eHumanities: Intelligent Analysis and Information System for Humanities and Culture (Extended Abstract)

Sven Becker, Marion Borowski, Melanie Gnasa, Kai Stalmann, Stefan Wrobel

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven
53757 Sankt Augustin

{sven.becker, marion.borowski, melanie.gnasa, kai.stalmann,
stefan.wrobel}@iais.fraunhofer.de

1 Introduction

While electronically available data have played an important role in the natural sciences for a long time, the impact of data and information on the humanities and culture has become clear only recently. Progress in computer science has made it possible to digitize the objects of research from these disciplines so that they can be made available electronically. This digitization assists scientific discourse as well as electronic networking. Key challenges arising in this connection derive from the type of material and the overwhelming amount of objects which need to be digitized. In the following, we discuss the main challenges, goals and solutions in an exemplary fashion for the so-called German Digital Library (DDB) project. This project, being part of the European Library initiative Europeana (www.europeana.eu) is aiming at making available the core objects of 30.000 German cultural and scientific institutions in an online version. We argue that the tasks of analysis, interpretation, and scientific networking can be supported considerably by adequately designed systems that meet the requirements of the humanities and the cultural sector.

The number of cultural and scientific institutions in Germany is estimated at around 30.000. They preserve more than 30 million cultural assets and scientific documents. The exact number of books, paintings, movies, musical compositions, historical monuments, archival documents, and others is not known. Right now, only an extremely small percentage of objects is made available in digital form. Using recent technical developments, however, digitization and indexing will open new opportunities to give access to cultural and scientific objects to a broader audience.

2 Related Work

The German Digital Library will add momentum to the current trend towards digitization by implementing a portal that aggregates, analyzes and retrieves digital data with its describing metadata. In doing so, a very high resolution digital reference copy of objects is stored in a decentralized way on the web sites of the individual libraries, archives, museums and various other types of public institutions, whereas a lower resolution version is stored directly on the DDB platform to enable a quick overview and approximate browsing. Today's digital cultural and scientific assets are distributed and often only accessible via different smaller portals. The German Digital Library's intention is to aggregate the various knowledge data in one retrieval and collaboration platform.

Libraries, scientists and 'Fachinformationszentren' (FIZ, centers of specialized information) have always had strong interests to drive digitization, e-publishing and information retrieval. Space and financial limitations are strong motivations for libraries. Archives have a need for exchanging data, but slightly differ from libraries, because they usually have an awarding authority paying for a particular archival duty. Scientists have the strongest need for direct and fast access to information that may be held in a library, archive or museum on the other side of the globe. These are some of the reasons why library and information science started embracing computer science, software engineering and computational linguistics. The FIZ in Germany, some of which have become technology and service providers with competitive computer centers and strong engineering teams, have combined efforts to launch vascoda.de[1] [Pi03], a portal making some 75 million records of scientific literature accessible. Vascoda, which is hosted at Technische Informationsbibliothek und Universitätsbibliothek Hannover, contributes to WorldWideScience.org [HJ08], a US maintained web gateway providing a federated search service on databases spread over the world with more than 200 million records. A first attempt to integrate sources from different cultural heritage institutions is the BAM-Portal [Ma02] (www.bam-portal.de). It enables the search in the collections of the participating libraries, archives, museums and other sources in Germany. To date this portal manages more than 45 million digital objects; about 1 million of those objects are available in digital form.

[1] It should be remarked that the funding of this project is expired and its future service is uncertain.

Libraries, archives, and museums grow and adapt themselves to a changing world, but these institutions also are the most important holders of cultural heritage. In 2004 Google started the Google Book Project (books.google.com) that aims at porting every book in every language into the Google Digital Library. In succession of this ongoing activity many publicly founded digitization projects and digital libraries have been set off. More recent examples are the Max Planck Digital Library (mpdl.mpg.de), or the Biodiversity Heritage Library (bhl.org). One of the first digital library projects ever, the Internet Archive (archive.org), is privately founded and dates back to 1996. National libraries followed up and started crawling digital content from websites. One of the most prominent European digital library projects, Europeana (europeana.eu), was initiated in 2005 and currently holds metadata from aggregators with links to about 6 million digital items from all over Europe.

3 Challenges

The German Digital Library was set out at the end of 2009 with the long-term goal of ingesting up to 300 million items from up to 30.000 national institutions within the next decades. The project is carried out by Fraunhofer IAIS in collaboration with the German National Library, a network of cultural and scientific institutions, the FIZ Karlsruhe, and other partners. When first released to the public, which is expected to happen in 2013, the library shall serve as a web portal and also as a platform that can contribute to other projects and libraries, for example to europeana.eu.

