. Yang, P. Keller, and P. Liggesmeyer. Visual approach facilitating the importance analysis of component fault trees. In *SAFECOMP Workshops*, pages 486–497, 2012.

# An Implementation for Fault Tolerance and Experimental Results

Frithjof Dau, SAP AG

**Abstract.** Even for small or medium sized contexts, the corresponding concept lattice soon becomes too large to be nicely displayed and understood, which renders a naive application of FCA for visual analytics challenging. Concept lattices depict all information of the underlying formal context. In certain settings, one can consider the context to be noisy or incomplete, and moderate and meaningful changes of the context such that the corresponding lattices significantly shrink in size and complexity are in those settings permissible. In this paper, such an approach is taken. The formal definitions for the approach are given, and using an implementation, several examples are provided which show the usefulness of the approach.

## 1 Introduction

CUBIST<sup>1</sup> is an EU funded research project with an approach that leverages BI to a new level of precise, meaningful and user-friendly analytics of data by following a best-of-breed approach that combines essential features of Semantic Technologies, Business Intelligence and FCA-based Visual Analytics. A main goal of CUBIST is the provision of novel Visual Analytics based on meaningful diagrammatic representations. The Visual Analytics part of CUBIST is complementing traditional BI-means by utilizing Formal Concept Analysis (FCA) for analyzing the data in a triple store: whereas traditional BI focuses on the visualization and analysis of numerical, quantitative data (“show me the numbers”), FCA is a well-known theory of data analysis which allows objects to be conceptually clustered with respect to a given set of attributes and then visualize the (lattice-ordered) set of clusters, e.g. by means of Hasse-diagrams, thus is best suited for visualization and analysis of structural dependencies in the data.

The starting point of FCA is a so-called formal context  $(O,A,I)$  consisting of a set  $O$  of formal objects, a set  $A$  of formal attributes, and an incidence-relation  $M \subseteq O \times A$  between the formal objects and attributes. The corresponding concept lattice preserves all the information of the formal context, and it might grow exponential in size compared to  $O$  and  $A$  (one might have  $2^{\max\{|O|,|A|\}}$  many formal concepts). Hence even for data sets of moderate size, the corresponding lattices become very soon too complex to be visualized and thus incomprehensible. This is a well-known problem of FCA and has already been addressed with different approaches. Iceberg-lattices, introduced by Gerd Stumme, show only the topmost part of a concept lattice, thus reducing the

---

<sup>1</sup> [www.cubist-project.eu](http://www.cubist-project.eu)

number of concepts shown to the user [8,9]. It is not only the number of concepts which often render the visualization of lattices complicated: Even with good layout algorithms, the Hasse-diagram of a lattice often contains a significant number of crossing edges, which is the main problem for the visual comprehension of graphs. Cassio Melo et al. present in [4,5] different metrics which allow for each node in a concept lattice to choose a uniquely given upper neighbor. By doing so, one can transform a lattice into a tree (with loss of information), which in turn can be better displayed compared to lattices. Finally, Boulicaut et al consider in [2,7] “noisy” formal contexts and directly change the incidence relation  $I$  in order to reduce the number of formal concepts. For CUBIST, in line with the approach of Boulicaut et al, Simon Andrews has provided in [1] an example with real-data of one the CUBIST use cases which clearly shows the advantage of this approach.

In this paper, we follow up the notion of considering contexts to be noisy and to extend the incidence relation in order to simplify the derived concept lattices. In the next section, we provide the mathematical definition on how we change the incidence relation. As described in the following section, this approach has been implemented into a small FCA-tool in order to show its effectiveness. A thorough, though artificial example is given in the next section, before we turn to examples based on real data out of the CUBIST project.

The authors want to stress that apart from providing the mathematical definitions, this paper focuses on the implementation and the experimental results. A thorough mathematical investigation is out of scope of this paper and will be provided in a different paper which is currently in preparation.

## 2 Formal Definition

Let a formal context  $(O,A,I)$  be given. Our basic idea is to understand the context to be noisy or incomplete, and we are aiming at adding crosses to  $I$  such that the derived lattice becomes smaller (hence, easier to read and understand). For an object  $o \in O$  and an attribute  $a \in A$ , our starting point is the well-known equation

$$oIa \Leftrightarrow o^{\text{II}} \subseteq a^{\text{I}} \Leftrightarrow a^{\text{II}} \subseteq o^{\text{I}} \Leftrightarrow o^{\text{I}} \rightarrow a \quad (1)$$

That is, we do not have  $oIa$  if some objects and/or some attributes violate (1). We define the sets of those objects and attributes as follows:

$$\begin{aligned} \text{Diff}_{\text{Obj}}(o,a) &:= \{ x \in O \mid x \in o^{\text{II}} \text{ and } x \notin a^{\text{I}} \} \text{ and} \\ \text{Diff}_{\text{Att}}(o,a) &:= \{ y \in A \mid y \in a^{\text{II}} \text{ and } y \notin o^{\text{I}} \} \end{aligned}$$

The straight-forward approach taken in this paper, is, roughly speaking, as follows: the bigger  $\text{Diff}_{\text{Obj}}(o,a)$  or  $\text{Diff}_{\text{Att}}(o,a)$ , the less we like to add an additional cross between  $o$  and  $a$ . In order to do so, it is possible to take only  $\text{Diff}_{\text{Obj}}(o,a)$  into account (this approach is reasonable in contexts having a small amount of attributes, but a large amount of objects, which is standard in traditional BI settings), or only  $\text{Diff}_{\text{Att}}(o,a)$ , or both. Moreover, we can either relate  $\text{Diff}_{\text{Obj}}(o,a)$  and  $\text{Diff}_{\text{Att}}(o,a)$  to the

overall sets of objects and attributes, i.e. taking a global approach, or relate them only to the concepts generated by  $o$  and  $a$ , i.e. taking a local approach. This gives rise to six different ways to measure “an approximate incidence measure” between  $o$  and  $a$  as follows:

$$\begin{aligned}
 G_{Obj}(o,a) &:= 1 - ( |Diff_{Obj}(o,a)| / |O| ) \\
 G_{Att}(o,a) &:= 1 - ( |Diff_{Att}(o,a)| / |A| ) \\
 G_{Obj,Att}(o,a) &:= 1 - 1/2 ( |Diff_{Obj}(o,a)| / |O| + |Diff_{Att}(o,a)| / |A| ) \\
 L_{Obj}(o,a) &:= 1 - ( |Diff_{Obj}(o,a)| / |o^{II}| ) \\
 L_{Att}(o,a) &:= 1 - ( |Diff_{Att}(o,a)| / |a^{II}| ) \\
 L_{Obj,Att}(o,a) &:= 1 - 1/2 ( |Diff_{Obj}(o,a)| / |o^{II}| + |Diff_{Att}(o,a)| / |a^{II}| )
 \end{aligned}$$

**Fig 1: Six approaches to measure the incidence between objects and attributes**

For any of these incidence measures  $S \in \{ G_{Obj}, G_{Att}, G_{Obj,Att}, L_{Obj}, L_{Att}, L_{Obj,Att} \}$  we obviously have for all  $o,a$ :

$$0 \leq S(o,a) \leq 1 \quad \text{and} \quad S(o,a) = 1 \Leftrightarrow oIa \quad (2)$$

So, for a given threshold  $t$  with  $0 \leq t \leq 1$  we can naturally set

$$J_{S,t} := \{ (o,a) \mid S(o,a) \geq t \} \quad (3)$$

Due to (2), we obviously have  $s \leq t \Leftrightarrow J_{S,t} \subseteq J_{S,s}$  and  $J_{S,1} = I$

Finally, the local measure for objects corresponds to the confidence of the according association rule, which is a nice rationale for that measure:

$$conf(o^I \rightarrow a) = \frac{|(o^I \cup \{a\})^I|}{|o^{II}|} = 1 - \frac{|o^{II}| - |(o^I \cup \{a\})^I|}{|o^{II}|} = 1 - \frac{Diff_{Obj}(o,a)}{|o^{II}|} = L_{Obj}(o,a)$$

We will now exemplify (3) for  $S := G_{Obj}$  (and  $L_{Obj}(o,a)$ , which is for this example the same relation) with the following simple context:

	a1	a2	a3
o1		X	X
o2		X	X
o3	X	X	
o4	X	X	
o5	X	X	
o6	X	X	
o7	X	X	
o8	X	X	
o9	X	X	
o10	X		X

Please note the following:

- For  $o_{10}$ , we have  $o_{10}^I \rightarrow a_3$  and  $o_{10}^I \rightarrow a_1$ , but not  $o_{10}^I \rightarrow a_2$ . This implication is violated by exactly one object, namely  $o_{10}$  itself. If we consider a threshold of 0.9 (that is, as we have 10 objects in total, we set a cross between  $o$  and  $a$  iff the condition  $o^I \rightarrow a$  is violated by at most one object), then we should add a cross between  $o_{10}$  and  $a_2$ . Note that the high amount of objects which have  $a_1$  as attribute (i.e.  $o_3 - o_9$ ) is not relevant.
- For  $o_1$ , we have  $o_1^I \rightarrow a_3$  and  $o_1^I \rightarrow a_2$ , but not  $o_1^I \rightarrow a_1$ . This implication is violated by exactly two objects, namely  $o_1$  itself, and  $o_2$ . If we consider a

threshold of 0.8 (that is, now we set a cross between o and a iff the condition  $o^I \rightarrow a$  is violated by two or less objects), then we should add a cross between  $o_1$  and  $a_1$

- A similar consideration applies of course to  $o_2$ .

The approximate incidence measure  $S := G_{Obj}$  and the concept lattices for the incidence relations  $I = J_{S,1}$ ,  $J_{S,0.9}$  and  $J_{S,0.8}$  is depicted in the next figure.

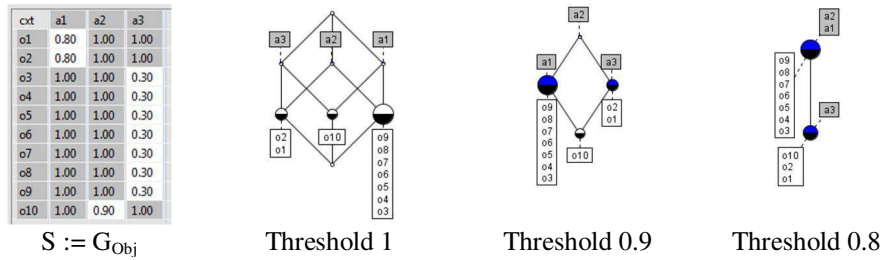


Fig 2: A simple example for  $G_{Obj}$

Generally, adding crosses to an incidence relation can decrease or increase the number of formal concepts, and one can hardly make statements on how the concept lattices change. For the herein defined extensions of the incidence relations, the situation is different. We start with a simple result which does not generally apply to extensions of incidence relations.

**Lemma:** Let  $(O,A,I)$  be a formal context, let  $S \in \{ G_{Obj}, G_{Att}, G_{Obj,Att}, L_{Obj}, L_{Att}, L_{Obj,Att} \}$  be an approximate incidence measure, let  $t$  with  $0 \leq t \leq 1$  be a threshold, and let  $J := J_{S,t}$ . Then, for all  $o_1, o_2 \in O$ ; we have  $o_1^I \subseteq o_2^I \Leftrightarrow o_1^J \subseteq o_2^J$ . Similarly, for all  $a_1, a_2 \in A$  we have  $a_1^I \subseteq a_2^I \Leftrightarrow a_1^J \subseteq a_2^J$ .

**Proof:** We only show the lemma for objects (i.e. the first implication), the proof for attributes is done analogously.

Let  $a \in A$  be an attribute. As we have  $o_1^I \subseteq o_2^I$  and hence  $o_2^{\text{II}} \subseteq o_1^{\text{II}}$ , we obtain  $\text{Diff}_{Obj}(o_2, a) \subseteq \text{Diff}_{Obj}(o_1, a)$  and  $\text{Diff}_{Att}(o_2, a) \subseteq \text{Diff}_{Att}(o_1, a)$ ; thus  $S(o_1, a) \leq S(o_2, a)$ . Particularly, we get  $o_1 J a \Leftrightarrow S(o_1, a) \geq t \Rightarrow S(o_2, a) \geq t \Leftrightarrow o_2 J a$ . Q.e.d.

