Date submitted: 22/06/2010

CONTENTUS – Towards Semantic Multimedia Libraries

Jan Nandzik
Acosta Consult
E-mail: jn@acosta-consult.de

Andreas Heß
German National Library
E-mail: a.hess@d-nb.de

Jan Hannemann
German National Library
E-mail: j.hannemann@d-nb.de

Nicolas Flores-Herr
Acosta Consult
E-mail: nf@acosta-consult.de

Klaus Bossert
Acosta Consult
E-mail: kb@acosta-consult.de

Meeting: **149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

The ever-growing amount of content and knowledge published online makes it possible for libraries to complement their own data and to present their collections in novel ways. Conceptually related information can be semantically linked so that users may benefit from richer data collections and novel search possibilities that capitalize on the inherent relationships between media, local metadata and external information sources.

This paper presents potential solutions for the fundamental challenges of integrating heterogeneous data sources and providing innovative semantic search approaches, as they are developed for libraries and multimedia archives within the CONTENTUS project.

Introduction

In Germany approximately 30.000 cultural institutions archive an incredible wealth of multimedia content stored on carriers like books, images, tapes and films. These cultural heritage organizations face the challenge to provide citizens with internet-based access to the knowledge contained within their vast multimedia collections. Novel semantic multimedia search services will be the future technological foundation for users to access digital collections. A prerequisite for the application of these technologies is the *integration of bibliographic metadata*, *automatically generated metadata* and *external information resources* into a knowledge base. This paper focuses on the related methodological and technical development carried out by the German National Library and partners in the context of the CONTENTUS project[1].

Searching Multimedia Collections: Challenges Faced by Cultural Heritage Organizations

In order to support semantic multimedia searches, large-scale multimedia collections need to be enriched with sufficient and appropriate descriptive metadata. However, to date the majority of multimedia assets are annotated and catalogued manually by information experts. As most multimedia resources, such as audiovisual media, contain a lot of information, manual metadata generation is very complex, cost-intensive and time-consuming.

In practice, when manually indexing large unstructured multimedia files either the asset can not be thoroughly described, or only fragments of the asset can be indexed in detail. This undesirable situation is commonly encountered in cultural heritage organizations and is due to the lack of human resources being able to cope with fast-growing multimedia collections.

With relevant and important resources on the internet (e.g., Wikipedia and Geonames) and collaboratively curated datasets like authority files, the task of indexing multimedia assets should not be restricted to describing their contents. As these external resources have the potential to semantically enrich collection items, multimedia assets as a whole or individual entities (e.g. persons, places, events) that are found within individual asset should be semantically connected to relevant internet-based datasets to complement the metadata available.

Nonetheless, multimedia indexing as well as linking entities within multimedia collection items to relevant external resources is difficult and far from being daily routine in cultural heritage organizations – this one of the challenges that is tackled by the project CONTENTUS.

The CONTENTUS Vision: Next Generation Multimedia Libraries

CONTENTUS is a research and technological development project led by the German National Library under the umbrella of the German government-funded THESEUS research initiative[2]. It provides a rich toolbox of solutions for cultural institutions and other content owners that facilitate a seamless transition from raw digital data to a semantic multimedia search environment [Bossert, Flores-Herr, Hannemann, 2009].

[1] http://www.theseus-programm.de/en-us/theseus-application-scenarios/contentus

[2] http://www.theseus-programm.de

The CONTENTUS framework and the methodologies and concepts that are being developed in the project yield a system that supports cultural institutions in providing end users with access to multimedia collections at a large scale. End-users benefit from innovative search options that are fuelled by the abundance of multimedia assets and metadata from various sources, including "traditional", intellectually compiled data, automatically generated information, and internet-based resources.

CONTENTUS works in close collaboration with ALEXANDRIA and Mediaglobe – THESEUS projects focusing on the development of the Web 3.0 and media asset technologies – to incorporate communities building and improving semantic knowledge networks. These collaborations will eventually produce open knowledge networks where multimedia assets of cultural institutions can be connected with resources of the social web: the *Next Generation Multimedia Libraries*

CONTENTUS aims to create an infrastructure for cultural heritage organizations that allows for the efficient processing of large multimedia collections and their linking with external metadata resources. The individual steps form a processing chain, as shown in **Figure 1**.

Figure 1: The CONTENTUS processing chain.

1.) **Digitization:** many archives still exist in analogue form, which makes an automated processing of the assets impossible. For such archives, mass digitization is the first step to opening the media to a comprehensive semantic search.

2.) **Quality Control:** automated quality analyses (e.g., book page scan quality check) and optimization procedures are essential to keep up with the speed of up-to-date digitization machines (e.g., book scanning robots). The goal here is to improve the media quality both for human consumption and content analysis (see next step).

