Supporting interoperability of distributed digital archives using authority-controlled ontologies

Alfons Ruch [(1)]

[(1)] ***University of Passau***
94030 Passau, Germany
EMail: Alfons.Ruch@uni-passau.de

ABSTRACT

In many application domains such as cultural heritage preservation or earth observation there exist numerous digital archives containing semantically related documents or general information. Asking queries across several archives is therefore often highly desirable but seldom straightforward and in many cases simply impossible. This is often due to the autonomous character of archives resulting in a semantical divergence. Based on ontologies with authority control, this paper presents an approach to allow archives to interoperate in a meaningful way, while still allowing them to remain autonomous.

Keywords: archive network, interoperability, authority-control, ontology

INTRODUCTION

Presently, most archives of monumental buildings are isolated in the sense that they serve as a stand-alone store of physical or in some cases digital documents. The *MonArch* project (Monumental buildings Archive network) [1] is a joint interdisciplinary research and development effort aiming at establishing an infrastructure to catalogue, inventory, and digitally secure information sources of monumental buildings and make this data available to the scientific community and general public[1].

MonArch-archives may form an integrated network, but manage their own data/metadata stocks autonomously, i.e., they can completely adapt them to their requirements as desired. This independence allows each archive to define arbitrary annotation schemes, such as tags, taxonomies or ontologies, to categorize archived information. Unfortunately, the independent annotation schemes tend to diverge significantly, making queries covering several archives virtually impossible.

One issue arises from the missing standard of document descriptions. To date no consistent set of metadata and no metadata model exist, either for the structural model of the building or for other descriptive categories such as material used, kind of damage observed, architectural category, cultural style. Therefore, asking queries across different archives and buildings is almost impossible, let alone combining several digital archives in a peer-to-peer network as proposed in recent research. This paper presents an approach based on authority control over the ontologies used to allow archives to interoperate in a meaningful way, while still allowing them to remain autonomous and to retain their individual metadata vocabularies.

To meet the challenge, we pursue a multistage approach. First, all networked archives begin with a common foundation of concepts, such as the standardized authority file of the German national library ("Schlagwortnormdatei der Deutschen Nationalbibliothek") [2] or the Library of Congress Subject Headings (LCSH) [10] or the Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU) [11]. A local archive may only extend this base ontology, i.e. by defining specializations of the provided concepts. The new concepts are restricted to the local namespace of the defining archive, in order to avoid conflicts between the ontologies of other archives. To answer a query covering the entire

[1] This work is funded by the German Research Foundation (DFG) under contract FR 1012/8-1.

archive network, the following cases must be considered: if the query was formulated using only concepts of the global base ontology, it can be answered directly. If, on the other hand, concepts only defined in the local ontology of the querying archive were used, the query must be restructured: (1) To retrieve the requested information from the archive network, the defined generalization hierarchy is used to infer the closest generic concept of the common ontology, (2) the networked archives return all information assigned to this generic concept or to local specialized concepts, (3) the querying archive displays the query results in a suitably annotated way, as they may belong to semantically different concepts of its local ontology.

To retrieve documents employing concepts that are semantically similar to local concepts, *competency queries* [9] can be used to ask the user about relationships between local concepts. By adding the answers to the local ontology as personalized extensions, the query results can be made more accurate.

USE CASE

The following running example is used to illustrated the concepts presented.

