Journal of Universal Computer Science, vol. 10, no. 6 (2004), 740-748
submitted: 26/1/04, accepted: 22/3/04, appeared: 28/6/04 © J.UCS

An Investigation into Sharing Metadata: "I'm not thinking what you are thinking"

Simone Stumpf
(University College London, United Kingdom
S.Stumpf@cs.ucl.ac.uk)

Janet McDonnell
(University College London, United Kingdom
J.McDonnell@cs.ucl.ac.uk)

Abstract: A small collection of metadata concepts has been jointly negotiated among a group of specialists to be relevant for classifying data used in their field. A series of comparisons are made to test levels of agreement between individuals when these concepts are used to tag data items. Inter-coder agreement measures are presented for a range of data sets and individuals with varying relationships to the data sets. The implications of the results for the use of metadata as a supporting mechanism for knowledge sharing are discussed.

Keywords: metadata, knowledge sharing, agreement
Categories: H.1, H.3, H.4

1 Introduction

Metadata is increasingly used to facilitate the management and sharing of information. Metadata usually refers to information about information; we will refer to metadata as a collection of concepts that describe the semantic content of a piece of information. Knowledge sharing benefits domain specialists by enabling richer structuring of their own knowledge. In our work, we are interested in how security specialists who investigate internal theft in the retail sector can learn from each other's experience of dealing with theft cases and how this can be supported.

Descriptions of content can be developed with or without human intervention. Automatically derived content descriptions based on word frequency and document structure underpin most web-based search engines. In this paper, we concentrate on content descriptions which have been elicited directly from domain specialists. Metadata – like ontologies to which it is related – can be developed by experts top-down, bottom-up or middle-out [Motta et al. 2000] and we focus on a collection of metadata concepts which have been derived in a middle-out fashion. In this approach, basic concepts of the domain are identified by looking at data items and then they are specialised or generalised.

It has been noted that sharing metadata, however it is produced, poses problems [Hameed et al. 2002] [Correa da Silva et al. 2002] and some strategies have been proposed to overcome these [Davies et al. 2003]. Here we report on a case study that explores issues related to the sharing of metadata. We investigate in detail the level of

agreement in applying content descriptions and highlight results from the analysis of consensus between and within user communities. Furthermore, we discuss approaches that can be employed to resolve differences in content descriptions.

2 An investigation into sharing metadata

2.1 Background

The work reported is part of a project concerned with the capture, representation and sharing of knowledge about dealing with theft by employees in retail organisations. Security specialists from nine organisations have taken part in knowledge acquisition sessions and provided data to support the development of a comprehensive profile of staff theft in the UK retail sector. The sharing of metadata was instigated to produce a common set of benchmark features relevant to counteracting staff theft.

Previously, we described the 'theorise-inquire' technique for expressing knowledge, theorising from it, identifying data suitable for testing theories, and the value to a business of the outcomes it produces [Stumpf and McDonnell 2003]. This procedure supports the validation of knowledge once it is expressed in a shareable form and draws attention to gaps in data and to information quality generally.

As part of this process, we characterise stereotypical situations of staff theft using the repertory grid technique [Kelly 1955]. As the technique is applied, individual situations that the expert has experienced are compared to draw out distinctions between them. Dimensions or ratings are associated with each distinction. In a completed grid, each exemplar experience is characterised by ratings associated with each of the distinctions identified.

The detection of conceptual differences, without the help of metadata, has received some attention as part of the repertory grid technique [Shaw and Gaines 1989] [Hill 1995]. In our work, we use metadata, firstly, as a structuring mechanism that allows domain specialists to form their own viewpoint on important features of staff theft; secondly, we use it as a communication aid to develop shared viewpoints on information amongst peers. Here we report on an investigation into the reliability of metadata application by domain experts, especially when contrasted to its application by researchers.

