Semantic Ageing of Complex Documents: A Case Study from Built Heritage Preservation

Christoph Schlieder and Peter Wullinger

Lehrstuhl für Angewandte Informatik in den
Kultur-, Geschichts- und Geowissenschaften
Otto-Friedrich-Universität Bamberg
christoph.schlieder@uni-bamberg.de

Abstract: At least two types of ageing processes affect digital records: media ageing and semantic ageing. We argue that the main challenges for digital preservation arise from semantic ageing, that is, the evolution of data formats. The article analyzes a type of document that is particularly vulnerable to semantic ageing: the digital maps produced by researchers on built heritage which combine spatial and thematic data. We describe a solution for format migration based on the ontological modeling of the thematic data.

1 Introduction

The preservation of digital records has been a concern of memory institutions such as libraries or museums ever since they admitted born-digital content into their collections [Kun97]. Around the year 2000, several research initiatives on digital long-term preservation started in the United States, in Europe, and in other parts of the world [Bor06], [Sol10]. Owing to these research efforts, a number of organizational and technological solutions have emerged for the preservation of digital texts and images. However, for more complex types of documents, the situation remains unsatisfactory. From a computational perspective, the most complex documents are not produced by the text-oriented disciplines which form the core of the "digital humanities" but by disciplines such as archeology or historic geography for which material records of cultural processes constitute a central data source. The digital maps which researchers on built heritage produce in order to document historic buildings provide an example of such complexity.

We argue that the computational challenge for long-term preservation of these documents is caused by semantic variability. This means that the semantics of the data is shared by only a small group of users and, additionally, it shows a high tendency to evolve. Any approach for the long-term preservation of complex documents has to somehow address this issue of semantic variability. The main contributions of this article are the following: We analyze the semantic ageing processes that affect the digital maps produced in the context of built heritage conservation. (section 2). Furthermore, we describe a solution for format migration based on ontological modeling (section 3). The article concludes with a comment about a recent discussion on semantic ageing.

2 Semantic Ageing Processes of Digital Records

Any concrete instance of a digital document is bound to a physical medium. The term *media ageing* denotes the process of physical deterioration that affects the medium until at some point the document's original bitstream cannot be recovered anymore. Although private computer users often experience media ageing, e.g. disk failures, as the unique source of their digital preservation problems, this is certainly not true for memory institutions. The standard approach for addressing media ageing consists in detaching the bitstream from the medium and in copying it onto another medium – a task that is easily handled by an adequate preservation planning scheme [Bor06].

A much more serious challenge arises from *semantic ageing*, that is, the evolution of data formats. Knowledge about the semantics of the data is quickly lost if it is not explicitly specified and maintained. The effects of semantic ageing become visible often too late, when the curators of a digital collection realize that no application software on the current platform is able to access a document via an import filter. Unfortunately, standardized data formats do not provide a long-term solution to semantic ageing [Sch10]. Over a period of 50 years even character formats evolve beyond any standard (ASCII, ISO 8559, Unicode). The solutions for the semantic ageing problem that the digital preservation initiatives found to be most reliable are based on two types of approaches: emulation or migration – or sometimes a combination of both.

Emulation recreates the runtime environment of the application software which generated the documents in the first place. Since it also permits to re-enact a user experience from the past, emulation is usually chosen for collections of interactive media such as video games or when the original look-and-feel of document manipulation is of particular importance. Archivists at Emory University, Atlanta, for instance, emulated the 20-year old software environment used by the writer Salman Rushdie to give literary scholars insight into the working conditions during the time when the writer was forced to live underground [Coh10].

While emulation guarantees a maximum of authenticity, it does not help to integrate past contents into the knowledge-based workflows of the present. Being able to run decade-old application software does not make it interoperable with today's technologies. Workflow integration requires more, namely a translation process that converts outdated data formats into currently supported formats, in other words, *migration*.

Documentation in built heritage preservation is based to a large extent on architectural drawings or maps that are generated by preservation scientists while inspecting the building [Sch02]. Digital workflows are supported by documentation systems running on a mobile computer, e.g. a TabletPC. An example for such a system is the Mobile Mapping System in its archiving edition (MMSarchive) which has been developed at the University of Bamberg and is currently used at several Middle European cathedrals and other monumental buildings ([Mat05], [Wul08], [Fre09]). The resulting digital documents include *inventory maps* describing the different parts of the building as determined by more or less extensive measurement procedures (tachymetry, laser scan) and *damage maps* which report the damages of the built structure.

A digital map associates spatial and thematic data as illustrated in Figure 1. The map consists of a collection of spatial objects, for instance, polygons representing individual ashlars, that is, the dressed stone blocks which make up the masonry. Each spatial object is associated with thematic data. Figure 1 shows an input dialog that permits to enter the preservation state, the construction phase, the surface variety, and the stone type.

