

Semantic Web Techniques for Multiple Views on Heterogeneous Collections: A Case Study

Marjolein van Gendt^{1,2}, Antoine Isaac^{1,2,3}, Lourens van der Meij^{1,2},
and Stefan Schlobach¹

¹ Vrije Universiteit Amsterdam

² Koninklijke Bibliotheek, Den Haag

³ Max Planck Institute for Psycholinguistics, Nijmegen
{mtvgendt, aisaac, lourens, schlobach}@few.vu.nl

Abstract. Integrated digital access to multiple collections is a prominent issue for many Cultural Heritage institutions. The metadata describing diverse collections must be interoperable, which requires aligning the controlled vocabularies that are used to annotate objects from these collections. In this paper, we present an experiment where we match the vocabularies of two collections by applying the Knowledge Representation techniques established in recent Semantic Web research. We discuss the steps that are required for such matching, namely formalising the initial resources using Semantic Web languages, and running ontology mapping tools on the resulting representations. In addition, we present a prototype that enables the user to browse the two collections using the obtained alignment while still providing her with the original vocabulary structures.

1 Introduction

Integrated access to multiple digital collections is a prominent issue within many research departments of Cultural Heritage (CH) institutions. These collections contain different kinds of objects, with different subjects, are described using different annotation schemes and controlled vocabularies and might be stored in and be accessible via different information systems: they are heterogeneous.

To access several such sources via one portal, one first needs to obtain syntactic interoperability by building a system that can get information from all sources simultaneously, using standard protocols or shared metadata schemes.

However, to maximally use the original resources, integrated systems should also tackle the hitherto unsolved *semantic interoperability* problem, i.e. properly take into account the conceptual similarities and differences between collections. Linking subject descriptors from the vocabularies used to annotate the different collections (e.g. between “birds” in one vocabulary and “flying beings” in another one) provides such interoperability. But it is also necessary to keep the original semantics of these vocabularies, such as found in the hierarchical relations between descriptors (e.g. “birds” as specialization of “animals”).

The general aim of our project, STITCH¹, is to determine to what extent *Semantic Web* (SW) techniques, such as ontology alignment, can solve these interoperability issues. As CH vocabularies are similar to ontologies, adapting this research to the CH sector seems promising.

Our first experiment and implementation aimed at providing integrated access to two heterogeneous collections, the Illuminated Manuscript collection² from the Dutch National Library (KB), and the ARIA Masterpieces collection³ from the Rijksmuseum in Amsterdam. In this paper, we describe the concrete steps of this experiment. First, a conversion to generic formats, such as RDF(s)⁴ and SKOS⁵, was required to provide integrated access to semantically linked CH collections. Second, we could align them using these computer-readable representations. We turned to two off-the-shelf ontology mappers (S-Match [8] and Falcon [11]) and evaluated their use for aligning CH controlled and structured vocabularies. Third, automatically found correspondences were used in a purpose-built interface for browsing different vocabularies and retrieving documents from several collections in parallel, based on the multi-faceted browsing paradigm.

As said, the goal of the research described in this paper is to evaluate the potential and limits of current Semantic Web technology for integrating multiple CH collections with heterogeneous vocabularies. Our main research questions are:

1. Are the current SW techniques suitable for solving this integration problem?
2. Are there specific CH problems that need particular efforts from the SW community?

The paper is structured as follows. In Section 2 we introduce our case study, by describing the two collections we aligned. In Section 3 we describe our solution to the problem from a practical perspective. In Section 4 we then discuss the relevance of our findings for both CH and SW practitioners, before we relate our work to existing work, and conclude.

2 Case Study: Illuminated Manuscripts and Masterpieces

The Illuminated Manuscripts and Masterpieces collections contain objects such as images, drawings, books and/or sculptures. Most interesting for us is the heterogeneity of the vocabularies used to describe these collections.