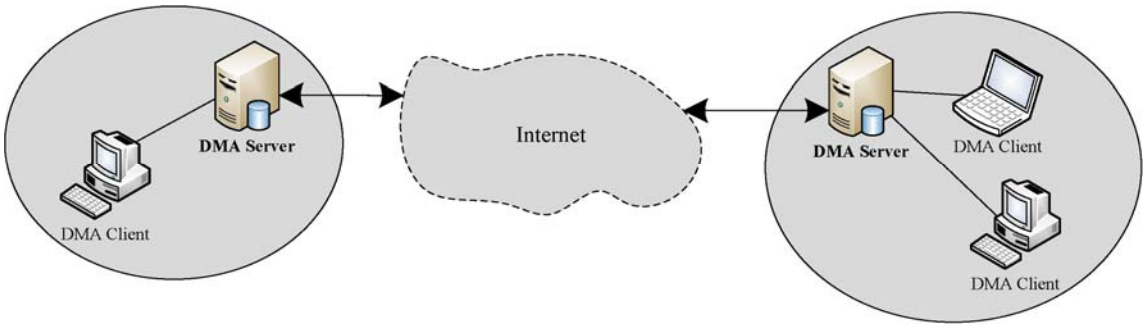

Figure 1: MonArch Architecture

Let a number of digital archives of architectural cultural objects (cathedrals, castles, etc.) be given that are technically connected over the internet (see Figure 1). Let each archive use the same common ontology, which in our example is generated from the authority file of the German national library for architecture (german: “Architektur”). Every archive can extend this authority-controlled ontology for architecture according to their requirements. Now, for example, in “Archive A” the derived concept *Front transept* and *Back transept* are added to the term *transept* (see Figure 2, left part) and in “Archive B” the more special concept *left transept* and *right transept* is added (see Figure 2, right part). If a user searches for documents that are related to the *front transept* or *back transept* in Archive A, all other relevant documents in the archive network should be found. In this case, documents in Archive B that are related to the *transept* and the *right transept* are of interest. The semantical relationships between these terms are best known to the domain expert. However, asking the “right” questions is not easy for an end user due to the simple fact that they can only be asked on the basis of domain knowledge, for instance by another domain expert. But if domain experts have to formulate the queries anyway, they may just as well define the entire semantical mapping among the domain terms on their own.

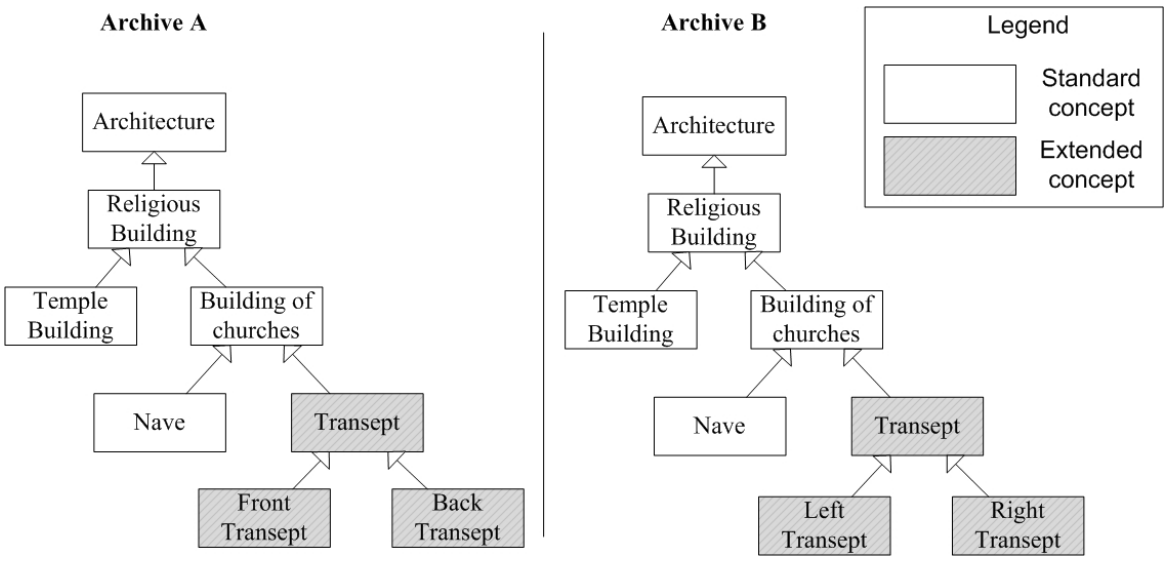

Figure 2: Example of extensions concepts

AUTHORITY CONTROL

Employing *Authority Control* means to use and maintain the forms of names and subjects, the use of uniform titles etc. consistently. Since this process creates a link between bibliographic records and the authority file, authority control provides the underlying structure of the catalog [3]. Authority control has their tradition in the field of library records with two main objectives: Name Authority Control and Subject Authority Control. Name Authority Control is the procedure serving to maintain a consistent use of the names of authors, composers, editors, etc.. In addition, the authority file may contain cross-references that will lead the user from the "wrong" to the "right" heading. The authority record will also contain a "sources found" field which is informative in identifying the work of an author when two or more "similar" headings are involved.

