Data Description and Archives for Scientific Research in the Future

Imaging & Media Lab
University of Basel

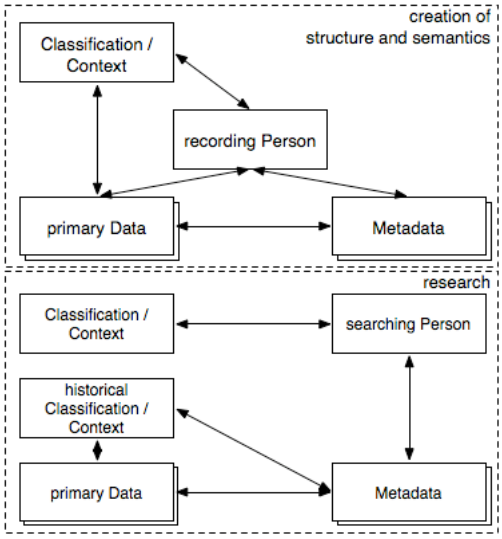
Simon Margulies, Ivan Subotic, Dr. Lukas Rosenthaler
[simon.margulies, ivan.subotic, lukas.rosenthaler]@unibas.ch
Bernoullistrasse 32, CH-4056 Basel/Switzerland
Phone +41 61 267 04 88, Fax +41 61 267 04 85

1. Introduction

At the last IS&T archiving conference 2005 the case has been made, that historical research of the future would change due to the growing availability of online resources of digital cultural heritage.[1]
Scientific research depends widely on a controlled tradition. Archives guarantee the latter by maintaining and providing information objects for the scientific research of the future. Universal access, independent of time or space, has been made possible by the interconnection of data collections by modern information technologies. The simplified access renders possible enquiries of much more source material, which has changed and will change the processes of scientific research.
Archives provide their primary data with various layers of metadata to guarantee the findability, readability and scientific interpretation of digital information objects. Mostly produced in XML this kind of data description makes a human- and machine-readable structuring of information objects possible. The underlying semantics of the structured description and thus the context of different information objects remain hidden to the machine. It can only be interpreted and used for further researches by humans conducting painstaking enquiries.
The following paper wants to point out the shared connections between data structure, data semantics, archiving and scientific research. Techniques will be presented, that provide archives with new possibilities and can help scientific research to handle the growing amount of source material.

2. Research in digital data collections

An archive is defined as an institution, which administrates and preserves an amount of documents (Archivgut) important to the historical coverage of the past of its sponsorship or a certain theme of the institution. For the future a growing interconnectedness between archives providing online access to digital databases is assumed.[2] Scientific research, especially historical research, depends strongly on context and linking between different source materials - also being held in different archives - and tries to derive and prove these contexts and links:



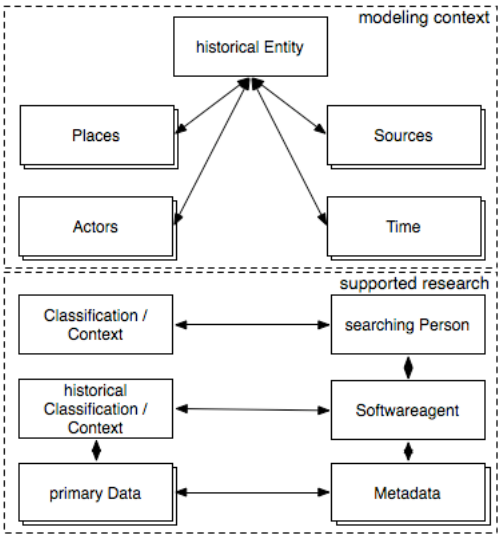
Such contexts are structured with a metadata-schema and described with the aid of a thesaurus by the archivist. The schema and its contents vary between different times, cultures, archives and people depending on their contexts and undertaken classifications. Future schemas will vary even more with a growing temporal and cultural distance between editors; already nowadays an agreement on a specific standard is unthinkable.[3]
If the data description, as it is common practice, is composed in XML, only keyword searching in the schema-specific digital data collection can be supported. Contexts among different source materials with different data descriptions and in different data collections can only be discovered by the human researcher and not by a machine or a software agent: Data gets highly structured by XML, but the underlying semantic entities of single parts and especially their context to other information objects remain hidden to the software agents. If a continuous growing amount of accessible source material is presumed, scientific research will become more