The Manuscripts collection contains 10.000 medieval illuminations which are, in addition to the standard bibliographical information, annotated by subject indices describing the content of the image. These indices come from the Iconclass classification scheme, a 25.000 element vocabulary with iconographical analysis as main purpose. An Iconclass *subject* consists of a *notation* – an alphanumeric identifier used for annotation – and a *textual correlate* – e.g. 25F9

¹ SemanTic Interoperability To access Cultural Heritage, <http://stitch.cs.vu.nl>

² <http://www.kb.nl/kb/manuscripts/>

³ <http://www.rijksmuseum.nl/collectie/index.jsp?lang=en>

⁴ <http://www.w3.org/RDF/>

⁵ <http://www.w3.org/2004/02/skos/>

mis-shapen animals; monsters. Subjects are organized in nine hierarchical trees. Other features are associative *cross-reference* links as well as mechanisms for subject specialisation, such as *keys* – e.g. 25F9(+33) would refer to the head of a monster. Additionally, subjects have simple *keywords* used for retrieving them: 25F9 is thus linked to “monster” and “shape”, amongst others. It is important to note that textual correlates are often in the form of glosses, e.g. Noah’s sacrifice; various animals are offered, possibly a lamb, a dove and a ram (often combined with the rainbow of the covenant).

The Masterpieces collection contains 700 objects such as paintings and sculptures and its subjects are indexed using the ARIA “catalogue”. This controlled vocabulary, conceived mainly as a resource for browsing, consists of about 500 terms and three sub-vocabularies. The first is intended for the layman, and contains subjects like Man, while the second is for more advanced users: it contains similar but finer-grained subjects like Male portraits. A third very small list – 6 types of objects, like Sculpture – is used as a high-level entry point to the system. The only “semantic” information found in this catalogue consists of specialisation links within the first two vocabularies, that can be interpreted as classical “Broader Than” relationships. The hierarchies are only two levels deep and there are occurrences of multiple inheritance.

3 Performing the Case Study

In this section we describe our approach for providing access to the integrated Illuminated Manuscripts and ARIA Masterpieces collections. Figure 1 shows our framework in a schematic way. In a first step we transform both collections and their respective thesauri into Semantic Web compliant representation languages. Secondly, we create an alignment between the two thesauri using existing mapping technology. Finally, we build a browser to access the linked collections.

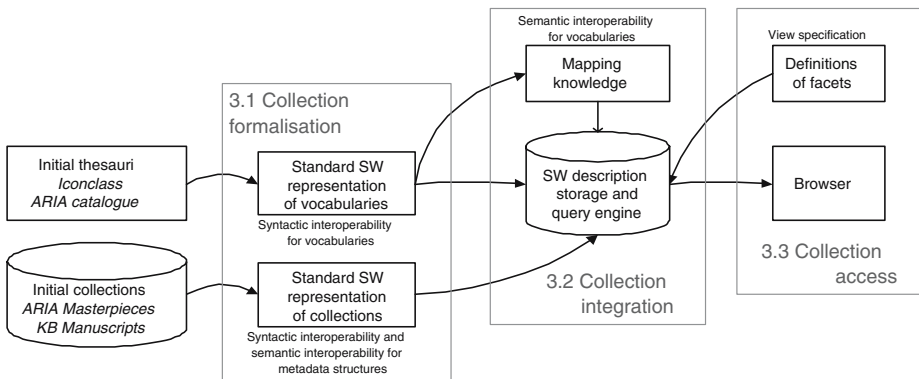


Fig. 1. The different steps of our experiment

3.1 Collection Formalisation

This case study supplies two types of CH resources that need transformation: the controlled vocabularies and the collections themselves.

Converting controlled vocabularies. There have been substantial methodological efforts concerning the conversion of CH vocabularies into SW formats. Similar to [1], we handle the knowledge acquisition process in two steps: first, analysing the sources our use-case provided, and second, formalising the knowledge they contain. This last step involves two consecutive conversions, to first get a standard representation and then an application-specific one.

Analysis. We had the controlled vocabularies as well as significant expert feedback at our disposal. As the vocabularies differ significantly in nature and use, we expected them to be difficult to represent using the same formal apparatus. The question was whether to take all peculiarities of the respective vocabularies into account, or to turn to some standard model. We opted for the latter, as we wanted to test a process – both for representation, alignment and exploitation – that could be generalized to a wider range of vocabularies.