3.) **Content Analysis:** manual metadata descriptions are often not sufficient to allow for an effective search for multimedia objects. Annotating text, audio and audiovisual content comprehensively, on the other hand, for example by transcribing speeches or indexing a news broadcast, takes enormous efforts in terms of manpower and financial resources. Within CONTENTUS, services that automatically analyze common multimedia resources like images, music recordings or videos are important to facilitate the generation of search-relevant information.

4.) **Semantic Linking:** to enrich the metadata available, automatically generated information can be linked with bibliographic metadata and internet resources. For example, the topic of a news documentary can be the author of a book, who can in turn be linked to a Wikipedia article or authority file entry. In addition, extracted entities such as places, persons, events etc. are disambiguated (i.e. to automatically discern apple as a fruit from apple as a corporation) and connected to the *Linked Open Data Cloud*, i.e., the entirety of the numerous sources of information available on the web as linked data.

5.) **Open Knowledge Networks:** in this step, multimedia assets can be further enriched by external communities with external resources.

6.) **Semantic Search:** CONTENTUS offers end-users innovative multimedia search functionality by combining searches for texts, images, audio and audiovisual content in a unified semantic user interface.

In this paper, we will focus on the challenges of data integration and approaches for facilitating semantic searches.

Data Integration

One of the key challenges in the CONTENTUS project is the need for integration of data and metadata from a variety of different sources. Typically, such data can comprise products of digitization efforts, born-digital documents, the contributions by the user community and more. Since we aim to make multimedia content from different sources accessible through a common interface, the challenge is to integrate and align the corresponding metadata. As opposed to traditional cataloguing systems we enrich our data with external sources, since we believe that the user can benefit from the additional information these sources contain. Even if the quality of the external metadata is sometimes (but not necessarily) lower than what can usually be expected from metadata created by librarians, it can complement existing data if it is either more detailed or contains aspects that have not been considered otherwise. For example, the catalogue of the German Music Archive (Deutsches Musikarchiv[3]) – an archive that hosts a central collection of sheet music and sound recordings in Germany and serves as the centre of bibliographic information related to music – does not list the individual songs or tracks of a recording. When linking the catalogue data to a music database, the user has access to more detailed information, such as the track lists.

For any data service that integrates information from different sources, we have to distinguish between two use cases:

1.) Linking

2.) Integration

In the first case, metadata from different sources are not stored in the same database, but only loosely coupled. Metadata from sources not under control of the service provider are only accessed when needed. This method has the advantage that (meta-) data that are presented to the service consumer are always as up to date as possible, even if they come from sources not under control of the service provider. The disadvantage is that the availability of external sources cannot be guaranteed.

In the second case, metadata from all sources are integrated into the same database or ontology store on the side of the service provider. The availability thus is only dependent on the service provider's own system. However, an update policy must be established in order to ensure that the

[3] The German Music Archive was founded in 1970 and it continues the activities of the Deutsche Musik-Phonothek, which existed from 1961 - 1969. The Musikarchiv is a department of the German National Library in Leipzig. See also: http://www.d-nb.de/eng/sammlungen/dma/samml_bestaende

data that are presented to the service consumer are not too outdated. Another issue that has to be considered is licensing, since data from other sources are not only accessed but copied.

In both scenarios the most important technical challenge is to find a mapping between the different schemas. This can either be done

1.) manually/intellectually or

2.) (semi-) automatically.

In CONTENTUS, we make use of both mapping approaches depending on the data sources. It should be noted that the metadata sources themselves could be either intellectually or automatically generated. For example, we make use of automated information extraction algorithms to find and disambiguate persons, organizations, places and topics from scanned texts, but we also incorporate intellectually generated authority file data.

In our current system, we integrate the following metadata sources:

- German National Library: authority files and catalog data
- Wikipedia: pictures of persons (planned: additional background information for persons and places)
- MusicBrainz: track listings for CDs
- Automatically extracted: persons, organizations, places and topics from text and audio, similarity between music tracks

The authority and catalog information serves as a reference. The mapping from Wikipedia to the authority file is maintained manually by volunteers. This mapping is already used in the German Wikipedia and the catalogue system of the German National Library. The mapping from track listings from MusicBrainz and the authority files is currently also manually generated.

Figure 2 illustrates the integration of data from different sources in CONTENTUS. Metadata stored in CONTENTUS can be seen as organized in a network, linking authors and works as well as additional information, for example about related places, topics or eras.

