Multimed Tools Appl
DOI 10.1007/s11042-011-0971-2

CONTENTUS—technologies for next generation multimedia libraries

Automatic multimedia processing for semantic search

Jan Nandzik · Berenike Litz · Nicolas Flores-Herr · Aenne Löhden · Iuliu Konya · Doris Baum · André Bergholz · Dirk Schönfuß · Christian Fey · Johannes Osterhoff · Jörg Waitelonis · Harald Sack · Ralf Köhler · Patrick Ndjiki-Nya

Abstract An ever-growing amount of digitized content urges libraries and archives to integrate new media types from a large number of origins such as publishers, record labels and film archives, into their existing collections. This is a challenging task, since the multimedia content itself as well as the associated metadata is inherently heterogeneous—the different sources lead to different data structures, data quality and trustworthiness. This paper presents the CONTENTUS approach

J. Nandzik (✉) · N. Flores-Herr
Acosta Consult GmbH, Zeißelstraße 15 HH, 60318 Frankfurt am Main, Germany
e-mail: jn@acosta-consult.de

N. Flores-Herr
e-mail: nf@acosta-consult.de

B. Litz · A. Löhden
Deutsche Nationalbibliothek, Informationstechnik, Adickesallee 1,
60322 Frankfurt am Main, Germany

B. Litz
e-mail: b.litz@dnb.de

A. Löhden
e-mail: a.loehden@dnb.de

I. Konya · D. Baum · A. Bergholz
Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

I. Konya
e-mail: iuliu.vasile.konya@iais.fraunhofer.de

D. Baum
e-mail: doris.baum@iais.fraunhofer.de

A. Bergholz
e-mail: andre.bergholz@iais.fraunhofer.de

D. Schönfuß
mufin GmbH, Büro Dresden, August-Bebel-Straße 36, 01219 Dresden, Germany
e-mail: dschoenfuss@mufin.com

Published online: 16 February 2012

towards an automated media processing chain for cultural heritage organizations and content holders. Our workflow allows for unattended processing from media ingest to availability thorough our search and retrieval interface. We aim to provide a set of tools for the processing of digitized print media, audio/visual, speech and musical recordings. Media specific functionalities include quality control for digitization of still image and audio/visual media and restoration of the most common quality issues encountered with these media. Furthermore, the CONTENTUS tools include modules for content analysis like segmentation of printed, audio and audio/visual media, optical character recognition (OCR), speech-to-text transcription, speaker recognition and the extraction of musical features from audio recordings, all aimed at a textual representation of information inherent within the media assets. Once the information is extracted and transcribed in textual form, media independent processing modules offer extraction and disambiguation of named entities and text classification. All CONTENTUS modules are designed to be flexibly recombined within a scalable workflow environment using cloud computing techniques. In the next step analyzed media assets can be retrieved and consumed through a search interface using all available metadata. The search engine combines Semantic Web technologies for representing relations between the media and entities such as persons, locations and organizations with a full-text approach for searching within transcribed information gathered through the preceding processing steps. The CONTENTUS unified search interface integrates text, images, audio and audio/visual content. Queries can be narrowed and expanded in an exploratory manner, search results can be refined by disambiguating entities and topics. Further, semantic relationships become not only apparent, but can also be navigated.

C. Fey
Institut für Rundfunktechnik GmbH, Production Systems TV,
Floriansmuehlstraße 60, 80939 München, Germany
e-mail: fey@irt.de

J. Osterhoff · J. Waitelonis · H. Sack
Hasso-Plattner-Institut für Softwaresystemtechnik GmbH,
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany

J. Osterhoff
e-mail: johannes.osterhoff@hpi.uni-potsdam.de

J. Waitelonis
e-mail: joerg.waitelonis@hpi.uni-potsdam.de

H. Sack
e-mail: harald.sack@hpi.uni-potsdam.de

R. Köhler
Technicolor - Corporate Research Division, Hanover Image Processing Lab, Deutsche
Thomson OHG, Karl-Wiechert-Allee 74, 30625 Hannover, Germany
e-mail: ralf.koehler@technicolor.com

P. Ndjiki-Nya
Fraunhofer-Institut für Nachrichtentechnik Heinrich-Hertz-Institut,
Einsteinufer 37, 10587 Berlin, Germany
e-mail: patrick.ndjiki-nya@hhi.fraunhofer.de

Keywords Automatic information extraction · Semantic Web · Workflow automation · Multimedia search and retrieval · Digital libraries · Content analysis

1 Introduction

Cultural institutions hold a vast amount of multimedia content stored on carriers such as books, images, tapes and films. Most of these cultural heritage organizations face the challenge of analog data deterioration through storage conditions and ongoing use of the media that can only be countervailed with a digitization of the data. Additionally, the institutions need to organize and manage their ever-growing collections and have to provide user access to the knowledge within.

While digitization today can be automated to a certain degree by employing mass digitization techniques, the necessary quality assessment still is a time-consuming manual task. Thus, quality assessment for digitization is performed only on random samples because of resource constraints. Errors in the digitized material not only occur because of problems with the digitization process but are also caused by defects within the analog material itself.

Another issue prohibiting user-friendly access is the manual knowledge acquisition bottleneck that leads to insufficient descriptive metadata for the media. Most digital multimedia data are not directly searchable as opposed to most websites or text files, as their meaning and semantics are not available as text. Existing metadata are often sparse, yet users begin to expect even more refined search possibilities than a full-text indexing can offer due to their experience with born-digital media created on the World Wide Web.

However, to date the majority of multimedia assets are annotated and cataloged manually by information experts, which is a complex, cost-intensive and time-consuming process. A lack of human resources for annotation leads to a multitude of assets that are either not thoroughly described, or only indexed fragmentarily.