Standard formalisation. The SKOS (Simple Knowledge Organisation System) initiative provides a standardized model to encode the most common knowledge organization schemes, such as thesauri or classification schemes, in SW languages. SKOS is an RDF vocabulary that is currently being developed within the W3C Semantic Web activity. ARIA proved almost fully compatible with the SKOS schema. We only managed to convert Iconclass subjects partly: SKOS could not cope with Iconclass idiomatic elements, such as keys.

Application-specific formalisation. Tools such as storage engines or browsers should interpret the SKOS files in accordance with their intended semantics. This often requires tweaking, e.g. to make our generic RDFS engine deal with the transitivity of the SKOS `broader` relation we had to interpret it as a sub-property of RDFS `subClassOf`.

Converting collection elements. Our main focus being description *vocabularies*, we just used the description *structures* as they were in the original collections, without enforcing a unified scheme like Dublin Core. From the two metadata schemes we constructed small metadata ontologies in RDF Schema.

3.2 Collection Integration

Having formalised our CH vocabularies in SW-compliant representations has the advantage that we can use existing ontology mapping tools to align them. We applied two state-of-the-art ontology mappers, Falcon and S-Match.

Falcon [11] is one of the best performing tools⁶ for aligning complex RDFS/OWL ontologies. It relies on a combination of lexical comparison and graph-matching techniques. First, it compares concepts based on the set of weighted terms derived from their lexical “environment”: their own identifiers, labels, comments, but also the ones of their immediate neighbors – parents, children – in the ontology. These similarities are used as input for the second step, which exploits a graph representation of the semantic information and matrix computation processes to finally return equivalence links between the concepts and relations of the compared ontologies.

S-Match. [8] has been developed for mapping vocabularies represented as trees. It has a modular approach where a *lexical* matching component, a background-knowledge component (“*oracle*”) and a *structure-based* mapping module all contribute to computing a mapping between the input trees. In S-Match default configuration, Wordnet⁷ is used as the background knowledge component.

S-Match is not a general ontology mapper, but specializes on hierarchical classification trees used to structure the access to documents. S-Match core mapping method exploits the fact that the meaning of a concept in such a tree is determined by the concepts in the path to the root. Based on the lexical component and the oracle, each concept is associated with a propositional formula representing all its “available meaning”. The mapping relations are then determined by the logical relations between the formulas for the concepts of the to-be-aligned classification trees.

Mapping results. In table 1 some good mappings produced by S-Match are shown, where the first mapping was produced mainly based on lexical mapping, the second using stemming, and the third making use of background knowledge.

Table 1. Some good S-Match mapping results

IC notation	Iconclass textual correlate	Relation	ARIA label
23L	‘the twelve months represented by landscapes’	Less General	‘Landscapes’
25A271	‘(map of) the North Pole’	Less General	‘Charts, maps’
23U1	‘calendar, almanac’	Less General	‘Publications’

Mapping thesauri proved to be difficult for both mappers, and the overall results were less than satisfactory. Evaluation measures for mapping results depend on their intended use. Regarding our intended browsing interface, precision is more important than recall, because we do not want to confront users with useless links. For S-Match a precision of 46% is obtained on a selected subset of Iconclass (1500 concepts) and the complete ARIA thesaurus (500 concepts); 46% of the mappings were correct. Falcon reached a precision of only 16%.

⁶ See the 2005 OAEI campaign, <http://oaei.ontologymatching.org/>

⁷ <http://wordnet.princeton.edu/>

3.3 Collection Access

We implemented a multi-faceted browsing (MFB) framework to evaluate and explore the results of our mapping effort. MFB involves constraining search criteria along – usually orthogonal – aspects of a collection called *Facets*. Here we tuned the MFB paradigm in an atypical way, since we used one category (the *subject* annotation) for defining several facets. Such a setting is possible because objects are often annotated by several subjects. So using one facet to search for “monkey” and another for “landscape” could retrieve pictures of a monkey in a landscape.

For searching through the integrated collections we explored three different views on integrated collections: *single*, *combined*, and *merged view*.

