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
Frank von Hagel, Institut für Museumskunde, Staatliche Museen zu Berlin, Stiftung Preußischer Kulturbesitz

Summary

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 or 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. Inasfar 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 long time. Otherwise the retrieval of information will become a matter of chance due to the limits of fragmented, knowledge producing social groups.

Keywords:
ontologies, metadata formats, authority files, CIDOC-CRM, portals, cultural heritage

On the following pages we shall present some standards and how they are used up to present. Subsequently we shall deal with European Archive-Library-Museum (ALM) projects and, in particular, the German ALM project, its implementation, its claims and its future targets.

Regarding standards we have to distinguish between syntax and semantics. Our 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 therefore be 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 don't determine the semantics. Therefore only statements about objects can carry meaning; those objects are the contents (text, image, and sound) offered via the 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 for example the thesauri of the Getty Foundation (TGN, AAT, ULAN) [2] or the subject headings of the national libraries (e. g. DE: SWD, FR: RAMEAU, subject headings used by the national libraries UK: LCSH) [3]. The advantage of using such authority files in linking different internet resources is the use of well-defined 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 the future of internet

[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] The *Schlagwortnormdatei der Deutschen Bibliothek* is not available online; see for basic information <http://www.ddb.de/eng/standardisierung/normdateien/swd.htm>; Subject Headings of the Library of Congress: <http://authorities.loc.gov/cgi-bin/Pwebrecon.cgi?DB=local&PAGE=First>; *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* der Bibliothèque National de France: <http://catalogue.bnf.fr/servlet/AccueilConnecte?Autorites=RAMEAU>.

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 former MACS[4] and the actual CrissCross[5] project is to coordinate – not to translate – those terms, in order to allow the retrieval of objects in different foreign language OPACs. 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[6]. 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.[7] Archival online platforms have partly realized functions like that, and museums could do that, too.[8] 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 are separated by tags or metadata structures. The Dublin Core Metadata Element Set (DCMES) is well known in that respect, but has lost some importance during the last years.[9] While in early 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 MuseoSuomi prefer a combination of ontologies and XML The metadata formats EAD and TEI are designed for the recording of archival documents and autographs.[10] 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).[11] The XML schema ABCD of the Taxonomic Databases Working Group (TDWG) of the Global Biodiversity Information Facility (GBIF) is relevant for natural history collections.[12] 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.[13]

All metadata formats have a rigid, rather inflexible structure. The links between metadata elements disappear because the elements are ordered parallelly and not hierarchically. Retrieval strategies must use complex and-combinations in order to get precise and/or complete results. The question e. g., which artists worked in certain

[4] Multilingual Access to Subjects: <https://ilmacs.uvt.nl/pub/>.
[5] CrissCross: <http://www.ddb.de/eng/wir/projekte/crisscross.htm>.
[6] For Information see: <http://www.w3.org/2001/sw/>.
[7] *subito* library document delivery service: <http://www.subito-doc.de/> engl. Version<http://www.subito-doc.com/>.
[8] cf. <www.bpk-images.de>.
[9] ISO 15836 (draft): <http://www.niso.org/international/SC4/n515.pdf>.
[10] Encoded Archival Description, supported from the Library of Congress: <http://www.loc.gov/ead/>; Text Encoding Initiative: <http://www.tei-c.org/>.
[11] <http://www.getty.edu/research/conducting_research/standards/cdwa/8_printing_options/definitions.pdf>.
[12] <http://www.gbif.org/>.
[13] <http://www.loc.gov/standards/mets/>.

places, cannot be answered. Information of a complex structure cannot be modelled appropriately by 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 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 Organization (ISO).[14] The CIDOC-CRM corresponds to the requirements of the Resource Description Framework (RDF(S)) and is compatible with object oriented and relational database models (R and OO DBMS) and with XML, especially by its infinitely nested structure. The CRM has been mapped to DCMES, EAD, TEI, and FRBR.[15] 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 varying 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 uniformity 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 is conceived 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

- 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.

[14] CIDOC-CRM in the version 3.4.9 from 2003: <http://cidoc.ics.forth.gr/docs/cidoc_crm_version_3.4.9.pdf>; intended to be ISO 21127 (draft).

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

Basically the issue is knowledge transfer from cultural heritage institutions.