2.2 Methodology

Nine separate repertory grids were elicited from retailers concentrating on features of stock theft or refund fraud. Each grid represented the way that cases of internal theft were characterised to support investigations by a particular retail company. Each specialist, or team of specialists, who contributed to the elicitation exercise provided their own exemplar situations and sets of distinctions for characterising them; hence, the classification system of conceptual differences [Shaw and Gaines 1989], which requires common exemplars or distinctions, is not suitable to compare these grids.

When we looked at sharing repertory grids between organisations directly to investigate the amount of overlap that experts themselves can identify between different organisational settings we found that the knowledge an organisation holds

cannot easily be shared across organisations although, in the case of our work, organisations were willing to learn from each other [Stumpf and McDonnell 2002].

To try to improve experience sharing across organisations a metadata framework consisting of six semantic content descriptions was developed in a middle-out, goal-directed manner as follows. Initially, a crude structuring of a limited number of data items (for purposes of offender profiling) were gathered from domain experts. This crude structure was refined and extended by abstracting semantic content from all remaining data items. Following this, a metadata framework relevant for loss prevention was negotiated and agreed by security specialists in a joint workshop session. The framework consists of concepts that describe data items in terms of offender profiles, offence characteristics, investigation processes, investigational outcomes, impact on the business and opportunities presented allowing staff theft to happen. The list of metadata concepts and their definitions are shown in table 1.

Metadata concept	Applies to data items that
Offence	Deal with behaviour by the perpetrator in committing the fraudulent activity within the retail company.
Offender	Give details of the perpetrator of staff theft.
Investigation	Specify details of the investigation process carried out by a company to counteract or detect a particular instance of staff theft.
Outcomes	Give the outcomes of an investigation of staff theft.
Impact	Contain details of the financial and other impacts for the retailer in terms of an instance of staff theft.
Opportunities	Deal with the company environment in a particular store or the company as a whole that provide opportunities for staff theft to occur.

Table 1: Metadata concepts and their definitions

Each organisation was then asked to classify their *own* data using the agreed, *common* metadata concepts. Independently from the domain experts, two researchers also applied the metadata concepts to each of the data sets. An example data set extract and its classification by domain experts and researchers is shown in figure 1.

Repertory grid data item			Metadata tag from domain expert (D8)	researcher R1	researcher R2
stayed the same	vs.	relocated / closed-reopened	Opportunity	Opportunity	Outcome
sells travel products	vs.	does not sell travel products	Opportunity	Opportunity	Opportunity
not London shop	vs.	London shop	Offence	Opportunity	Opportunity
disciplinary proceedings against staff	vs.	no discipline problems	Offender	Opportunity	Offender
low turnover	vs.	high turnover	Opportunity	Opportunity	Opportunity
frequent stock take	vs.	infrequent stock take	Opportunity	Opportunity	Opportunity
airport shop	vs.	not airport shop	Opportunity	Opportunity	Opportunity
busy non-stop	vs.	busy at certain times	Opportunity	Opportunity	Opportunity
adequate staffing levels	vs.	inadequate staffing levels	Opportunity	Opportunity	Opportunity
Continued theft	vs.	Opportunist theft	Offence	Offence	Offence
Full-time employee	vs.	Part-time employee	Offender	Offender	Offender
Suspect in charge of store at time of offence	vs.	Suspect not in charge of store at time of offence	Offender	Offender	Offender
Close of play	vs.	Trading hours	Offence	Offence	Offence
Detained at time of offence	vs.	Detained after offence	Offender	Outcome	Outcome
Offence witnessed	vs.	Offence not witnessed	Offence	Investigation	Investigation
Suspect mid 30s	vs.	Teenagers	Offender	Offender	Offender
Store manager	vs.	Deputy manager	Offender	Offender	Offender
Suspect in charge of the store	vs.	Not in charge of the store	Offender	Offender	Offender
own use	vs.	selected items stolen to order	Offence	Offence	Offence
Low value	vs.	High value	Offence	Offence	Offence
Low volume (£100)	vs.	Large volume (£15K)	Offence	Offence	Offence
Uniformed police	vs.	CID	Investigation	Investigation	Investigation
Long service 10 years or more	vs.	New to business less than six months	Offender	Offender	Offender
Well respected	vs.	Disliked	Offender	Offender	Offender
Admission of guilt	vs.	No admission	Offender	Outcome	Outcome
CCTV evidence	vs.	No CCTV evidence	Investigation	Investigation	Investigation
Keyholder involved	vs.	No keyholder involvement	Offence	Offender	Offender
Uniformed guards	vs.	No uniformed guards	Opportunity	Opportunity	Opportunity
small store	vs.	Larger store	Opportunity	Opportunity	Opportunity
Weekend staff	vs.	Weekday staff	Offender	Offender	Offender
Previous history of staff theft in store	vs.	No previous history in store	Offence	Opportunity	Opportunity