Therefore, the transition of analog media to the digital domain cannot stop at a mere digitization step. Novel search services utilizing the results of an automated semantic media analysis will be the technological foundation for users to access digital collections [38].

Since the beginning of digital media processing in the 1980s there has been a vast variety of research dealing with the automated extraction of information from multimedia data. The first works adressed limited datasets while recent research focuses on more robust mass processing, often with data sets from social media websites such as Flickr or Youtube [39, 85, 96]. Since the beginning of the last decade, larger intellectually created test sets as, e.g. TRECVID and MAMMIE [7, 67, 77] have become increasingly popular and nowadays form a well comparable benchmark throughout the research community.

At the same time, Tim Berners Lee coined the term *Semantic Web* [13], which describes an open, machine-understandable web of data. In his vision, knowledge is linked and combined in order to create a greater benefit than can be gained by the sum of its parts.

Archives and other content holders only gradually begin to realize the added value of integrating external data sources and user generated knowledge. Using external datasets, users can for example search for media assets on the country of *France*, even if their origin has not been explicitly annotated. One mention of "*Toulouse*" or "*Avignon*" within the full text is sufficient, once these locations are linked with their corresponding Geonames entries. But since there can be different locations (or entities in general) relating to the same label, the linking algorithm has to include the entities' context for disambiguation. This context is often sparse, and thus the disambiguation can only claim a sub-par reliability. But not only the automatic linking of existing content to external sources poses problems, also their range of trustworthiness is very diverse. Thus, any serious system has to preserve information about provenance and reliability of external sources and has to take them into account for its output interfaces.

Within CONTENTUS [20], we thus identified that the following challenges need to be addressed in order to implement a successful solution for digital cultural heritage archives:

- The deterioration of analog media.
- The deficient quality of existing data and the loss of quality through digitization.
- The knowledge acquisition bottleneck.
- The consolidation of heterogeneous metadata.
- The search, discovery and navigation of relevant content.

This paper focuses on the technical developments carried out in the context of the CONTENTUS project. We develop a modular and automated end-to-end processing chain for digitized audiovisual, audio, text and still image data that feeds into a unified semantic multimedia search interface [63]. Some of the algorithms or solutions used in CONTENTUS are not entirely new, instead, we strived for minimizing necessary user interaction by coupling all of our processing steps together and the broadest possible applicability for heterogeneous data. Obviously this meant some trade-off between maximized performance of specialized algorithms and necessary robustness with regards to the different data sources.

The CONTENTUS engine should allow unattended processing of all source data types from their ingest to availability through the search interface. The high degree of automation also allows user generated content to be seamlessly integrated into the media corpus, while keeping all information editable by its users.

The process includes modules for quality control in digitization, restoration of damaged media or errors due to imperfect digitization, content analysis and external metadata integration and incorporates these into a scalable and flexible workflow engine. Thus, the CONTENTUS project merges the techniques of classical information retrieval and the Semantic Web.

The processing modules within CONTENTUS are optimized specifically on handling digitized archive media—born-digital resources, such as HTML websites or digital audio/visual and image data were not in focus of the project. Many analysis modules are tuned to work with information in German language such as speech recognition and Named Entity Recognition. While they might be applied to other languages as well, that would at least require adapted training material. Also, we did not work on three-dimensional representations of media that are slowly gaining importance in museums and other collections of three-dimensional artifacts.

This paper is structured as follows: Section 2 describes the media processing chain and the overall architecture of the CONTENTUS system. Section 3 puts the CONTENTUS approach into perspective and discusses semantic search in general, search in digital libraries and in multimedia collections. Section 4 gives an overview of the *media specific* processing steps quality control and content analysis. Section 5 describes *media independent* processing steps such as Named Entity Recognition, disambiguation, indexing and data linking. Section 6 presents our semantic multimedia search interface. Finally, Section 7 contains our conclusion and outlook.

2 The CONTENTUS project

The CONTENTUS project, implemented as a *Public Private Partnership*, is one of five use cases of the THESEUS [82] research program funded by the German Ministry of Economics and Technology (BMWi). Wherever technologies could benefit other use cases, additional solutions are developed by centralized development units pooled in the *Core Technology Cluster (CTC)* of the THESEUS program. The CONTENTUS project has a lifespan from 2007 until 2012.

To incorporate end-user and expert communities and improve semantic knowledge networks CONTENTUS closely collaborates with two other THESEUS projects: the collaborative knowledge engine for end-users ALEXANDRIA [2] and the digital film archive Mediaglobe [55].

2.1 The processing chain

The process from the ingest of media to their accessibility through the search interface can be divided into a series of six processing steps. These individual steps form a processing chain, as shown in Fig. 1.

1. **Digitization:** Mass digitization is the first step to counteract deterioration and to prepare online access of the media. To help organizations with their digitization efforts two digitization guidelines are compiled—one for printed and one for audio/visual media (in cooperation with the Mediaglobe project). These currently undergo a review process and will be published shortly.
2. **Quality control:** This step incorporates automated quality analysis and quality optimization. Automated quality checks are necessary to keep up with the speed of current digitization machines like book scanning robots and film reel

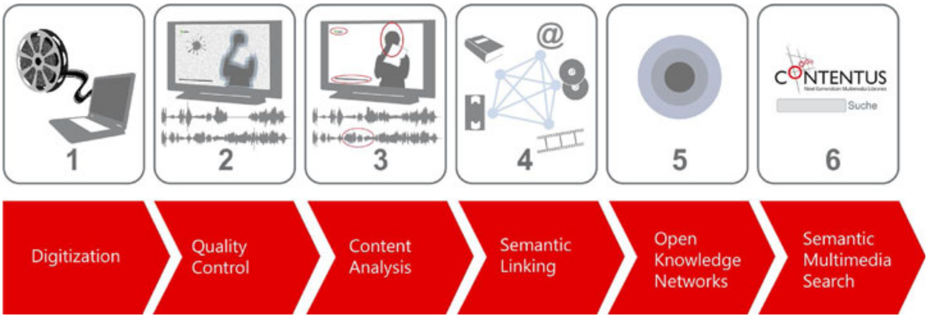


Fig. 1 The CONTENTUS processing chain

digitizers. Where possible, digitization errors (like oblique scans, too much background) and media faults (like video dropouts, scratches in film material) are corrected. The goal of this restoration step is to improve the media quality both for human consumption and automatic content analysis. These steps are described in Section 4.

