Convergence of internet services in the cultural heritage sector – the long way to common vocabularies, metadata formats, ontologies

Jörn Sieglerschmidt
Bibliotheksservice-Zentrum Baden-Württemberg, Konstanz

Abstarct

Since several years it has been observed that information offered by different knowledge producing institutions on the internet is more and more interlinked. This tendency will increase, because the fragmented information offers on the internet make the retrieval of information difficult as even impossible. At the same time the quantity of information offered on the internet grows exponentially in Europe – and elsewhere - due to many digitization projects. Insofar as funding institutions base the acceptance of projects on the observation of certain documentation standards the knowledge created will be retrievable and will remain so for a longer time. Otherwise the retrieval of information will become a matter of chance due to the limits of fragmented, knowledge producing social groups.

On the following pages I shall present some standards and how they are used up to present. Subsequently I shall deal with European ALM projects and the German project, its implementation, its claims and its future targets.

Regarding standards we have to distinguish between syntax and semantics. My use of these terms doesn't or rather only partly correspond to the concepts of information science.[1] Data carry the content of a communication. It can be thus analyzed semantically. The structure of data cannot be completely separated from the content, because application rules influence the content as – reversely – the content or rather the intended communication influences the application rules. Grammar and - in the narrower sense - syntax preform, but doesn't determine the semantics. Therefore only statements about objects can carry meaning; those objects are the contents (text, image, sound) offered via internet.

Texts – and only these, as long as image or sound retrieval devices are not reliable - can be indexed with the help of authority files. Well-known authority files of that kind are the thesauri of the Getty Foundation (TGN, AAT, ULAN) [2] or the subject headings of the national libraries (e. g. DE: SWD, FR: RAMEAU, GB: LCSH) [3], to list only few examples. The advantage of using such authority files lies in linking different internet resources by using the ID numbers assigned to every term. A museum object like an airplane (German: Flugzeug) could be linked to literature of libraries by using this number (DE: SWD: 4017672-1; FRBNF11931002). Furthermore the user could get information about similar objects and literature by using the semantic net that the terms of a thesaurus offer (synonyms, related, narrower, broader terms etc.). Multilingual tools are indispensable for

[1] Sieglerschmidt,J.,Metadaten: http://www2.bsz-bw.de/cms/service/museen/publ/metadaten-js-2002.pdf.

[2] Getty Vocabularies: http://www.getty.edu/research/conducting_research/vocabularies/.

[3] *Schlagwortnormdatei der Deutschen Bibliothek*: not online; Subject Headings of theLibrary of Congress: http://authorities.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First; *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* of the Bibliothèque National de France: http://catalogue.bnf.fr/servlet/AccueilConnecte?Autorites=RAMEAU.

the future of internet retrieval. The German term Fahrzeug (broader term of Flugzeug, SWD-ID 4016320-9) has the same meaning as the French véhicules (FRBNF11975775) and the British/American vehicles (US/GB: LSCH). The aim of the MACS project is to coordinate – not to translate – those terms, in order to allow the retrieval of objects in different foreign language OPACs. [4] Thus it will be possible to use multilingual semantic nets for internet retrieval – similar to the semantic web, where terminological combinations are used for linking customer services.[5] Back to our example: The reverse path could be followed viz. the reference from literature found in an OPAC to museum objects. The efforts of catalogue enrichment aim at services like that. The example might be sufficient in order to show that here m:n-relations are concerned, a semantic space that has been widened by the use of authority files and that could be expanded by supplementary dimensions. Going beyond such attempts the semantic web offers help designing typical scenarios of social action, i. e. workflows that are usually tied together. Libraries and their online services have realized such workflows already. Thus a book that could not be found in the library nearby will be ordered by interlibrary loan. Or an article can be ordered as pdf-file using the *subito*-portal.[6] Archival online platforms have partly realized functions like that, and museums could do that, too.[7] Basically the task is to model repetitive workflows (administrative processes) and to extract the relevant information using the model structure and to offer services linked to it.

