Semantic Excavation of the City of Books

Anna Tordai
VU University
1081a De Boelelaan
Amsterdam, Netherlands
atordai@cs.vu.nl

Borys Omelayenko
VU University
1081a De Boelelaan
Amsterdam, Netherlands
b.omelayenko@cs.vu.nl

Guus Schreiber
VU University
1081a De Boelelaan
Amsterdam, Netherlands
schreiber@cs.vu.nl

ABSTRACT

As the Semantic Web gains momentum, so grows the interest in making knowledge kept in various repositories available. In this paper we describe a case study using a methodological approach for porting cultural repositories to the Semantic Web. The approach consists of thesaurus conversion, meta-data schema mapping, meta-data value mapping, and thesauri alignment. It is derived from our experience collected in a number of conversions we have performed for the E-Culture project, and in this paper we apply it to a collection of data about images related to book printing.

1. INTRODUCTION

In this work we present a case study based on the four activities presented as a poster in [5] at K-Cap 2007. These activities are necessary for converting cultural heritage data into RDF/OWL. The context of this work is the MultimediaN E-Culture project [4][1], a leading Semantic Web project that won the Semantic Web Challenge in 2006. The objective of this project is to create a large virtual collection of cultural heritage objects that supports semantic search. Metadata and vocabularies are represented in RDF/OWL. The project demonstrator (see the demonstrator at the project website) includes multiple vocabularies which are partially semantically aligned.

This paper builds on earlier conversions of metadata and thesauri and their commonalities. There are currently 5 collections and 6 thesauri that are part of the E-Culture demonstrator. Among them are the collections from the Royal Tropical Institute (KIT)[2] in Amsterdam and the National Museum of Ethnology (RMV)[3] in Leiden. The thesauri include three from Getty[4]: the Art and Architecture Thesaurus (AAT), the Thesaurus of Geographical Names (TGN) and the United List of Artist Names (ULAN), as well as the Dutch Ethographic Collection Foundation (SVCN)[5] thesaurus. These form "standard" vocabularies in the cultural heritage field, meaning various institutions have agreed upon, and approved their usage. "Local" thesauri or vocabularies on the other hand are often created or maintained by a single institution or person.

The objective of the present work is to describe the conversion of the Bibliopolis[6] collection (Latin for city of books) and its alignment to existing vocabularies performed within the E-Culture project. We follow the four-step process described in [5] to convert the thesaurus and metadata such that these become an interoperable part of the virtual collection. The Bibliopolis collection consists of images related to book-printing, and range from photographs of publishing houses to illustrations of the printing process and a local thesaurus of keywords. It is a good example of the range of data we come across when dealing with cultural heritage collections and vocabularies.

To represent the collections the project uses a specialization of Dublin Core (DC) for visual resources (all objects in the virtual collection are required to have an image as their data representation) as the guiding metadata scheme. This Dublin Core specialization is named the Visual Resources Association Core (VRA)[7] scheme which follows the Dublin Core dumb-down principle (i.e. it is a proper specialization and does not contain extensions). Likewise, we model collection-specific metadata schemes as specializations of VRA.

For the representation of thesauri the project uses the SKOS Core Schema[8]. It was designed to support vocabulary interoperability and is currently undergoing standardization by the World-Wide Web Consortium (W3C). SKOS has already been adopted by large organizations such as NASA.

This paper is organized as follows. We discuss related work in Section 2. We present our approach in Section 3 followed by a short presentation of the Bibliopolis data in Section 4. Next, we devote four sections to describe the case study based on the following four activities: thesaurus con-

[1] http://e-culture.multimedian.nl
[2] http://www.kit.nl/
[3] http://www.rmv.nl/
[4] http://www.getty.edu/research/conducting_research/vocabularies/
[5] Acronym for Stichting Volkenkundige Collectie Nederland http://www.svcn.nl/thesaurus.asp
[6] http://www.bibliopolis.nl/
[7] http://www.vraweb.org/
[8] http://www.w3.org/TR/swbp-skos-core-guide/

version, metadata schema mapping, metadata mapping and thesaurus alignment. Finally, we conclude this paper with a discussion in Section 9.

2. RELATED WORK

In the area of thesaurus conversion Miles et. al. [3] propose guidelines for migrating thesauri to the Semantic Web using the SKOS Core schema. They distinguish between standard and non-standard thesauri, and propose to preserve all information in the thesaurus by using sub-class and sub-property statements where necessary.