Figure 1: Digital map with recorded facade damages, St. Stephan, Passau

It is important to note that the thematic data differs considerably between preservation sites. Semantic heterogeneities arise not just by the fact that different sets of properties are recorded but also from an incompatible logical structuring. For instance, while the data model for the Passau cathedral has the stone type encoded in the value of an attribute, the data model for the Vienna cathedral uses explicit mapping objects for each particular stone type. Semantic ageing occurs because the requirements of the documentation task evolve. New types of objects, properties and relations are introduced others become obsolete. Sometimes, the modeling is just restructured to reflect not new concepts but a different understanding of their logical organization. In this process of format evolution, however, the data does not become invalid. There is a need to reuse digital maps in scientific workflows decades after they have been generated.

Digital preservation has studied semantic ageing mainly for single media documents although approaches for preservation planning for mixed media documents have been proposed [Hun06]. The digital maps of historic buildings are more complex than the typical mixed media document which is an aggregate of different medias (e.g. an MPEG2 container encapsulating a video stream and an audio stream encoding). Fortunately, metadata for built heritage maps possess two properties that can be leveraged to find suitable solutions for format migration: first, built heritage map metadata can be represented as formal ontologies [Mat05], [Wul08]. This enables us to utilize technologies that deal with ontology change, most importantly techniques for ontology evolution and ontology matching. Secondly, because of their origin in

datasheet oriented workflows, built heritage maps share a simple, yet expressive meta-structure. Representing built heritage metadata as formal ontologies gives us the ability to represent translations between different document formats as formal translation rules between different ontologies.

3 Format Migration as Ontology Change Problem

[Flo08] identify various subfields of ontology change: Methods from ontology evolution, debugging and versioning may be employed in the design and re-design phases, when an existing document format is adapted to new requirements and correctness as well as traceability of the changes need to be guaranteed. To find translation rules between two different but similar ontologies, it is possible to employ methods from the subfields of ontology matching and mapping. Ontology matching [Euz07] is the process of finding correspondences between matching elements of two different ontologies. Ontology mapping consists in applying the mappings to transform information modeled in the source ontology into the target ontology.

Various methods for ontology matching and mapping have been developed. Unfortunately, most of these methods are still limited to simple correspondences. A simple correspondence can only map single entities (concepts, roles, instances) from the source ontology onto entities from the target ontology. Simple correspondences cannot capture translations that require to consider more than one ontological element at the same time. Only limited research is available with regard to the derivation of complex mappings between ontologies [Stu08], [Švá09]. *Model based refinement* is a novel approach developed by the second author to detect complex ontology mappings. Starting with the initial mapping of two core concepts, model based refinement tentatively applies a set of refinement rules to obtain more complex mappings between source and target models. This is done by heuristically exploring the set of possible source and target models to find suitable matching models. Automated description logic reasoning is used to limit the set of detected models to only consistent models.

Figure 2 shows two snippets from built heritage ontologies. The snippets represent different modeling variants for a facade stone (ashlar). In both representations, two features of an ashlar are recorded: the integration epoch of the stone and the stone type. The modeling variant seen left in Figure 2 represents these features as a simple flat taxonomy using direct subclasses only. The ontology shown right in Figure 2 evolved from the first. It differs in a number of respects.

Some information is represented differently and an extension has been performed. The integration epoch is represented not as a subclass but as a scalar-valued attribute (i.e. a description logic data property). The values of the scalar attribute now refer to the start of the appropriate century and not the ordinal of the century. The second ontology also features stone *Type* as a standalone concept connected to *Ashlar* via a concept-concept relationship (a description logic object property). Additionally, a third stone type variant *Nürnberger* has been introduced which has no appropriate representation in the first ontology. To determine a mapping between the two ontologies represented in Figure 2,

we start with an initial mapping: *Ashlar* in the first ontology is most likely equivalent to *Ashlar* in the second ontology. Such simple mappings may be obtained using a traditional ontology mapper. Such a mapping corresponds to a list of matching models. In the initial case, a single individual of type *Ashlar* from the source ontology is mapped onto a single individual from the target ontology of the same time.