As stated in the introduction, a thorough investigation of the approximate incidence relations will be the subject of a different paper. Here we only provide a remark that even though we can prove statements about the relationship between  $I$  and  $J_{S,t}$  which do not hold for  $I$  and any extension  $J \supseteq I$ , the dependencies between  $I$  and  $J_{S,t}$  have to be investigated further. For example, the author conjectured that we have  $P^J = P^{\text{III}}$  for all sets of objects  $P \subseteq O$ , but the context in Fig 3, together with the approximate incidence measure  $G_{Obj}(o,a)$ , shows that this equation does not generally hold. To see this, consider a threshold  $t=0.8$ ,  $P := \{o_0, o_1, o_2\}$ : we then have  $P^{\text{III}} = \{o_1, o_2, o_3, o_4\}$  and  $P^J = \{o_1, o_2, o_3, o_4, o_6, o_7, o_8\}$



cxt	a1	a2	a3	a4	a5	a7	a8
o0	.	X	.	.	.	.	.
o1	X	X	X	X	.	.	X
o2	X	X	X	.	.	.	.
o4	.	.	.	.	.	.	X
o5	X	X	X	X	.	X	.
o6	X	X	.	.	.	.	.
o7	.	.	.	.	X	.	.

cxt	a0	a1	a2	a3	a4	a5	a6	a7	a8
o0	1.00	0.85	1.00	0.85	0.85	0.85	1.00	0.85	0.85
o1	0.85	1.00	1.00	1.00	1.00	0.85	1.00	0.85	1.00
o2	0.57	1.00	1.00	1.00	0.85	0.57	0.71	0.71	0.71
o4	0.71	0.85	0.85	0.85	0.85	0.71	0.85	0.71	1.00
o5	0.85	1.00	1.00	1.00	1.00	0.85	0.85	1.00	0.85
o6	0.42	1.00	1.00	0.85	0.71	0.42	0.57	0.57	0.57
o7	1.00	0.85	0.85	0.85	0.85	1.00	0.85	0.85	0.85

Fig 3: Counterexample for  $P^J = P^{IJ}$

### 3 Implementation

We have extended the tool presented in [3] in order to implement the approach presented in this paper. The tool, initially developed to transform SPARQL-queries into formal contexts, can now load formal contexts (independently from SPARQL) and transform them according to (3). A partial screenshot is provided below.

cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.90
O3	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O4	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00
O5	1.00	0.90	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00
O6	1.00	0.80	1.00	0.60	0.90	0.90	0.90	0.80	0.80	
O7	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40
O8	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40
O9	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70
O10	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70

Fig 4: screenshot of a tool which implements approximate incidence relations

The tool allows to load a formal context (as a Burmeister file) and computes all six approximate incidence measures. Either the original context or one of the measures can be selected for display, as the next screenshot shows.

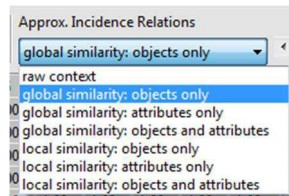


Fig 5: The different metrics as they appear in the tool

The slider allows to adjust the threshold  $t$ . A cell for an object  $o$  and an attribute  $a$  is marked grey iff  $S(o,a) \geq t$  (where  $S$  stands for the selected measure). One can now either store only the single derived context with the incidence relation  $J_{S,t}$ , or a set of derived contexts, where all thresholds 0.9, 0.91, ..., 0.99 (these thresholds are manually chosen and so far hardcoded) and all six measures are used (thus, 60 Burmeister files are stored).

## 4 An thorough, artificial example

In this section, we exemplify all measures and different thresholds with an artificial example. Let us consider the following formal context and its concept lattice:

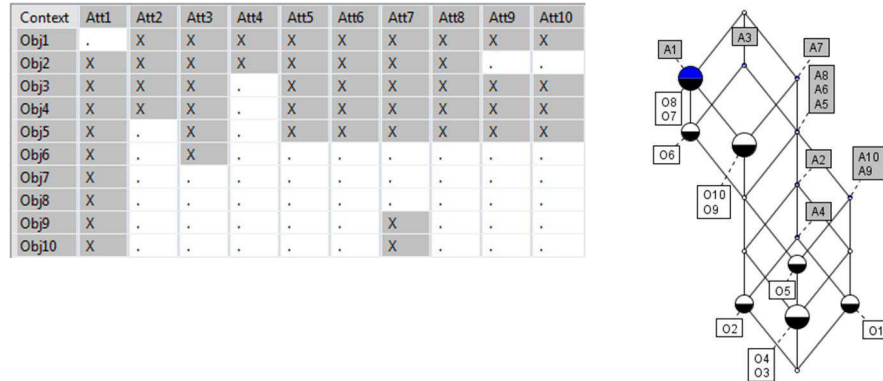


Fig 6: Example context and lattice for approximate incidence relations

For this context, we obtain the following values for the six different measures:

		Global										local										
Objects	cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
	O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.90	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00
	O3	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
	O4	1.00	1.00	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
	O5	1.00	0.90	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.66	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
	O6	1.00	0.80	1.00	0.60	0.90	0.90	0.90	0.90	0.80	0.80	0.80	1.00	0.60	1.00	0.20	0.80	0.80	0.80	0.80	0.60	0.60
	O7	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40	0.40	1.00	0.33	0.55	0.11	0.44	0.44	0.66	0.44	0.33	0.33
	O8	1.00	0.40	0.60	0.20	0.50	0.50	0.70	0.50	0.40	0.40	0.40	1.00	0.33	0.55	0.11	0.44	0.44	0.66	0.44	0.33	0.33
	O9	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70	0.70	1.00	0.50	0.66	0.16	0.66	0.66	1.00	0.66	0.50	0.50
	O10	1.00	0.70	0.80	0.50	0.80	0.80	1.00	0.80	0.70	0.70	0.70	1.00	0.50	0.66	0.16	0.66	0.66	1.00	0.66	0.50	0.50
Attributes	cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
	O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.80	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.71	0.71	
	O3	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00
	O4	1.00	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00
	O5	1.00	0.90	1.00	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.83	1.00	0.71	1.00	1.00	1.00	1.00	1.00	1.00
	O6	1.00	0.50	1.00	0.40	0.60	0.60	0.90	0.60	0.40	0.40	0.40	1.00	0.16	1.00	0.14	0.20	0.20	0.00	0.20	0.14	0.14
	O7	1.00	0.40	0.90	0.30	0.50	0.50	0.90	0.50	0.30	0.30	0.30	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	O8	1.00	0.40	0.90	0.30	0.50	0.50	0.90	0.50	0.30	0.30	0.30	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	O9	1.00	0.50	0.90	0.40	0.60	0.60	1.00	0.60	0.40	0.40	0.40	1.00	0.16	0.00	0.14	0.20	0.20	1.00	0.20	0.14	0.14
	O10	1.00	0.50	0.90	0.40	0.60	0.60	1.00	0.60	0.40	0.40	0.40	1.00	0.16	0.00	0.14	0.20	0.20	1.00	0.20	0.14	0.14
objs and atts	cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	cxt	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
	O1	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
	O2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.85	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.35	0.35	
	O3	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00
	O4	1.00	1.00	1.00	0.85	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.42	1.00	1.00	1.00	1.00	1.00	1.00
	O5	1.00	0.90	1.00	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.75	1.00	0.35	1.00	1.00	1.00	1.00	1.00	1.00
	O6	1.00	0.65	1.00	0.50	0.75	0.75	0.90	0.75	0.60	0.60	0.60	1.00	0.38	1.00	0.17	0.50	0.50	0.40	0.50	0.37	0.37
	O7	1.00	0.40	0.75	0.25	0.50	0.50	0.80	0.50	0.35	0.35	0.35	1.00	0.16	0.27	0.05	0.22	0.22	0.33	0.22	0.16	0.16
	O8	1.00	0.40	0.75	0.25	0.50	0.50	0.80	0.50	0.35	0.35	0.35	1.00	0.16	0.27	0.05	0.22	0.22	0.33	0.22	0.16	0.16
	O9	1.00	0.60	0.85	0.45	0.70	0.70	1.00	0.70	0.55	0.55	0.55	1.00	0.33	0.33	0.15	0.43	0.43	1.00	0.43	0.32	0.32
	O10	1.00	0.60	0.85	0.45	0.70	0.70	1.00	0.70	0.55	0.55	0.55	1.00	0.33	0.33	0.15	0.43	0.43	1.00	0.43	0.32	0.32

Fig 7: All six approximate incidence measures for the example

For the thresholds 0.9 (upper half of table) and 0.8 (lower half of table), the next figure provides for each measure S the corresponding concept lattice.

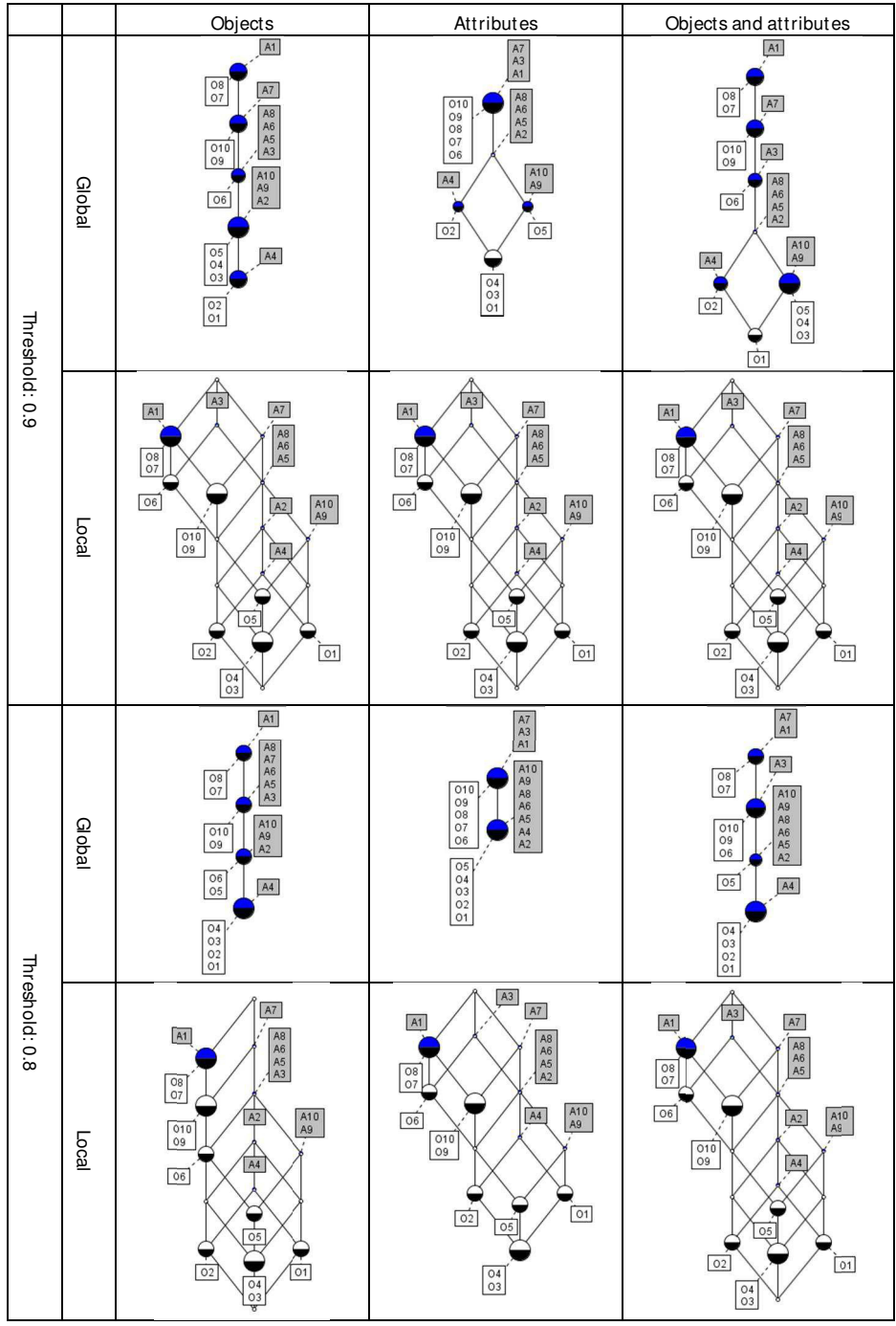
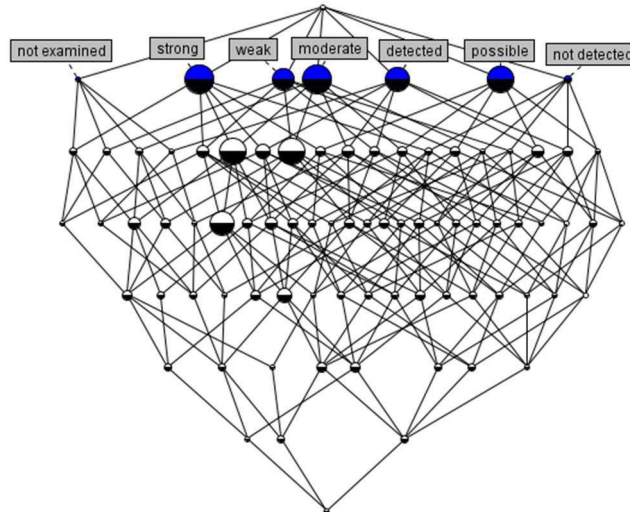


Fig 8: Concept lattices for thresholds 0.8 and 0.9 and all six measures

## 5 Heriot-Watt University Example

The example of this section is based on the Heriot-Watt University use case in the research project CUBIST, in particular, it is built upon a real data set, namely - embryo gene expression data. Gene expression information describes whether or not a gene is expressed (active) in a location. There is a fixed set of levels of expressiveness: A gene can in some location be detected, or more one can more precisely state that it is weakly, moderate or strongly detected. Moreover, it can be stated that a gene is not detected, that it is possible for the gene to be expressed, or that it is not examined. Based on public available data, we have generated a context where the objects are (names of) mouse genes (we have a total of 6613 formal objects in the context), the attributes are the seven different levels of expressiveness, and a cross is set between a gene and a level if that gene is detected in some location with that level of expressiveness. The resulting lattice, having 81 formal concepts, is depicted below.



