Semantic Web 1 (2010) 143–147
DOI 10.3233/SW-2010-0013
IOS Press

Digital heritage: Semantic challenges of long-term preservation

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Pascal Hitzler, Wright State University, USA
Open review(s): Krzysztof Janowicz, Pennsylvania State University, USA

Christoph Schlieder
University of Bamberg, Germany
E-mail: christoph.schlieder@uni-bamberg.de

Abstract. The major digital preservation initiatives are about as old as the idea of the Semantic Web but the research areas only had little effect upon each other. This article identifies connections between the two research agendas. Three types of ageing processes are distinguished which affect digital records: media ageing, semantic ageing, and cultural ageing. It is argued that a period of 100 years constitutes an appropriate temporal frame of reference for addressing the problem of semantic ageing. Ongoing format migration constitutes currently the best option for temporal scaling at the semantic level. It can be formulated as an ontology matching problem. Research issues arising from this perspective are formulated that relate to the identification of long-term change patterns of ontologies and the long-term monitoring of ontology usage. Finally, challenges of cultural ageing are discussed.

Keywords: Digital preservation, format migration, ontology matching, ontology change

Introduction

Scalability is widely considered a core objective of the Semantic Web, but it is mainly looked at from a quantitative data perspective, considering, for instance, the number of RDF triples that can be handled. In digital preservation, the focus lies on finding solutions that scale well along the temporal dimension [1]. Memory institutions such as museums care very much whether the documents and/or the data they publish will be accessible in 50 years from now. Considerable effort was invested, for instance, at Emory University, Atlanta, to make the 20-year old digital material of the writer Salman Rushdie accessible by recreating the software environment he used via emulation [3]. Increasingly, less famous individuals ask what sort of digital legacy they will be able to leave with today's technologies. The problem of digital preservation has moved from a concern of specialists to mainstream awareness [15]. In the following, we will explore the challenges of temporal scalability.

In the pre-digital world, the preservation of written records over long periods of time depended on several prerequisites which are rarely made explicit. Firstly, the record needs to be preserved physically. Secondly, the semantic capabilities to read and interpret the records have to persist. A reader of a clay tablet, for instance, has to master a particular form of cuneiform writing and the Acadian language. Thirdly, there must be a community that (still) shows interest in the record. Only an interested community will mobilize the resources required to teach its members complex semantic skills or to even engage in the deciphering of extinct languages. In many respects, the preservation of digital records faces similar problems.

1570-0844/10/$27.50 © 2010 – IOS Press and the authors. All rights reserved

The ageing of digital records

One reason for which digital preservation can fail is media ageing. Any medium that carries a digital encoding will physically deteriorate until it is no longer possible to recover the original bit stream. This process of media ageing has received much attention from memory institutions but it seems less critical for the Web with its capacity to easily replicate data.

Like written records before the computer, digital content is affected by *semantic ageing*, that is, the evolution of data formats and the fact that knowledge about data semantics quickly disappears if not specified explicitly. Finally, there is a process which may be called *cultural ageing*. This process is rarely discussed in connection with digital preservation. Gradually, the community loses interest in some particular content. The corresponding documents are no longer retrieved, the data is no longer used in inferences. Knowledge about the semantics of digital records may persist for a while after the community loses interest in their content. However, as the semantic knowledge is not maintained and transmitted any more, its loss is almost unavoidable.

Choosing a temporal frame of reference

Before identifying the semantic challenges of digital preservation it is important to determine at which temporal scale to address the issue. At the short-term end there is the time frame which the legal regulations of many countries provide for the preservation of business documents, namely 10 years. The market offers a number of archiving solutions which handle digital preservation at this scale by using archiving formats with are maintained for at least a decade. The preservation problem may be considered solved at this scale.