The work of Van Assem et. al. [6] is based on these guidelines, and they propose a three step method consisting of the analysis of the thesaurus, mapping to the SKOS schema and the creation of the conversion program. The case studies do show however that non standard thesauri are more difficult to convert completely as some features cannot be mapped to the SKOS schema.

The problem of interoperability between two collections has been discussed in [1]. Within the SIMILE project Butler et.al. report on the conversion and linkage of a visual works dataset and learning object dataset using XSLT. The first dataset was converted using the VRA schema and the second using Dublin Core, although non standard properties were created as extensions. Issues discussed range from the creation of URIs to dealing with hierarchical terms.

In [2] Hyvönen et. al. describe the MuseumFinland project encompassing multiple collections and ontologies. The collections of various Finnish museums and additional ontologies were converted into RDF/OWL. The metadata of the collections was transformed using a common term ontology, while the additional ontologies form an additional semantic link between the collections and were further enhanced by manual editing and enrichment.

3. APPROACH

The process developed within the E-Culture project for converting datasets to an interoperable Semantic Web format was presented in [5]. Once again, our goal is syntactic and semantic integration of data. In achieving this goal we are driven by the practical needs of the E-culture project: the need to integrate multiple collections. Accordingly, we follow a practical bottom-up approach where we enrich real-world data with a thin layer of semantics to achieve interoperability. This approach may be seen as an alternative to the top-down approach that is very common in the Semantic Web community. With the top-down approach we would first need to develop a conceptual model of the cultural heritage world in order to be able to perform semantic enrichment of the data. This ontology development effort has not been started yet and such efforts would take several years to be finished. However, there are a number of thesauri available at the moment which are widely used by the cultural communities. In our approach we perform syntactic integration and take the first step towards semantic integration by performing terminological integration. The task of integrating collections and vocabularies from both a structural and terminological perspective has evolved into four activities which are summarized in Fig. 1:

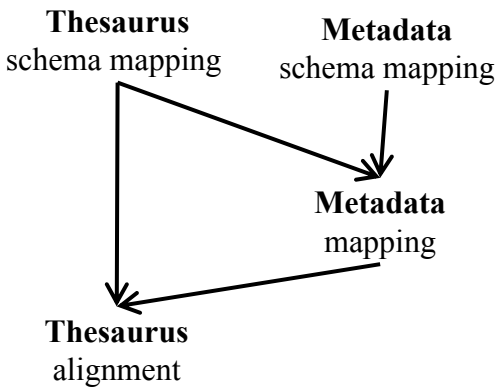


Figure 1: The four activities for converting a collection.

- Thesaurus conversion, including thesaurus schema mapping. This step is a relatively well-researched area, e.g. [6], with SKOS being the default option for thesaurus schema.
- Metadata schema mapping. Here we are looking at generic schemas like Dublin Core and its specializations to the cultural domain, such as VRA.
- Metadata conversion. At this step the data values are converted and looked up in the local thesaurus or external vocabularies using information extraction techniques. Data interpretation is also common here, especially for data that does not directly fit the standard vocabularies.
- Thesaurus alignment. Here we align the thesaurus to external (standard) vocabularies with ontology alignment techniques.

Structural integration is performed during thesaurus schema mapping for vocabularies, and metadata schema mapping for collections. The terminological integration performed during metadata mapping and thesaurus alignment is dependent on the schema mapping activities, which we denote with vertical arrows. As vocabularies tend to be used in collection metadata making this link explicit is part of the semantic enrichment process. Collection metadata in turn may contain implicit vocabularies hidden in data values that are candidates for thesaurus alignment.

4. BIBLIOPOLIS DATA

The Bibliopolis data from the Koninklijke Bibliotheek (KB), the National Library of the Netherlands, consists of two XML files: collection and thesaurus. The collection file contains the metadata of 1,645 images related to the printing of books and book illustrations. The thesaurus contains 1,033 terms used as keywords for indexing images. These two files are a part of the Bibliopolis website. Both the thesaurus and the metadata are bilingual (English and Dutch).