**Fig 9: Genes and Levels of Expressiveness without approximation**

The lattice clearly shows a well-known challenge for FCA. Most of possible combinations of attributes are indeed instantiated objects in the context, often by few only, (the clarified context still contains 77 objects), which leads to an explosion of the number of concepts. Amongst the concepts, we have some formal concepts with large number of objects, but most concepts are rather small and render the lattice cluttered and hard to read. As we have many objects and only few attributes, it is sensitive to apply the global object-based measurement  $G_{Obj}$  to the context. With a (quite high) threshold of 97 already, the concept lattice is reduced significantly, as shown in Fig 10.

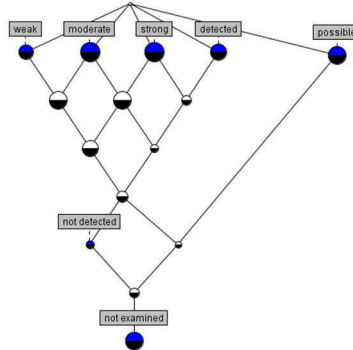


Fig 10: Genes and Levels of Expressiveness with approximation, threshold 0.97

## 6 Summary and next steps

In this paper, we have introduced a fault tolerance approach for FCA and shown promising experimental results.

Of course, first of all, this approach has to be thoroughly investigated from a formal point of view. It has to be proven that for each approximate incidence measure, the number of formal concepts decreases when the threshold is decreased. More specifically, it has to be investigated how for a given measure and two thresholds  $t_1 \geq t_2$  how the concept lattice for  $t_1$  can be mapped onto the concept lattice for  $t_2$ . It has to be scrutinized how the different measures relate to each other. And it has to be investigated how the approach taken in this paper relates to other research, most importantly the work of Boulicaut et al.

From a conceptual point of view, it has to be worked out on how derived concept lattices are to be understood, particularly how the relationship of derived lattices and association rules of the origin lattice is.

Next, from an implementational point of view, we have to elaborate on the algorithm which computes the measures, i.e. scrutinizing its complexity and possibly improving its performance (this is e.g. important for big data sets and real-time applications).

Finally, from an user interface point of view, it would be desirable to have a lattice visualization with a slider to adjust the threshold for a given measure such that adjusting the threshold with the slider leads to animated transitions between the derived lattices. Here the mappings which have been mentioned in the section about the formal investigations are likely to be helpful, and further mathematical investigations might be needed.

In the long run, assuming that the steps mentioned above have been undertaken, the approach taken in this paper is envisioned to bring FCA closer to business intelligence applications and visual analytics, where one needs easy-to-understand and in-

teractive visualizations for large amounts of data, and where a certain and controlled loss of information for the sake of simplified diagrams is permissible.

**Acknowledgement** This work is part of the CUBIST project (“Combining and Uniting Business Intelligence with Semantic Technologies”), funded by the European Commission’s 7th Framework Programme of ICT, under topic 4.3: Intelligent Information Management.

## 7 References

1. Andrews, S. and McLeod, K.: Gene Co-Expression in Mouse Embryo Tissues. In: Dau, F. (ed.) 1st CUBIST Workshop, at ICCS 2011, Derby, UK. CEUR Workshop Proceedings, Vol. 753, pp. 1-10. ISSN: 1613-0073
2. Besson, J., Robardet, C., Boulicaut, J.F.: Mining a New Fault-Tolerant Pattern Type as an Alternative to Formal Concept Discovery. Scharfe et al. ICCS. LNAI 4065. Springer (2006)
3. Dau, F.: Towards Scalingless Generation of Formal Contexts from an Ontology in a Triple Store In: Dau, F and Andrews, S: Proceedings of the second CUBIST workshop 2012. KULeuven press, 2012
4. Melo, C.A., Aufaure, M.-A., Le Grand, B. and Bezerianos, A.: Extracting and Visualizing Tree-like Structures from Concept Lattices In: 15th International Conference on Information Visualization (IV 2011). London, UK, 2011.
5. Melo, C.A., Aufaure, M.-A., Bezerianos, A. and Le Grand, B.: Cubix: A Visual Analytics Tool for Formal Concept Analysis. In: 23ième Conférence Francophone Sur l’IHM (IHM 2011) – Demo. Sophia-Antipolis, France, 2011.
6. Pensa, R. G., Boulicaut, J-F.: Towards Fault-Tolerant Formal Concept Analysis. In: Bani-dini, S., Manzoni, S. (eds.) AI\*IA 2005, LNAI 3673, pp. 212{223, Springer-Verlag, Berlin Heidelberg (2005
7. Pensa, G.R.; Boulicaut, J.F.: Towards fault-tolerant formal concept analysis. In: Congress of the Italian Association for Artificial Intelligence AI\* IA, 2005
8. Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhil, L.: Computing iceberg concept lattices with Titanic. Data & Knowledge Engineering, Volume 42, Issue 2, August 2002, Pages 189–222
9. Stumme, G., Taouil, R., Bastide, Y.: Conceptual clustering with iceberg concept lattices. In: Proc. of GI-Fachgruppentreffen Maschinelles Lernen’01, Universität Dortmund, 2001

# REA analysis of SAP HCM; some initial findings

Richard Fallon and Simon Polovina

Communication and Computing Research Centre (CCRC)

Sheffield Hallam University, UK S1 2NU

`richard.l.fallon@student.shu.ac.uk`

`S.Polovina@shu.ac.uk`

**Abstract.** This paper explores further the claim that the Transaction-Oriented Architecture (TOA) based on the principles of Resources, Events, Agents (REA) can enhance Enterprise Resource Planning (ERP) systems by providing a principled theoretical basis that can underpin ERP business process implementations. We provide details of some of our initial findings of the REA/TOA analysis which we carried out on the SAP Human Capital Management (HCM) module. Given that SAP is recognized as the dominant ERP system with over 50% of the market share, this technology is viewed as the representative case study technology for exploring the theory of REA in actual ERP systems. In particular O’Leary’s and Dunn et al.’s works are expanded upon, substantiating O’Leary’s findings that SAP was found to be consistent with REA in its database, semantic and structure orientations. Using SAP’s HCM module as the exemplar, two notable discoveries are made. These are namely (i) identifying that several anomalies exist in the underlying data model, and (ii) that there are many more REA entities than previously discovered by Dunn et al. Through the SAP HCM exemplar it is shown that REA adds value to modelling business processes in ERP systems.

**Keywords:** SAP A.G., ERP, Design Patterns, REA, HCM, TOA (Transaction-Oriented Architecture), Semantics, Ontology, Combining and Unifying Business Intelligence with Semantic Technologies (CUBIST)

## 1 Introduction

In this paper we explore the claim that TOA can be used following the principles of REA to enhance business process modelling in ERP systems, by providing a tool that can be used to increase the system design and understanding of the business process implementation and the underlying data model.

Fundamental to REA/TOA is the concept of a design pattern, since REA is defined in terms of an object design pattern. A design pattern is a recognized, named solution to a common design problem [1]. Catalogs of design patterns have been produced by the Gang of 4 as the solution to commonly found object oriented software design problems [2]. The concept of the Transaction Model (TM) was introduced to provide a method of encapsulating the REA model using CG concepts and thus allowing for the capture of organizational transactions by providing abstract constructs [3]. TOA offers the possibility of providing the tools (TM, TrAM, MAS) and concepts required to model the structure and transactions of an organization and provide purpose and direction to Service-Oriented Architecture (SOA) [4].

There are many vendors of ERP software, the top five vendors are SAP, Peoplesoft, Oracle, J.D. Edwards, and Baan. SAP is recognized as the dominant ERP system with over 50% of the market share [5]. Due to the clear commercial importance of the SAP solution, it was considered a logical step to use this implementation as an exemplar for ERP systems.

We provide evidence that shows how REA can be successfully used for modelling SAP business processes (in SAP HCM) and how SAP can be considered in part as complying with REA theories. However, the results of the research also indicate that through non-compliance with the REA ontology, how data is lost or stored again (repeated) within the SAP database. Confirming one of McCarthy's [6] original theories that led to the REA ontology, since he identified that using conventional data storage techniques (such as double entry), would lead to inconsistency of data, information gaps and overlaps in data or data spread.

This paper proceeds as follows. Section 2 contains the core of this paper, by making an REA analysis on one (SAP) business domain, SAP HCM and one business process within this domain, labor (labour) requisition. Section 3 provides a final summary and outlook.

## **2 REA analysis of SAP HCM**

### **2.1 HR business process**

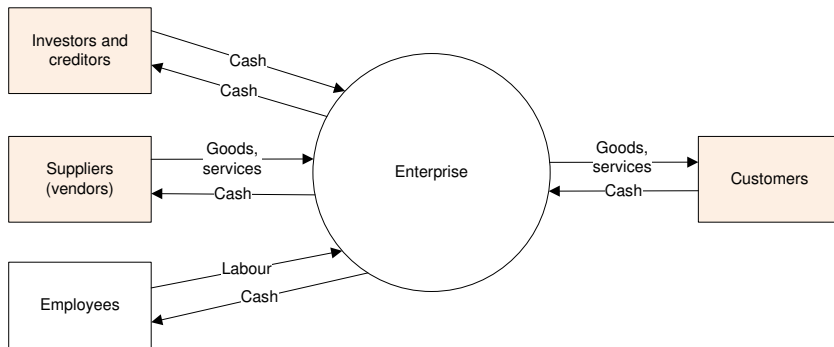
The HR business process is defined by Dunn et al. [7] as encompassing all that is required to acquire and then pay for employee labor. The HR business process is commonly separated into two separate sub-processes, where one sub-process; (i) personnel is responsible for hiring, training, evaluating and terminating employees and the other sub-process; (ii) payroll is responsible for the time management and subsequent payment of the employee's services [7].

### **2.2 REA Enterprise Value System**

In the REA Enterprise Value System the HR business process is defined as the point of contact between the enterprise and its employees [7]. In this sense the employees are seen as external suppliers (external agents) or business partners providing labor to



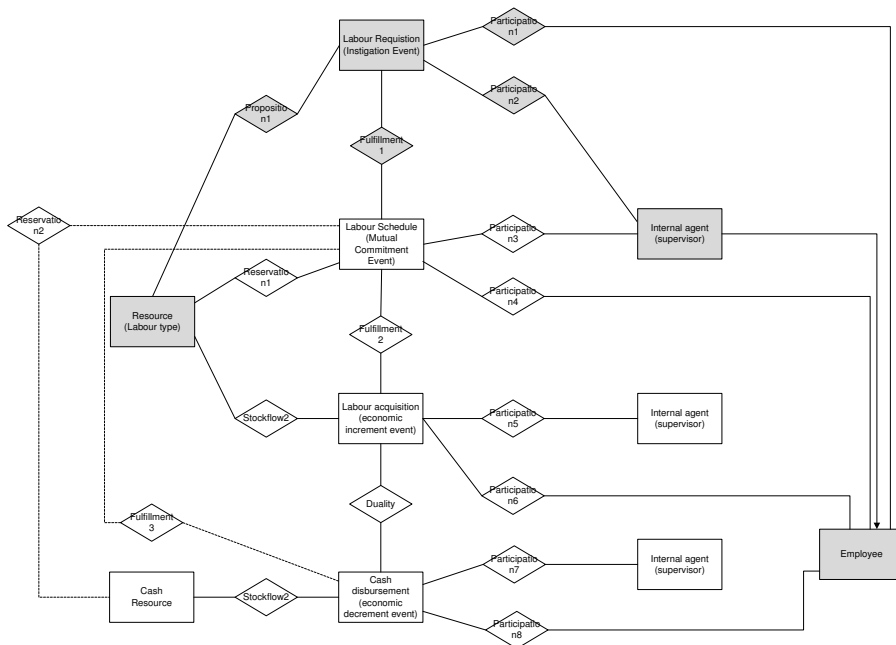
the organization in return for cash. The HR business process in the Enterprise Value System is identified in **Fig. 1** below.