Subject Authority Control represents the concepts used for the subject heading of the records. These concepts are structured by *narrower concep*t und *broader concept* relations. It is modeled using the Web Ontology Language (OWL) [4], where *Concept* is the main class and *related* and *broader* are object properties among the *Concept* classes.

Important authority files are the standardized authority file for subjects of the German National Library ("Schlagwortnormdatei der Deutschen Nationalbibliothek") [2], the Library of Congress Subject Headings (LCSH) [10] and the Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU) [11].

Transforming an authority file into an ontology

Let us have a closer look at a sample authority file and its use in our context. The information in the authority file of the German National Library is encoded in the *machine readable exchange format for libraries* "Maschinelles Austauschformat für Bibliotheken" (MAB) [5]. In MAB, among other attributes, every concept has a unique identifier as well as lists of related concepts and broader concepts. Listing 1 shows an MAB entry for the *architecture* (German: "Architektur") concept.

```
### 00865nM2.01200024 s
001 4002851-3 (Identifier)
070aDNB
800sArchitektur (name of the concept)
830sBaukunst (the same concept)
860s|Baudenkmal (related concept)
860s|Bauweise (related concept)
860s|Innenarchitektur (related concept)
```

Listing 1:Transforming an authority file into an ontology

From a given authority file an OWL Ontology can be automatically generated. The transformation algorithm is provided with an authority file as an input parameter. For each concept in this authority file a Universal Resource Identifier (URI) is generated. A new concept class is added to the ontology using the URI as its identifier and the original concepts term (e.g. the literal *architecture*) as a data type property. In the next step, the relations for this concept are added to the ontology. All *related*, *broader* and *narrower* concepts have to be added to the ontology. Afterwards, the appropriate relationships have to be added as roles. Figure 3 shows a fragment of the ontology generated for the *architecture* concept found in the input authority file.

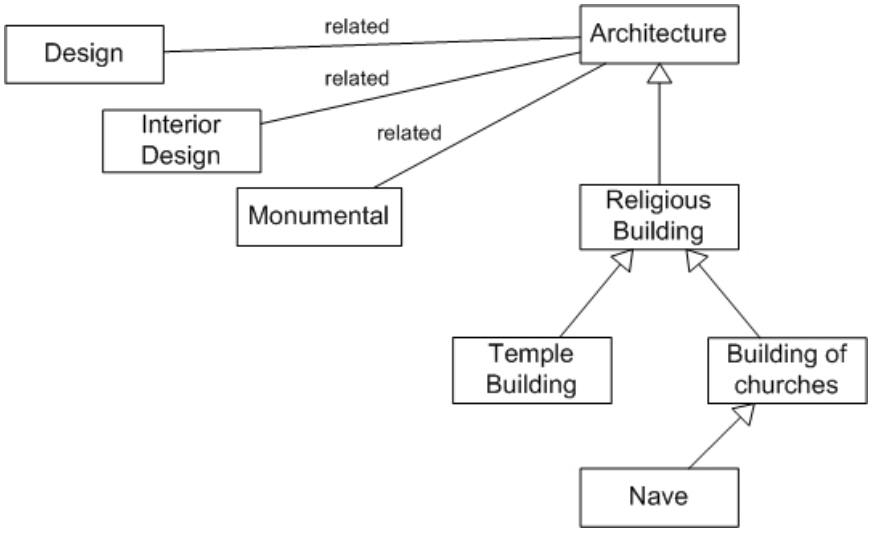

Figure 3: Simplified Concept Model

Local Extension of the authority-controlled Ontology

After generation, the authority-controlled ontology is common to all archive systems. It can only be edited by an archive with special access privileges. Each archive has the autonomy to extend the base ontology with new concepts. By convention, an archive may only add new concepts to the authority file that have a narrower meaning. The benefit of this approach is that a *lowest common subsumer* for every extension can be found in the common ontology. The set of founded subsumers can then be used for query retrieval and to calculate the similarity between the locally added concepts.