Probably, the most ambitious temporal frame of reference considered for digital preservation is the formidable period of 10.000 years promoted by the Long Now Foundation [2]. Without doubt, it is intellectually challenging to look at ten millennia, the relevant unit of analysis for a number of global problems such as climate change. It is less clear, however, in what way such a very long-term perspective fosters the emergence of technological solutions (e.g. format repositories) radically different from those currently discussed in digital preservation.

For the purpose of this article a much more modest frame of reference is chosen, a period of 100 years, which is more accessible to empirical evaluation as well as closer to personal experience. Centering this frame of reference upon the present sets a double agenda: (1) finding strategies to access digital contents from the past 50 years in spite of media ageing and semantic ageing, (2) planning the preservation of currently accessible digital content for future use during the 50 years to come.

A major consequence of this specific planning horizon consists in the fact that the problem of semantic ageing cannot be solved anymore by simply agreeing on a standard format for digital archiving. Half a century is just plenty of time for requirements to evolve beyond any standard. This holds even for plain text as the chronology of character formats illustrates. The year 1963 witnessed the first edition of the ASCII standard which ceased to evolve with a last update in 1986. In the same year, the Latin-1 character set was published which became part of the ISO 8859 series of standards. ISO ceased maintenance of these standards in 2004 to concentrate its resources on Unicode. The example shows that even for data of little semantic complexity only a sequence of standards was able to bridge a period of almost 50 years. Note also that the end of a standard´s evolution does not imply the end of its usage.

Digital preservation and the Semantic Web

Digital preservation has been a very real concern of memory institutions who addressed the problem long before the problem of an impending “digital dark age” [10] became known to a wider audience. Public funding of the major research initiatives started around the turn of the millennium, notably the National Digital Information Infrastructure and Preservation Program (NDIIPP) established in 2000 by the US congress and comparable European research initiatives. In other words, the mainstream of digital preservation research is about as old as the Semantic Web. Unfortunately, both strands of research have only interacted in rather limited ways so far.

The digital preservation initiatives basically explored two families of approaches for the problem of semantic ageing: migration and emulation – as well as combinations of both. *Migration* is especially interesting for document-centered workflows, in-

cluding those used in the humanities and in cultural-historic research. The ideal target format for migration is published under an open source license, comes with an explicit account of its semantics, and possesses a large community of users.

Emulation constitutes the best solution for archives of highly interactive media, e.g. interactive art or video games. Emulation strives for authenticity, for a reenactment of a user experience from the past. However, being able to run the software which created the data does not per se make it interoperable with present day technology. Migration, on the other hand, aims at the integration of past content into future knowledge-based workflows. Because of the focus on data and interoperability, migration seems to blend more easily with the different flavors of the Semantic Web – definitely with the idea of a Web of semantically interoperable knowledge bases but to a certain extent also with the more recent idea of a Web of Linked Data.

At least two levels can be distinguished at which migration strategies are currently supported by Semantic Web technologies: the preservation planning level and the semantic transformation level. A major result of the digital preservation initiatives was to conceive preservation as an ongoing process based on an appropriate digital curation lifecyle model (e.g. [4,16]). Preservation planning is a central element of such a life cycle model. At the *preservation planning level*, migration strategies are implemented by services that monitor data formats and data access mechanisms on the one hand and available migration tools on the other hand. Emerging risks are assessed (obsolescence detection) and recommendations for migration pathways are generated. A first link between the world of digital preservation and the world of the Semantic Web exists at this planning level. Preservation services have been described as Semantic Web services, for instance, in the PANIC system [9] which uses an extension of the OWL-S ontology to describe preservation-specific services and computes semantic matches to support service discovery.

Services such as format transformations are based on the preservation metadata that comes with the digital records. This works best for atomic single media records but becomes more difficult for composite multimedia records. For highly complex data objects such as those produced in the architecture, engineering, and construction (AEC) industry by special-purpose CAD systems ready-to-use migration services simply do not exist [5]. In such cases, before addressing migration at the preservation planning level, it has to be implemented at the *semantic transformation level*. For specialized domains such as architectural drawings, archival formats start to emerge although they have some limitations from the point of view of digital preservation [14].