3. **Content analysis:** CONTENTUS services that automatically analyze still image, text, music and audiovisual assets play an important role in the generation of search-relevant information. These modules include e.g. segmentation of videos, music and scanned pages, OCR transcription, speech and speaker recognition and music and text classification. The content analysis steps belong to the category *media specific* as described in Section 4.
4. **Semantic linking:** Entities such as places, persons, events etc. are extracted from textual transcripts, disambiguated and linked with catalog, user generated and Internet resources to augment the available information. For example, a person as a 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.
5. **Open knowledge networks:** CONTENTUS allows users to upload media and edit existing and automatically generated knowledge to extend the knowledge base. The data model used allows for preserving and exploiting both provenance information such as user role or Internet source and algorithmic confidence ratings.
6. **Semantic multimedia search:** CONTENTUS offers end users an innovative faceted multimedia search functionality by combining searches for and within texts, images, audio and audiovisual content in a unified semantic user interface.

2.2 CONTENTUS system architecture and data corpus

The CONTENTUS system uses mostly web services for communication between the different processing modules as shown in the architecture overview in Fig. 2. This makes it easy to replace single modules or extend the functionality by using external tools as the interface uses standardized methods.

Initially we considered to have a single web service based processing engine for all analysis and restoration steps. This was not favored as we would have suffered from insufficient data bandwidth to transfer high resolution audio/visual data over the web. Thus, media processing within CONTENTUS is carried out on two separate processing clusters—one for restoration of audio/visual media (the *A/V platform*) and one for all other media types, the CONTENTUS *service platform*. After restoration of the audio/visual media, presentation copies are transferred to the *service platform*, where all following analysis steps are executed. All other media types are solely processed on the *service platform*.

The CONTENTUS *service platform* utilizes cloud computing technologies to ensure scalability and flexibility with regards to managing huge media collections. It is able to dynamically allocate and request resources such as processing nodes. On the technical layer, we extended a Web Service-Business Process Execution Language (WS-BPEL)-based [94] orchestration to allow the orchestration of stateful Web Services Resource Framework(WSRF)-based [95] cloud services. Each media type has its specific processing chain that handles the whole process from ingest of data and corresponding metadata files to updating the repository and search index.

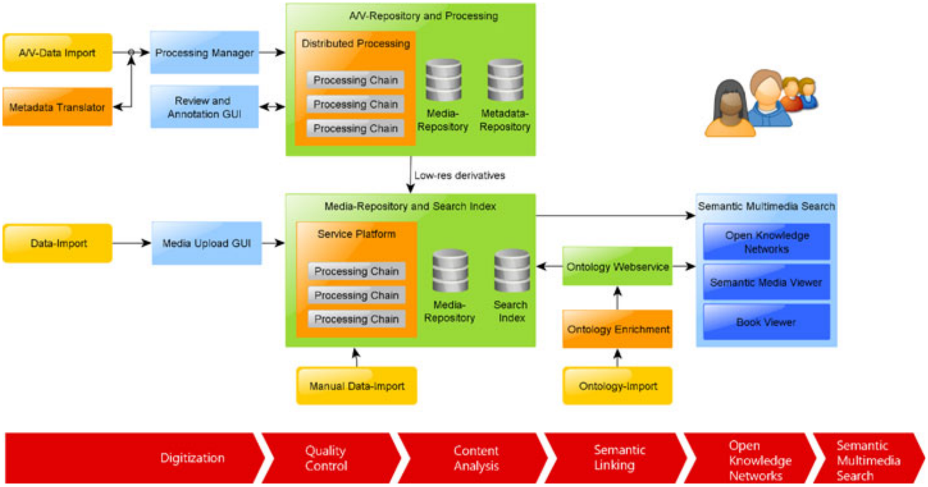


Fig. 2 The CONTENTUS system architecture

The CONTENTUS media repository is based on an Apache Jackrabbit implementation, while the index (described in greater detail in Section 6.3) is an out-of-the-box installation of the Apache Solr Server.

In our initial test-runs its current installation on a 16-node cluster the service platform is able to process about 15.000 to 20.000 scanned pages per day if all functionalities (quality control, segmentation, OCR, Named Entity Recognition, data linking and updating of the index) are enabled. On a similar hardware the audio/visual processing time still is significantly slower than the real-time if all restoration and analysis steps are performed. In both cases, the biggest part of the processing time is taken up by transfers of the digitized data, which shows that more effort has got to be spent on optimizing data flow and thus minimizing the transmission overhead between the cluster nodes. After that, further benchmarking has got to take place.

2.3 Test set

In order to show the possibilities of the CONTENTUS semantic analysis and linking methods we were in need of a sufficient set of real-world media for analysis. The media set had to incorporate a significant amount of potential semantic relations and had to originate from a relatively closed domain.