Interoperability

When seeing authority files as defining just the *upper part* of an applicable ontology, individual users or local archives can define subcategories of descriptive terms rather freely. Queries can be restricted to the

standardized terms as defined by the authority file. Provided query answering is run in subsumption mode, i.e. including subtrees (see Figure 4, left part). The result is still compatible with other archives but includes archival documents assigned to the more fine-grained terms. It is well-known that this kind of query is supported by all existing essential inference services in the field of ontology-based reasoning. As long as every node of a network uses the common metadata, even queries spanning more than one digital archive can be accommodated. The answer set is simply computed by running the same query on every local archive and combining the partial answers using set union (see Figure 4). Similarly, query answers can be restricted to only those documents having terms common to all archives accessed.

To answer a query covering the archive network, the following cases must be considered: if the query was formulated using only concepts of the global base ontology, it can be answered directly as previously shown.

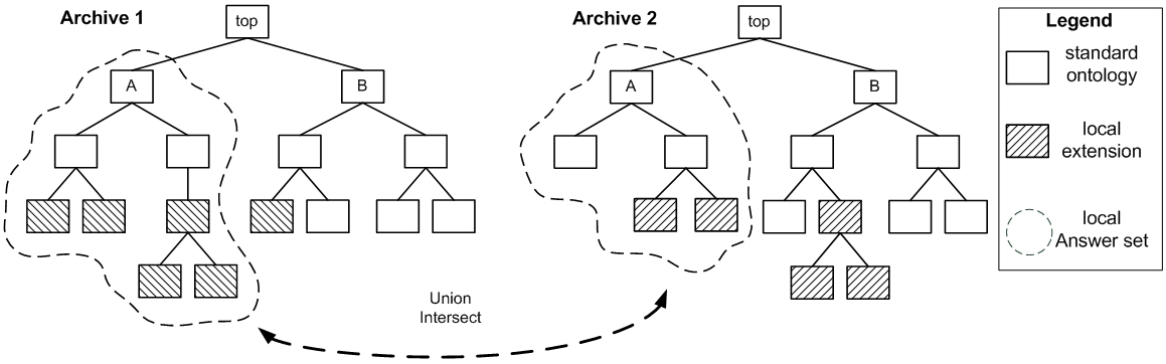

Figure 4: Using Authority Files for Semantical Interoperability

If, on the other hand, concepts only defined in the local ontology of the querying archive were used, the query must be restructured. To retrieve the requested information from the archive network, the defined generalization hierarchy is used to infer the closest generic concept of the common ontology. Then the networked archives return all information assigned to this common concept or to local specializations of this concept. This result may contain some documents with a semantically different meaning, which is often the case if the extended subtrees are rather deep. To exclude such documents, an interoperability extension as defined below is used:

First, we introduce a measure of similarity between concepts. Similarity measures rely on either the edit distance by Levenshtein [6], or the semantic relations as in WordNet [7] or the information content [8]. Applying similarity measures, relations among the concepts can be determined. For example, the edit distance between *front transept* and *right transept* is 5. Similarly, for *transept* and *right transept* the edit distance is 6. By comparing these values, one can infer that *front transept* and *right transept* are more closely related to each other than *transept* and *right transept*.

One weakness is common to all similarity measures: they are only semi-automated, meaning that a human has to check if the matching was correct. One way for improvement is to ask the user directly about the relations among the concepts [9].

An appropriate user interaction consists of two parts: one part is the formulated question itself, the other part are the answer possibilities. To follow our running example, Table 1 shows examples for formulated competency questions. There are questions to determine the relations among terms: is the meaning of the terms *the same*, *narrower* or *broader*.

Competency Question
Is *front transept* **the same as** *right transept*?
Is *front transept* **a narrower term than** *right transept*?
Is *front transept* **a broader term than** *right transept*?

Table 1: Example for competency questions

For the answer possibilities a Likert Scale [12] has been chosen to offer different choices to the user. In our example, there are five possible answers: *Strongly disagree*, *Disagree*, *Neither agree nor disagree*, *Agree* and *Strongly agree*. Each answer is assigned a percentage value to determine its importance for the matching (see Table 2).

Answer	**Percentage**
Strongly disagree	0 %
Disagree	25 %
Neither agree nor disagree	50 %
Agree	75 %
Strongly agree	100 %

Table 2: Answer option with assigned relevancy percentages

There is a one further challenge, which has not been mentioned yet. The most difficult step is to determine which questions are the most relevant to ask, because a user is only willing to answer a limited number of questions. To solve this, the similarity measures mentioned earlier can be used. By calculating the similarity between the concepts, the most relevant concepts for the queries can be determined (i.e. the higher ranked one).