Thesaurus. The thesaurus contains core terms, augmented with their synonyms in plural, and variants of these terms

```
<inm:Record>
    <inm:NUM>2</inm:NUM>
    <inm:TWOND>academiedrukkers</inm:TWOND>
    <inm:TWVAR>academiedrukker</inm:TWVAR>
    <inm:TWVAR>universiteitsdrukker</inm:TWVAR>
    <inm:DEF>aan een universiteit verbonden...</inm:DEF>
    <inm:TWRT>academische geschriften</inm:TWRT>
    <inm:TWRT>overheidsdrukkers</inm:TWRT>
    <inm:ENG>university printer</inm:ENG>
    <inm:INVOERDER>emo</inm:INVOERDER>
    <inm:INVDAT>12/13/01</inm:INVDAT>
    <inm:TWSYN>universiteitsdrukkers</inm:TWSYN>
    <inm:TWBT>drukkers</inm:TWBT>
    <inm:TWNT/>
    <inm:TWOND_EN>university printers</inm:TWOND_EN>
    <inm:TWVAR_EN>university printer</inm:TWVAR_EN>
    <inm:TWVAR_EN>academy printer</inm:TWVAR_EN>
    <inm:TWVAR_EN>academic printer</inm:TWVAR_EN>
    <inm:DEF_EN>a printer appointed by...</inm:DEF_EN>
    <inm:TWSYN_EN>academy printers</inm:TWSYN_EN>
    <inm:TWSYN_EN>academic printers</inm:TWSYN_EN>
</inm:Record>
```

Figure 2: Thesaurus record for term `University printer`

in singular along with a descriptive note. Each record may also contain related, broader and narrower terms. Additionally, it contains some administrative data: initials of the record creator, the date of entry, and the date of modification. A sample XML element for the term `university printer` is shown in Fig. 2.

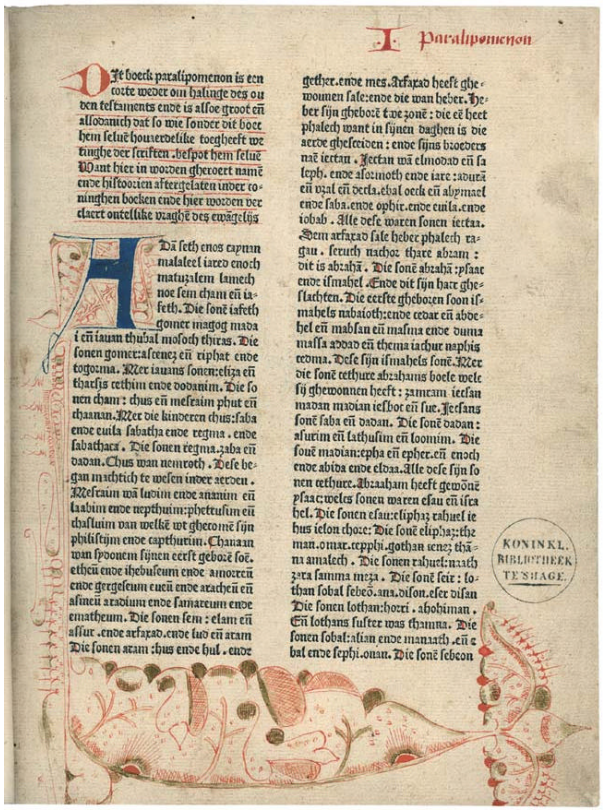
Metadata. The metadata forms the description of images related to book printing. The data consists of titles and descriptions of the objects, names of their creator(s) with signatures of their roles, such as *a* for *author*. The works are also classified according to the technique used, their type, and a library classification of the subject matter. The metadata includes copyright information, measurements and other administrative information. An example collection object plus corresponding metadata is shown in Fig. 3.

5. THESAURUS CONVERSION

Thesaurus schema mapping and conversion is a relatively well-researched area. In our work we used the method for thesauri conversion proposed by van Assem [6]. As for the thesaurus schema, we use SKOS within the E-Culture project.

Mapping the Bibliopolis thesaurus turned out to be relatively straightforward as it fit the SKOS template. Table 1 shows the details of the mapping of the thesaurus representation in Fig. 2 to SKOS. Two XML elements were not converted, as they contained bookkeeping information and are not meant for public consumption. One XML element (see last column in the table) turned out to be a duplicate piece of information and was therefore omitted. It should be noted that this conversion was guided by the requirements of the project which does not include complete conversion of the data.