The syntax of the data offered on the internet is equally important for linking information, because it defines where and which type of data can be found using certain orthographical notation rules. All kinds of databases use such syntactical rules (fields, types of fields, rules for data entry). Usually different data like that are separated by tags or metadata structures. The Dublin Core Metadata Element Set (DCMES) is well known in that respect, but is losing importance during the last years.[8] While in early Archive-Library-Museum (ALM) projects like the German BAM project the DCMES seemed to be an appropriate metadata format for all three types of documentation structure, current projects like the Finnish prefer a combination of ontologies and XML.[9] The metadata formats EAD and TEI are designed for the recording of archival documents and autographs. The Getty Foundation developed metadata standards for the museums: Cataloguing Cultural objects (CCO) and Categories for the Description of Works of Art (CDWA,CDWA lite).[10] The XML schema ABCD of the Taxonomic Databases Working Group (TDWG) of the Global Biodiversity Information Facility (GBIF) is significant for natural history collections.[11] The Metadata Encoding and Transmission Standard (METS) supported by the Library of Congress likewise is a XML schema aiming at the description and linking of different internet resources.[12]

All metadata formats have a rigid, rather inflexible structure. The links between the metadata elements disappear because the elements are ordered parallelly and not hierarchically. Retrieval strategies must use complex and combinations in order to get

[4] Multilingual Access to Subjects: https://ilmacs.uvt.nl/pub/.

[5] For information see: http://www.w3.org/2001/sw/.

[6] subito documents from libraries: http://www.subito-doc.de/, especially articles.

[7] cf. www.bpk-images.de.

[8] ISO 15836 (draft): http://www.niso.org/international/SC4/n515.pdf.

[9] Encoded Archival description, von der Library of Congress getragen: http://www.loc.gov/ead/; Text Encoding Initiative: http://www.tei-c.org/.

[10] http://www.getty.edu/research/conducting_research/standards/cdwa/8_printing_options/definitions.pdf.

[11] http://www.gbif.org/.

[12] http://www.loc.gov/standards/mets/.

precise and/ or complete results. The question e. g., which artists worked in certain places, cannot be answered. Information of complex structure cannot be modelled appropriately by the flat metadata structures.

Ontologies offer a solution for such problems because they are very flexible regarding the adaptation to new requirements. They can model enormously complex structures without producing confusion. The here cited example of the Conceptual Reference Model (CRM) of the Comité International pour la documentation (CIDOC), one of the committees of the International Council of Museums (ICOM), does just this. The CIDOC-CRM is evaluated as draft by the bodies of the International Standards Organisation (ISO).[13] The CIDOC-CRM corresponds to the requirements of the Resource Description Framework (RDF(S)) and is compatible to object oriented and relational database models (R and OO DBMS) and to XML, especially by its infinitely nested structure. The CRM has been mapped to DCMES, EAD, TEI, and FRBR.[14] These mappings show that the CRM is very powerful and more consistent than other standards. The CRM could therefore be recommended as a general reference model for specific metadata and database formats of all kinds. The CRM offers a model for all knowledge producing institutions, because even complex scholarly contents can be handled.

The CRM distinguishes classes (84 *entities* so far) and relations (141 *properties* so far). The CRM has a syntactical structure with (historical) events as main reference points. Classes have content only if a reference is given. Classes can be repeated infinitely. Classes and properties are organised in a hierarchical structure of supraclasses, subclasses, and elementary classes. Supraclasses are apart from the term class itself:

- place,
    - temporal entities,
    - physical stuff,
    - actors,
    - conceptual objects.
- Primitive values, i. e. classes without relations to others, are:
    - number,
    - time primitive and
    - string.
- Detailed information can be downloaded from the internet site of the CIDOC-CRM.

Historical knowledge is incomplete, not finite, and not conclusive. Its elements show a variable statistical stability regarding the revision of knowledge by current discourse, existence in real life, identity and relations to other elements of knowledge. The CRM wants to describe the structure of knowledge and its change. In order to enhance monotonicity it structurally and conceptionally prefers transitions instead of fixed states, opinions instead of propositions. It avoids the repetition of relationships and excludes exceptions.

The CRM conceives itself as a tool of knowledge organization, of knowledge administration and of scientific documentation in a narrower sense. That doesn't mean to exclude a wider, non-academic public, but strives to satisfy the demands of scholarly research through precision and detailed information. Areas of application are

[13] CIDOC-CRM version 3.4.9 of 2003: http://cidoc.ics.forth.gr/docs/cidoc_crm_version_3.4.9.pdf; intended to become ISO 21127 (draft).

[14] Functional Requirements for Bibliographic Records: http://www.ifla.org/VII/s13/frbr/frbr.pdf of the International Federation of Library Associations and Institutions (IFLA).

- the integration of information of cultural heritage institutions,
- the improvement of the understanding of this information,
- the representation of this information, e. g. in the form of storytelling.