Within the collection of the German National Library we identified the collection of the Music Information Center of the former German Democratic Republic (MIZ) as a suitable candidate for digitization and processing. It consists of historical printed books, press clippings, program brochures, audio tapes and both shellac and vinyl records. Since audio/visual material is not part of the collection, we added historical film material and selected erroneous A/V archive material from the film archive DEFA Spektrum and the public Bavarian Broadcast (Bayerischer Rundfunk) for restoration purposes and German news broadcasts for speech analysis and segmentation.

Within the CONTENTUS project we used the test set as shown in Table 1.

Table 1 Summary of CONTENTUS data set presenting the type of media, its amount and storage size

Media	Amount	Storage
Books	1.600/385.471 pages	7.42 TB
Historical press clippings	154.240 pages	6.38 TB
Current newspapers	30.000 pages	0.90 TB
Program brochures	208.775 pages	3.51 TB
Photo	6.680	0.42 TB
Sheet music	64.112 pages	0.12 TB
Audio tapes	9.245 tapes	1.15 TB
Shellac records	4.774	0.18 TB
Vinyl records	9.788	1.54 TB
"Heute" news	373, ca. 200 h	20 GB (Web quality)
Other broadcasts	160, ca. 130 h	36 GB (Web quality)
Historical film material	14, ca. 10 h	102 GB

3 Related work

Many systems and research projects are striving towards using Semantic Web technology for searching multimedia archives. Given the substantial amount of literature and projects dealing with search systems as well as individual underlying technologies [6, 22, 23, 48, 79, 80, 93, 97] we will first try to narrow down our understanding of semantic multimedia search and then compare CONTENTUS to a number of related projects.

The term *semantic search engine* is used in various contexts and often denotes rather different technologies. Thus, research projects, products or internet search services mentioned in this section do not exactly match the understanding of semantic multimedia search (SMMS) within CONTENTUS.

In our definition the term *semantic multimedia search engine* is not equivalent to basic querying languages for RDF such as SPARQL and their implementations. Instead, a semantic multimedia search engine consists of a user interface and backend that allow for finding and browsing content and make use of the meaning (semantics) of extracted and integrated metadata. We therefore propose that a combination of criteria should be used to compare related work:

1. **Media**—Audio, audio/visual and still image media as well as textual media from different types and sources (born digital and digitized media),
2. **Semantic features**—The use of URIs, a formally structured knowledge base, a formal description of resources, a formal description of properties and , relations between resources, the usage of a Semantic Web query language like SPARQL, reasoning and disambiguation,
3. **Metadata**—Data automatically extracted from the media, catalog data, user generated knowledge,
4. **Linked data**—The incorporation of external sources for extending the knowledge base,
5. **Result access**—Direct access to results or references to resources through the search interface, faceted filtering.

This section highlights the most relevant projects that focus on semantic search, multimedia search or both. A comparison of the search interface developed in CONTENTUS and other projects has already been published by Waitelonis et al. [89]. Other work that is related to single processing steps or algorithms within CONTENTUS and not semantic search as a whole is mentioned in the respective sections (cf. Sections 4 and 5). We have compiled a tabular overview of related research projects in Table 2.

A search engine project employing information extraction and ontology learning from multimedia data is BOEMIE (e.g. [66]). The objective of the project was the automatic extraction of metadata from multimedia content (audio, video, still images, text, and compound media like web pages or text overlay on video). The metadata were, on the one hand, used to enrich an existing domain ontology by populating it with instances (in case of BOEMIE, an athletics ontology was used), and, on the other hand, to extend and develop the ontology itself in a semi-automatic process, using multimedia instances which could not be classified as instances of one of the ontology's existing concepts. In CONTENTUS, the system's ontology is similarly populated. However, the CONTENTUS ontology itself is not automatically extended by new concepts, as in the projects context it was considered to be important to manually curate the system's knowledge base. Still, we do make use of new concepts delivered by the named entity recognition as search filters.

Ding and Sølvberg present an approach for semantic search in digital libraries [28]. The project is similar to CONTENTUS as it also finds solutions to cope with heterogeneous metadata records. Ontologies are utilized for storing metadata records and for broadening queries with related terms in order to create more relevant results. Disambiguation is advanced with a lexical database modeling terms with their meanings and relationships. The resulting system serves three purposes. It can be applied as a framework for re-processing metadata records into a semantic collection. Further, it can be used as a platform for searching over heterogeneous collections and for exploiting ontologies in information searching.

The open source software suite Greenstone [92] is created for building and distributing digital library collections. It helps to organize information and to publish it on the Internet in the form of a searchable, metadata-driven digital library. In an extended edition semantic digital library modules are available [37]. With the help of the semantic modules, a disambiguation of search terms is possible. This approach differs from CONTENTUS as disambiguation within CONTENTUS is part of the automated entity extraction (from texts or speech transcripts) using ground truth from Wikipedia. The most striking difference to CONTENTUS is that feature-based similarity search (e. g. for music) is not part of the Greenstone digital library system.

Guha et al. presented two semantic search systems that augment and improve traditional text search results by Google [35]. *Activity Based Search* includes a semantic search for domains, including musicians, athletes, actors, places and products. *W3C Semantic Search* provides semantic search for the homepage of the World Wide Web Consortium. Due to the lack of sufficient data on the Semantic Web at the time of publication, the requisite portions of data were modeled. The aim of both search systems is to apply semantics for a disambiguation of searched entities. This is also an important issue in CONTENTUS as the automatically obtained data includes ambiguous entities. Instead of an ontology, CONTENTUS utilizes information from Wikipedia for disambiguation. The advantage is that we can employ the extensive data at hand and do not need to create knowledge manually.