However, many projects in the AEC industry have documentation needs that require some sort of application specific semantic modeling. It is near at hand to use ontological modeling languages such as OWL for describing those application ontologies and for relating them to the domain ontology [8]. This is a second link that has been established between the world of digital preservation and the world of the Semantic Web.

Once that data semantics is captured by ontological modeling, the problem of migrating from one data format to another can be described as an ontology matching problem which transforms a source ontology into a target ontology [6]. Format migration is thus closely related to ontology change as defined in [7]. In adopting such an approach, we must, however, be aware of the general limitations of ontological modeling. While many aspects of data semantics are easily captured by modeling formalisms such as description logics, some aspects of the semantics of natural languages are difficult to render. The same holds for epistemic drifts that are not reflected by a change of the logical modeling of data.

The challenges of semantic ageing

Within this setting – digital preservation based on ongoing format migration modeled as a sequence of ontology changes – a number of challenges arise. In one way or the other, they are all related to the issue of how well solutions scale over the chosen temporal frame of reference of 100 years, or rather 50 years, if only forward preservation is considered.

Identifying long-term patterns of ontology change

Only by looking at periods that are significantly longer than the 10 years handled by current technology, it can be determined how the changes in the ontologies which cause semantic ageing distribute over time. There seem to be change processes with an almost constant rate of change. However, in many cases, changes occur in bulk. Open research questions relating to ontology change include:

- What different change patterns are there? Do they depend on the type of ontology (top-level, domain, task, application)?
- Is it possible to predict impending bulk changes by analyzing the time series of changes and the structural complexity of the ontology?
- Media ageing is studied by artificial ageing processes. Can similar simulation approaches be designed for semantic ageing?

Monitoring long-term usage of ontologies

Software tools with rich functionality (e.g. special-purpose CAD systems), tend to generate data with complex semantic relationships. Often, however, only a fraction of the functionality is used to actually create a digital record (e.g. a CAD document with only 2D geometries). Migration at the semantic transformation level would be greatly simplified if for a collection of digital records it is known whether there are parts of the source ontology that are not used by any of the records, or only used by very few. The long-term evolution of ontology usage has not been studied so far. Issues to be addressed in this context include:

- How does the population of classes with instances change over long time intervals?
- Which instances are actually used in queries and inferences? Do usage patterns change over time?
- How can information about ontology usage patterns help to improve ontology matching?
- How is ontology usage monitoring best integrated into preservation life-cycle management?

The challenges of cultural ageing

The creation of meaning by communities is an ongoing process which is inevitably accompanied by an antagonistic process in which meaning is lost. Pre-computer history is full of examples for this process of cultural ageing which affects natural languages and their writing systems as well as complex belief systems such as religions. Cultural ageing has at least a technical benefit. Only the records that a community still shows interest in will be migrated which reduces the semantic translation workload by orders of magnitude. The downside is also evident. Processes of cultural renewal ("renaissances") which generate interest in content that was considered uninteresting for generations are not possible.

At present, cultural ageing does not constitute a focus of research on digital preservation. Probably, this is due to the fact that there are sufficiently many other problems that seem more pressing. On the other hand, it is difficult to imagine a satisfactory solution to digital preservation which does not take the mechanisms of cultural ageing into account. This is not so much a matter of trying to prevent cultural ageing – a hopeless task – rather than to monitor the community's access to the digital records and to identify content that will become vulnerable to semantic ageing because of the community's loss of interest.

Web archiving provides a good example of how the digital version of cultural ageing operates. Defining a selection policy for the Web sites that are going to be preserved constitutes the crucial first step in the design of any Web archive [13]. Different computational methods have been proposed to determine the relevance of a web page such as Google's PageRank or the HITS algorithm [11]. The underlying problem – measuring visibility in a large-scale communicative processes – has been studied from many disciplinary perspectives including social network analysis and bibliometrics. Simulation studies can show how visibility evolves over time [12].