Basically the issue is knowledge transfer from cultural heritage institutions.

Portals trying to combine such information in one interface have been designed since the end of the nineties, when the EU increasingly demanded the convergence of information from the cultural heritage institutions archive, library, and museum. Initially the requirements of the EU had an effect on the development of national initiatives during the last years. A few examples of such initiatives will now form the final part of this article. The following portals will be presented:

- Norway: Arkiv, Bibliotek og Museum, ABM utvikling Statens Senter;[15]
- Denmark: Nordjyllands kulturhistoriske søgebase (NOKS);[16]
- Netherlands: Cultuurwijzer Nederland;[17]
- Germany: BAM, Portal für Bibliotheken, Archive, Museen.[18]

All portals are monolingual. Some offer foreign language versions of the project description. So ABM utvikling e. g. offers an English and a German version. All portals are in the stage of development and offer therefore a more or less restricted program regarding the range of retrievable objects, although the claim – to be realized in the future - is to offer a complete regional/ national portal. The Cultuurwijzer expresses it generally:

> *De Cultuurwijzer is een platform dat het publiek toegang geeft tot het culturele erfgoed in Nederland. Het biedt in wisselende thema's een schat aan informatie. Deze thema's geven toegang tot de onderliggende informatie.*

All portals – with the exception of ABM utvikling - offer content information (text, image, sound) in addition to the structured description of items (title, author/creator etc.) for the retrieval of the digital resources. The projects make efforts to implement national and international standards, e. g. NOKS: *at fremme koordinering og samordning af eksisterende standarder*. Different retrieval strategies are offered: either a full text retrieval or a retrieval structured by a - sometimes facetted - classification. Thus retrieval can be restricted to certain institutions or types of institutions, to types of media, or provenance of objects. The technical tools employed are not described in detail, so that technical solutions cannot be compared. Such a technical description exists only for the Finnish museums portal *Museusuomi*.[19]

The BAM portal follows the principles expressed above, too: the development of the BAM portal into a cultural heritage portal that is able to cooperate in a national and international framework, the development of appropriate organizational and administrative structures for the maintenance of the services after the end of funding by the German Research Association, the definition of conditions of participation, the further development of technical functions. The BAM portal is formed by a consortium:

- Stiftung Preußischer Kulturbesitz (Institut für Museumskunde der Staatlichen Museen zu Berlin),
- Bundesarchiv Koblenz/ Berlin,
- Landesarchiv Baden-Württemberg,

[15] http://www.abm-utvikling.no/om/english.html.

[16] http://www.noks.dk/.

[17] http://www.cultuurwijzer.nl/asp/page.asp?alias=cultuurwijzer.nl.

[18] http://www.bam-portal.de/.

[19] Hyvönen, Eero e.a.: A cultural community portal for publishing museum collections on the semantic web. In: http://www.seco.tkk.fi/publications/2004/hyvonen-saarela-et-al-a-cultural-community-2004.pdf; cf. also iid.: http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html.

- Landesmuseum für Technik und Arbeit Mannheim,
- Bibliotheksservice-Zentrum Baden-Württemberg.

The project will be financed by the German Research Association (DFG: Deutsche Forschungsgemeinschaft) in a third project phase until the middle of 2007. The main targets are: the improvement of the technical performance and the functionalities, the steady and even growth of content of all three participating types of institution. The main problem is that of different stages of digitization, so that a balanced information offer seems impossible for the next years.

The BAM portal collects the data on a central server like the Finnish project, because only thus the search engine employed, Lucene, will be able to be effective technically (figure 1). Linguistic procedures and the use of authority files (e. g. SWD) are used during the retrieval process in order to widen the semantic space of a searched term. A sufficient performance is only possible with this technical design of the portal. Only thus a user friendly performance can be achieved. Devices of distributed retrieval will be tested, if the now implemented devices are running reliably.