Portals trying to combine such information in one interface were designed since the late nineties, when EU funding mechanisms increasingly demanded the convergence of information from the cultural heritage institutions archive, library, and museum. The requirements of the EU had a visible 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 contribution. The following portals present digital heritage in the internet:

- Norway: Arkiv, Bibliotek og Museum, ABM utvikling Statens Senter;[16]
- Denmark: Nordjyllands kulturhistoriske søgebase (NOKS);[17]
- Netherlands: Cultuurwijzer Nederland;[18]
- Germany: BAM, Portal für Bibliotheken, Archive, Museen.[19]

Only some of these portals offer foreign language versions of the project description, while the bulk of information is only available in one language. ABM utvikling, e. g. offers brief English and German versions. All portals are currently in the stage of development and are therefore more or less restricted with regard to the range of retrievable objects, although the claim – to be accomplished in the future - is to offer a complete regional/national coverage. 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, images, sound) in addition to structured descriptions of specific items (title, author/creator etc.). The projects make an effort implementing national and international standards, e. g. NOKS: *at fremme koordinering og samordning af eksisterende standarder*. Different retrieval strategies are possible: either full-text retrieval or structured retrieval, by – sometimes facetted – classification schemes. Retrieval operations can be restricted to certain institutions or types of institutions, to types of media, or to the provenance of objects. Technical tools employed are not described in detail, so that technical solutions cannot be compared. Such technical description exists only for the Finnish museums portal *Museusuomi*.[20]

The main objective of the BAM project is to present the variety of cultural and scientific traditions and to develop a nation wide cultural heritage portal that presents information from archives, libraries and museums on the level of individual items (record, book, object) and of collections, in order to show the relations between and to connect the different strands of cultural knowledge. As a means to this end it is intended

- to cooperate with other relevant initiatives on a national and international level,
- to develop appropriate organizational and administrative structures for the sustainable maintenance of the service after public funding has run out,
- to define conditions of participation, and to further develop technical functions,
- to coordinate and unify the relevant metadata,
- to support the use of metadata standards and authority files.

The BAM portal is formed by a consortium:

- Stiftung Preußischer Kulturbesitz,
- Bundesarchiv,
- Landesarchiv Baden-Württemberg,

[16] <http://www.abm-utvikling.no/om/english.html>.

[17] <http://www.noks.dk/>.

[18] <http://www.cultuurwijzer.nl/asp/page.asp?alias=cultuurwijzer.nl>.

[19] <http://www.bam-portal.de/>.

[20] 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 <http://www.archimuse.com/mw2004/papers/hyvonen/hyvonen.html>.

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

The collections presented are (May 2006):

- the South West German union catalogue (Südwestdeutscher Bibliotheksverbund) with about 12 million titles,
- the Northern German union catalogue (Gemeinsamer Bibliotheksverbund) with about 22 million titles,
- the Prussian Cultural Heritage Foundation with the library catalogues of about 10 million titles and the online content of the image archive with about 15.000 items,
- the national union catalogue for post-medieval handwritten materials, Kalliope, with about 1 million items,
- the regional archives of Baden-Württemberg with 2500 online inventories,
- the national archives with 500 inventories,
- the Historical Museum of the city of Leipzig (Stadtgeschichtliches Museum) and four South Western State Museums (Landesmuseum für Technik und Arbeit Mannheim, Badisches Landesmuseum Karlsruhe, Staatliche Kunsthalle Karlsruhe, Württembergisches Landesmuseum Stuttgart) with a few thousand objects.

The project is financed by the German Research Council (DFG: Deutsche Forschungsgemeinschaft) in its current third project phase that will last until the middle of 2007. The main objectives of this third project phase are: to improve the technical performance, the functionalities, and the steady growth of content of all three participating types of institutions. The idea of the portal is to produce results from all three types of institutions, whatever topic you are searching for. Therefore the amount of content – and hopefully a balanced one – is crucial for satisfying this expectation of users.

The BAM portal offers to the participating institutions a common, cross-institutional access to their digital catalogues, repertories, and inventories. Scholars, students, and other users are provided with a first open and free of charge access to cultural heritage information via the internet. Institutions without an own online catalogue will be able to host their data on the BAM server. It is possible that such institutions implement a search form on their websites for the presentation of the specific digital catalogue of the institution, while the data remains within the BAM-Portal database. Supplementary functionalities for the content delivering institutions (e. g. long term preservation) and the users (e.g. e-learning, e-scholarship) are planned, but will not be realized in this phase of the project.