**Fig. 1.** Payroll Human Resource Process in the Enterprise Value System [7]

Within the HR business process Dunn et al. [7] identify two key forms of resources that of human capital, the labor provided by the employees and the cash paid by the organization to the employee in return for the labor which was provided.

In REA terms the HR business process is identified (**Fig. 2**) as a special case of the acquisition/payment cycle, consisting of four key business events; labor requisition, labor schedule, labor acquisition and cash disbursement [7].

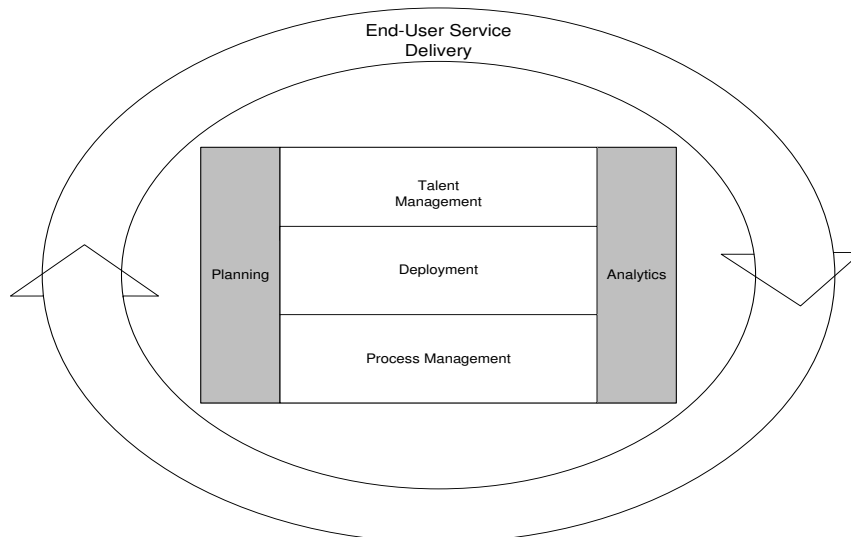


**Fig. 2.** Payroll Cycle Extended REA Ontology Database Design Pattern [7]

The ideas from Dunn et al. [7] and their initial REA diagram (**Fig. 2**) were used as a basis, from which this paper provides a more detailed investigation into the labor requisition business event which is shaded in grey in **Fig. 2**. These investigations have shown how it is however possible to provide further detail (than that provided by Dunn et al. [7]) of the labor requisition event and subsequently identify new REA entities.

### 2.3 SAP Human Capital Management (HCM)

The HR module is identified within SAP as Human Capital Management (HCM). SAP HCM consists of three separate sub modules which are identified as Talent Management, Workforce Deployment and Workforce Process Management. These three HCM sub-modules are then surrounded by; Workforce Planning and Analytics as detailed below in **Fig. 3**.



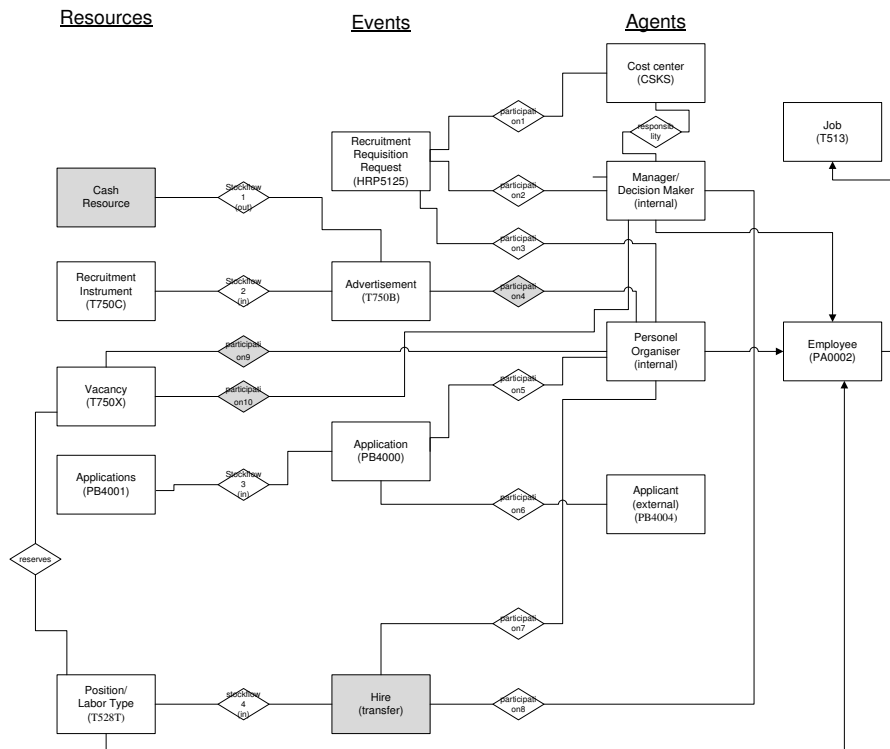
**Fig. 3.** SAP ERP Human Capital Management (HCM) [8]

### 2.4 Labor Requisition

The labor requisition event is defined by Dunn et al. [7] as the identification of a need for labor. Supervisors are usually responsible for determining this need through monitoring either one or all of enterprise growth (or the lack of), production plans, sales forecasts, employee turnover and other indications of labor requirements.

The labor requisition event can be aligned with the recruiting process within Talent Management in SAP HCM. Through our REA analysis of SAP HCM we have gone a step further than Dunn et al. [7] and identified four further (sub) events within the main labor requisition event, these four (sub) events are; requisition, advertisement,

application and hire (detailed in **Fig. 4** below). For each independent box (Resource, Event and Agent) the corresponding SAP table has been identified and is shown in brackets. The Resources, Events and relationships which are shaded in grey have been found to be non-REA compliant and will be discussed below, together with the other REA entities which were identified.



**Fig. 4.** Labor Requisition

**REA entities.**

Dunn et al. [7] state that in a valid REA design, each Resource, Event or Agent entity can be found stored within a separate database table. Using this criteria we have produced the results detailed in the Tab. 1-5 which show that we have; (i) identified many more REA entities than those defined by Dunn et al. [7], together with numerous relationships, (ii) that the tables for all the entities (except for cash resource and hire event) could be adequately accounted for within the (SAP) REA data model. The new entities discovered are detailed below;

(sub) Events.

Events are defined as ‘a class of phenomena which reflect changes in scarce means [economic resources] resulting from production, exchange, consumption, and

distribution', Yu and Yu quoted by McCarthy [6], the following REA (sub) events were identified as detailed in Tab. 1 below.

**Table 1.** Labor Requisition Events

<b>Event</b>	<b>Description</b>	<b>REA compliant</b>	<b>Comments</b>
Recruitment Request	a request from within the organization for new personnel	yes	
Advertisement	placing of an advertisement for a new position	no	Value of PCOST stored only in this table and not in a cash resource table
Application	the receipt of an application from an applicant	yes	
Hire	the point at which an applicant becomes a new employee	no	No separate table for this event, data transferred instead from application directly to employee

Resources.

McCarthy [6] originally defined a resource as equivalent to an asset in accounting terms and subsequently a resource was further defined by Dunn et al. [7] as something with or without substance that are provided or used (consumed) during an organizations business activities thus the following resources were identified in the labor requisition event as detailed below in Tab. 2.

**Table 2.** Labor Requisition Resources

<b>Resource</b>	<b>REA compliant</b>	<b>Comments</b>
Cash	no	No table found for this resource
Recruitment Instrument	yes	
Vacancy	yes	
Position/labor type	yes	
Applications	yes	

Agents.

McCarthy [6] defined agents as persons or agencies that participate in economic events or are responsible for subordinates that participate in these events. The following agents were identified in the labor requisition event as detailed below in Tab. 3.

**Table 3.** Labor Requisition Agents

<b>Agent</b>	<b>REA compliant</b>	<b>Internal/External</b>	<b>Comments</b>
Cost centre	Yes	Internal	
Manager/decision maker	Yes	Internal	
Personnel organiser	Yes	Internal	
Applicant	Yes	External	

**Relationships.**

The following relationships were identified in the labor requisition event as detailed below in Tab. 4. The table details each relationship together with the corresponding Resource/Event/Agent which the relationships connect.

**Table 4.** Labor Requisition Relationships

<b>Relationship</b>	<b>REA compliant</b>	<b>Direction</b>	<b>From</b>	<b>To</b>	<b>Comment</b>
Participation1	yes		Recruitment Request	Cost centre	

Participation2	yes		Recruitment Request	Manager/decision maker	
Participation3	yes		Recruitment Request	Personnel organiser	
Participation4	no		Advertisement	Personnel organiser	No entry for personnel organizer found in the advertisement table
Participation5	yes		Application	Personnel organiser	
Participation6	yes		Application	Applicant	
Participation7	yes		Hire	Personnel organiser	
Participation8	yes		Hire	Manager/decision maker	
Participation9	no		Vacancy	Personnel organiser	Agent found in the resource table NOT in the event table
Participation10	no		Vacancy	Manager/decision maker	Agent found in the resource table NOT in the event table
Stock-flow1	no	out	Cash	Advertisement	Cash resource not defined correctly
Stock-flow2	yes	in	Recruitment Instrument	Advertisement	
Stock-flow3	yes	in	Applications	Application	
Stock-flow4	no	in	Position/labor type	Hire	Hire not defined correctly, Historical data will be lost
Responsibility	yes		Cost centre	Manager/decision maker	
Reserves	yes		Vacancy	Position/Labor type	

### **Hire Event .**

#### **Duality.**

A known problem was encountered when modelling the hire event with respect to the duality relationship, since the REA ontology fails to explicitly specify whether the duality relationship should be seen as a property of the type level (relevant for the

modelling phase) or of the instance level (of a running system) [9]. Since if the 'conceptual modelling support' function was to be used which states that, when duality is used as the criterion of a valid model then; 'all types of a valid model must be coupled in duality relationships with other events' [9]. In the case of the Hire Event we have interpreted this duality relationship as belonging to the instance level of a running system, which means that there is no way to identify which named event type this particular hire event should be paired with. There was no evidence found within the SAP system of any event type which could correspond with the duality properties of the hire event. As observed by Borch and Stefansen [9] both possibilities are acceptable and they suggest that 'it is possible for the user of the ontology to make his or her own interpretation'. Our interpretation of this event, that it is belonging to the instance level, corresponds with the fact that the REA model was defined given the SAP implementation and not that the implementation followed an REA design.

#### Data storage.

As previously stated Dunn et al. [7] assert that in a valid REA design, each Resource, Event or Agent entity can be found stored within a separate database table. However within SAP HCM the hire event is not stored within a unique table, instead when this event (an applicant is hired) occurs the information about the applicant is moved from the applicant table directly to the employee table. Thus data is lost at this point since it is not possible to trace back directly to historical details of the event. There are of course practical implications which must be taken into account when a system is implemented, such as the necessary storage requirements when each and every event is stored. O'Leary [10] identifies this same issue and states that 'an events accounting system is a theoretical ideal which realistically would never be implemented'. He then draws the same conclusion that unless storage became costless and abstracting detail was 'painless', there would never be full event histories. The assumption can therefore be made that the designers at SAP made the decision to reduce data storage requirements by storing this event and the relevant data in this way.

#### **Cash Resource.**

For the business process labor requisition, the storage of cash resource does not follow REA principles, since the value of placing an advert (a cash resource) is stored within the Advertisement Event in the SAP table T750B in the column PCOST. However, this value does not find duality within the system, since it does not at any point within the business processing get transferred to a Cash Resource table or subsequently to a general ledger (resource) table. The value PCOST is used later by a SAP reporting process to determine how many applicants are received through a specific advert and thus determine a cost per advert per application. But at no point is the value PCOST booked against any cash accounts. There is no stock-flow in, in terms of an advert that has been placed, but no stock-flow out in terms of a cash resource, the payment for the advert which has been placed. Therefore in the processing of labor requisition, the SAP system does not conform with REA theory, which leads to information been lost or repeated (at a later date) in the database. It is

our assumption that this value (PCOST – cost of advertising) must at a later date be deducted from the general accounts ledger, however no evidence could be found to confirm this assumption.