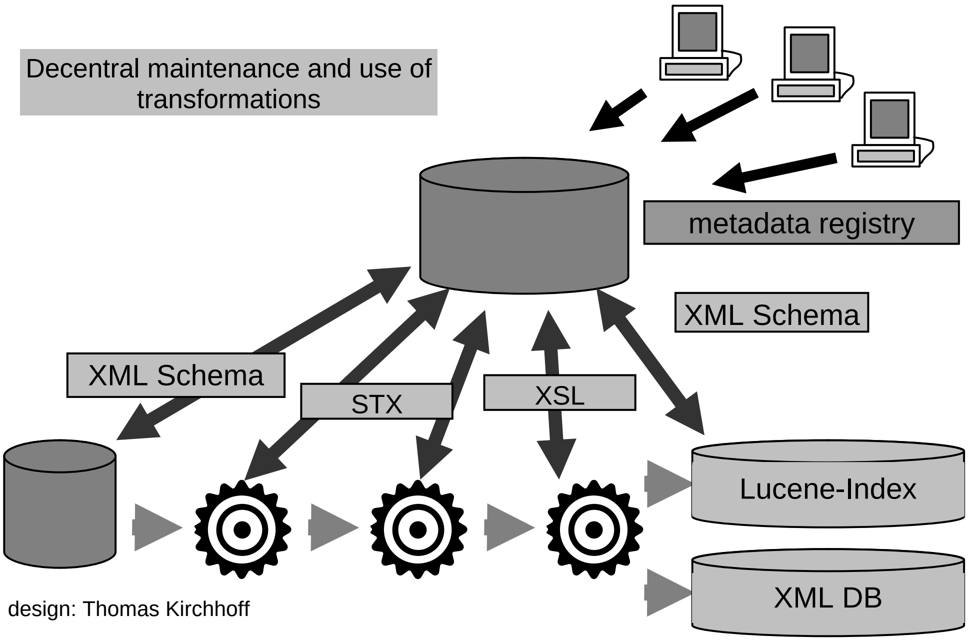


Figure 1: Formatin of the index in BAM

The BAM portal offers the participating institutions a common, suprainstitutional access for their digital catalogues, repertories, and inventories. Scholars, students, and other citizens get a first online access to cultural heritage information. Also institutions without an own online catalogue get the chance to host their object information on the BAM-server. It is possible that these institutions implement a search form on their websites for the presentation of a digital catalogue of the institution, but based on the BAM-Portal database.

The simple Google slot or an expanded search form enable a simultaneous and complex search over digital collections of heterogeneous provenance and structure. The resulting list is linked to the specialized information systems of the institutions (figure 2). If the user wants more detailed information, she/he is directed to the website of the institution holding the object. There can be found digital reproductions of the object, if they exist.

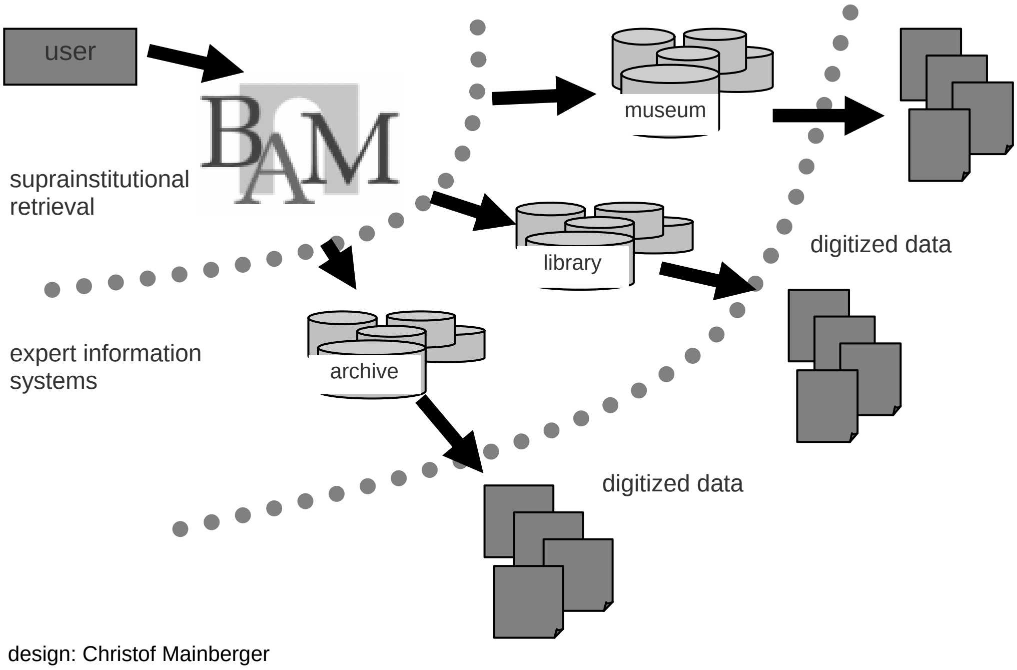
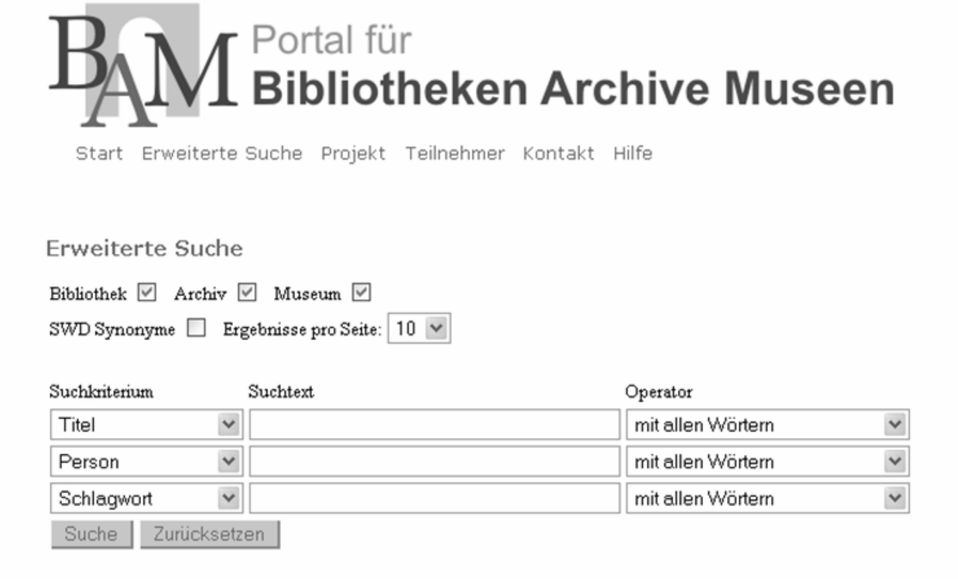