Figure 1: Extract from a data set showing metadata assignments

Metadata classifications of repertory grid data – datasets S1 to S11 – have been obtained from two researchers – R1 and R2, respectively – and nine domain experts – D1 to D9, respectively. Table 2 shows the data sets considered in the analysis, their size and from whom they originated.

Data set	Number of data items in data set	Original source
S1	69	D1
S2	67	D2
S3	44	D3
S4	55	D4
S5	70	D5
S6	57	D6
S7	26	D7
S8	61	D8
S9	44	D9
S10	14	Common subset of data items in S1 and S2
S11	19	Common subset of data items in S4 and S5

Table 2: Data set description

The results from applying metadata tags to data sets was then analysed for agreement between the users employing the Kappa statistic. The Kappa statistic is proposed as a measure of inter-coder agreement on category placement [Carletta 1996]. The Kappa coefficient is given by

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times the coders agree and P(E) is the maximum number of times that one would expect them to agree by chance [Siegel and Castellan 1988]. This measure does not take into account any weighting in favour of experienced or naïve coders, or indeed any indication of the severity of disagreement. Kappa scores below 0.4 signify poor agreement, between 0.4 and 0.6 fair agreement, between 0.6 and 0.75 high agreement and above 0.75 signifies excellent agreement [Fleiss 1981].

An analysis of Kappa intra-coder reliability and inter-coder agreement was carried out. The results of this analysis are reported by examining the level of agreement of metadata tagging as follows: 1) of a single researcher over time 2) between researchers and domain specialists 3) between researchers, and 4) between domain specialists.

2.3 Results

The first condition evaluates if a researcher consistently applied the metadata to the same data, measuring the intra-coder reliability – this addresses the question "Am I still thinking in the same way as I thought before?". To this end, two different subsets of data, S10 (a common subset of S1 and S2 containing 14 data items) and S11 (a common subset of S4 and S5 containing 19 data items) were tagged with metadata by researchers R1 and R2 on two separate occasions; each set was therefore seen twice by the same researcher. The results of intra-coder reliability for S10 and S11 are presented in table 3.

Coders	Data set	K	Assessment
R1, R1	S10	0.78	Excellent
R1, R1	S11	0.80	Excellent
R2, R2	S10	0.80	Excellent
R2, R2	S11	0.55	Fair

Table 3: Kappa scores for a researcher over time

It can be seen that, mainly, the researchers apply the metadata consistently. Where only fair consistency was achieved further investigation showed that researcher R2 provided metadata descriptions that partially matched (the Kappa statistic only takes into account full matches). If these partial matches are taken into account, R2's consistency is higher than reported (K = 0.80).

The second condition – the question "Are domain specialists thinking in the same way as researchers?"– tested the agreement between domain specialists from whom the data originated and researchers. Kappa scores were again calculated; these results are listed in table 4.