The creation of the URI deserves special mention. When creating a URI we derive it from the real term identifier followed by the disambiguation signature and the thesaurus version. For example, in the Bibliopolis case the real identifiers are stored in field TWOND (and not NUM that contains



©Koninklijke Bibliotheek (`http://www.kb.nl/`)
Den Haag, Koninklijke Bibliotheek, 169 E 56

```
<inm:Record>
    <inm:NUMMER>6</inm:NUMMER>
    <inm:TITEL>Delftse Bijbel...</inm:TITEL>
    <inm:TITEL_EN>Delft Bible...</inm:TITEL_EN>
    <inm:MAKER>Yemantszoon, Mauricius : d</inm:MAKER>
    <inm:OBJECT>tekstbladzijde</inm:OBJECT>
    <inm:TECHNIEK>boekdruk</inm:TECHNIEK>
    <inm:DATERING>10 jan. 1477</inm:DATERING>
    <inm:CLASSIFICATIE>D</inm:CLASSIFICATIE>
    <inm:ORIGINEEL>Bijbel. Oude
        Testament...</inm:ORIGINEEL>
    </inm:REPRODUCTIE>
    <inm:TWNAAM/>
    <inm:TWOND>typografische vormgeving</inm:TWOND>
    <inm:TWOND>bijbels</inm:TWOND>
    <inm:TWGEO>Delft</inm:TWGEO>
    <inm:OMSCHRIJVING>Eerste bijbel die in het
        Nederlands verscheen...</inm:OMSCHRIJVING>
    <inm:OMSCHRIJVING_EN>The first Bible to
    appear in the Dutch language...</inm:OMSCHRIJVING_EN>
    <inm:AFMETINGEN>27 x 20 cm</inm:AFMETINGEN>
    ...
</inm:Record>
```

Figure 3: A fragment of a real XML record depicting a Delft Bible dated 10 January 1477, originated from Delft, classified with category 'bibles'. (Certain fields may be empty)

Data Item	Function	Activity	Source and Target Property/Class
NUM	Internal identifier	Create literal	*source:* 2 *target:* vra:location.refId "2" ;
TWOND	Preferred term in Dutch	Create URI, literal and language tag	*source:* academiedrukkers *target:* bp:academiedrukkers rdf:type skos:Concept ; skos:prefLabel "academiedrukkers"@nl ;
TWSYN	Synonym in Dutch	Create literal and language tag	*source:* universiteitsdrukkers *target:* skos:altLabel "universiteitsdrukkers"@nl ;
TWVAR	Term in singular form in Dutch	Create literal and language tag	*source:* academiedrukker *target:* skos:altLabel "academiedrukker"@nl ;
DEF	Definition in Dutch	Create literal and language tag	*source:* aan een universiteit verbonden... *target:* skos:definition "aan een universiteit verbonden..."@nl ;
TWBT	Broader term	Look up concept URI and add URI	*source:* drukkers *target:* skos:broader bp:drukkers ;
TWNT	Narrower term	Look up concept URI and add URI	*source:* narrower term *target:* skos:narrower bp:narrower_term ;
TWRT	Related term	Look up concept URI and add URI	*source:* overheidsdrukkers *target:* skos:related bp:overheidsdrukkers ;
TWOND_EN	Preferred term in English	Create literal and language tag	*source:* university printers *target:* skos:prefLabel "university printers"@en ;
TWSYN_EN	Synonym in English	Create literal and language tag	*source:* academy printers *target:* skos:altLabel "academy printers"@en ;
TWVAR_EN	Term in singular form in English	Create literal and language tag	*source: university printer* *target:* skos:altLabel "university printer"@en ;
DEF_EN	Definition in English	Create literal and language tag	*source:* a printer appointed by... *target:* skos:definition "a printer appointed by..."@en ;
ENG	English translation of term	Not converted; duplicate information	*source:* university printer
INVOERDER	Entered by	Not converted: not part of requirements	*source:* emo
INVDAT	Date of entry	Not converted: not part of requirements	*source:* 12/13/01

Table 1: Mapping thesaurus data to SKOS

a file-specific index rather than the real term identifier), they are unambiguous, and we have a single version.

6. METADATA SCHEMA MAPPING

In this activity we map the original record fields (see Fig. 3) to a metadata schema. In the E-Culture project we use the VRA Core scheme which is a specialization of Dublin Core[9] for visual resources (our target type of resources).