Table 2 Related work—a comparison of existing research projects with CONTENTUS

Name	Media	Semantic features	Metadata	Linked data
Projects				
Boemie	Image, video, audio, text	Ontology	Catalog data	–
CONTENTUS	Image, video, audio, text	Ontology, entity disambiguation	OCR, NER, speech transcripts, video and text classification, catalog data	Wikipedia, dpPedia, PND
Greenstone	Image, (video), audio, text	–	Catalog data	–
Informedia	Video, audio, images, text	Similarity, QbE, entity disambiguation	Speech transcripts, video and text classification, catalog data	–
MEDIAMILL	Video	Similarity, QbE	Content, context and style analysis	
MESH	Image, video, audio, text	Ontology, reasoning, summarization	OCR, NER, speech transcripts, video and text classification + catalog data	–
MULTIMATCH	Video, audio, images, text	Document and query translation, similarity, QbE	Speech transcripts, catalog data	WWW data (crawled)
PHAROS	Video	Similarity, QbE, dynamic facets	Video and audio analysis, user generated knowledge	–
Rushes	Video, audio, images, text	Reasoning	Video analysis, user generated knowledge	–
VERGE	Video	Similarity, QbE	Video classification, speech transcripts, catalog data	–
VIDI-Video	Video	Ontology	Speech transcripts, video classification + catalog data	–
VITALAS	Video, audio, images, text	–	Video and image classification, user generated knowledge + catalog data	–
YOVISTO	Video	–	Catalog data, user generated knowledge	–

The Informedia projects [42, 43] by the Carnegie Mellon University combine analysis of video images, audio and text for a broad number of search scenarios. The project Informedia II in particular focuses on summarization and visualization across multiple video documents and textual material using related keywords as well as in-depth geographic and temporal breakdown and mapping of search results. Similar to CONTENTUS extracted entities are disambiguated—however, their identifiers are not used for linking to entities of a metadata catalog or external knowledge sources such as DBpedia. Both Informedia projects do not incorporate parts of the Semantic Web stack.

The MediaMill [56] is a semantic video search engine by the Intelligent Systems Lab Amsterdam of the University of Amsterdam. The technologies incorporated in the search engine originate from various fields of research such as machine learning, computer vision, image and video processing, language technology and information visualization. One of the engines' primary goals are the development of automated visual concept detection and novel interfaces for video media retrieval. Although CONTENTUS also has a focus on user interface, its design is centered on a traditional result list combined with a faceted search functionality. In contrast to CONTENTUS, MediaMill does provide search possibilities for high-level visual concepts such as "Apple" or "Boat" in audiovisual content.

The project Mesh [57] develops technologies to extract, compare and combine content from multiple multimedia news sources, automatically create advanced personalized summaries on the basis of extracted semantic information and thereby facilitating search for multimedia content in the domain of news. Comparable technological approaches to those of CONTENTUS are automatic semantic content analysis for images, videos, audio and text as well as the use of a combined search approach using SPARQL queries for ontology-based retrieval in combination with a vector-space information retrieval model.

The research project MULTIMATCH [4] aims at creating a semantic search engine for cultural heritage work, using document content, metadata, context and occurrence of relevant terms and concepts in the content, extracting such information automatically from the content where possible. The project combines proprietary data from content providers with data publicly available, converting from several metadata formats (e.g. MPEG-7, MARC21) to RDF where necessary. This way of combining different metadata resources is similar to the one in CONTENTUS. A multimodal, multilingual interface provides search-term related results as portlets in a web portal, providing explorative access to similar and related content.

PHAROS [26] is a framework for an audio/visual content-based search engine. It is characterized by automated extraction of features of audio and video (e.g. speaker turns, shot detection) and personalization. Concerning these topics the focus of CONTENTUS and PHAROS overlap. The PHAROS framework has a modular concept: content analysis and metadata generation, search in metadata and feature-based similarity search. Apart from the limit of media types, the project gives some interesting impulses for the work in CONTENTUS. However, there is little overlap with one of CONTENTUS main focuses on integrating external knowledge and catalog data with automatically extracted features to use for semantic search.

The project RUSHES [74] aims at indexing, accessing and conveying raw, unedited audiovisual footage using semantic analysis. That includes automatic low-level metadata generation and content indexation as well as learning and reasoning technologies that are based on multimodal analysis of the raw footage in combination with semantic inference. In comparison to CONTENTUS the project focuses on topics within the field of audiovisual analysis research that are particular to the envisaged usage scenario: context reuse and access to raw footage. The project CONTENTUS in turn, processes and searches for finalized audiovisual material such as news, documentaries or movies. Interestingly, the RUSHES metadata model utilises Semantic Web technologies (ontology APIs and reasoners).

The video retrieval system VERGE [86] from the Informatics and Telematics Institute (Greece) combines visual similarity search—that is, a feature vector-based similarity analysis comparing query image and dataset images—textual keyword information related to shots derived from speech analysis and machine translation, as well as retrieval using high-level visual concepts (such as landscapes, animal, outdoor). The extraction of high-level visual concepts is based on extracted low-level features combined with training and classification. These above mentioned three main features of VERGE are not part of CONTENTUS search or media analysis except for the use of textual transcripts for retrieval. Next to CONTENTUS incorporating more media types, the main difference between VERGE and CONTENTUS related to media analysis and search lies in the entity extraction and identification steps prior to linking them to catalog metadata and external resources and the usage of Semantic Web stack technologies by the latter project. In turn, CONTENTUS has not implented search for high-level visual concepts.

The VIDI-Video [87] project based audiovisual search engine comprises technologies from machine learning, audio event detection, and video processing. Similar to CONTENTUS metadata, keyword annotations, audiovisual data, speech, and explicit knowledge (in the form of ontologies) from different sources are incorporated to improve video search. One notable difference is the usage of a video annotator to produce the ground truth for the project. This annotator enables the linking of frames to given elements of an ontology.