Coder	Data set	K	Assessment
R1, D1	S1	0.21	Poor
R2, D1	S1	0.15	Poor
R1, D2	S2	0.32	Poor
R2, D2	S2	0.29	Poor
R1, D3	S3	0.77	Excellent
R2, D3	S3	0.77	Excellent
R1, D4	S4	0.57	Fair
R2, D4	S4	0.47	Fair
R1, D5	S5	0.67	High
R2, D5	S5	0.59	Fair
R1, D6	S6	0.77	Excellent
R2, D6	S6	0.77	Excellent
R1, D7	S7	0.37	Poor
R2, D7	S7	0.46	Fair
R1, D8	S8	0.62	High
R2, D8	S8	0.63	High
R1, D9	S9	0.50	Fair
R2, D9	S9	0.40	Fair

Table 4: Kappa scores for agreement between researchers and domain specialists

The results show that agreement between researchers and domain specialists show a great deal of variance, ranging from poor to excellent. Data sets S1, S2 and S7 in particular generated low Kappa scores between researchers and domain specialists.

We can also compare the agreement between the researchers themselves to address the question, "Is a researcher thinking in the same way as another researcher?". Agreement figures between researchers' descriptions were then calculated and are presented in table 5.

Coder	Data set	K	Assessment
R1, R2	S1	0.4	Fair
R1, R2	S2	0.49	Fair
R1, R2	S3	0.88	Excellent
R1, R2	S4	0.79	Excellent
R1, R2	S5	0.63	High
R1, R2	S6	0.76	Excellent
R1, R2	S7	0.85	Excellent
R1, R2	S8	0.87	Excellent
R1, R2	S9	0.84	Excellent

Table 5: Kappa scores for agreement between researchers

In general, very high Kappa figures can be achieved between researchers; in comparison, the agreement between researchers and domain specialists presented in table 3 is comparatively lower, i.e. the researchers readily agree with each other on metadata application but not with the domain specialists from which the data originated. In the case of data set S7 the researchers agree with each other to a considerable amount; this is in contrast to results in table 3 which shows that researchers do not agree with the domain specialist in the application of metadata to a great extent. It should be noted that the application of metadata to data sets S1 and S2 still generates low agreement figures, even between researchers, i.e. they do not agree with the domain specialist in the application of metadata nor with each other.

Finally, the agreement between domain specialists in applying metadata to the same data set was investigated, answering the question "Is a domain specialist thinking in the same way as another domain specialist?". Two subsets of distinctions, S10 and S11 were analysed; for each data set, one of the domain specialist was the original source of the data items, the other was eager to adopt them. Both parties claimed that these data items made sense to them and that they are relevant to staff theft. Data set S10 comprises distinctions which were originally elicited from domain specialist D2 and then taken up by D1 into his repertory grid; data set S11 includes distinctions elicited from D4 and received by D5 into her repertory grid. Hence, S10 is a common subset of data items contained in S1 and S2 and S11 is a common data subset of S4 and S5. Each domain specialist applied the metadata independently to their data sets. The results of the analysis are presented in table 6; data sources are marked with *.

Coder	Data set	K	Assessment
D1, D2*	S10	0.00	Poor
D4*, D5	S11	0.54	Fair

Table 6: Kappa scores for agreement between domain specialists

The results show that there is no agreement between domain specialists for metadata application in the instance of data contained in data set S10, whereas there is fair agreement on metadata descriptions in the case of data set S11. It should be noted that S1 and S2 already generated low levels of agreement in other conditions and this is also reflected in data set S10, which is a common subset of S1 and S2.

3 Discussion of results and potential solutions

The results that have been presented here point of a variety of causes of disagreements in sharing metadata; there are a number of reasons why people do not think what other people are thinking.

One of the reasons for disagreement can simply be traced back to issues with the metadata concepts and definitions themselves. This accounts for the diversity and low level of Kappa figures seen in the agreement between researchers and domain specialists in table 4. Researchers' interpretation of the metadata may be different from that of specialists who have direct experience of the domain. Hence, one

possible strategy to adopt to overcome this problem is to explicitly designate ownership of a classification. This strategy assumes that whoever is the source of data is always right in their application of metadata to its description. Hence the focus shifts from sorting out problems with the classification – enforcing one view on the domain – to keeping track and matching differing classifications – supporting sub-views on the domain.