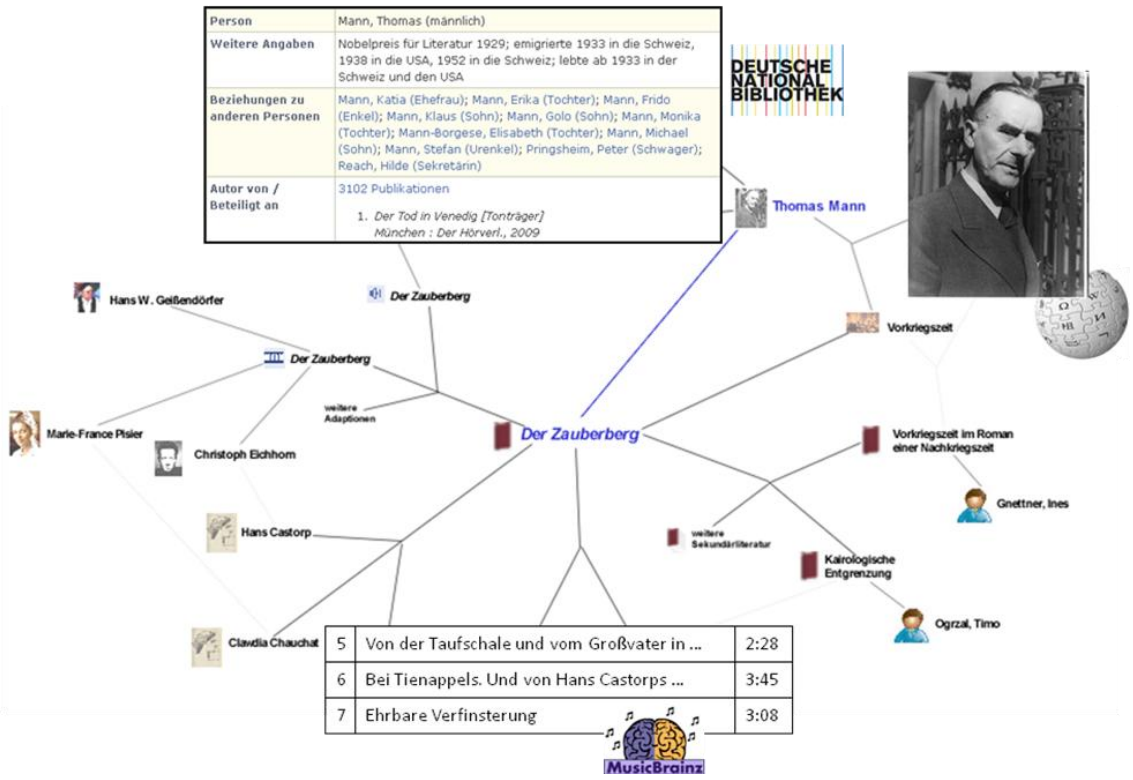

Figure 2: Integration of data from different sources regarding the work "Der Zauberberg" from German author Thomas Mann

Using URIs

According to Linked Open Data principles it is recommended to use de-referenceable and persistent URIs as identifiers. URIs allow us to tie together different sources of information regarding entities of interest (persons, places, organizations, etc.). As such, they are central to our efforts of aligning data sets and integrating information from heterogeneous data sources into the CONTENTUS knowledge base.

Disambiguating Persons

In order to improve search functionality it is important to link media, such as text documents, with other information, such as data from authority files. This is currently implemented for persons, corporate bodies, places, etc. One of the challenges here is to determine which person is actually meant, if more than one person with the same name exists. Also, persons' names are sometimes not at first glance distinguishable from other words that appear in a dictionary. In CONTENTUS, not only the authors of documents or other persons mentioned in metadata are linked to authority files but also persons mentioned in the extracted text of the documents themselves. This requires the use of automatic algorithms for extracting names of persons as well as for disambiguating them.

The approach used in CONTENTUS is that by Pilz and Paaß [Pilz and Paaß, 2009]. To disambiguate a person's name, a comparison is made between the context in which the name appears and a reference document where the true identity of the person is known. In our concrete example the text is compared to the Wikipedia entries of the persons in question. The person, whose Wikipedia

article is most similar to the relevant text of the document, is then identified as the person mentioned in the original document.

Instance and Schema Mapping

Creating mappings between heterogeneous data sets is one of the oldest problems in information science. We have to distinguish between two parts here: The mapping of instances or objects and the mapping of data structures. The problem of automatically mapping instances is equivalent to duplicate detection. It is common to use string distance metrics such as the well-known Levenshtein [Levenshtein, 1965] or Jaro-Winkler [Winkler, 1999] metrics and/or phonetic similarity measures such as Soundex [Russell, 1918] as a means to detect identical instances. In practice, modern matching algorithms use a combination of metrics; see e.g. [Johnston and Kushmerick, 2004].

The problem of automatically matching schemas has been addressed in research and literature since databases have been introduced [Melnik et al., 2002] and has been revisited for XML schema matching, and recently also for ontology matching [Shvaiko et al., 2009; Heß, 2006]. Algorithms typically use both structured and lexical similarities and some also exploit when instances are known to be represented in both schemas.