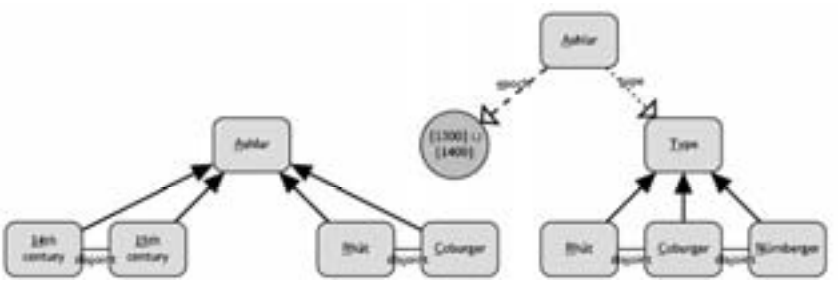

Figure 2: Ontology evolution which restructures the conceptual modeling

A stepwise refinement of the initial mapping is computed by expanding the matched models on both sides. In the first ontology, we refine the initial match by introducing subclasses. Using existing information about concept disjointness, a total of four different refinement models may be obtained: (*A*, 14th, *R*), (*A*, 15th, *R*), (*A*, 14th, *C*), (*A*, 15th, *C*). With regard to the second ontology, there are multiple refinement choices – to introduce the *epoch* attribute or introduce a new concept *Type* linked to *Ashlar*. Proceeding with this stepwise refinement process, we can additionally introduce a new *type* link at the target side and introduce appropriate subclasses of *Type*. This yields a total of six models at the target side: (*A*, 1300, *s*, *S*, *R*), (*A*, 1400, *s*, *S*, *R*), (*A*, 1300, *s*, *S*, *C*), (*A*, 1400, *s*, *S*, *C*), and (*A*, 1300, *s*, *S*, *N*), (*A*, 1400, *s*, *S*, N*)*. The desired mapping is now easily extracted.

Model based refinement is well suited to match axiomatized ontologies arising from ontology evolution because it is possible to use a description logic reasoner to eliminate inconsistent refinement early on. Additionally, comparing expanded description logic models is much simpler than comparing abstract axiomatic descriptions.

3 Discussion and Conclusions

We presented and analyzed a type of documents that is particularly vulnerable to semantic ageing – the digital maps produced by researchers who are studying built heritage. Since the maps combine spatial and thematic data, they are more complex than the mixed media documents that current preservation management approaches do handle. Furthermore, the conceptualization of the thematic data tends to evolve because the requirements of the documentation task change over time. We have shown that an novel ontology mapping technique, model based refinement, can be used to compute the format transformations needed to implement the migration of formats which semantic ageing requires.

Recently, [Ros10] has questioned whether format evolution constitutes a real threat because of market mechanisms that would prevent the obsolescence of formats. However, there exist a number of document-centred workflows in the digital humanities that are supported by special purpose software only such as the task of documenting built heritage. [Ros10] speaks of "immature markets" because the software is adopted just by a small community of expert users in which case, he admits, format obsolescence is very common. We have shown that in such cases there exist technological means to compute the necessary transformations from obsolete data formats into formats which can be accessed by current application software.

References

[Bor06] Borghoff, U., Rödig, P., Scheffczyk, J., and Schmitz, L.: Long-term Preservation of Digital Documents: Principles and Practices, Springer, Berlin, 2006.

[Coh10] Cohen, P.: Fending Off Digital Decay, Bit by Bit. The New York Times, C1. March 16, 2010

[Euz07] Euzenat, J. & Shvaiko, P.: Ontology Matching, Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2007

[Flo08] Flouris, G.; Manakanatas, D.; Kondylakis, H.; Plexousakis, D. & Antoniou, G.: Ontology change: classification and survey, The Knowledge Engineering Review, Cambridge Univ Press, 2008, 23, 117-152

[Fre09] Freitag, B. & Schlieder, C.: MonArch--Digital Archives for Monumental Buildings, Künstliche Intelligenz, 2009, 4/2009, 31-35

[Hun06] Hunter, J. and Choudhury, S.: PANIC – An Integrated Approach to the Preservation of Composite Digital Objects using Semantic Web Services. International Journal on Digital Libraries, 6(2), 2006, pp. 174–183

[Kun97] Kuny, T. (1997) A Digital Dark Ages? Challenges in the Preservation of Electronic Information. In: 63rd IFLA Council and General Conference, Workshop on Preservation and Conservation, http://archive.ifla.org/IV/ifla63/63kuny1.pdf (2 Apr 2010).

[Mat05] Matyas, S. & Schlieder, C.: Generating Content-related Metadata for Digital Maps in Built Heritage, Proceedings of the First on-Line conference on Metadata and Semantics Research (MTSR'05): Advances in Metadata Research, Rinton Press Inc., 2005

[Ros10] Rosenthal, D.: Format Obsolescence: Assessing the Threat and the Defenses, Library Hi Tech, 28, 2, 2010, pp. 195-210

[Sch10] Schlieder, C.: Digital Heritage: Semantic Challenges of Long-term Preservation, Semantic Web Journal, 1, 2010

[Sch02] Schuller, M. Building Archaeology, Monuments and Sites: VII. International Council on Monuments and Sites (ICOMOS), 2002.

[Sol10] Solvberg, I. and Rauber, A.: Digital Preservation in Solvberg, I. and Rauber, A. (eds), Digital Preservation, European Research Consortium for Informatics and Mathematics, 2010, pp. 12–13.

[Stu08] Stuckenschmidt, H.; Predoiu, L. & Meilicke, C.: Learning Complex Ontology Alignments: A Challenge for ILP Research, International Conference on Inductive Logic Programming (ILP2008), 2008

[Švá09] Šváb-Zamazal, O. & Svátek, V.: Towards Ontology Matching via Pattern-Based Detection of Semantic Structures in OWL Ontologies, Proceedings of the Znalosti Czecho-Slovak Knowledge Technology conference, 2009

[Wul08] Wullinger, P. & Schlieder, C.: Digital Maps of Historical Buildings: Preservation Issues and Solutions, Proceedings of IS&T's Archiving 2008 Conference, 2008