However, problems with metadata concepts do not fully explain the differences in the level of consensus. Firstly, researchers were able to use the metadata reliably (see table 3); this suggests that the metadata is – for them – stable and unambiguous, as it can be applied consistently over time. Secondly, it appears that problems arise with only certain data sets, such as data sets S1 or S2 (see table 4, table 5 and table 6). These generated lower agreement levels across the board whereas other data sets achieved good agreement between all communities applying metadata. A further possible cause of disagreement may therefore be attributable to the quality of data items to which meta-tags are applied. Whilst distinctions in a repertory grid can be readily understood by the originating domain specialists, it is important to refine data items that do not provide enough contextual information to the point where other coders can apply metadata consistently if the purpose is to arrive at a shared understanding and consistency of thinking. However, as table 6 shows, even where data items are claimed to be commonly understood and of relevance, sharing of metadata is problematic.

Lastly, applying metadata to data items appears to be strongly influenced by previous experience with coding. Coding is a skill; those with a coding background – i.e. researchers – tend to use metadata consistently in the same way as each other (table 5). This is highlighted by results from applying metadata to data set S7, where researchers' agreement is very much higher than that between researchers and domain specialists. Lack of skill in applying metadata may be one cause of low agreement between user communities; this means that it may be that neither the metadata nor the data items to which metadata is applied are at fault, instead it points to the lack of a skill which needs to be developed among domain specialists.

References

[Boose 88] Boose, J.: "Uses of repertory grid-centred knowledge acquisition tools for knowledge-based systems"; International Journal of Man-Machine Studies, 29 (1988), 287-310

[Carletta 1996] Carletta, J.: "Assessing agreement on classification tasks: the kappa statistic"; Computational Linguistics, 22(2) (1996) 249-254

[Correa da Silva et al. 2002] Correa da Silva, F.S., Vasconcelos, W.W., Robertson, D.S., Brilhante, V., de Melo, A.C.V., Finger, M., Agusti, J.: "On the insufficiency of ontologies; problems in knowledge sharing and alternative solutions"; Knowledge-Based Systems, 15 (2002) 147-167

[Davies et al. 2003] Davies, J., Duke, A., Sure, Y.: "OntoShare – An Ontology-based Knowledge Sharing System for Virtual Communities of Practice"; Proceedings of I-KNOW'03, Graz, Austria, July 2-4 (2003)

[Fleiss 1981] Fleiss, J.L.: "Statistical methods for rates and proportions"; Wiley, Chichester, UK (1981)

[Hameed et al. 2002] Hameed, A., Sleeman, D., Preece, A.: "Detecting mismatches among experts' ontologies acquired through knowledge elicitation"; Knowledge-Based Systems, 15 (2002) 265-273

[Hill 1995] Hill, R.: "Content analysis for creating and depicting aggregated personal construct derived cognitive maps", Advances in Personal Construct Psychology, 3 (1995) 101-132

[Kelly 1955] Kelly, G. A.: "The Psychology of Personal Constructs"; WW Norton & Company, New York, (1955)

[Motta et al. 2000] Motta, E., Buckingham Shum, S., Domingue, J.: "Ontology-driven document enrichment: principles, tools and applications"; International Journal of Human-Computer Studies, 52 (2000) 1071-1109

[Shaw and Gaines 1989] Shaw, M.L.G., Gaines, B.R.: "Comparing conceptual structures: consensus, conflict, correspondence and contrast"; Knowledge Acquisition, 1, 4 (1989) 341-363

[Siegel and Castellan 1988] Siegel, S., Castellan Jr, N.J.: "Nonparametric Statistics for the Behavioural Sciences"; MacGraw-Hill, New York (1988)

[Stumpf and McDonnell 2002] Stumpf, S.C., McDonnell, J.T.: "Is there an argument for this audience?"; Proceedings of the 5th Conference of the International Society for the Study of Argumentation, Amsterdam, 25-28 June, (2002) 981-984

[Stumpf and McDonnell 2003] Stumpf, S., McDonnell, J.: "Data, Information and Knowledge Quality in Retail Security Decision Making"; Proceedings of I-KNOW'03, Graz, Austria, July 2-4 (2003)