Mapping Locations

We could rely on intellectually generated mappings between schemas or instances in some cases (see above), because they were – in case of the Wikipedia mapping – collaboratively created or – in case of the MusicBrainz mapping – very easy to create. However, for larger mapping tasks it is crucial to have reasonably accurate automatic mapping algorithms.

For the future development of the semantic search in CONTENTUS, we are planning to include novel graphical controls (see next section). In order to be able to display geographical information about locations that are for example found in the full text of media documents or that is connected through metadata. The goal is to include mappings to a geographical database such as GeoNames.

We are planning to use a combination of heuristics and similarity metrics to achieve this task. In the authority file that serves as a basis for the mapping, information about the country and (if existent) federal state or province in which a city is located is usually available. This information can be exploited for disambiguation, if a city's name is not unique (e.g., Paris in Texas, USA vs. Paris in France). Similar approaches have been used successfully to disambiguate other authority file information in the context of the German National Library's first linked open data project [Hannemann et al., 2010].

Search and Navigation:

The search engine developed within the CONTENTUS project combines two information sources: a traditional full-text index of OCR and audio transcripts, as well as semantic information held within an ontology. The underlying media of this "Semantic Multi-Media Search" (SMMS) comprises audiovisual and audio material, scanned print media and born-digital text documents.

The CONTENTUS search aims to grant access to all these information sources through a unified interface. Consequently, the main design challenges for the user interface (UI) were:

1.) The transparent combination of different data sources

2.) The seamless integration of multi-media data and associated metadata

3.) A user-friendly access to semantic search features

Using semantic information for searches imposes three main advantages over traditional search engines:

1.) Users can browse information in an explorative way by following semantic links between media assets and information sources

2.) Individuals and keywords can be disambiguated by their meaning

3.) Relationships between search results become apparent

The CONTENTUS approach to UI design

In order for end users to fully utilize the integrated metadata sources and the different media, it is essential to provide a search interface that is both intuitive and provides novel semantic search capabilities. The CONTENTUS project has produced two working, web-based prototypical iterations of its Semantic Multi-Media Search. We have gathered user feedback on the usability of the two prototypes since 2008 through demonstrations at trade fairs like the Frankfurt Book Fair 2008 and 2009 and the International Broadcasting Conference (IBC) fair in Amsterdam 2009.

Before the design phase for the third demonstrator (currently under development), we held two paper prototyping (see e.g. [Maaß, 2008]) sessions at the Institut für Rundfunktechnik (Broadcast Technology Institute) in Munich in 2010 to reconfirm the previous (positive) feedback from trade fair visitors, this time with users from an archival and library background.

We decided against confronting the test user group with our existing prototypes of our web based search engine. Instead, in a first pass we presented the participants a set of predefined search tasks and asked them for their ideas on how a UI could most easily solve these. In the second pass we then showed our test group a set of control elements to get feedback on how they would be understood and which kind of interaction the test users expected.

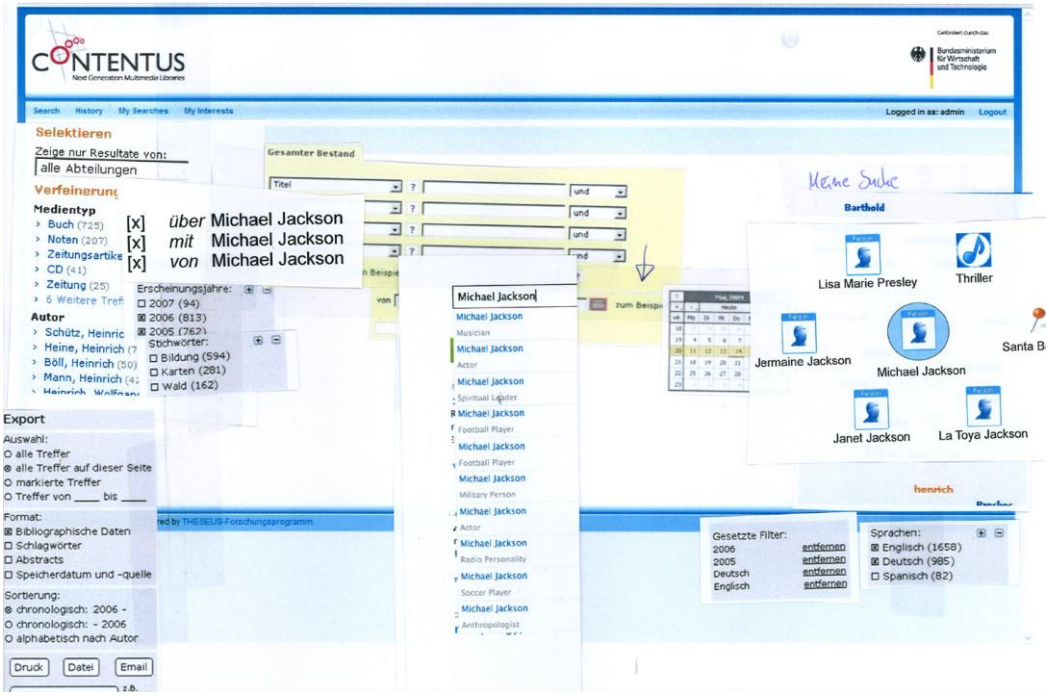

Figure 2: A CONTENTUS paper prototyping example. Users could freely select and arrange predefined controls in a second pass of our prototyping sessions in 2010