The cultural heritage profiles of the participating institutions will also be presented on the website. This includes information about the particulars, history, collection focus, etc. of the content delivering institutions. In this field there are close relations to the MICHAEL project.[21] Direct access to data from a type of institution and even to data from a specific institution will be offered by the extended search form and by an guided access to institutions mentioned in the results retrieved.

The simple Google slot or an extended search form allows for searches over digital collections of heterogeneous provenance and structure. Results are linked back to the originating catalogues of the data in question where detailed information, in as far as it is available, may be accessed. (Fig.1). Digital reproductions of the object, if these exist at all, are also available.

It has to been noted that the cataloguing of objects is done with different levels of granularity in the three participating types of institutions, which results in very different numbers of records available. Museums e. g. have to invest a lot of effort in the

[21] <http://www.michael-culture.org/index.html>.

cataloguing of their unique resources, while libraries may even take over entire bibliographic records from other libraries. The list of results therefore is somehow unbalanced and will be so in the future. The ranking procedures have to provide for a better ranking of museums and archives, so that these get a chance to be present on the top of the list.

The search engine Lucene, software that collects local data and inserts it into a common database, is able to work effectively by using pre-converted data in formats like EAD, MAB2 and individual museum formats and mappings (Fig. 2). The usability of the CRM for these mappings will be tested during the project phase. By using authority files (see beneath) and stemming tools for the retrieval process search terms will be expanded by either synonyms or additional words that are related to the term searched. Issues of performance do no longer exist within this technical design. Distributed retrieval functionalities was tested during the first phase and can be reactivated when the now implemented procedures will run effectively.

The authority file mainly used in the BAM project is the Schlagwortnormdatei (SWD; Subject Headings Authority File of the German National Library). The SWD provides a controlled vocabulary for all areas of knowledge and will be expanded in the future with special regard to the needs of museums. 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)[22]. The SWD is compiled by Die Deutsche Bibliothek and six regional German library organizations.

On the one hand, the SWD traditionally is a tool of librarians. This means that archival and museum jargon is not represented in all respects. Regarding the museums the vocabulary of material and everyday culture is represented poorly. On the other hand the SWD contains a lot of everyday speech that users of the BAM portal are likely to use, because the norms of the determination of terms intend to find – as many as possible – descriptors that are part of everyday speech. Of course, the creation of the SWD is an ongoing process. There will be ample opportunities for archivists and museum specialists to add their specific vocabularies to it.

The SWD may be used in the search option of the BAM Portal (Fig.3). Applying this functionality, users will expand their query with synonyms with hierarchical top-terms being available for retrieval. But the use of authority files is not yet common in German archives and museums. A lot of the information that the BAM project deals with has no relation to the SWD. However, 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.

The BAM project, like the other ALM projects mentioned, fits well into the plans of the EU. We are strongly interested in cooperation with others, in order to develop the idea of a common portal. First steps are made by projects in the different types of institutions and in different countries.[23] Photo archives and other institutions holding cultural heritage objects present a lot of information online. Universities and educational institutions offer much of their knowledge online, too.[24] They should be integrated into a