**Vacancy Resource.**

The vacancy resource is stored within the SAP table T750X. The table contains foreign keys to the HR personnel organizer (participation9) responsible for this vacancy and the line manager (participation10) to which this vacancy has been assigned. The table also contains a foreign key (reserves relationship) to the position/labor type table that provides details of the position which is vacant. The structure of this table does not conform directly with REA theory, since the agents (involved in the event) should not be assigned directly to a resource (table) but should in fact be assigned to the event taking place [7].

**Relationship participation4.**

The advertisement (event) table does not contain the details of the personnel organizer responsible for this advert, shown in **Fig. 4**. Labor Requisition as relationship participation4. This does not conform to REA theory, which states that each agent which participates or is responsible for an economic event should be identified.

**2.5 REA compliance**

From the REA entities identified in SAP HCM and detailed in Tab. 1, we have produced the following table Tab. 5, which shows what percentage of the REA entities identified can be defined as been REA compliant.

**Table 5.** REA compliance

Entity	Number found	REA compliant	Compliance
Resources	5	4	80%
Events	4	2	50%
Agents	4	4	100%
Relationships	15	11	73%

**3 Summary and Outlook**

When examining the results as detailed in Tab. 1-4 and more specifically at the REA compliance of each of the entities discovered Tab. 5, we can concur with the results of O'Leary [10], in that we have underpinned how SAP's business processes (in SAP



HCM) can be effectively modeled using REA techniques. However we go further in two significant areas;

The results have shown (in detail) how REA can be used for modelling a business process; Human Resources. The detailed evidence shows one database table (-and several smaller anomalies) where SAP is not REA compliant, and resulting from this non-compliance, also shown how data is lost or repeated in the SAP database.

We have also confirmed a further statement from O'Leary [10] that 'SAP was found to be consistent with REA in its database, semantic, and structure orientations. However, there were some implementation compromises in the structuring and semantic orientation of the SAP data model'.

With regards to modelling business processes such as HR, O'Leary [10] makes the statement;

'For many real-world settings, REA is underspecified. For example, if we want to know how a human-resources process works, REA provides no direct insights. However, given a human resources model or system, we can map it to REA to try to understand it better or we can build a system using REA as a guide to the underlying data model, etc.'

This statement is reiterated by Geerts and McCarthy [11], however in section 2 we have shown how a business process in SAP HCM can in fact be adequately modeled using REA techniques and thus be subsequently represented in REA templates.

It is unlikely that SAP was implemented following a generic template model, since SAP has been implemented over a successive period of development over many years and thus is likely to contain many artifacts from the past such as the classical general ledger system of accounting. Moreover SAP was clearly not originally implemented following an REA paradigm [10], so it is also clear that the differences between REA and SAP can be interpreted as modelling compromises from an REA perspective. The same conclusion is made by Hesselund [12], who then suggests that this difference in interpretation will provide (a positive) feedback to the ontology development process and (can) be used as inspiration for further extensions of the core ontology.

The REA designs detailed by Dunn et al. [7] were a useful starting point, from which we have shown how REA can be used as a useful tool that can be used to increase the system design and understanding of the business process implementation. Through using REA analysis we have shown how our theoretical REA designs can be mapped directly to a real world (SAP) implementation.

A recognized limitation to REA modelling was encountered when analyzing SAP HCM, in that the REA model identifies only a structural view of the system, with the result that all behavioral aspects of the model must then be identified and documented using techniques such as data-flow diagrams [13]. The recognized solution to this problem is the use of unified modelling language (UML) for object-oriented modelling, previously identified by Booch et al. [14]. This again emphasizes the need noted by others [10, 13] for further research that will lead to a set of tools and procedures which will allow REA designs to be used for the entire development life cycle.

The data identifying the REA compliance of the REA entities discovered Tab. 5, would appear to indicate that in the critical area of defining event entities, SAP has

the most problems with REA compliance. Through this non-compliance with the REA ontology, we have shown how data is lost or stored again (repeated) within the SAP database. This confirms McCarthy's [6] original theory which led to REA, since he identified that using conventional data storage techniques (such as double entry), it would lead to inconsistency of data, information gaps and overlaps in data or data spread.

We have corroborated the findings exactly as foreseen by [10], namely that in two significant areas i.e. (i) SAP could benefit from an REA approach to business process engineering, since this would avoid data loss, and (ii) REA could benefit from an analysis of a real ERP system. Notably in this respect we have identified many more entities than those defined by Dunn et al. [7].

## 4 References

- [1] J. M. Bieman, G. Straw, H. Wang, P. W. Munger and R. T. Alexander, "Design patterns and change proneness: An examination of five evolving systems," in Software Metrics Symposium, 2003. Proceedings. Ninth International, 2003, pp. n/a.
- [2] E. Gamma, Design Patterns: Elements of Reusable Object-Oriented Software. Boston: Addison-Wesley, 1995.
- [3] S. Polovina and R. Hill, "A transactions pattern for structuring unstructured corporate information in enterprise applications," International Journal of Intelligent Information Technologies (IJIT), vol. 5, pp. 33-47, 2009.
- [4] S. Polovina, "The transaction concept in enterprise systems," in The 2nd CUBIST Workshop, 2011, pp. 43.
- [5] V. B. Gargeya and C. Brady, "Success and failure factors of adopting SAP in ERP system implementation," Business Process Management Journal, vol. 11, pp. 501-516, 2005.
- [6] W. E. McCarthy, "The REA accounting model: A generalized framework for accounting systems in a shared data environment," The Accounting Review, vol. 57, pp. 554-578, 1982.
- [7] C. L. Dunn, J. O. Cherrington and A. S. Hollander, Enterprise Information Systems: A Pattern-Based Approach. Boston: McGraw-Hill, 2005.
- [8] R. Haßmann, C. Krämer and J. Richter, Personnel Planning and Development using SAP ERP HCM. Boston: SAP PRESS, 2010.
- [9] S. E. Borch and C. Stefansen, "Evaluating the REA enterprise ontology from an operational perspective," in Proceedings of the CAiSE, 2004, pp. n/a.

[10] D. E. O'Leary, "On the relationship between REA and SAP," *International Journal of Accounting Information Systems*, vol. 5, pp. 65-81, 2004.

[11] G. L. Geerts and W. E. McCarthy, Eds., *Modeling Business Enterprises as Value-Added Process Hierarchies with Resource-Event-Agent Object Templates*. Berlin: Springer, 1997.

[12] A. Hesselund, "Modeling issues in REA," in 2006b, pp. n/a.

[13] U. Murthy and C. Wiggins Jr, "OOREA: An object-oriented resources, events, agents model for enterprise systems design," in 2004, pp. n/a.

[14] G. Booch, J. Rumbaugh and I. Jacobson, "The unified modeling language," *Unix-Review*, vol. 14, pp. 41-44,46,48, 1996.

# Evaluating and Analyzing Inconsistent RDF Data in a Semantic Dataset: EMAGE Dataset

Nwagwu Honour Chika

Cultural Communication and Computing Research Institute (C3RI)  
Faculty of Arts, Computing, Engineering and Sciences  
Sheffield Hallam University, United Kingdom

Honour.C.Nwagwu@student.shu.ac.uk

**Abstract.** This paper explains how to evaluate and analyse inconsistent Resource Description Framework (RDF) data by using EMAGE semantic (RDF) dataset as its use case. The author exploits the sub graph matching powers and mathematical functions of SPARQL query in evaluating inconsistent RDF data in a semantic dataset. He also proposes a mathematical method for calculating the amount of inconsistency in RDF data through a graph search approach. Finally, He analyzed the evaluated inconsistent RDF data.

**Keywords:** Triples, RDF data, Inconsistent data, Ontology, SPARQL queries

## 1 Introduction

EMAGE is a database of in situ gene expression data in the mouse embryo and an accompanying suite of tools to search and analyze the data (<http://www.emouseatlas.org/emage/>). EMAGE publishes in situ gene expression data for the developmental mouse. Its data is collected through a scrutinized process which involves assessing and tabulating of Biologist's experimental reports. These data include reports on gene expressions in mouse experiments which are reported elsewhere [11], the gene expression database (GXD), and laboratory reports among others. The Biologist's experimental report determines the strength of the expressed gene in a tissue of a mouse at a particular Theiler Stage. The Theiler stages correspond to a 28 days period associated with the developing mouse denoted by TS01 to TS28. More information about EMAGE datasets and mouse experiments can be found at the Edinburgh Mouse Atlas Project (EMAP) website [10, 12].

EMAGE's dataset can serve as a platform for Biologists to find solutions to the causes of abnormalities in organisms. Biologists can suggest answers to the cause of abnormalities in organisms through comparing the data indicating the strength of expressed gene in a healthy organism with that of unhealthy organism [9]. Neverthe-

less, data from some of the experiments which provide Biologists with this needed information can sometimes be inconsistent and these inconsistencies could be as a result of experimental error or simply a slight variation in experimental conditions [8]. Also, the accuracy of a dataset with inconsistent information can be increased through deleting the inconsistent data but at the cost of an increase in the incompleteness of the dataset. This cost can be avoided or minimized by properly evaluating and analyzing the degree of the inconsistency in the dataset. The author has explained how the inconsistency of RDF data can be identified, evaluated and analyzed. He has achieved this by explaining what RDF data model is in section 2.0, Identifying inconsistent RDF data in EMAGE dataset in section 3.0, Evaluating and analyzing inconsistent RDF data in section 4.0 and finally, the author presents his approach on how inconsistent RDF data can be evaluated and analyzed in section 5.0.

## 2 RDF data model

Information in semantic dataset is represented by RDF data in the form of triples and stored in a triple store. A triple consists of subject, predicate and an object. An illustration of a RDF triple is as shown in figure 1 below.

```
<http://www.cubist_project.eu/HWU#tissue_EMAP_42>  
<http://www.w3.org/2001/01/rdf-schema#label> "embryo" .
```

Figure 1: A triple in EMAGE dataset

Each triple in RDF dataset represents a statement of a relationship between the entities denoted by the nodes that it links. RDF data can contain one or more triples. Each triple is composed of a subject, predicate and an object. In RDF data, each subject of a triple is represented by a Universal Resource Identifier (URI) or blank node, each predicate is represented by a URI and each object node is represented by a URI, a blank node or a literal. For example in figure 1, the subject of the triple is a URI “http://www.cubist\_project.eu/hwu#tissue\_EMAP\_42”, the predicate is a URI “http://www.w3.org/2000/rdf-schema#label”, and the object node is a literal “embryo”. The author adopted turtle serialization format (<http://www.w3.org/TR/turtle/>) in this example. RDF data has other serialization formats for representing its data such as N-Triple, N3, RDF/XML and RDFa.

### 3 Identifying inconsistent RDF data in EMAGE dataset: SPARQL Query Language

Inconsistency exists in RDF data when the data does not conform to the rules governing their design. This is evident when there is a contradiction in the RDF data such that the RDF data contains both  $A$  and  $\neg A$ .

Inconsistency in EMAGE dataset is identified through identifying data which do not conform to EMAGE's textual annotation rules. These rules include the general "detected somewhere in" and "not detected everywhere in" rules which are used to propagate gene expression levels up and down the hierarchical structure of a particular EMAP anatomy. In addition, the expression level of a gene in a particular structure of a given Theiler stage in EMAGE dataset is reasoned through propagation approach. Through propagation approach, the associated level of gene expression in tissues that exhibit "is\_part\_of" relationship with other tissue(s) within a particular structure are propagated up or down the given structure in line with the chosen level of gene expression of that structure. As a consequence, gene expression levels could be inconsistent. This can be as a result of positive propagation (expressions propagated up the anatomy) that contradicts with an experimental result or negative propagation (expressions propagated down the anatomy) that contradicts with an experimental result. Also, gene expression can be completely contradictory (two experiments on the same tissue in which a gene is stated as detected in one experiment and not detected in the second experiment) or partly contradictory (two experiments on the same tissue in which the genes detected have different expression levels). Also, Inconsistency in EMAGE datasets has been categorized and defined [9] as either binary inconsistency: gene that is both expressed and not expressed in a given tissue of a Theiler stage and analogue inconsistency: involving varied strength levels of a particular gene in a given tissue of a Theiler stage.

In other to identify inconsistent RDF data, a subset of EMAGE RDF model dataset was stored in OWLIM-SE triple store (<http://www.ontotext.com/owlim>). The investigated dataset has 1,216,277 triples. The author applied appropriate SPARQL queries as to retrieve inconsistent data from the stored RDF dataset. He was able to detect binary inconsistency in the investigated dataset in some tissues which have "is\_part\_of" relationship with other tissues of the same hierarchical annotation structure. In these tissues, a gene is specified as "detected" and also specified as "not detected" in their related tissue. An example of EMAGE hierarchical annotation structure is shown in the figure 2 below. The SPARQL query in figure 3 identifies RDF data with binary inconsistency from Theiler stage 15 of the investigated HWU RDF model dataset. It can also be applied to any other Theiler stage by changing the Theiler stage number in the statement under label #3 of the query. Table 1 displays the result set. The author used the hash key (#) together with a unique number in the SPARQL query to identify comments that explain the SPARQL statement(s).