Our user test results show that the average user indeed prefers a classical "Google-approach" for his search entry: a search slot and a textual list representation of search results. However, we suspect that one reason for this preference is that many users are not familiar with more innovative or unusual user interface elements and are thus reluctant to use them.

Since we consider the explorative possibilities as one of the strongest advantages of semantically assisted search interfaces, we consequently had to choose an interface that encourages users to utilize the "added semantic value" and at the same time does not overstrain and disorientate them with unfamiliar interaction possibilities. Most users preferred a faceted search interface for narrowing their initial result lists with disambiguated keywords over a dedicated query language and over disambiguation as a type-ahead search feature.

A Sample Use Case

The current user interface allows for the following sample interaction:

A user is looking for books written by a journalist called *Michael Jackson*. Accordingly he enters the term "Michael Jackson" into the search application. However, *Michael Jackson* also is the name of a very popular singer and musician. Similarly to a conventional search engine, the SMMS first returns plain text matches against the search index, since there is no way for the application of guessing which of the two individuals the user might have meant.

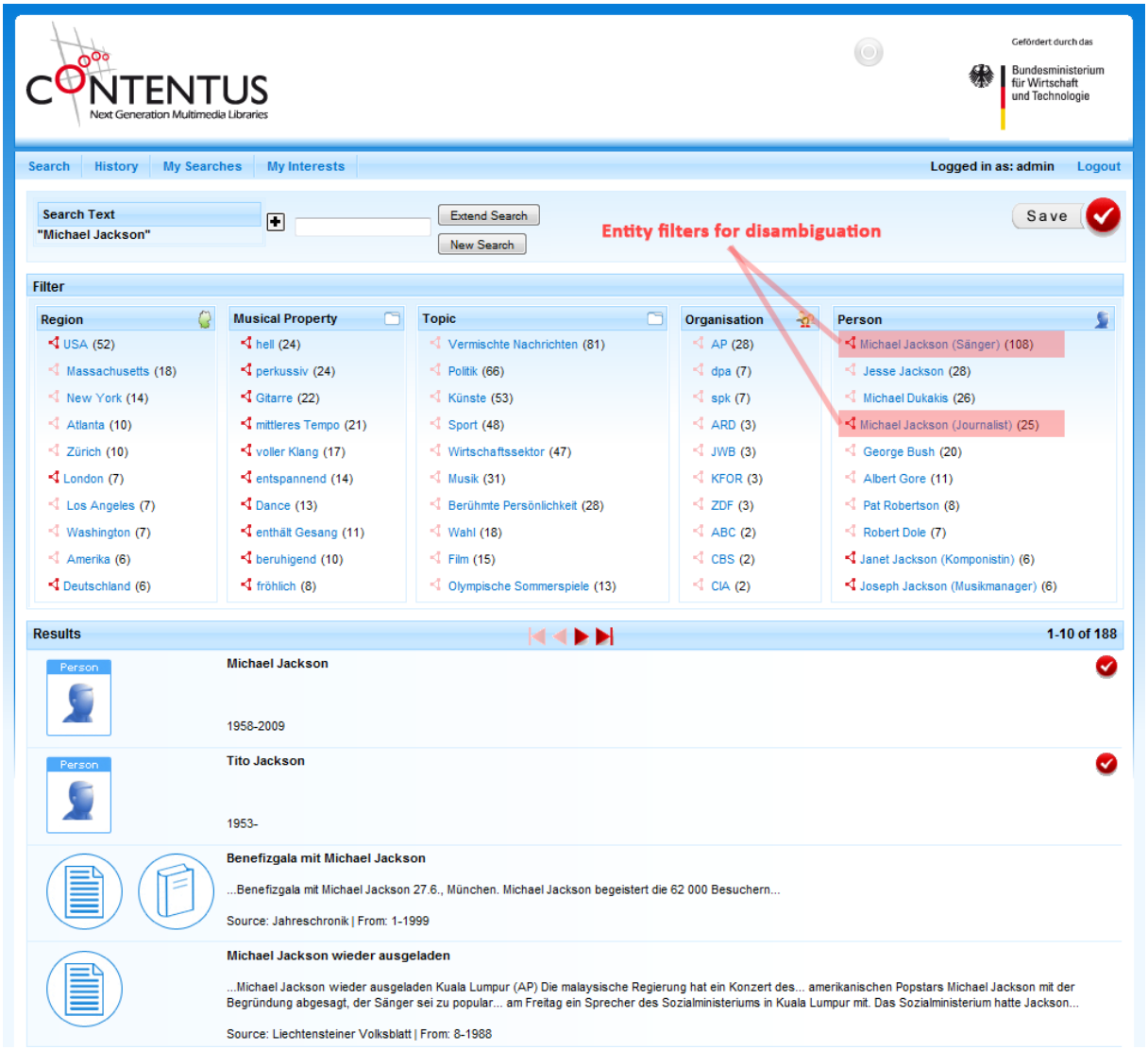

Figure 4: The CONTENTUS result set for the search term "Michael Jackson" prior to filtering and disambiguation of persons

Due to the similarity of names the application returns a result list containing a mix of wanted and unwanted search results across all media types. Most of these are related to the artist Michael Jackson (and not to the journalist) and are thus not of interest to the user.