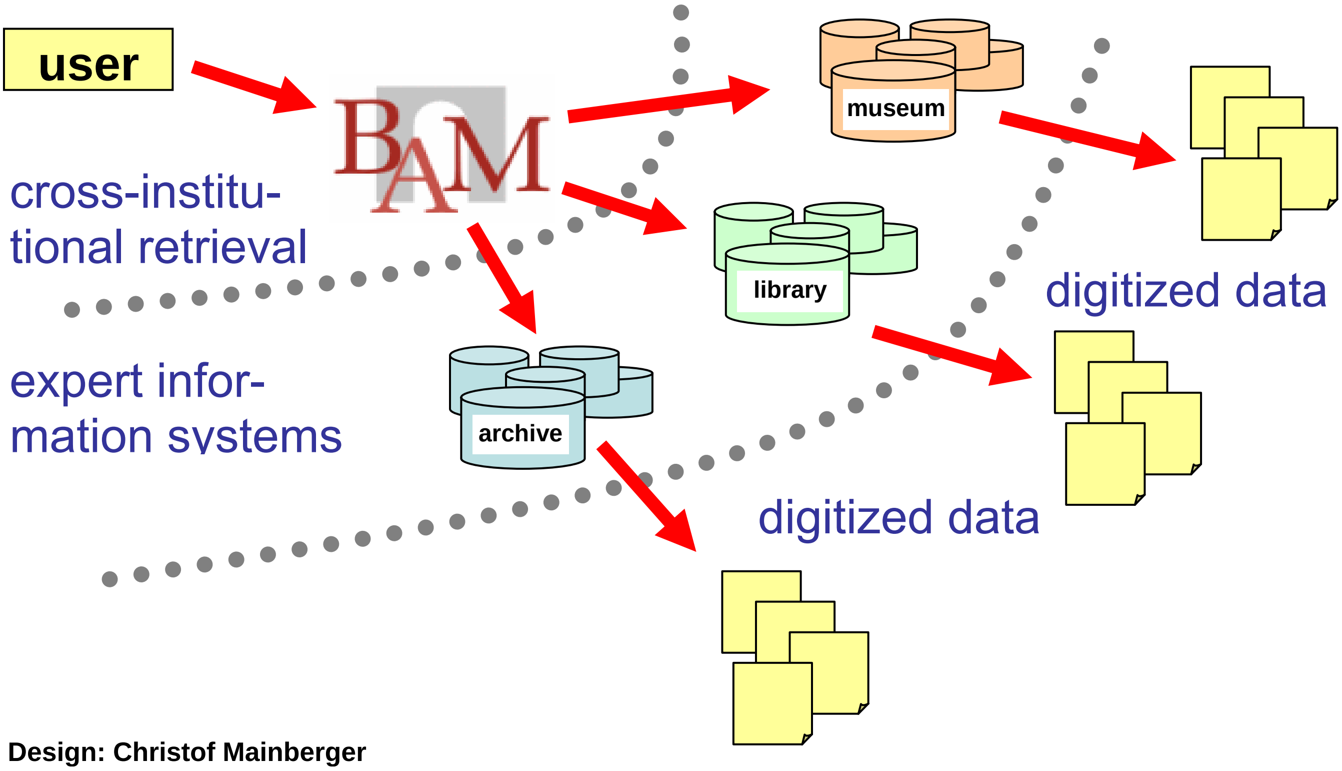
[22] <http://support.ddb.de/swd/listeNSW/index.htm>.

[23] Beside the above mentioned projects there are many projects of the national libraries, of the national museums and archives: e. g. British Library: <http://www.bl.uk/collections/toppage.html>, the National Archive of Germany: <http://www.bundesarchiv.de/bestaende_findmittel/bestaendeuebersicht/index_ frameset.html> or Joconde of the Ministère de la culture de France: <http://www.culture.gouv.fr/ documentation/joconde/fr/apropos/presentation-joconde.htm>.

[24] E. g. Basel Mission Picture Archive: <http://www.bmpix.org/>, ECHO project: <http://echo2.mpiwg-berlin. mpg.de/home> or the portal of university collections Universeum: <http://www.universeum.de/>, to mention only but a few examples.

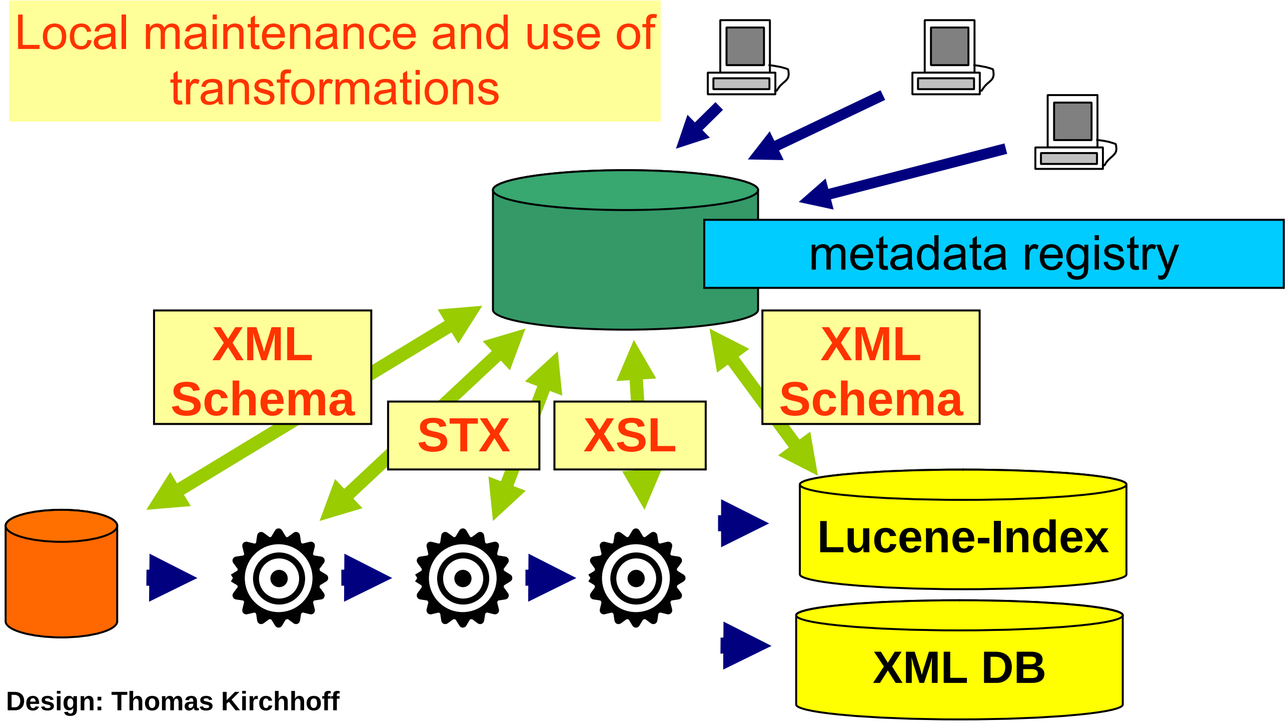
European portal presenting the diversity and the common traditions of European cultural heritage in all branches of knowledge.