Before mapping to the schema we analyze the metadata (including examination of any additional documentation, websites, and interviews with experts). The meaning of the fields needs to be understood to find a correct correspondence within the target schema. The first impression of the meaning of a field might be misleading. For example, the TWGEO field was initially mapped to vra:location, *i.e.*, the DC/VRA element indicating where the work was created. However, the documentation showed that the field actually gives information about the location related to the subject, and not the creation place. We finally used the VRA Core v4 element vra:subject.geographicPlace, which gives the correct interpretation. This element is a subproperty of DC/VRA subject.

An important additional consideration is that certain records or fields may contain confidential or administrative information such as acquisition or bookkeeping information. For example, the amount for which an object is insured should not be publicly visible. This situation did not occur with the Bibliopolis data.

[9] `http://dublincore.org/`

Table 2 shows an overview of the mapping from the XML record fields to a VRA metadata schema with examples. Here we face two situations. First, in the simplest case, there is a exact semantic match between an original field and a VRA field. Second, if this is not the case, the field should be specified as a specialization of an existing VRA element. In the Bibliopolis case this occurs with the ORIGINAL[10], REPRODUCTION and CLASSIFICATION fields. The first two are specific "titles", the third one is a specific "subject" description. In Table 2 we see that the RDF/OWL specification contains property definitions in the Bibliopolis namespace (`bp:`) paired with a statement about the subproperty relationship with a VRA element.

One field requires some deeper study. The MAKER field not only contains the creator of the work, but also a character indicating the role that the person played in creating the work. As shown in the example record in Fig. 3 the MAKER field has the value `Yemantszoon, Mauricius : d`, where "d" stands for "drukker", Dutch for "printer". To preserve the roles of the creators we specialize the VRA property vra:creator with the properties that correspond to the roles found in the Bibliopolis data. This resulted in a set of RDF/OWL definitions such as:

```
bp:drukker rdfs:subPropertyOf vra:creator
bp:origineel rdfs:subPropertyOf vra:title
bp:reproductie rdfs:subPropertyOf vra:title
bp:classificatie rdfs:subPropertyOf vra:subject
```

[10] For readability we use the English in the text, in cases where it is close to the Dutch equivalent ("original" vs. "origineel")

(The example uses the RDF N3 notation).

Dublin Core has excellent general coverage. In all collections we tackled sofar, we were able to find for each field a Dublin Core / VRA which was either an equivalent, or could act as superproperty of a local specialization. This characteristic makes Dublin Core a powerful tool for metadata interoperability.

7. METADATA VALUE CONVERSION

After the schema is created the data values of the fields have to be converted. As discussed in [5] we have two kinds of fields: those that contain free-text literal values, such as a description field, and those that contain values from (implicit) vocabularies, such as the fields for keywords or geographic places. In the latter case we distinguish between three kinds of vocabularies to which the field value can be converted:

1. A local vocabulary.
2. A vocabulary that is implicitly present in the field values.
3. Terms that may belong to a vocabulary.

In the Bibliopolis dataset we had the following situations for metadata value mappings:

Converting to a local vocabulary concept. Option 1 is exemplified by the values of the field TWOND which represent thesaurus concepts. This relationship is explicitly present in the source data and is preserved during the metadata value conversion. We create the RDF/OWL representations and use the corresponding URIs of these entries in the Bibliopolis thesaurus. Once again, these URIs are composed of text as the records refer to the (unique) Dutch text label of the concept and not to the concept identifier. This is relevant information for the choice of the URI naming scheme for vocabulary concepts (cf. Section 5).

Converting to an implied vocabulary concept. In this case we map field values to resources which form new vocabularies implicitly present in the data. In the Bibliopolis data there were two fields whose values formed an implicit vocabulary.

In Table 2 we see the value "D" in the field CLASSIFICATIE. Further analysis revealed that these single-letter values actually represent a small vocabulary for library-type classifications of the subject. This information is not part of the XML data, but is only shown on the website of Bibliopolis. This classification vocabulary has also some broader/narrower relations. We represented this vocabulary using the SKOS template and mapped the field values to concepts from this vocabulary.