In addition to the media result list the search interface also provides several dynamic filter lists (facets), which are automatically generated from the search result sets. These comprise the most relevant concepts and named entities within the set of results and are compiled from both intellectually prepared catalogue meta-data and information recognized by the automatic content analysis modules of CONTENTUS.

The relevance of facets is not only based on their frequency within the result set, but the most effective reduction of the result set size - facets that occur within all (or most) of the results are omitted as they offer no substantial filtering possibilities.

The filter facets are grouped into a fixed set of classes:

- Musical concepts
- Locations
- Topics
- Organizations
- Persons

The user can now use these filter facets to narrow his search - this internally adds the corresponding term or disambiguated entity with a logical "and" to the original search term. Each filter facet features a coloured icon that represents data provenance (see Figure 4) - this allows for a distinction between disambiguated persons contained in the libraries' authority files and generic named entities found by statistical analyses of the text material.

As the majority of the search results in our example have a connection to the *artist* Michael Jackson, many of the topics and entities also have a relation to music. But we also see topics like "Beer" and "Whisky" which are common for the works of the *journalist* Michael Jackson. The filter list of persons shows both Michael Jacksons within our person database, as well as related persons like the siblings of the pop singer. One click on the journalist's facet reduces the result set to all media relevant to the user – no longer are search results related to the singer shown.

Interestingly the filter facets for the search term *Michael Jackson* also show topics and organizations (like KFOR, the Kosovo Force) that have nothing to do with the most obvious two persons, the journalist and the artist. While some users were confused by this and discarded these entries as non-relevant noise, many explored further and found out about a third Michael Jackson, a general for the NATO forces - a result not anticipated but nevertheless useful.