[1] Clifford Lynch. Archiving, Stewardship, Curation: From the Personal to the Global Sphere.
[2] E.g. by distributed archiving systems [1].
[3] In this regard Dublin Core [2] embodies an example of the greatest possible common denominator. It remains unquestioned to small for an adequate data description for preservation of digital source material of all kind.

difficult, because much more material must be sighted and contexts analyzed without being able to access the support of automated software agents.[4]

3. Machine-readable structure and semantics (ontologies)

To confront the growing amount of accessible digital source material and to benefit from it, future scientists need support by software agents. Such software agents can point out contexts between different source materials [3]. To be able to link different material, a software agent needs a machine-readable structure and semantic of the single information object. In computer science such descriptions are expressed in 'ontologies'. Ontology is defined as a formal and explicit specification of a shared and common conceptualization [4]. A conceptualization defines an abstract model of the basic context and rules of an entity in the real world. The relevant concepts of such an entity must have been identified by a group that agrees upon the rules of these concepts. The formal and explicit specification guarantees the machine-readability of such specifications. During an enquiry a software agent can consequently carry out deductive reasoning of contexts and propose links between different information objects:



To formally express ontologies the W3-Consortium has defined RDF, RDFS and OWL [5][6]. These technologies are based on XML and offer the possibility to easily integrate already existing parts of XML-metadata into ontologies.
For the ontological description of cultural heritage various projects have been promoted [7] and reference models published [8]. Future implementations of archiving solution should examine and integrate these technologies for the obvious reasons mentioned above [9].

[4] This kind of research is equal to the usual enquiry in common search machines on the WWW resembling rather the look for a needle in a haystack than a scientific research.

Not only a software-supported research gets possible by an ontology oriented data description of digital source material. Further valuable information for the future research is preserved: an explicit formal conceptualization of a certain domain furnishes a future historian with more evidence to draw conclusions about passed contexts of the source material and to gain more information about the past.

4. Conclusion

The growing amount of data and access to digital source material will make enquiries for scientific research more difficult: The more heterogeneous data and different data description are present and used, the bigger and the more complex the amount of data to be handled gets. To get advantage of the growing amount and access to data, researchers need support from intelligent software agents, which can deduce contexts from different source materials. Therefore digital information objects should be described by and integrated into ontologies. Existing Metadata-standards do not meet these demands only providing a machine-readable structure of information objects but not their underlying semantics.

5. Literature

[1] http://www.distarnet.ch/
[2] http://dublincore.org/
[3] Signore, Oreste. Ontology Driven Access to Museum Information. In: Proceedings of CIDOC 2005, Zagreb, 2005. http://www.w3c.it/papers/cidoc2005.pdf
[4] Benjamins, V.R., Contreras, J et al. Cultural Heritage and the Semantic Web. In: The Semantic Web: Research and Applications. Ch. Bussler, John Davies, Dieter Fensel Rudi Studer (Eds.). First European Semantic Web Symposium, ESWS 2004. Berlin, 2004.
[5] http://www.w3.org/2001/sw/
[6] http://www.w3.org/2004/OWL/
[7] http://www.esperonto.net/
[8] http://cidoc.ics.forth.gr/
[9] Subotic, Ivan. Margulies, Simon. Rosenthaler, Lukas. DISTributed ARchiving NETwork - Distarnet. Poster proposal for IS&T Archiving 2006. http://www.distarnet.ch/papers/is_tposter2006.pdf

6. Authors

Simon Margulies studied History and Computer Sciences at the University of Zurich. He is working on his Ph.D. in History in the field of Archiving digital Data. Together with Ivan Subotic he develops Distarnet, a DISTributed ARchival Network [1]. He analyzes the impact of digital data as source material for the historical research, develops Distarnet and advises various projects and companies on data modelling, metadata and data retrieval.