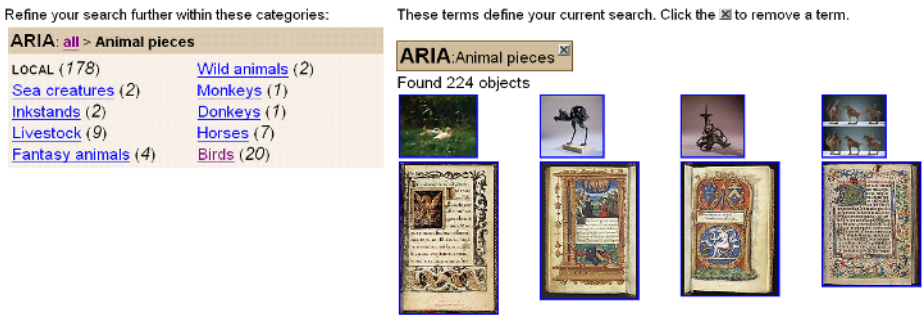


Fig. 2. Single View: Using the ARIA thesaurus to browse the integrated collections

The *Single View* presents the integrated collections from the perspective of only one of the collections. The elements of the other collection are made accessible by means of the correspondences between their subject annotations and the concepts of the current view. In figure 2 the first four pictures come from the Rijksmuseum, the others are Illuminated Manuscripts. Browsing is done solely using the ARIA Catalogue, i.e. these illuminations have been selected thanks to the automatically extracted mapping between ARIA concept “Animal Pieces” and Iconclass “25F:animals”.

The *Combined View* provides simultaneous access to the collections through their respective vocabularies in parallel. This allows us to browse through the integrated collections as if it was a single collection indexed against two vocabularies. In figure 3 we made a subject refinement to ARIA “Animal pieces”, and narrowed down our search with Iconclass to the subject “Classical Mythology and Ancient History”. Only three Manuscripts matched these criteria. Notice that we browse according to a “biological” criterion using ARIA, and a “mythological” one from Iconclass to come to our results.

The *Merged View* provides access to the collections through a merged thesaurus combining both original vocabularies into a single one, based on the links found between them in the automatic mapping process. For figure 4 we made the same selection as for the “single view” case. But notice that the “merged



Fig. 3. Combined View: Using ARIA and Iconclass to browse the integrated collection

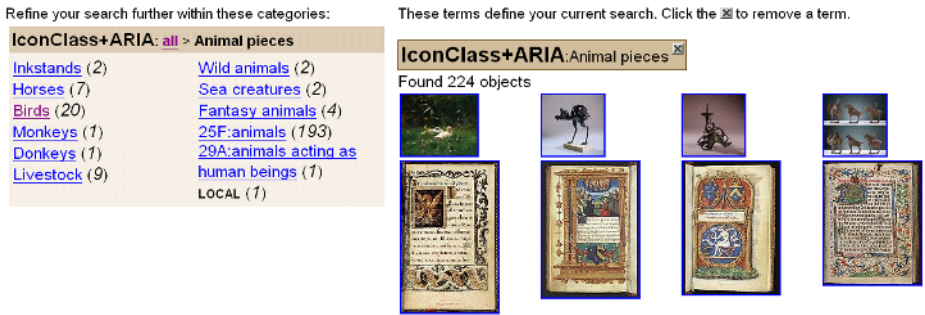


Fig. 4. Merged View: Using a merged thesaurus to browse the integrated collection

view” now provides both ARIA concepts such as “Birds” and an Iconclass concept “29A:animals acting as human beings” for further refining our search. The mapping primitives determine the merging: two concepts that are identified to be equivalent are merged into one new concept, and if the mapping determined that a concept from one scheme is broader than a concept in the other scheme, the second concept is added as a child of the first.

Prototype details. The design of our browser was inspired by the Flamenco search interface framework [9]. It is implemented in SWI-Prolog and uses the Sesame RDF repository⁸ for storage and querying.

4 Lessons Learned

The main goal of our research was to find out to what extent SW techniques can solve heterogeneity issues when integrating multiple CH collections.