A notable project that does not use Semantic Web stack technology but makes use of machine learning algorithms is VITALAS [88]. The search engine offers query-by-example functionality. After entering a search text the user may refine the search by choosing so-called multimedia concepts which are automatically extracted. In contrast to CONTENTUS, the research is focused on extraction of high-level visual features from videos and images, not textual contents derived from OCR or speech-to-text transcripts. Furthermore, the project does not consider the use of Semantic Web technologies or or complex formal representations of knowledge like ontologies.

YOVISTO [90] is a platform for uploading and content-based searching academic audiovisual content such as lectures. It offers Web 2.0 functionalities like collaborative tagging and commenting. It is possible to search within a video using an automatically generated full-text index. Similar to CONTENTUS, YOVISTO features search features like faceted filtering, search in speech transcripts (using automated speech recognition applied on a subset of videos) and automated temporal segmentation of videos. In comparison to CONTENTUS YOVISTO does not account for other media types than audiovisual formats.

A quantitative evaluation of the CONTENTUS media analysis and search functionalities with regards to recall and precision in comparison to the research projects mentioned above has proven to be very difficult. Every project is working on a different and in most cases specific media corpus while the depth and focus of media indexing greatly varies. Single processing steps like audiovisual restoration or Named Entity Recognition could in principle be compared across projects, if common benchmark test sets were used. But since most projects—CONTENTUS is no exception—deploy domain-specific adaptations in their algorithms and do not offer the possibility of replacing the underlying corpus, an overall comparison is nearly impossible. Additionally, a common ground truth for evaluation of semantic multimedia search as described above could not be found.

A comparison against well-known web search engines (such as Google, Bing and Yahoo) is impractical as well, as their indices and media corpora contain mostly web sites and born-digital material such as videos and images. The extremely wide variety of content indexed by these web search engines limits the possible application of Semantic Web technologies in practice. While most of the web search engines nowadays use thesauri and thus offer a basic query expansion for many search terms, we know of no application of unique identifiers for disambiguation of entities or resources. Also none of these search portals seem to make use of transcriptions of the content itself, until now they rely mostly on indexing its accompanying metadata (The vertical video search portal Blinkx [14] claims to use speech-to-text technology though it neither shows nor allows for searching in transcripts). Thus, the focus of most general web search engines differs greatly from more specialized semantic search engines. A quantification of improvements in search engine user satisfaction by using Semantic Web technologies still remains to be investigated.

4 Quality control and content analysis: media specific analysis

During the ingest into the CONTENTUS system, media from all supported types undergo automated *media specific* analysis steps for quality assessment, quality improvement and information extraction. These processing steps are a preparation for entity extraction, the data linking step and presentation through the user interface. Every media source (books, press clippings, photos, film, video, musical and speech recordings) thereby has its specific process chain, a combination of processing modules and parameterization, depending on the type and content of the material. All analysis functionalities developed or improved within the project are described in this section—of course we also use existing state-of-the-art processing in many areas (e.g. for optical character recognition (OCR), dewarping scanned print documents and face detection in audio/visual media to name a few).

4.1 Print/Still Image

For the Print/Still Image processing the quality analysis is described first, followed by the restoration and the semantic analysis.

4.1.1 Quality analysis

During the digitization of the CONTENTUS print materials we encountered only very few problems that were not related to human handling or quality issues of the originals. Low color accuracy due to changing lighting conditions (in case of camera digitization) was by far the most frequent issue found in random samples during the ongoing digitization, while sharpness, lens aberrations, sensor artifacts could be ruled out by an accurate initial set-up of the digitization machines.

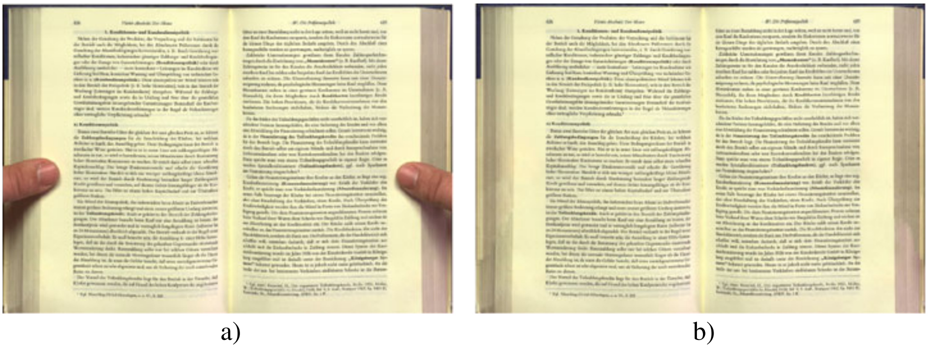
Still Image quality assessment is based on a standardized color target, the XRite ColorChecker. The intention thereby being to distinguish between scanner-related artifacts and intrinsic image distortions or defects. The ColorChecker fidelity is assessed based on a full-reference measure called delta E [51] that basically corresponds to the perceived distance between a current color patch and a corresponding anchor. Significant deviations trigger an alarm within the service platform that indicates the need for rescanning the corresponding media. Images under evaluation can also be characterized by further quality measures as brightness, contrast, sharpness, blocking and overall quality. These will be developed into more detail in Section 4.2.1. A quantitative evaluation of our work can be found at [53].