user
BAM
suprainstitutional retrieval
museum
library
archive
digitized data
expert information systems
digitized data
design: Christof Mainberger

Figure 2: Concept of information layers

One main focus of the project is the semantic standardization of the metadata used. Therefore the employment of metadata standards is supported. The employment of linguistic features and of authority files will be examined, in order to improve the retrieval results.

The authority file mainly used in the BAM-Projekt is the Schlagwortnormdatei (Subject Headings Authority File, SWD). The SWD provides a standardised vocabulary, i.e. a controlled terminology. The sources used for this are to be found in the Liste der fachlichen Nachschlagewerke zu den Normdateien (List of the reference works for authority files). The SWD is compiled by Die Deutsche Bibliothek and six regional German library organizations. On the one hand, archives and museums are missing parts of their specific vocabulary in thus file. On the other hand, the SWD contains large parts of the everyday language that the users of the BAM-Portal are likely to use. The technical language of the archivists and museologists can also get part of the SWD like the missing subjects.

The SWD will be used in the BAM-Portal in the expanded search by the user (figure 3). With this option, the search string also retrieves the related and the narrower terms of the search term in the same search run.



Erweiterte Suche

Bibliothek ☑ Archiv ☑ Museum ☑

SWD Synonyme □ Ergebnisse pro Seite: 10

Suchkriterium	Suchtext	Operator
Titel		mit allen Wörtern
Person		mit allen Wörtern
Schlagwort		mit allen Wörtern

Figure 3:Expanded BAM Search

The SWD is also a tool that is used during the automatic indexing of large text portions not related to authority files of the BAM Project. The use of authority files is not yet common in the work in German archives and museums. In this case, a lot of the information that the BAM-Project deals with has no relation to the SWD. So we have tested the automatic indexing of these data and relate the most important results (for roots, composite terms) of this process to the SWD.

Finally subject matter relations between the different collections are investigated, in order to enable links between objects of different provenance. Classification tools can thus be offered as an alternative access point of retrieval. The cultural heritage institutions will be presented as a whole with emphasis on their profile, their history, and main focus of collection in the BAM portal beside the access to the individual objects.

The use of the portal will show whether the offer will be widely accepted in this form, but not before the amount of information has grown, especially regarding the museums and the archives. The consortium of the BAM portal supports the principles of the open access initiatives; it wants to be open to all citizens and free of charge. Initially the addressees of the portal are the much more limited number of scholarly and culturally interested people like students and scholars. Supplementary functionalities should enhance the attractiveness of the portal and give an incentive for institutions to deliver content. But it will depend on further funding, whether offers – and efforts - can be made regarding functionalities like long term preservation, e-learning, e-science, etc.