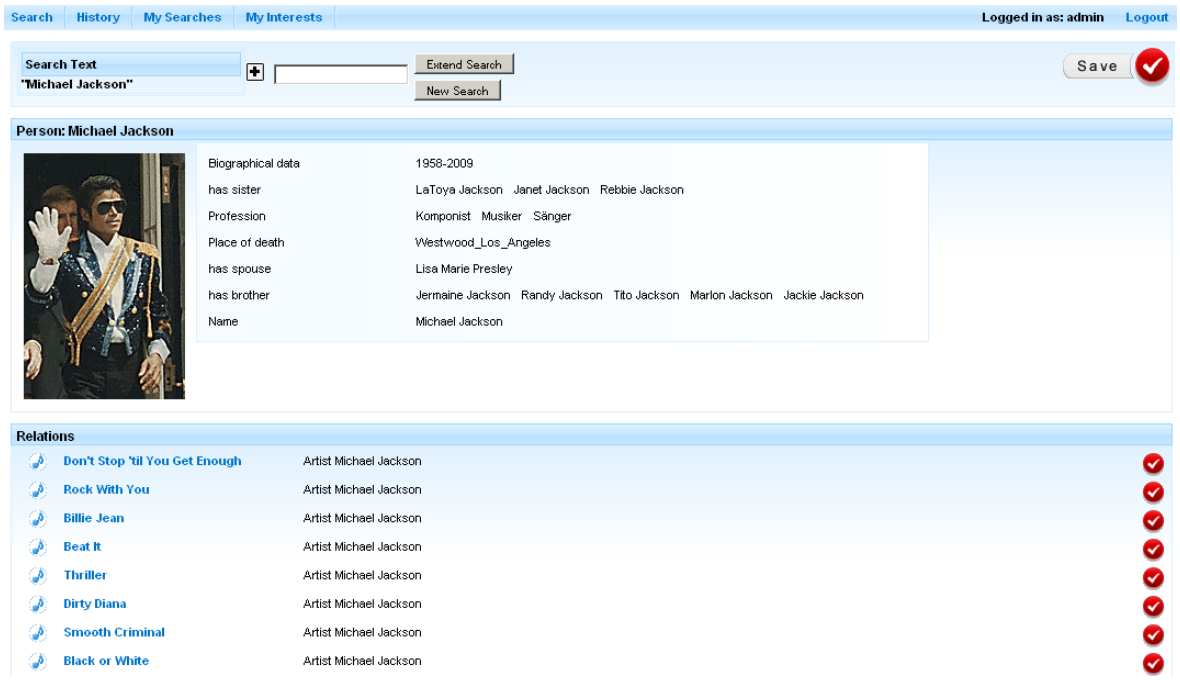

Figure 5: the entity page for the singer Michael Jackson

Every disambiguated person also has an *entity page* that can be invoked by clicking on the persons' entry in the result list. Here, users are provided with all of the assigned semantic information like relatives of persons, their works as creators, dates and locations of birth and so on. The entity pages are further enriched with images, bibliographical information and text from Wikipedia. Figure 5 shows the entity page for the *singer* Michael Jackson.

From the entity page users can trigger a new search by clicking on any of the linked entities, topics, locations and so forth, thus enabling a true semantic browsing experience that seamlessly blends with the relatively conventional look and feel of the interface.

Novel Search Interface Elements

Our user tests showed that *interaction with graphical representations of semantic graphs* was in most cases not fully understood or deemed impractical. While users realized the meaning behind a graphical view for relations between persons, they did not grasp the idea of interacting with the visualizations.

An *interactive timeline control* has proven as widely acceptable to most of our paper prototyping test group. Narrowing the result set by marking a time frame on the control seemed to be an intuitive way of searching data and is universally applicable on most knowledge domains.

Filter facets with a hierarchy (of hypernyms or hyponyms like plant -> flower -> rose) were also proposed by some users, so we will test their usability in a future prototype as we already have parts of the authority file subject headings in a hierarchical order.

An interactive map, a graphical visualization of locations within search results has been positively evaluated by many of the test persons. Users will be able to confine their result set by marking a geographical area on the map.

Interface Test Results

Our user tests have shown that:

- Semantic search features greatly help to reduce the effort of locating relevant matches in large multi-media archives.
- It is crucial for users to understand *how and why* any search hits made it into the result set. Otherwise the semantic layer can be confusing, especially if we include farther connections like relatives of a matching person into the result set.
- Users are reluctant to use novel visualizations as a sole search entry. They expect a traditional search slot, but accept interactive visualizations as a search refining tool.
- An explorative search is used mostly as a secondary step after entering one or more keywords. None of the users proposed pure exploration as their preferred method for answering our question set, but all were, on the other hand, reacting favourably to the browsing facilities offered by our prototypes, especially the entity pages.

Planned Additions to the UI

During the time frame of the project we will add at least the following functionalities to our interface:

- *Roles for entities in the filter facets:* our test users equivocally stressed the necessity of being able to differentiate between filtering for e.g. media written by person A and for media having person A as the subject.
- *Improved explanation of results and facets:* the result list should reflect why any element has made it into the result set, especially for results that only have an indirect semantic relation to the search term.
- *Interactive timeline visualization control:* users should be able to narrow their result set by marking a time range on the visualization so that only results within that time frame are shown.
- *Interactive map:* users should be able to restrict their results to a freely selectable area on the map.

Conclusion

Not only in Germany, but also on the level of the European Union, huge efforts are being undertaken to ensure the availability of digitized cultural heritage material, whether it is for long term archival or for the creation of digital libraries like *Europeana* or the *Deutsche Digitale Bibliothek*. Therefore, more and more libraries and archives are faced with the challenge of integrating assets from various digitization projects, local metadata and external data sets. Unfortunately, there is still a lack of tools that facilitate an uncomplicated yet comprehensive supply of the assets and metadata for library systems and catalogues.

On the other hand, there is a strong demand by the users of libraries and archives to access digital media in a context organized in an equitable manner across all media types. Audio and video content, according to contemporary habits of media use, is expected to be directly integrated in information objects and therefore should also be part of search engines in libraries and archives. This demand often leads to the need of integrating 3rd-party tools and data sources. CONTENTUS is in the process of developing technologies and concepts that will address these challenges and significantly simplify the production, the provision and the usage of digital media collections.

With the already realized two web-based demonstration systems we have shown that this kind of aggregation and presentation of media assets and metadata is feasible and – more importantly – also valuable for users of libraries and other archives. Knowledge access and discovery through semantically assisted searches will doubtlessly grow in importance and we believe that the outcomes of CONTENTUS can be important building blocks for next generation digital library systems.