4.1.2 Restoration

In the digitization process of books and magazines, the problem of distorted or erroneous regions often occurs. In Fig. 3a, two digitized book pages with unwanted objects (two thumbs) are exemplarily depicted. Manual or semi-automatic removal of such objects is very time consuming and inadequate for large data sets. A fully automatic inpainting algorithm is thus proposed to erase unwanted regions seamlessly (cf. Fig. 3b). Dominant, not necessarily linear structures are preserved through structure aware filling methods, while smooth transitions between original and synthetic textures are enhanced through photometric correction methods [47]. A quantitative evaluation is only possible subjectively as no appropriate measure exists. Figure 3 is provided to allow a subjective performance assessment of the proposed approach.

4.1.3 Semantic analysis

The document image analysis system in CONTENTUS processes scanned and optimized document images and produces a hierarchical representation of the logical entities (such as articles, Chapters, sections down to text lines and single characters) for each document page. As most relevant information in printed documents is located in text areas (as opposed to drawings or halftone images), our system is specially tuned to maximize the amount of textual information extracted. The document



a) b)

Fig. 3 Inpainting results for a test image with thumbs. Original picture (*left*), proposed approach for unwanted object removal (*right*). Original picture by courtesy of Imageware

image processing chain employed in CONTENTUS consists of the following independent modules:

1. **Pre-processing:** Greyscale or color input images are binarized by either employing a global optimal binarization [65] or an adaptive binarization step [49], depending on the scan quality and available contrast in the original media. Some of the CONTENTUS print media (hectographical copies) offered light blue text on a yellowish background that could not be handled by a global binarization threshold. The amount of data passed to the following processing steps is dramatically reduced by binarization in order to allow the application of more complex layout segmentation algorithms (which would otherwise require a much higher amount of processor running time).
2. **Separator detection:** For solid separator detection a general-purpose method was obtained from the combination of two recent research algorithms [33, 98]. The Gatos et al. algorithm is used for improving the quality of the vertical and horizontal separators, followed by the extraction of the DSCCs as described by Zheng et al. Logical separators (white empty spaces) are determined using the method proposed in [15]. This algorithm employs a branch-and-bound search for finding a set of maximally empty rectangles given a set of obstacles (bounding boxes of foreground connected components). A subsequent triage of all separators is performed by using information about the dominant character size on the page.
3. **Robust page segmentation** is achieved as a result of an improved version of the algorithm introduced by Jain and Yu [44], capable of segmenting documents with Manhattan layouts exhibiting moderate amounts of noise. The original bottom-up method was enriched with top-down information in the form of the logical column layout and dominant character size on the page, resulting in a robust hybrid segmentation algorithm. The raw textual regions are afterwards further refined and merged. To this end several font characteristics (such as stroke width, x-height, italic property) are computed for each text line and used to derive the text regions with similar characteristics.
4. **Logical layout analysis:** A compound distance measure between any two text blocks is computed as a weighted mean of the Euclidean distance between their bounding boxes and a value directly related to the logical distance between the two text blocks. While most logical distance weights are quite similar between different publishers, there also exist layout-specific weightings which allow the adaptation of the logical layout analysis to different document styles.
5. **Reading order** detection is performed using a topological sorting on a set of block precedence pairs as proposed in [16], enriched with information regarding the layout columns present in the scanned image. The precedence relationships between blocks are detected using a script-specific algorithm, e.g. for roman script a standard left-to-right, top-to-bottom procedure. Although in the CONTENTUS use case we only have documents featuring Roman script as input, it is important to note that the script detection and implicitly the selection of the appropriate algorithm may be done completely automatically for any script. The most plausible reading order computed together with the segmented logical entities is used for logical region merging in order to obtain the final sorted list of articles/sections on a document page. The contents of text regions are subject

to an optical character recognition (OCR) module so as to convert the scanned image areas into editable text.

Additionally, two algorithms providing guidance through global information are employed as sub-parts for various modules throughout the system:

- *Dominant character size* is computed from the smoothed histograms of the heights, respectively widths of the connected components located within the foreground regions of the document page [49].
- Determination of the *logical column layout* of the document is done by means of dynamic programming using the lists of separators and the (rough) page regions.

In the recent ICDAR 2009 Page Segmentation Competition [5] the system clearly outperformed renowned competitors such as *Google* and *Abbyy* (Fig. 4).

4.2 Video and film

A film, often also called motion picture, is a sequence of analogue photographic images that can be stored on different transparent support materials: nitrate, acetate and polyester film.

Video is the electronic counterpart of film as it stores sequences of still images representing scenes in motion. It was first designed for television systems, but has been further developed in many formats: digital and analog. The carrier materials such as magnetic tape, optical disk, are not expected to have a longer life expectancy than film—as low as 20 years according to experts. The only way of preserving video material is through digitization or constant recopying.

The CONTENTUS processing of A/V material consists of quality analysis, restoration and semantic analysis. Both Film and Video material may contain a parallel audio

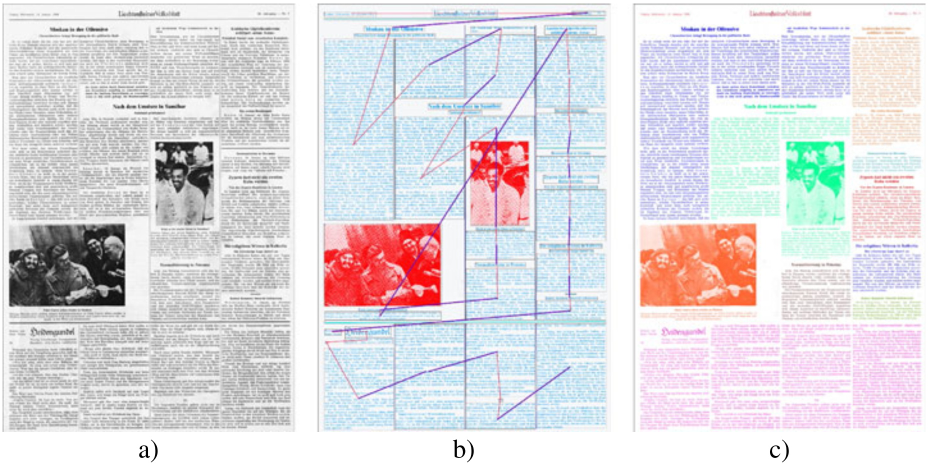


Fig. 4 **a** Original greyscale document image; **b** Graphical visualization of page segmentation result with detected reading order superimposed; **c** Final logical layout segmentation result (detected articles)

recording. The analysis of the audio part (wherever applicable) is described in the audio specific Section 4.3.