The general conclusion is positive: in a relatively short time we managed to implement an integrated browsing environment that was built purely on accepted standards for representing data, and which used existing tools for storage,

⁸ Available on <http://www.openrdf.org>

querying and mapping. However, there is more to be learned for CH collection managers and developers of SW tools alike. In this section, we first try to answer questions concerning the practical relevance of chosen techniques and tools: to which extent can CH use-cases be successfully addressed by such solutions? We then explore the problems raised by our experiment from the point of view of SW researchers. Is our approach methodologically and technologically sound?

4.1 A Cultural Heritage Perspective

Conversion Process. Implementing a realistic process for going from CH resources to SW-compatible formats was successful, but often non-trivial.

Conversion and standards for CH vocabularies. CH vocabularies often rely on complex models that are non-standard, which can hinder the conversion process. Especially for Iconclass some modeling decisions had to be made. For example, for *notations* we used the SKOS `prefLabel` property to enforce the necessary uniqueness constraint, even though notations like 25F9 definitely miss the lexical flavor to make them proper *terms* e.g. `mis-shapen animals; monsters`. Even worse, some features could not be represented at all, like *keys* or the additional network of *keywords*. Potentially interesting information had to be sacrificed for the sake of generality, which illustrates the trade-offs of using standards.

Ontology mapping vs. thesaurus mapping. For our case study we applied off-the-shelf SW ontology mappers. However, CH controlled vocabularies have features that make them really different from ontologies, e.g. glosses for describing concepts instead of simple terms. Here we describe the repercussions these peculiarities have on alignment quality.

Mapping poorly structured schemes. Most ontology mappers rely on structure-based comparison using ontology semantics: subsumption relations, properties, etc. However, thesauri have less strictly defined semantic relations and their consistency is not always enforced. Because of this and the loss of information in the formalisation step, the only usable structural information present in our thesauri is the broader and narrower term hierarchy.

Falcon heavily exploits structure components usually present in expressively modelled ontologies. An analysis of the few correct results from Falcon shows that the lexical mapping works fine, but that the reliance on graph-based techniques usually contributes negatively to the overall process.

S-Match produces much better mapping results, as it was purpose-built for tree-like structures and uses the extensive lexical background information found in Wordnet. Nevertheless, the influence of the difference of the depth levels in both thesauri has unfortunate consequences: the fact that S-Match uses the full path of a classification tree for the mapping implies that its output almost always consist of specialisation links *from* Iconclass concepts *to* ARIA concepts. For browsing, this is very damaging, as it constrains the way a user can specialize her queries: once she is browsing Iconclass subjects, she cannot find ARIA specialisations anymore.

Gloss features and concept matching. The gloss features of concepts cause two anomalies to occur: 1) natural language meaning of a sentence is not interpreted, and 2) the meaning of single terms is not disambiguated by the remainder of their gloss, and thus interpreted as if denoting concepts on their own.

Table 2. Some bad S-Match results

IC notation	Iconclass textual correlate	Relation	ARIA label
23H	‘seasons of the year represented by concepts other than [...] landscapes [...]’	Less General	‘Landscapes’
29D	‘natural forms in stones, wood, clouds’	Less General	‘Jewellery’

An example of a bad match caused by lack of natural language interpretation is the first mapping in table 2: S-Match does not interpret “other than”, which causes 23H to wrongly match **Landscapes**.

Using Wordnet as background knowledge sometimes also leads to finding irrelevant links based on comparing single words, which could have been disambiguated by the other words found in the glosses. In table 2, ‘Jewellery’ would legitimately map to precious stones, but the other tokens in 29D should have provided enough information to disambiguate between the different kinds of stones. An option for improvement would be to focus on smaller but more relevant pieces within Wordnet, e.g. taking only closest siblings into account.

4.2 A Semantic Web Perspective

Generalizability. The *Semantic Web* claims to provide generic solutions. Therefore, the question arises whether it would be easy to reproduce what we did with new collections. Surely, we would benefit from the experience we gained in this case study, and the SW frameworks proved to be flexible enough to cope with different representational choices. But the transformation and mapping process would remain case-study dependent in at least two ways: First, the conversion effort depends on the technical and functional requirements implied by the choice of specific *tools* and *tasks*. Second, both conversion and alignment processes are dependent on the CH *resources*. Take for example the influence of the structure of the vocabularies on the mapping process we discussed in the previous section.