Like Europeana, the German Digital Library is conceived of supporting networking activities of the partners involved, but other than Europeana, the German Digital Library is deemed to hold not only metadata but content also. Furthermore, it shall endeavor to contribute to the semantic web and offer value added services based on metadata and content. One of the virtues of these projects is that the ingested data could be rich and of high quality. But in fact quality may vary: certain collections contain metadata annotations at item level, others at collection level only. Page level metadata is currently hardly provided at all. This task is addressed by the CONTENTUS [Pa09] project which is an application scenario of German THESEUS research program [Th10]. It aims at digitizing text and multimedia collections in order to annotate them semantically. Besides the heterogeneity of metadata and content in terms of quality and formats, the challenges for projects like CONTENTUS, Europeana or the German Digital Library are manifold. From a usability perspective the most demanding problems are precision and recall, name disambiguation, clustering of (near) duplicates, and multilinguality of metadata and queries. First results from the CONTENTUS project for name disambiguation [Pi09] and the detection of higher level ontology concepts for words and phrases [PR09] are already available and can be applied to the German Digital Library. A custodian might for good reasons emphasize those requirements that ensure an adequate representation of objects provided. Scientific use on the other hand does not necessarily rely on easy usability and beautiful representation, but requires the ability to safely identify, localize, and evaluate items. Sharing objects in a distributed environment or workspace for collaborative work is highly desirable for scientists but may also fit well into a web2.0 and semantic web savvy user's world.

Linked data has to be maintained as certain providers come up with rich metadata that has inner links and also may refer to external targets like authority files and other resources on the web. The key challenge is the metadata mapping, partly because of the heterogeneity of the input received by providers, but also because a one-to-one mapping and harmonization for metadata and ontologies is close to impossible. Experience and best practices must be applied when transforming and consolidating formats into an internal knowledge model which is needed for clustering and reasoning. Besides from a state of the art multilingual and faceted search, not only scientists would profit from queries including transitions, like "Which person is related to event (x)?" or "When was (x) first used?" Queries like these that run against billions of triples on a public platform and that may end up with more than 1.000 parallel users are not common technical requirements.

The fact that user generated content (objects, links, valuation) can seriously enrich a platform has already been demonstrated by other projects, for example by the portal of the Australian newspapers (http://newspapers.nla.gov.au). Commercially exploitable services and products will be offered at a later stage but require an architecture well suitable for the integration of services implementing e-commerce functionality and payment systems. A cornerstone towards monetization is that IPR and other regulations are strictly respected. Sealing of digital objects, a technique firstly applied in ancient times to physical objects using wax, may be needed to ensure the authenticity of cultural heritage objects.

More advanced technical features are related to data mining and text/image/audio processing, like automatic metadata extraction from textual or binary content, clustering based on raw content and automatic linking. One of the preconditions is to tightly cooperate with those institutions that offer services like persistent identifiers and URN resolvers, that maintain authority files and vocabularies, or already do work on platforms like Fedora [PL98] and eSciDoc [Ra09], on frameworks and tools. While decentralization is part of the challenge, it also bears a clue for the solution. A lesson learned from earlier projects is that mapping and visualization of the cultural heritage objects can only be done in an adequate way by using a decentralized approach: It is the provider who has the best knowledge of the digital objects. In addition, domain specific consulting is needed to recognize and establish human task processes, patterns of usage and templates for visualization and mapping. Thus the acceptance of the project stands or falls with the usability that end users experience, but this experience is heavily reliant on understanding how cultural heritage is managed in traditional modern knowledge managing institutions.

Acknowledgement

The work presented here was funded by the Federal Government Commissioner for Culture and the Media.

References

[HJ08] Hitson, B. A.; Johnson, L. A.: WorldWideScience.org: Bringing Light to Grey. Tenth International Conference on Grey Literature: Designing the Grey Grid for Information Society, 8-9 December 2008.

[Ma02] Maier, G.: Gemeinsames Internetportal für Bibliotheken, Archive und Museen– BAM-Portal. World Library and Information Congress: 68th IFLA Council and General Conference, 2002.

[Pa09] Paaß, G. et.al.: Text Mining and Multimedia Search in a Large Content Repository. In: Proceedings of the Sabre Conference on Text Mining Services, TMS 2009, Leipzig, 2009.

[Pi03] Pianos, T.: Vascoda - a Portal for Scientific Resource Collections Created by German Libraries and Information Centres. World Library and Information Congress: 69th IFLA General Conference and Council, 2003.

[Pi09] Pilz, A., et. al.: Entity resolution by kernel methods. In: Proceedings of the Sabre Conference on Text Mining Services, TMS 2009, Leipzig, 2009.

[PL98] Payette, S.; Lagoze, C.: Flexible and Extensible Digital Object and Repository Architecture (FEDORA). Research and Advanced Technology for Digital Libraries, Volume 1513, p. 517, 1998.

[PR09] Paaß, G.; Reichartz, F.: Exploiting semantic constraints for estimating supersenses with crfs. In: Proceedings of Ninth SIAM International Conference on Data Mining 2009.

[Th10] Theseus. http://theseus-programm.de/theseus-basic-technologies.html. 2010

[Ra09] Razum, M. et. al.: eSciDoc Infrastructure: A Fedora-Based e-Research Framework. Research and Advanced Technology for Digital Libraries. Volume 5714, p227-238, 2009.