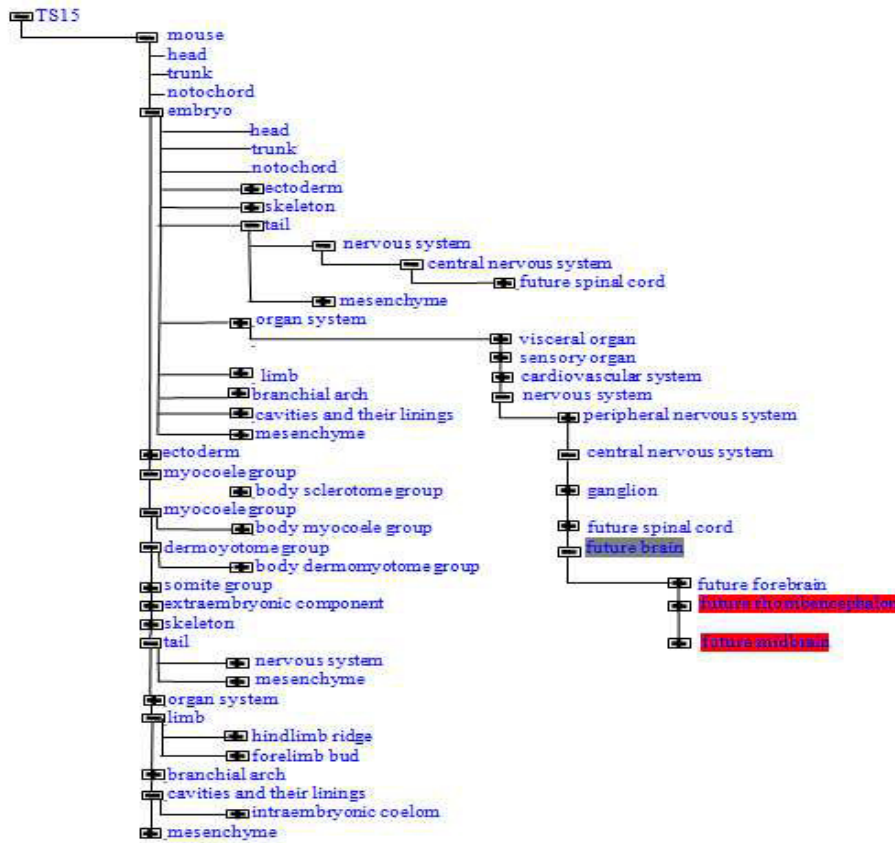


Figure 2: A subset of the Anatomy Ontology of Thiler stage 15 (drawn from <http://www.emouseatlas.org/emap/ema/home.html>)

To illustrate the different types of inconsistent data in the investigated dataset, the author used instances from Thiler stage 15. Figure 2 shows a subset of EMAP anatomy of Thiler stage 15.

Table 1: Binary inconsistent tissue experiments of Thiler stage 15

Gene_label	T_label	T_Experiment_label	Gene_strength	T2_label	T2_Experiment_label	Gene_strength2
Pax2	future midbrain	EMAGE:3530	hwu:level_detected	future brain	EMAGE:984	hwu:level_not_detected
Pax2	future midbrain	EMAGE:3879	hwu:level_detected	future brain	EMAGE:984	hwu:level_not_detected
Pax2	future rhombencephalon	EMAGE:3879	hwu:level_detected	future brain	EMAGE:984	hwu:level_not_detected

```

#1 Declare URI namespace
prefix hwu: <http://www.cubist_project.eu/HWU#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#

#2 Select variables whose bindings are returned as solutions of the query
SELECT DISTINCT ?gene_label ?t_label ?t_Experiment_label
?gene_strength ?t2_label ?t2_Experiment_label ?gene_strength2
where { {

#3 Select a set of triple pattern that depicts the investigated RDF data: Set 'A'
?x rdf:type hwu:Textual_Annotation ; hwu:belongs_to_experiment
?y ; hwu:in_tissue ?z ; hwu:has_involved_gene ?g ;
hwu:has_strength ?gene_strength .
?z hwu:has_theiler_stage hwu:theiler_stage_15 ; rdfs:label
?t_label .
?y rdfs:label ?t_Experiment_label .
?g rdfs:label ?gene_label
}
}
OPTIONAL #4 SPARQL key word which enables optional match
{
#5 Select optional variables contradicting set 'A' in another set: set 'B'
?b rdf:type hwu:Textual_Annotation ; hwu:belongs_to_experiment
?y2 ; hwu:in_tissue ?ztissue2 ; hwu:has_involved_gene ?g ;
hwu:has_strength ?gene_strength2 .
?ztissue2 rdfs:label ?t2_label .
?y2 rdfs:label ?t2_Experiment_label .

#6 Stipulate the relationship between set 'A' and set 'B'
?z hwu:is_part_of ?ztissue2 .

#7 Stipulate the necessary condition that can ascertain any
#7 possible contradictory values between set 'A' and set 'B'
Filter(?gene_strength = hwu:level_detected && ?gene_strength2
= hwu:level_not_detected ) } }

#8 Aggregate values of variables to be returned
group by ?gene_label ?t_label ?t_Experiment_label
?gene_strength ?t2_label ?t2_Experiment_label ?gene_strength2

#9 Restrict expected results to allow only the output of contradictory values
having ( (round((count(?t2_label)/(count(?t_label))*100)) > 0)

#10 Establish the order for the result set
order by ?gene_label

```

Figure 3: Query to identify binary contradictory RDF data in Theiler Stage 15



The result set in table 1 above, shows identified binary inconsistent RDF data in Theiler stage 15. As an example, some tissues (future midbrain and future rhombencephalon of experiments EMAGE:3530 and EMAGE:3879 respectively) with involved gene “Pax2” whose expression level are specified as “level\_detected” were identified. Future midbrain and Future rhombencephalon have the same involved gene “Pax2” and a “is\_part\_of” relationship with the tissue “Future brain” whose expression level is specified as “level\_not\_detected” in EMAGE:984. These expression levels of Pax2 as specified in these experiments contradict each other and do not abide with the semantics of the word “is\_part\_of” as utilized by EMAP. In addition, analogue inconsistency was detected in the investigated dataset in some tissues which have “is\_part\_of” relationship with other tissues. The identified analogue inconsistent data involve a gene with varied strength levels such as “strong” and “moderate” in tissues that have “is\_part\_of” relationship with other tissues. Analogue inconsistency in RDF data from Theiler stages 15 of the investigated dataset was identified by substituting the filter condition under label #7 of figure 3 with the below filter condition:

```
Filter(?gene_strength = hwu:level_strong && ?gene_strength2 =
hwu:level_weak || ?gene_strength = hwu:level_moderate &&
?gene_strength2 = hwu:level_weak || ?gene_strength =
hwu:level_strong && ?gene_strength2 = hwu:level_moderate)
```

Table 2: Analogue inconsistent tissue experiments of Theiler stage 15

Gene_label	T_label	T_Experiment_label	Gene_strength	T2_label	T2_Experiment_label	Gene_strength2
Fkbp3	branchial arch	EMAGE:5349	hwu:level_strong	embryo	EMAGE:5349	hwu:level_weak
Fkbp3	limb	EMAGE:5349	hwu:level_strong	embryo	EMAGE:5349	hwu:level_weak
Msx1	2nd branchial arch mesenchyme	EMAGE:5411	hwu:level_moderate	2nd branchial arch	EMAGE:3839	hwu:level_weak
Nav2	epithelium	EMAGE:6026	hwu:level_strong	otocyst	EMAGE:6026	hwu:level_weak
Pax1	3rd branchial pouch endoderm	EMAGE:61	hwu:level_moderate	3rd branchial pouch	EMAGE:246	hwu:level_weak
Pax1	3rd branchial pouch endoderm	EMAGE:61	hwu:level_moderate	3rd branchial pouch	EMAGE:3938	hwu:level_weak
Smarcb1	branchial arch	EMAGE:5062	hwu:level_strong	embryo	EMAGE:5062	hwu:level_weak
Smarcb1	limb	EMAGE:5062	hwu:level_strong	embryo	EMAGE:5062	hwu:level_weak

The result set in table 2, shows the identified analogue inconsistent RDF data in Theiler stage 15. As an example from the table, some tissues (Branchial arch and limb of experiment EMAGE:5349) with involved gene “Fkbp3” whose expression levels are specified as “level\_strong” have been identified from the investigated dataset. Branchial arch and Limb have “is\_part\_of” relationship with the tissue Embryo. Yet, Fkbp3 has a level of expression “level\_weak” in Embryo in the same experiment. These expression levels of “Fkbp3” as specified in the experiment contradict each other and do not abide with the semantics of the word “is\_part\_of” as utilized by EMAP. Examples from other EMAGE inconsistency types include the inconsistency from positive propagation: Gene “Pax2” was “detected” in Future midbrain in EMAGE:3879 and “not detected” in Future brain in EMAGE:984 (Table 1). Future midbrain is part of future brain and it is located at a lower part to future brain in the

anatomy structure of Theiler stage 15 (figure 2). Future brain should unavoidably have the same gene expression as future midbrain if gene expression is to be propagated up the anatomy. The strength level of future brain is therefore contradicted by not fully propagating *Future midbrain's* gene expression level up the anatomy and this result to 'an inconsistency of positive propagation'. On the other hand, Future midbrain should unavoidably have the same gene expression level as future brain if gene expression is to be propagated down the anatomy. The strength level of future midbrain was contradicted by not fully propagating the gene expression level in future brain down the anatomy and this result to 'an inconsistency of negative propagation'. Figure 2 shows the tree illustrating the hierarchical structure of future midbrain and future brain in Theiler stage 15.

#### **4 Evaluating and analyzing inconsistent RDF data**

There are two main methods of dealing with inconsistent data in a dataset: to diagnose and repair it, and reasoning with the inconsistency [3]. Also, various approaches such as [7, 8] have been proposed on reasoning with the inconsistent data. The act of addressing inconsistent data through identifying the inconsistency with the aim of repairing it through deleting the inconsistent data will inevitably increase the incompleteness of the dataset. More so, the use of various reasoning approaches on inconsistent dataset would produce varied result sets for a given approach on the dataset. These lapses can be addressed through measuring and detailing of the inconsistencies in the retrieved information from an inconsistent dataset.

Obviously, measuring inconsistency has been proven useful in analyzing diverse range of information types such as news reports [4]. However, there are a few approaches [1, 2] for measuring the inconsistencies of semantic datasets. There are other publications which verify and validate the RDF data held within a database [5, 6] but these works do not measure and analyze the amount of inconsistency in inconsistent information retrieved from the database. Consequently, the author assesses the amount of inconsistency in inconsistent information from a graph based approach. He achieves this through adopting the sub graph matching powers of SPARQL queries.

#### **5 Approach**

The amount of inconsistency in an investigated RDF data can be measured by evaluating the amount of contradiction in the RDF data against the likelihood of the contradiction to occur. This amount is assessed herein by calculating their ratio as a fraction of 100. The result educates us on how large/small the embedded contradiction in the RDF data is. As stated above, the amount of inconsistency in EMAGE's data from a graph based approach is herein assessed through adopting the mathematical and sub graph matching powers of SPARQL queries. This approach can be applied to all RDF dataset formats. It necessitates proper SPARQL query skills and adequate knowledge of the dataset by the dataset analyst. The amount of contradictions in the data under investigation against its total possibility to occur in the dataset is calculated as follows:

**X<sub>m</sub>** = A RDF graph pattern in a RDF dataset

**X<sub>k</sub>** = Contradictory sub graph of X<sub>m</sub>

The interest is in calculating the amount of **X<sub>k</sub>** in **X<sub>m</sub>** such that

$\sum X_k$  = Total number of contradictions in **X<sub>k</sub>**

$\sum X_m$  = Total number of occurrence of **X<sub>m</sub>** in the dataset

$$\text{Amount of Inconsistency in } X_m = \frac{\sum X_k}{\sum X_m} * \frac{100}{1}$$

In this investigation, the question “what amount of binary or analogue contradiction is present in the expression levels of the genes in each tissue experiment of Theiler stage 15” is answered. The amount of Binary or analogue inconsistency in RDF data from any of the Theiler stages of the investigated dataset is identified by adding the following SPARQL statement before label #2 of figure 3.

```
Select ?gene_label ?t_Experiment_label
round((count(?gene_strength2)/(count(?gene_strength))) * 100)
as ?amount_of_inconsistency)
{
```

And also substituting the aggregation statement under the label #8 of the query with the below statement:

```
Group by ?gene_label ?t_Experiment_label
```

The result set of the administered query on Theiler stage 15 is as displayed in table 3 and 4 below.