Role of standards. In our approach the role of SKOS was crucial. Such a standard helps to integrate the different components of a framework. It also contributes to improving the extendability of the framework: for example, an additional SKOS-encoded thesaurus could be integrated easily in our tools.

The lack of *de facto* standards for alignment tools was a prominent problem. S-Match takes as input indented trees, which caused an important loss of information. Falcon does better, as it admits expressive standard RDF/OWL ontologies. For output things are even worse: Falcon outputs links in a standardized syntax, but its semantics are unclear. Again, S-Match was less generic, as its output is an ad-hoc non-standard format.

Methodological process guidance. The SW community already got concerned with conversion and deployment of CH vocabularies, and has proposed methodological guidelines. Van Assem et. al. [1], for example, advocate three conversion steps. In the first step, the original vocabulary is translated into an RDFS/OWL model that mirrors the original structure as precise as possible. In the second step, one interprets the model so that intended semantic properties can be explicitly assigned to the RDFS/OWL representation. Finally, one can represent the vocabulary using a standard model like SKOS.

In our experiment we took this process as a guidance, although, focusing on generality and implementation matters, we only applied its last two steps. However, for mapping purposes, the process itself might be questionable. On the one hand, using a standard model only, as described in [2], can help aligning vocabularies: a basic part of the integration process is partly dealt with by conversion. On the other hand, in order to give alignment tools more information for mapping, a conversion step specific to each controlled vocabulary could be beneficial.

Scalability. SW solutions are often criticized for their performing poorly against massive data sets, which are common in the CH world. Indeed, as Falcon uses a complex algorithm, it was practically impossible to have it run on complete Iconclass. Some division had to be done beforehand. However, S-Match performed better: it took five hours to achieve a complete alignment, which is not a problem since our application does not need to compute mappings at runtime.

5 Related Work

Our case study has been influenced by portal projects like The European Library⁹ and the Memory of the Netherlands¹⁰. But these do not use correspondences between vocabularies, though this problem has already been identified in the Digital Library DL field [6]. Some DL projects like MACS [3] or RENARDUS¹¹ have used mappings, but they relied on manual alignment, costly and possibly imprecise. We wanted to explore the use of automatic alignment of concept schemes, like currently done in the SW community. This community produced a number of dedicated tools [12], sometimes inspired by previous database integration efforts [5]¹². However, automatic alignment methods usually lack concrete experiments that would assess the feasibility of integrating them in deployed applications, even when they explicitly focus on the thesaurus field [4].

Our approach is thus closer to settings like [10] or [7] that try to apply SW techniques to concrete (CH) cases, except for our focus on automatic alignment.

⁹ <http://www.theeuropeanlibrary.org>

¹⁰ <http://www.geheugenvannederland.nl>

¹¹ <http://www.renardus.org>

¹² We could have tried to directly turn to such techniques. But while they naturally focus on the structure of data – as encoded in database schemas – we focus on the semantics of descriptors that come in unstructured subject annotations.

Actually [10] also implements faceted browsing; we both were inspired by the Flamenco framework [9]. We could have tried to re-use these solutions; however, availability problems and our need for flexible experiments with various setups made us decide to build our own prototype.

6 Conclusion

In this paper we have presented a case study aiming at solving the semantic interoperability problem in the context of CH resources, using automatic alignment processes between their vocabularies to avoid heavily labour-intensive and ambiguous manual alignment work.

This study provides interesting insights regarding the use of SW techniques in a CH environment. We have seen that the conversion of vocabularies using standardised formats is possible, and helps their deployment. We have also shown that based on such representations and automatically found mappings, an operational interface for browsing heterogeneous collections in an integrated way *can* be implemented.