The user answers can then be ranked by their percentage values. By this ranking the relevant answers are chosen and added to the ontology as new relations. The document retrieval can use these new relations and return more accurate results for queries.

Related Work

Falquet et al. [13] presented an ontology-based interface to access a library of virtual hyperbooks (a hyperbook in this paper represents the content information of a book). The hyperbooks build the overall ontology in the system in a bottom up way. Our approach uses authority files to generate the ontology in a top down manner which then can be used as a predefined structure that can be restructured by the application domain needs.

The aim of the FAST project (Faceted Application of Subject Terminology) [14] is to adapt the Library of Congress Subject Headings (LCSH) in a faceted schema with a simplified syntax. The main difference to our approach is our usage of ontologies allowing to use reasoning techniques which is not possible in [14].

The Virtual International Authority Files project [15] is aimed at building a linked system that connects existing authority files. Tillet's approach uses ontologies as an external information source to create links between these authority files. In our approach the authority files themselves are transformed into ontologies and can be used directly for the reasoning with other ontologies.

Conclusion

This paper has shown an approach to increase the interoperability of autonomous archives based on authority-controlled ontologies. The first step was to generate an authority-controlled ontology which is distributed in the archive network, where each archive is allowed to extend the ontology by new specialized concepts. To answer queries covering the archive network, the common ontology is applied. If local concepts have been used, accurate documents can be found by using the lowest common subsumer in the common ontology and asking the user automatically generated competency questions. Using the user answers, new relations can be added to the ontology and relations between the concepts in the different archives can be inferred.

REFERENCES

[1] – MonArch Team. The MonArch Project. http:// www.monarch-project.eu.

[2] – Deutsche Nationalbibliothek. Schlagwortnormdatei (SWD). available at http://www.dnb.de/standardisierung/normdateien/swd.htm.

[3] - Yoojin Hong, Byung-Won On, Dongwon Lee: System Support for Name Authority Control Problem in Digital Libraries: OpenDBLP Approach. ECDL 2004:134-144.

[4] – W3C. Web Ontology Language (OWL), Semantics and Abstract Syntax: http://www.w3.org/2004/OWL/, 2004

[5] - Deutsche Nationalbibliothek: Maschinelle Austauschformat für Bibliotheken (MAB). available at http://www.d-nb.de/standardisierung/formate/mab.htm

[6] - Vladimir I. Levenshtein: Binary codes capable of correcting deletions, insertions, and reversals. In: Doklady Akademii Nauk SSSR, 163(4) S. 845–848, 1965.

[7] - Feiyu Lin and Kurt Sandkuhl. A survey of exploiting wordnet in ontology matching. In Max Bramer, IFIP AI, volume 276 of IFIP, 341–350, Springer, 2008.

[8] - Philip Resnik: Using Information Content to Evaluate Semantic Similarity in a Taxonomy, In Proceedings of the 14th International Joint Conference on Artificial Intelligence, 448–453, 1995.

[9] - Michael Grüninger and Mark S. Fox: The Role of Competency Questions in Enterprise Engineering, Proceedings of the IFIP WG5.7 Workshop on Benchmarking, 1994.

[10] - Library of Congress, Subject Headings (LCSH). available at http://authorities.loc.gov/

[11] - Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU). available at http://rameau.bnf.fr/index.htm

[12] - Likert, Rensis. A Technique for the Measurement of Attitudes. Archives of Psychology 140: 1–55, 1932.

[13] - Gilles Falquet, Claire-Lise Mottaz Jiang, Jean-Claude Ziswiler: Ontology Based Interfaces to Access a Library of Virtual Hyperbooks. ECDL 2004, 99–110, 2004.

[14] - O'Neill, Edward T., et al,. FAST: Faceted Application of Subject Terminology. World Library and Information Congress, 69. IFLA General Conference and Council, Berlin, 2003.

[15] - Barbara B. Tillett, A Virtual International Authority File, IME ICC3, Cairo, 2005.

Alfons Ruch is a Research Assistant at the chair of Information Management at the University of Passau. His research activities are in the areas of information systems. Currently, his work concentrates on distributed digital archives using methods from information integration and knowledge representation