The RDF example in Fig. 4 shows the SKOS specification of a subset of such classification subjects, including the **D** concept. The **M** concept ("secondary subjects") has a hierarchical substructure.

```
bp:A rdf:type skos:concept .
bp:A skos:prefLabel @en
    "General works" .

bp:D rdf:type skos:concept .
bp:D skos:prefLabel @en
    ‘‘History of the art of printing" .

bp:M rdf:type skos:concept .
bp:M skos:prefLabel @en
    "Secondary subjects" .

bp:M1 rdf:type skos:concept .
bp:M1 skos:prefLabel @en
    "Philosophy, psychology" ;
    skos:broader bp:M

bp:M4 rdf:type skos:concept .
bp:M4 skos:prefLabel @en
    "language and literature" ;
     skos:broader bp:M .

bp:M41 rdf:type skos:concept .
bp:M41 skos:prefLabel @en "English" ;
     skos:broader bp:M4 .

bp:M41 rdf:type skos:concept .
bp:M41 skos:prefLabel @en "German" ;
     skos:broader bp:M4 .
```

Figure 4: RDF specification (in N3 notation) of some sample classification concepts. The "M" concept is the top concept of a BT/NT hierarchy

The other implicit vocabulary present within the data is that of roles. The field MAKER contains the name of the creator along with its role (eg: `Yemantszoon, Mauricius : d` where d stands for printer) which is one of the 14 roles. We create RDF representations of these terms as SKOS concepts.

Converting into a typed resource. Again, we create new RDF resources from field values that are potentially part of some vocabulary. We create a unique URI by adding the field name to the field value. For example, for values of the field TECHNIQUE this results in `&bp;techniek_boekdruk`, which is part of the `bp:` namespace. The reason for this is that the values of TECHNIQUE and OBJECT sometimes coincide, for example, `foto` is a technique as well as an object type. This vocabulary can be an existing standard vocabulary such as the AAT in which case an alignment between the new resource and the vocabulary has to be performed. In the Bibliopolis data a number of values of the fields TECHNIQUE, OBJECT and TWGEO can be aligned to the AAT and TGN. There were a small number of unmapped values of field TECHNIQUE (13) and of field OBJECT (5) as can be seen in Table 3. These terms can be added to the AAT by extending it. The alignment and extension is further discussed in Section 8.

We also create resources from field values where the vocabulary the values belong to is unknown or the mapping is not performed. This allows for the option of creating future semantic extensions, although as a result we have a number of resources we do not use. In general, these may be names of organizations or persons, places, cultures or historical periods. In Bibliopolis the values of MAKER and TWNAAM contain person names. These names can possibly be linked to the ULAN vocabulary. We create resources out of these

Data Item	Function	Activity	Source and Target Property/Class
NUMMER	Record Id	Create URI and additional project specific triples (&vra;Work)	*source:* 6 *target:* bp:6 rdf:Type vra:Work .
TITEL	Title in Dutch	Create literal and language tag	*source:* Delftse Bijbel... *target:* vra:title "Delftse Bijbel..."@nl ;
TITEL_EN	Title in English	Create literal and language tag	*source:* Delft Bible... *target:* vra:title "Delft Bible..."@en ;
MAKER	Creator and his marker for role	Extract name and role marker, create URI and label for name and convert marker to role, create role as subproperty of vra:creator	*source:* Yemantszoon, Mauricius : d (d stands for drukker meaning printer) *target:* bp:drukker bp:Yemantszoon_Mauricius ; bp:Yemantszoon_Mauricius rdf:type ulan:person ; rdfs:label "Yemantszoon Mauricius" .
OBJECT	Object type	Map to AAT or create local extension to AAT and mapping	*source:* tekstbladzijde (text page) *target:* vra:type bp:object_tekstbladzijde ; bp:tekstbladzijde rdf:type skos:concept . skos:prefLabel "tekstbladzijde"@nl ; skos:broader AAT:pages ;
TECHNIEK	Technique used	Map to AAT or create local extension to AAT and mapping	*source:* boekdruk (book printing) *target:* vra:technique bp:techniek_boekdruk ; bp:boekdruk rdf:type skos:concept . skos:prefLabel "boekdruk"@nl ; skos:broader AAT:printing ;
DATERING	Date	Interpret and filter data	*source:* 10 jan. 1477 *target:* vra:date "10-01-1477"
ORIGINEEL or REPRODUCTIE	Title of the original or reproduction (book) containing the image	(The title, author, date, place and page number can be extracted)	*source:* Bijbel. Oude Testament... *target:* bp:origineel "Bijbel. Oude Testament..."@en ;
CLASSIFICATIE	Classification of the work in librarian terms using a code	Interpret code, Create URI with code, use interpretation as label keep identifier and create resource	*source:* D (code interpreted as History of book printing) *target:* bp:classificatie bp:D ;
TWNAAM	Person used as subject for work	Interpret name and create URI	*source:* John Do *target:* vra:subject.personalName bp:John_Do ; bp:John_Do rdf:type ulan:person ; rdfs:label "John Do" .
TWOND	Thesaurus term used as subject	Create mapping to thesaurus	*source:* typografische vormgeving *target:* vra:subject bp:typografische_vormgeving ;
TWGEO	Place used as subject for work	Create mapping to TGN where possible or keep literal with language tag	*source: Delft* *target:* vra:subject.geographicPlace tgn:7006804 ;
OMSCHRIJVING or OMSCHRIJVING_EN	Dutch or English description	Create literal and language tag	*source:* Eerste bijbel die... *target:* vra:description "Eerste bijbel die..."@nl ;
AFMETINGEN	Size of the work	Create literal	*source:* 27 x 20 cm *target:* vra:measurements.dimensions "27 x 20 cm" .