Lessons learned

A few guiding principles could be established that have proven to be useful for achieving the project's goal. The lessons we have learned so far:

- *Modular design is essential.* As not all libraries and archives are alike in terms of their needs, some might not require digitization techniques, and others may already offer search interfaces and simply require technologies to generate and combine metadata for their assets. Consequently, the design of the CONTENTUS solutions is intentionally modular. For each of the different procession steps (see Introduction), independent solutions exist that can be used by interested institutions individually or together.
- *Open standards and interfaces are important.* In order to facilitate the aforementioned integration of CONTENTUS technologies, we focused on open standards, interfaces and data formats. For elements of the semantic multimedia search in CONTENTUS, for example, we employ a service-oriented architecture (SOA). The interaction of the different modules via *Web Services* takes the need of a modern library infrastructure into account and offers most flexibility for the integration of different data sources may they be in-house or provided by a 3rd-party service provider. For integrating external information sources, typical formats of linked data collections (XML/RDF) make it easy to utilize such metadata.
- *URIs are valuable for semantically linking assets, concepts and information sources.* See „Using URIs“, above.
- *Users prefer simple, well-structured, yet powerful interfaces.* This is especially true when it comes to novel functionalities as provided by semantic searches. Graphical representations, e.g., of concepts and relations, have to be kept intuitive and easy to use, even across knowledge domains. Extensive personal configuration options for the user interface are strongly demanded –particularly by professional users.

Future Work and Vision

Currently, the project is upholding its yearly iterative release cycle and thus rapidly approaching the third web-based demonstration system, which will be presented to the professional audience at the exhibition of the International Broadcaster Conference (IBC) in September 2010. The new demonstrator comprises a reworked user interface as well as an extended semantic facet engine and better handling of multimedia content. By the end of the year the new CONTENTUS SMMS demonstrator will also be presented at the stationary demo centre of the THESEUS research program in Berlin and at selected events of the library and archive community.

Subsequent development efforts will concentrate on extending semantic capabilities by integrating a *semantic media viewer* to allow for a better interaction with named entities recognized by the system. Another big challenge will be the extension of personalization and community features, which will form a novel way of *cooperative information exploitation.* It will make it possible for users to comprehensively interact with the information assets whether for personal use or in cooperation with co-workers and user groups. Last but not least we will integrate more valuable data sources from the linked open data cloud.

One vision of CONTENTUS is to demonstrate its concepts for metadata integration and semantically assisted search within the context of a real historic collection. For this purpose large parts of the archive of the *Musikinformationszentrum* (MIZ) of the former German Democratic Republic (GDR) have been digitized. The various media assets of this secluded collection will be integrated into the final CONTENTUS demonstrator which will be available in early 2012. We believe that this content is very suitable for showing the advantages of our system in a specific knowledge domain and that it will lead to new insights about the musical life in the former GDR.

References

Bossert, Klaus and Nicholas Flores-Herr and Jan Hannemann. *CONTENTUS: Technologien für digitale Bibliotheken der nächsten Generation*. Dialog mit Bibliotheken, Bd. 21, p. 14-20. ISSN 0936-1138. German National Library, 2009

Hannemann, Jan and Jürgen Kett. *Linked Data for Libraries.* In: Proceedings of World Library and Information Congress: 76th IFLA General Conference and Assembly (IFLA 2010), Gothenburg, Sweden

Heß, Andreas, 2006. An *Iterative Algorithm for Ontology Mapping Capable of Using Training Data.* In: Proceedings of the 3rd European Semantic Web Conference (ESWC 2006), Budva, Montenegro

Johnston, Eddie and Nicholas Kushmerick, 2008. *Web Service aggregation with string distance ensembles and active probe selection.* Information Fusion 9(4): 481-500 (2008)

Levenshtein, Vladimir I., 1965. *Binary codes capable of correcting deletions, insertions, and reversals.* In: Doklady Akademii Nauk SSSR. 163, Nr. 4, 1965, S. 845–848 (In Russian. English translation in: Soviet Physics Doklady, 10(8) S. 707–710, 1966)

Maaß, Christian and Elica Savova, 2008. *Paper Prototyping in der Softwareentwicklung.* In: Das Wirtschaftsstudium, 11/2008 (In German)

Melnik, Sergey and Hector Garcia-Molina and Erhard Rahm, 2002. *Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching,* In: Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose CA, USA

Pilz, Anja and Gerhard Paaß, 2009. *Named Entity Resolution Using Automatically Extracted Semantic Information.* In: Proceedings of workshop Lernen, Wissen, Adaptivität (LWA 2009), Darmstadt, Germany

Russell, Robert C., 1918. United States Patent 1261167, application filed Oct. 25, 1917, patented Apr. 2, 1918.

Shvaiko, Pavel and Jérôme Euzenat and Fausto Giunchiglia and Heiner Stuckenschmidt and Natasha Noy and Arnon Rosenthal (Editors), 2009. *Ontology Matching (OM-2009), Papers from the ISWC Workshop*. October 2009.

Winkler, W. E., 1999. *The state of record linkage and current research problems*. Statistics of Income Division, Internal Revenue Service Publication R99/04, 1999.