Monitoring cultural ageing

A similar type of selection has been effective in the pre-computer era archives. What is new, is the quantitative scale, that is, the number of records for which choices need to be made. The choice is not necessarily one of inclusion or exclusion but rather a decision about the quality level at which semantic ageing is dealt with. An online journal with a high impact factor will probably enjoy a premium migration process involving human intervention which even preserves, for instance, interactive 3D-models and HD videos while the automatic standard migration process for less visible journals is going to concentrate on preserving just text, tables, and images. A number of research problems are connected to the selection process triggered by cultural ageing.

- Which is the appropriate way to quantify the visibilty of documents and/or data in the relevant communicative processes, e.g. in scientific communication in the humanities?
- What is an adequate formal model of the loss of semantics triggered by cultural ageing?
- How can digital preservation reflect the plurality of interests that different communities show?
- What type of preservation planning will permit or even encourage the rediscovery of documents or data?

Conclusions

Although the digital dark age is a menace of the present, the processes of media ageing, semantic ageing, and cultural ageing have been effective since pre-computer times. In a world of distributed digital data, however, semantic ageing constitutes a bigger problem than media ageing. The best way to overcome the effects of semantic ageing is by migrating digital records into new formats. We have seen how Semantic Web technologies support migration at the preservation planning level as well as at the semantic transformation level. Research challenges have been formulated for semantic ageing and for cultural ageing.

Long-term preservation constitutes an application field that forces the Semantic Web research community to adopt a much longer temporal frame of reference. By doing so it places the study of ontology change under a new perspective which focuses, among others, on the changes of ontology use and on the changing population of classes with instances (categorization patterns of instance data).

Taking cultural ageing seriously means to abandon the idea that digital preservation operates like a time capsule. The picture of content that is enclosed in a digital capsule to be opened at some moment in the future is misleading because it is not the past that sends messages to the future. Rather, it is the present that makes choices, selecting content from the past and linking to it. This ongoing process of linking from the present into the past makes up digital heritage.

References

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

[2] Brand, S. (2000) *Clock of the Long Now: Time and Responsability*, Basic Books, New York, NY.

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

[4] Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., et al. (2009) DCC&U: An Extended Digital Curation Lifecycle Model. *The International Journal of Digital Curation*, **4**(1), Article 3.

[5] Doyle, J., Viktor, H., and Paquet, E. (2009) A Metadata Framework for Long Term Digital Preservation of 3D Data. *International Journal of Information Studies*, **1**(3), 165–171.

[6] Euzenat, J. and Shvaiko, P. (2007) *Ontology Matching*, Springer, Berlin.

[7] Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., and Antoniou, G. (2008) Ontology Change: Classification and Survey. *The Knowledge Engineering Review*, **23**(2), 117–152.

[8] Freitag, B. and Schlieder, C. (2009) MonArch – Digital Archives for Monumental Buildings. *KI*, **23**(4), 30–35.

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

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

[11] Langville, A. and Meyer, C. (2006) *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ.

[12] Malsch, T., Schlieder, C., Kiefer, P., Lübcke, M., Schmitt, M., and Stein, K. (2007) Communication Between Process and Structure: Modelling and Simulating Message Reference Networks with COM/TE. *Journal of Artificial Societies and Social Simulation*, **10**(1), Article 9.

[13] Masanès, J. (2006) *Selection for Web Archives in Web Archiving*, Masanès, J. (ed), Springer, Berlin, pp. 71–90.

[14] Smith, M. (2009) Curating Architectural 3D Models. *The International Journal of Digital Curation*, **4**(1), Article 8.

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

[16] Strodl, S., Becker, C., Neumayer, R., and Rauber, A. (2007) How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure in *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries. JCDL 2007; Vancouver, British Columbia, Canada, June 18–23, 2007*, Rasmussen, E. and Larson, R. (eds), ACM Press, New York, NY, pp. 29–38.