Table 3: Amount of binary inconsistency in tissue experiments of Theiler stage 15

<b>Gene_label</b>	<b>T_Experiment_label</b>	<b>Amount_of_inconsistency (%)</b>
Pax2	EMAGE:3530	50
Pax2	EMAGE:3879	33

Table 3 above, gives a more clarifying result set of each inconsistent experiment in Theiler stage 15 of the dataset than table 1. Rather than listing inconsistent experiments singly (like in table 1), the amount of its occurrence in the RDF data with the stipulated pattern is measured. These measures inform us of the amount of inconsistent assays in each tissue experiment of a particular Theiler stage in the dataset. As an example in EMAGE:3530, it can reliably be stated that half (50%) of the assays are binary inconsistent. While in EMAGE:3879, less than half (33%) of the assays results

are binary inconsistent. Consequently, decisions by Biologists to carry out further test or to remove existing experimental results from the dataset can be made.

Table 4: Amount of analogue inconsistency in tissue experiments of Theiler stage 15

Gene_label	T_Experiment_label	Amount_of_inconsistency (%)
Fkbp3	EMAGE:5349	20
Mxcl	EMAGE:5411	8
Nzv2	EMAGE:6026	2
Paxl	EMAGE:61	22
Smarb1	EMAGE:5062	22

Figure 4 below, depicts a flowchart for measuring inconsistency of RDF dataset.

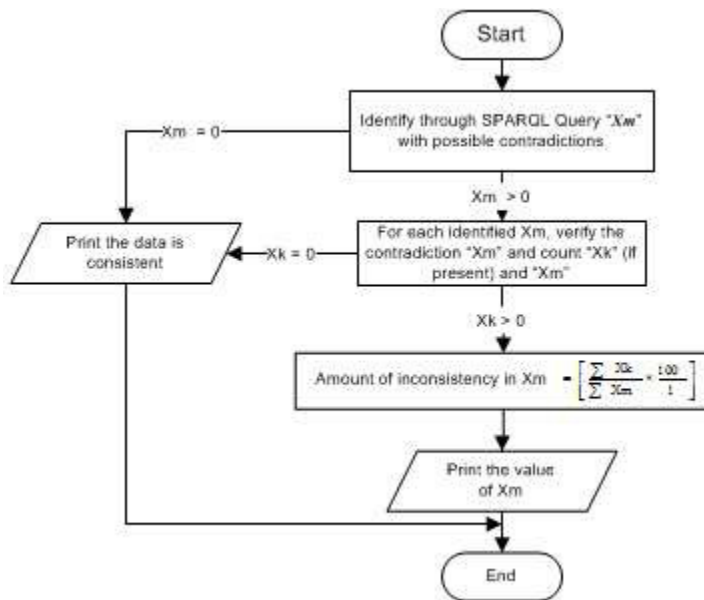


Figure 4: Flowchart for measuring inconsistency in RDF data

A tissue experiment can have several assays. The author's approach identifies the amount of these inconsistent assay(s) in their corresponding tissue experiment. For example, in Theiler stage 15 of the investigated dataset, there are 6 assays on EMAGE:3879, and 2 of them are binary inconsistent thus the amount of inconsistency in the experiment is calculated by dividing 2 with 6 and multiplied the result by 100. The importance of identifying the amount of inconsistency in a tissue experiment is to identify how valid the assay results of a particular experiment are.

## 6 Conclusion

Evaluating and analyzing inconsistent RDF data of a RDF model dataset is a field yet to be explored. Interestingly, it has been shown in this paper that the measure and analysis of inconsistent RDF data gives an insight to the soundness of the information under investigation. Nevertheless, the author hopes to improve on this research by automating these processes of identifying, evaluating and analyzing inconsistent RDF data.

The author acknowledges the partners of CUBIST project especially Heriot-Watt University and Sheffield Hallam University for their support and provision of his research datasets. He also acknowledges his two PhD supervisors "Simon Andrews" and "Simon Polovina" for their invaluable contributions and review of this work.

## Reference

1. Grant, J., and Hunter, A. (2006). Measuring inconsistency in knowledgebases. *Journal of Intelligent Information Systems*, 27(2), 159-184.
2. Grant, J., and Hunter, A. (2008). Analysing inconsistent first-order knowledgebases. *Artificial Intelligence*, 172(8), 1064-1093.
3. Huang, Z., van Harmelen, F., and ten Teije, A. (2006). Reasoning with inconsistent ontologies: Framework, prototype, and experiment. *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*, 71-93.
4. Hunter, A. (2002, July). Measuring inconsistency in knowledge via quasi-classical models. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE* (pp. 68-73). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.
5. Jerven, B., Sebastien, G., and the UniProt Consortium. Catching inconsistencies with the semantic web: a biocuration case study
6. Jupp, S., Parkinson, H., and Malone, J. *Semantic Web Atlas: Putting Gene Expression Data Into Biological Context*.
7. Lembo, D., Lenzerini, M., Rosati, R., Ruzzi, M., and Savo, D. (2010). Inconsistency-tolerant semantics for description logics. *Web Reasoning and Rule Systems*, 103-117.
8. McLeod, K., and Burger, A. (2007). Using argumentation to tackle inconsistency and incompleteness in online distributed life science resources. In *Proceedings of IADIS International Conference Applied Computing* (pp. 489-492).
9. McLeod, K., and Burger, A. (2011). WP7 requirement document of CUBIST Consortium 2010-2013. Available at [http://www.cubist-project.eu/fileadmin/CUBIST/user\\_upload/Deliverable/CUBIST\\_D7.1.1\\_HWU\\_v1.0.pdf](http://www.cubist-project.eu/fileadmin/CUBIST/user_upload/Deliverable/CUBIST_D7.1.1_HWU_v1.0.pdf)
10. Richardson L, Venkataraman S, Stevenson P, Yang Y, Burton N, Rao J, Fisher M, Baldock RA, Davidson DR, Christiansen JH. EMAGE mouse embryo spatial gene expression database: 2010
11. Suda, Y., Hossain, Z. M., Kobayashi, C., Hatano, O., Yoshida, M., Matsuo, I., and Aizawa, S. (2001). Emx2 directs the development of diencephalon in cooperation with Otx2. *Development*, 128(13), 2433-2450.
12. Theiler, K. (1989). *The house mouse: atlas of embryonic development* (p. 168). New York: Springer-Verlag.

# FCAWarehouse, a prototype online data repository for FCA

Constantinos Orphanides and George Georgiou

Conceptual Structures Research Group  
Communication and Computing Research Centre  
Faculty of Arts, Computing, Engineering and Sciences  
Sheffield Hallam University, Sheffield, UK  
c.orphanides@shu.ac.uk georgios.georgiou@student.shu.ac.uk

**Abstract.** This paper presents FCAWarehouse, a prototype online data repository for FCA. The paper explains the motivation behind the development of FCAWarehouse and the features available, such as the ability to donate datasets and their respective formal contexts, the ability to generate artificial formal contexts on-the-fly, and how these features are also available through a set of web-services. The paper concludes by suggesting future work in order to enhance its usability.

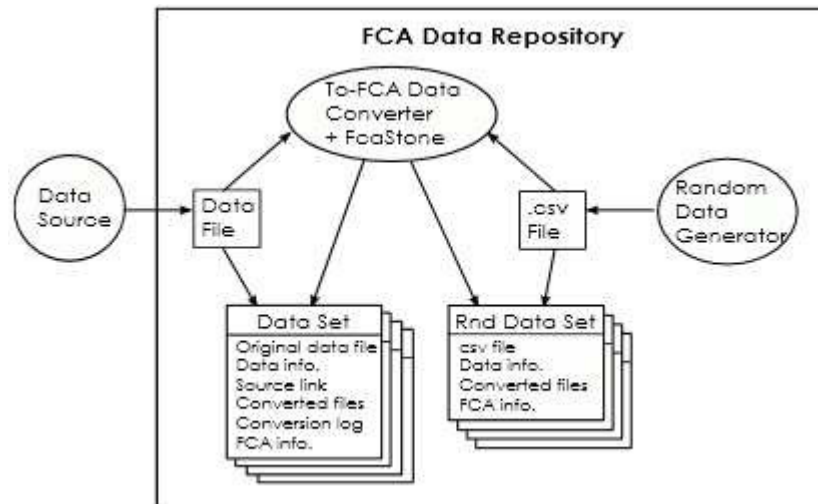
## 1 Introduction

In recent years, different means of archiving empirical data have been considered and implemented to build a space where data collection could be exploited and analyzed accordingly, if required. The different archiving techniques not only introduced communities an alternative approach of maintaining the quality and accessibility of data, but also improved the indexing of available data collections with proper categorization. An example of such an archiving technique are online data repositories.

Online data repositories are digital libraries which allow the comprehensive collection, management and preservation of digital content and the ability to offer it to targeted user communities [10,7,8]. An example of such a repository is the UCI Machine Learning Repository, a repository used by the machine learning community for the empirical analysis of machine learning algorithms [6]. Data repositories exist for numerous targeted communities; however, a centralized, public resource of data in FCA formats for the FCA community has not been implemented to date.

The idea of an FCA online data repository, to provide FCA practitioners with a resource of public datasets in FCA formats, was proposed in [2] at 2009. The proposal envisioned the automatic conversion of uploaded traditional datasets into formal contexts, as well as the ability to generate artificial datasets in CSV format, to be then converted into a formal context with the user being able to determine the number of objects, number of attributes and the density. The automatic conversion of datasets to formal contexts would be handled by a

“Data-to-FCA” converter and the tool FcaStone<sup>1</sup> would be used to convert formal contexts from one FCA format to another. Both of these tools would be incorporated in the repository as components. The architecture of the proposed repository is shown in Figure 1.



**Fig. 1.** Proposed FCA Data repository system architecture in [2].

The “Data-to-FCA” converter of the proposed repository was the most complex component to implement with regards to the effort and time needed to develop it. However, since the publication of the aforementioned paper, the “Data-to-FCA” converter has been developed as a standalone desktop application called FcaBedrock<sup>2</sup>, a formal context creator for FCA with the ability of converting datasets in various formats to formal contexts in the Burmeister (.cxt) or FIMI (.dat) formats [5,4]. The formal contexts generated by FcaBedrock can be then loaded in In-Close<sup>3</sup>, a fast formal concept miner, to count the number of formal concepts in a formal context and produce, if necessary, smaller sub-contexts based on the well-known notion of *minimum support* [3,1]. Subsequently, the motivation behind implementing an FCA data repository became stronger and resulted in the development of a prototype data repository for FCA, named FCAWarehouse.

<sup>1</sup> <http://sourceforge.net/projects/fcastone>

<sup>2</sup> <http://sourceforge.net/projects/fcabedrock>

<sup>3</sup> <http://sourceforge.net/projects/inclose>

## 2 FCAWarehouse

FCAWarehouse<sup>4</sup> is a prototype data repository for FCA developed as part of a BSc Computing final year project [9] at Sheffield Hallam University (SHU). It provides the ability of donating datasets and their respective formal contexts, browsing and searching datasets, an administration back-end for librarians to manage donated datasets, creating artificial formal contexts, as well as providing all of its functionality through a set of ReSTful<sup>5</sup> web services to make FCAWarehouse interoperable with third-party FCA tools. A diagram of its architecture is shown in Figure 2 and an explanation of its features is given in the following sections.

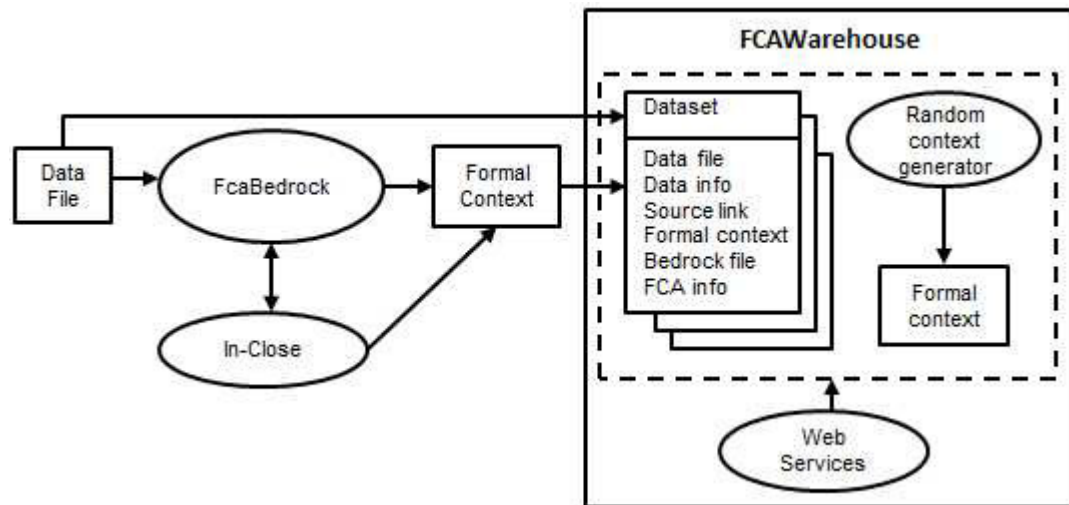


Fig. 2. System architecture of FCAWarehouse.