Table 2: Part of the Bibliopolis metadata with examples, function and RDFS property/classes

names with URIs in the `bp:` namespace removing invalid characters and spaces. The concepts are of type ulan:person and the human readable label contains the name.

Converting to a literal. Finally, pieces of text such as titles and descriptions are converted to literals. In Bibliopolis the values of TITLE and DESCRIPTION fields were converted into literals with language tags as the title and description of works is both in English and in Dutch.

8. THESAURUS ALIGNMENT

The local thesaurus and the resources containing techniques, object types and locations extracted from the data during the metadata conversion process need to be aligned with standard vocabularies.

We aligned the Bibliopolis thesaurus to AAT by syntactically matching the Dutch skos:prefLabel to the Dutch translation of AAT preferred terms and mapped 209 concepts out of 1033 as presented in Table 3.

Then, we need to identify the relation between the matched terms. The OWL owl:sameAs relation is typically an over-

Source Data	Vocabulary	Terms		Instances	
		Mapped	Total	Mapped	Total
Thesaurus	AAT	209	1033	-	-
Metadata technique	AAT	15	28	1332	1468
Metadata object type	AAT	14	19	978	1507
Metadata subject place	TGN	32	69	349	480

Table 3: Mappings between the Bibliopolis data and other vocabularies

statement that we try to avoid, as ambiguity is quite common. The SKOS Mapping Vocabulary specification[11] was created for the purpose of linking thesauri to each other. It specifies relationships such as skos:exactMatch, skos:broadMatch, skos:narrowMatch and more for aligning vocabularies. For this alignment the mappings are still based on the lexical match of term labels, that corresponds to the relation skos:exactMatch.

[11] `http://www.w3.org/2004/02/skos/mapping/spec/`

The field TWGEO contains geographic names which were mapped to TGN. As the values of this field are in Dutch we extended TGN by adding the Dutch label terms to the proper concept. For example, the value `Parijs` is the dutch label of `Paris` in TGN. Such extensions had to be performed manually, while the mapping of values to cities in the Netherlands could be performed automatically as the labels in TGN contain the Dutch language version. We used syntactic matching for finding appropriate mappings along with some additional techniques to reduce ambiguity, such as restricting the search to cities instead of provinces and the use of background knowledge like the vernacular names of cities. We only automatically mapped unambiguous terms, manually mapping ambiguous terms. Background knowledge of the collection data helped in solving ambiguity as it restricted the places the data could be associated to.

The values of the fields TECHNIQUE and OBJECT were also aligned with AAT using syntactic matching and once more use skos:exactMatch relation. As can be seen in Table 3, a number of terms were not mapped. We extend the AAT by adding the leftover terms to some part of the vocabulary if possible. For instance, the technique `boekdruk` (book printing) is not part of AAT but is a special kind of printing technique, therefore the AAT concept `printing` is selected as broader term. We use the SKOS template to represent the extension.

From Table 3 we can see that a large number of resources are created without being linked to vocabularies. Such resources might be seen as an unnecessary overhead but they can be used in the future when new vocabularies are added or mapped manually. Almost 80 percent of the thesaurus terms were not mapped to AAT and while a number of terms could be linked with skos:broadMatch, this would require additional manual work which could take up a significant amount of time while yielding few matches. This is not the case for the values of TECHNIQUE, OBJECT and TWGEO fields where by manually aligning 13, 5 and 37 terms respectively would yield complete alignments. For OBJECT linking 5 terms would yield an alignment of another 500 occurrences of the term in the metadata which is one third of the total occurrences and well worth the manual effort.