Fig. 1: Concept of information layers



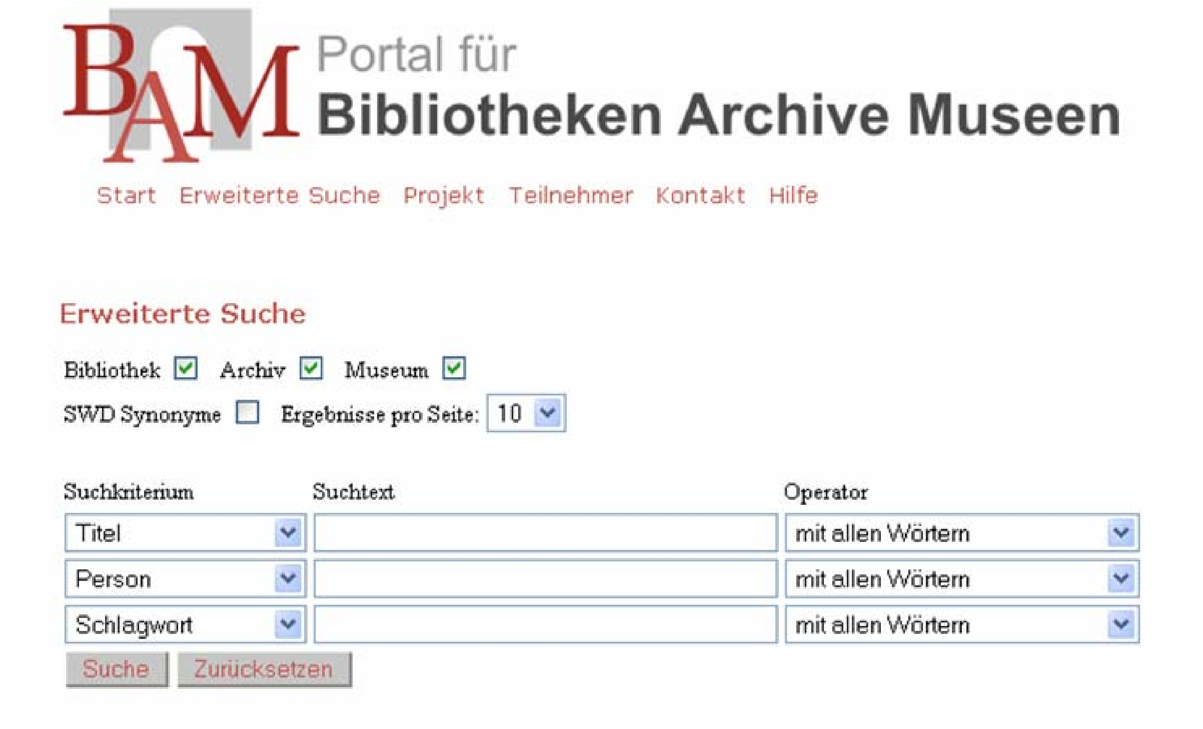
Design: Christof Mainberger

Fig. 2 Formation of the index in BAM



Design: Thomas Kirchhoff

Fig. 3: Extended BAM Search



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