If all collections and thesauri were available in standard formats (SKOS, RDF) or when automatic conversion is feasible so that translation steps would not be needed anymore, our framework would provide a very easy way of integrating heterogeneous collections. However, there still are problems to solve before this ideal situation occurs:

- we have to overcome the loss of semantics when translating the thesauri into SW standards, for instance by providing more expressive standards,
- ontology mapping tools should be compliant with the SW standards concerning input and output formats, and
- specifically for CH controlled vocabularies, it would be preferable to have a SKOS standard inference engine instead of an RDF(s) one¹³.

Furthermore, all tools (mappers, inference engines) should be scalable for handling the enormous amount of data present in CH.

Concerning the use of ontology mappers for our CH case, we learned that available ontology alignment techniques need to be tuned to be of use for e.g. thesaurus mapping. Most mappers use resources that are absent from thesauri, e.g. properties, and refrain from (properly) using all information found in thesauri, e.g. synonyms. S-Match mapping quality (46%) is a lot higher than Falcon one (16%), but must still be improved to be useful for browsing purposes. Typical features such as gloss descriptions and poor structuring should be taken into account when constructing a thesaurus mapper. So, to perform semantic integration of CH collections the way we envision, automated mapping techniques are indispensable, but should absolutely be adapted.

Finally, our interpretation of Multi-Faceted Browsing provides multiple views or access points for a same set of data. This way users can choose the vocabulary

¹³ Note the discrepancy between this point and the first: the use of standards limits the amount of transferable information, but provides generalizability.

they are most comfortable with and thus personalised access is granted. We encourage readers to try our browser at <http://stitch.cs.vu.nl/demo.html>.

Acknowledgements

STITCH is funded by NWO, the Netherlands Organisation for Scientific Research. We would like to thank KNAW and RKD institutes for Iconclass resources, and Gerda Duijfjes-Vellekoop for helping with these. We also owe the Rijksmuseum and CHIP project people (especially Lloyd Rutledge) for the ARIA material, the researchers from Trento University (especially Pavel Shvaiko) for their help with S-Match, and Henk Matthezing for his constant advice at the KB.

References

1. van Assem, M., Menken, M. R., Schreiber, G. et al.: A Method for Converting Thesauri to RDF/OWL. Int. Semantic Web Conference, Hiroshima, Japan, 2004.
2. van Assem, M., Malaise, V., Miles, A., Schreiber, G.: A Method to Convert Thesauri to SKOS. 3rd European Semantic Web Conference, Budva, Montenegro, 2005.
3. Clavel-Merrin, G.: MACS (Multilingual access to subjects): A Virtual Authority File across Languages. *Cataloguing and Classification Quarterly* 39 (1/2), 2004.
4. Constantopoulos, P., Sintichakis, M.: A Method for Monolingual Thesauri Merging. ACM SIGIR Conference, Philadelphia, USA, 1997.
5. Doan, A. and Halevy, A.: Semantic Integration Research in the Database Community: A Brief Survey. *AI Magazine, Special Issue on Semantic Integration*, 2005.
6. Doerr, M.: Semantic Problems of Thesaurus Mapping. *Journal of Digital Information*, 1 (8), 2004.
7. Gasevic, D., Hatala, M.: Searching Web Resources Using Ontology Mapping. K-CAP Workshop on Integrating Ontologies, Banff, Canada, 2005.
8. Giunchiglia, F., Shvaiko, P., and Yatskevich, M.: Semantic Schema Matching. 13th International Conference on Cooperative Information Systems (CoopIS 2005).
9. Hearst, M., English, J., Sinha, R., Swearingen, K. and Yee, P.: Finding the Flow in Web Site Search. *Communications of the ACM*, 45 (9), 2002.
10. Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A. et al.: MuseumFinland - Finnish Museums on the Semantic Web. *Journal of Web Semantics*, 3(2), 2005.
11. Jian, N., Hu, W., Cheng, G., and Qu, Y.: Falcon-AO: Aligning Ontologies with Falcon. K-CAP Workshop on Integrating Ontologies, Banff, Canada, 2005.
12. Kalfoglou, Y., Schorlemmer, M.: Ontology Mapping: The State of the Art. *The Knowledge Engineering Review Journal*, 18(1), 2003.