9. DISCUSSION

Interoperability is becoming one of the key issues in the open Web world. Many research programs, such as the IST program of the EU, have interoperability high on the agenda. However, real interoperability between collections is still scarce. Until now, many approaches have focused on interoperability as a problem between two collections.

In this paper we take a different approach. We assume a multitude of collections will become part of the interoperable space; the activities we present can to a large extent be carried out by studying an individual collection. Mapping to existing other vocabularies requires knowledge of other components, but there is no need for these to be complete. For vocabulary alignment the adage "a little semantics goes a long way"[12] holds. Also, one should not view this as a one-shot thing. Metadata and vocabularies change, so extensions will take place at regular intervals in time. This also means that tool support should be in place to support this process, allowing updates to be generated semi-automatically, similar to the AnnoCultor[13] that is being currently developed within the E-Culture project.

For the E-Culture virtual collection we have now carried out this process a number of times. This paper should be viewed as a post-hoc rationalization of this work. Our goal is to provide a set of methods and tools that allow collection owners (museums, archives) to carry out this process. Cultural-heritage institutions are now often bound to closed content management systems; the "three-O" paradigm (open access, open data, open standards) is gaining support, but we have to provide the owners of collections with the necessary support facilities.

We see two potential weaknesses of this work. Firstly, our process still requires much more tool support. In particular for vocabulary alignment we need to explore how existing tools, such as the ones participating in the OAEI contest, perform on this data set. Our current work is still to much based on manual work and only uses simple syntactic tools.

Secondly, the use of Dublin Core as "top-level ontology" for the structure as metadata can also be perceived as a risk. What if the collection has metadata fields that fit with none of the DC elements? However, this was not a problem in either of these six collections. For the moment it seems Dublin Core is indeed a key resource in information interoperability. However, it is a challenge to construct reasoners that make use of the collection-specific specializations.

This article does not show the actual added value of the converted collection content. For this the readers are encouraged to visit the E-Culture online demonstrator, which contains the Bibliopolis data.

10. ACKNOWLEDGMENTS

We are grateful to our colleagues from the Multimedian E-Culture team: Alia Amin, Lora Aroyo, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Marco de Niet, Annelies van Nispen, Marie France van Orsouw, Jacco van Ossenbruggen, Annemiek Teesing, Jan Wielemaker and Bob Wielinga. We would also like to thank Mark van Assem for his input. The project is a collaboration between the Free University Amsterdam, the Centre of Mathematics and Computer Science (CWI), the University of Amsterdam, Digital Heritage Netherlands (DEN) and the Netherlands Institute for Cultural Heritage (ICN). The MultimediaN project is funded through the BSIK programme of the Dutch government.

We are especially thankful to Marieke van Delft of the Koninklijke Bibliotheek (National library of the Netherlands) for her cooperation in the Bibliopolis case.

11. REFERENCES

[1] M. H. Butler, J. Gilbert, A. Seaborne, and K. Smathers. Data conversion, extraction and record linkage using xml and rdf tools in project simile.

[12] quote from J. Hendler

[13] `http://annocultor.sourceforge.net/`

Technical report, Digital Media Systems Laboratory and HP Laboratories, August 2004.

[2] E. Hyvönen, E. Mäkelä, M. Salminen, A. Valo, K. Viljanen, S. Saarela, M. Junnila, and S. Kettula. Museumfinland–finnish museums on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 3(2-3):224–241, October 2005.

[3] A. J. Miles, N. Rogers, and D. Beckett. Migrating thesauri to the semantic web - guidelines and case studies for generating rdf encodings of existing thesauri.

[4] G. Schreiber, A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, and B. J. Wielinga. Multimedian e-culture demonstrator. In I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, editors, *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 951–958. Springer, 2006.

[5] A. Tordai, B. Omelayenko, and G. Schreiber. Thesaurus and metadata alignment for a semantic e-culture application. 2007.

[6] M. van Assem, V. Malaisé, A. Miles, and G. Schreiber. A method to convert thesauri to SKOS. In Y. Sure and J. Domingue, editors, *ESWC*, volume 4011 of *Lecture Notes in Computer Science*, pages 95–109. Springer, 2006.