### 2.1 Datasets


Each dataset in FCAWarehouse is comprised of metadata such as name, additional information, original source as well as the original data and its respective formal context (Figure 3).

All of the files and metadata are provided by the donator during the donation process. In cases where the formal contexts are created using FcaBedrock, a

<sup>4</sup> <http://www.fcawarehouse.com>

<sup>5</sup> Representational State Transfer (ReST). <http://www.ibm.com/developerworks/webservices/library/ws-restful/>



Photo	
Name	Adult
Concepts	68872
Attribute	Other
Area	Social Sciences
Data Type	Multivariate
Task	Classification
Format	Other
Description	Extraction was done by Barry Becker from the 1994 Census database, using the following conditions: ((AGE>16) && (AGI>100) && (AFNLW<50000)) whether a person makes over 50K a year.
Missing Value	<input checked="" type="checkbox"/>
Source	<a href="http://archive.ics.uci.edu/ml/datasets/Adult">http://archive.ics.uci.edu/ml/datasets/Adult</a>
Acknowledgements	Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science.
email	c.orphanides@shu.ac.uk
Inserted	4/16/2013 1:50:40 PM
Updated	Apr 16 2013 1:51PM
Download	<a href="#">Click Here</a>

**Fig. 3.** Example of a dataset entry in FCAWarehouse (partial screenshot).

*Bedrock* file (**.bed**) is also provided which contains the metadata of the conversion so that users can load the original data and the *Bedrock* file in *FcaBedrock* and have a look at, or modify, the parameters and conversion criteria set for each attribute in the original dataset (Figure 4).

The files donated for each dataset are stored physically on the server, with the database only holding the URL to each dataset. As the prototype is currently hosted on a server with limited storage capabilities, all files uploaded for a dataset during the donation process are automatically compressed as **.zip** files.

After a user has donated a dataset, the librarians are notified about the new submission and have the ability of accepting or declining the submitted dataset. Consequently, only datasets marked as ‘Accepted’ by the librarians are visible to end-users.

Attributes				
	Attributes	Type		Convert?
▶ 0	age	Continuous	▼	<input checked="" type="checkbox"/>
1	workclass	Categorical	▼	<input checked="" type="checkbox"/>
2	fnlwgt	Categorical	▼	<input type="checkbox"/>
3	education	Categorical	▼	<input checked="" type="checkbox"/>
4	education-num	Categorical	▼	<input checked="" type="checkbox"/>
5	marital-status	Categorical	▼	<input checked="" type="checkbox"/>
6	occupation	Categorical	▼	<input checked="" type="checkbox"/>
7	relationship	Categorical	▼	<input checked="" type="checkbox"/>
8	race	Categorical	▼	<input checked="" type="checkbox"/>
9	sex	Categorical	▼	<input checked="" type="checkbox"/>
10	capital-gain	Continuous	▼	<input checked="" type="checkbox"/>
11	capital-loss	Continuous	▼	<input checked="" type="checkbox"/>

**Fig. 4.** FCAWarehouse’s Bedrock (.bed) file of the Adult dataset loaded in FcaBedrock (partial screenshot).

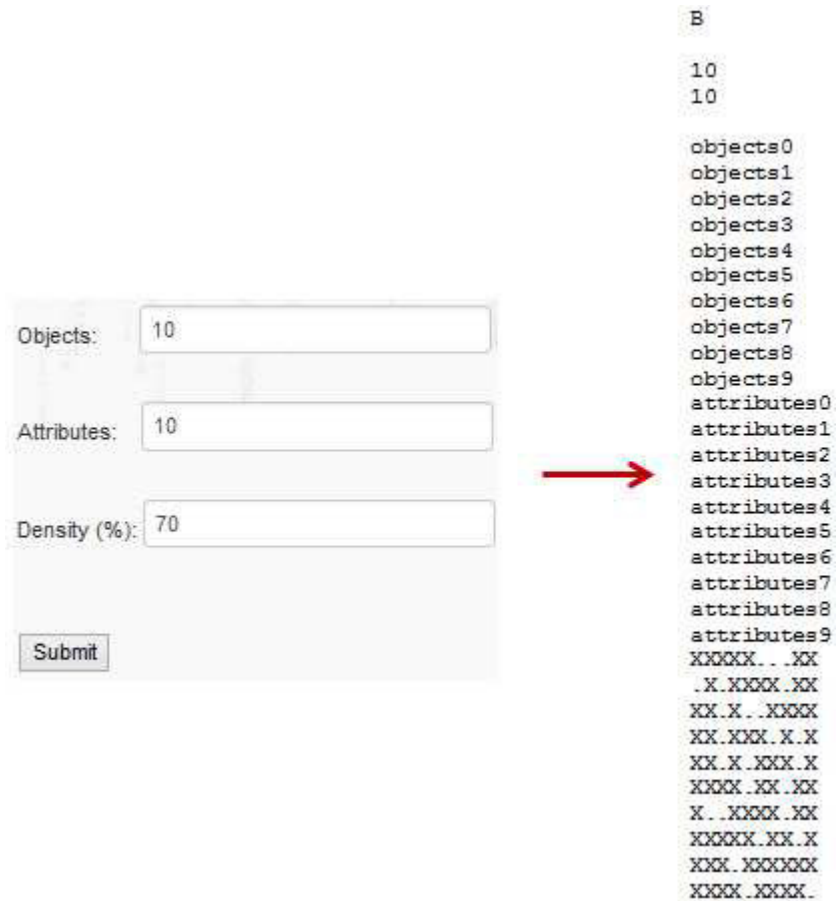
## 2.2 Artificial formal contexts

FCAWarehouse provides the ability of generating artificial formal contexts in the Burmeister (.cxt) format by predefining the number of formal objects, number of formal attributes and the density. Developers of tools such as the formal concept miners In-Close<sup>6</sup> and FCbO<sup>7</sup> can use this feature to generate formal contexts of size of their choice in order to test the limits and efficiency of their algorithms.

The generated artificial formal contexts are randomly generated by predefining the number of objects, number of attributes and density of the formal context. The density is defined as a percentage indicating the minimum amount of crosses (formal attributes) that each row (formal object) should contain in the formal context. Setting for example number of objects to 10, number of attributes to 10 and density percentage to 70% will result in a formal context with 10 formal objects and 10 formal attributes where *at least* 70% of each row is consisted of crosses (Figure 5). Setting the density to 0 will generate a random artificial formal context with no density criteria.

<sup>6</sup> <http://sourceforge.net/projects/inclose>

<sup>7</sup> <http://sourceforge.net/projects/fcalgs>



**Fig. 5.** Example of an artificial formal context with 10 objects, 10 formal attributes and density set to 70%.

### 2.3 Web services

FCAWarehouse implements a set of web services for the interoperability of FCAWarehouse with third-party FCA applications. At the moment there are a total of four web services, namely:

- **GenerateContext:** Accepts as input the number of objects, number of attributes, density percentage and outputs a corresponding artificial Burmeister formal context in XML.
- **GetAllDataSets:** Returns all datasets, along with their metadata, in XML. Only datasets marked as ‘Accepted’ by the librarians are returned.

- **Get10MostRecentDatasets:** Returns the 10 most recent datasets along with their metadata, in XML. Only datasets marked as ‘Accepted’ by the librarians are returned.
- **SearchDataset:** Accepts a string as input and returns any datasets which contain the given string in their name, along with their metadata, in XML. Only datasets marked as ‘Accepted’ by the librarians are returned.

### 3 Further Work and Conclusion

While useful in its current state, FCAWarehouse is still in a prototypical state; a number of improvements can be implemented to enhance its usability. For example, the feature of artificially generating formal contexts can be extended to simulate formal contexts that real datasets would produce. This can be achieved by defining sets of adjacent columns in the formal context to represent a single attribute. Assuming, for example, a categorical attribute (with each of its values being a formal attribute in the formal context) and each formal object only being able to have one of its values (quite common in real datasets), will result in a formal context where no more than one cross will exist in the columns representing that attribute. The same logic can be applied to various type of attribute: boolean attributes could be interpreted as single formal attributes in the formal context and continuous attributes could be grouped using ranges rather than one formal attribute for each numerical value. In this way, FCAWarehouse can act as a benchmarker for the comparison of tools and algorithms by providing citable random data as well as converted real datasets.

A feature which could prove quite useful in the future, with some modifications and adjustments, are FCAWarehouse’s web services. Interesting use-cases can emerge from this feature; for example, the formal context creator FcaBedrock could feature a “Create formal context from FCAWarehouse” option. By using the provided web services, a list of the available datasets could appear in FcaBedrock. Selecting one of the datasets could automatically download the dataset and start auto detecting its values to create a formal context. Creating a formal context in FcaBedrock requires inputting a dataset, defining its metadata and then creating the formal context; considering, however, the “Generate Context” web service, the ability of creating artificial formal contexts in FcaBedrock without requiring initial data as input could be made possible with minimal effort.

Further work also includes the incorporation of the tool FcaStone, to convert data between FCA and non-FCA formats, in FCAWarehouse to offer the power of FCA to those currently outside of the FCA community.

The final vision of FCAWarehouse is of an online FCA data repository which facilitates the creation, conversion and donation of datasets for FCA, providing a useful collection of real and artificial datasets in a wide variety of FCA and non-FCA formats to open the way for the wider use of FCA.

## References

1. Andrews, S. (2011). *In-Close2, a High Performance Formal Concept Miner*. In: Proceedings of the 19th International Conference on Conceptual Structures (ICCS) 2011. LNAI 6828. Berlin: Springer-Verlag. pp. 50–62.
2. Andrews, S. (2009a). *Data Conversion and Interoperability for FCA*. In: Proceedings of the 4th Conceptual Structures Tool Interoperability Workshop (CS-TIW) 2009, held in conjunction with the 17th International Conference on Conceptual Structures (ICCS) 2009: “Leveraging Semantic Technologies”. pp. 42–49.
3. Andrews, S. (2009b). *In-Close, a Fast Algorithm for Computing Formal Concepts*. In: Rudolph, Dau, Kuznetsov (Eds.): Supplementary Proceedings of ICCS’09, CEUR WS 483.
4. Andrews, S. and Orphanides, C. (2012). *Knowledge Discovery Through Creating Formal Contexts*. In: International Journal of Space-Based and Situated Computing, **2**(2). pp. 123–138.
5. Andrews, S. and Orphanides, C. (2010). *FcaBedrock, a Formal Context Creator*. In: “Conceptual Structures: From Information to Intelligence”, Proceedings of the 18th International Conference on Conceptual Structures (ICCS) 2010, LNAI 6208. pp. 181–184.
6. Bache, K. and Lichman, M. (2013) *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.
7. Candela, L., Athanasopoulos, G., Castelli, D., El Raheb, K., Innocenti, P., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G. and Ross, S. (2011) *The Digital Library Reference Model*. Last accessed 01 April 2013 at: <http://bscw.research-infrastructures.eu/pub/bscw.cgi/d222816/D3.2b%20Digital%20Library%20Reference%20Model.pdf>
8. DSpace Cambridge (2012). *Digital Repositories*. [online]. Last accessed 18 April at <http://www.lib.cam.ac.uk/dataman/pages/repositories.html>
9. Georgiou, G. (2013). *FCAWarehouse, an Online FCA Data Repository*. BSc (Hons) Computing final year project, Sheffield Hallam University, Sheffield, UK, 2013.
10. Scherle, R., Carrier, S., Greenberg, J., Lapp, H., Thompson, A., Vision, T. and White, H. (2008). *Building Support for a Discipline-Based Data Repository*. In: Third International Conference on Open Repositories 2008, Southampton. Last accessed 20 March 2013 at: [http://pubs.or08.ecs.soton.ac.uk/35/1/submission\\_177.pdf](http://pubs.or08.ecs.soton.ac.uk/35/1/submission_177.pdf)