4.2.1 Quality analysis

Automatic video quality assessment is one of the most challenging tasks in image processing, as properties of the human visual system that are relevant for quality perception are not well understood to date. Subjective quality perception still is the most reliable quality measure but cannot be seen as a realistic option to handle large multimedia archives. Therefore, CONTENTUS proposes automatic no-reference quality measures to be used for audio/visual material, as no original material typically is available to compare against.

The following algorithms for determination of quality have been developed so far:

1. **Brightness:** Our algorithm is based on an estimation of global brightness. Brightness strongly depends on the context and is often used as an artistic element. Therefore, brightness cannot be seen as a quality measure in its own right in all cases but rather as a descriptive feature.
2. **Contrast:** Similar to brightness, video contrast can be seen as a descriptive feature that does have an impact on the perceived image quality. The proposed measure is based on global and local luminance estimation. Image contrast is inferred based on further brightness statistics.
3. **Sharpness:** is a reliably quantifiable quality feature. The proposed algorithm is based on the calculation of the slope of the power density function, the so-called *Depth of Field*, and the *Just Noticeable Blur* approach [31]. The sharpness measure integrates perceptual aspects of the human visual system and thus allows the prediction of perceived quality.
4. **Blocking artifacts:** are generated by lossy compression during encoding. No-reference measures for JPEG and H.26x/MPEG coding artifacts were developed. The CONTENTUS blocking artifacts metric, which was evaluated with the JPEG dataset from the CSIQ database, resulted in a Pearson and Spearman correlation coefficients of 0.94684 and 0.93257, respectively. Its performance is better than PSNR and comparable to SSIM.
5. **Overall quality:** In order to allow a reliable overall quality prediction, a dedicated measure was developed. It consists of a linear combination of the above-mentioned measures with optimized weights. The overall quality metric was trained with the TID2008 database and evaluated with the CSIQ database. It performed better than PSNR for most distortions and comparably to SSIM.

4.2.2 Restoration

European libraries, broadcast and film archives collect assets in many obsolete analog and digital formats threatened by physical deterioration. Digitization and digital preservation—if done properly—just allow to freeze the deterioration in its current state and restoration done on the digital data can never bring back any information lost. Still, the consumption of digitized media benefits greatly from a reduction of the most common visibly distracting artifacts as drop-outs in video or scratches and dust on film material. As manual restoration is very costly and time-consuming (manual restoration of one hour of film material takes four hours to four days), restoration methods are required that can automatically restore defective audio/visual content.

In CONTENTUS, a procedure for efficient drop-out detection and restoration is presented. This artifact class is one of the most frequent ones in video archives.

Drop-outs are caused by momentary loss of tape contact with the playback head or by flaws on the tape. Some videotape recorders integrate a circuit that detects dropouts and replaces them with information from the previous scan line.

This phenomenon can also occur when the path followed by the read head of the recorder does not correspond to the location of the recorded track on the magnetic tape. Mistracking can occur in both longitudinal and helical scan recording systems. The read head must capture a given percentage of the track in order to produce a playback signal. If the head is too far off the track, record information will not be played back.

The proposed detection algorithm is a two-pass approach, where frames of the potentially deteriorated video sequences are classified into *valid* and *suspect*, based on global color statistics of the images. Suspect pictures are further submitted to local, quad-tree-based analysis for refined evaluations. This yields a subset of identified damaged pictures with accurately localized defects. Detected defective frames are restored using a motion-compensation-based approach. A quantitative evaluation can be found in [45].

Scratches, dust, dirt, stains, abrasion and some more often affect film content as mentioned above. These usually come from the technical process of developing, handling, storing, and screening or scanning the film footage. In some rare cases static objects may already be induced during capturing, for example fluff within a lens or dirt on a scanner glass. In the following, all these defects will be simply referred to as *scratch and dirt*.

To remove these image deteriorations, usually a manual adjustment of certain parameters is needed to fine-tune detection ratios etc., sometimes individually for each scene. Within CONTENTUS, we have developed new scratch and dirt detection and removal algorithms that do not require any manual parameter adjustment. Although not completely perfect, they are capable of reducing the number of objects significantly while leaving other image regions completely untouched, so that the results can be classified sufficient for many purposes without user interaction. Quantitative evaluation of the automatic scratch detection can be found in [62]. In critical situations, where perfect restoration is required, automatic results can be improved through graphical user review of metadata for each object. This is achieved by a new metadata driven workflow.

The workflow for automatic film restoration is shown in Fig. 5 and consists of the following steps:

1. Automatic detection of objects
2. Interactive review of detected objects and quality reporting
3. Automatic removal of objects

The first step, detecting scratch and dirt objects automatically is challenging due to a number of varying factors as, e.g. noise, grain, flicker, jitter, and motion. Detection algorithm design generally aims to reduce misdetection and undetection rate by separating likelihoods around the detection threshold. Though for real world algorithms the probability density functions (pdf) of detected scratch and dirt objects

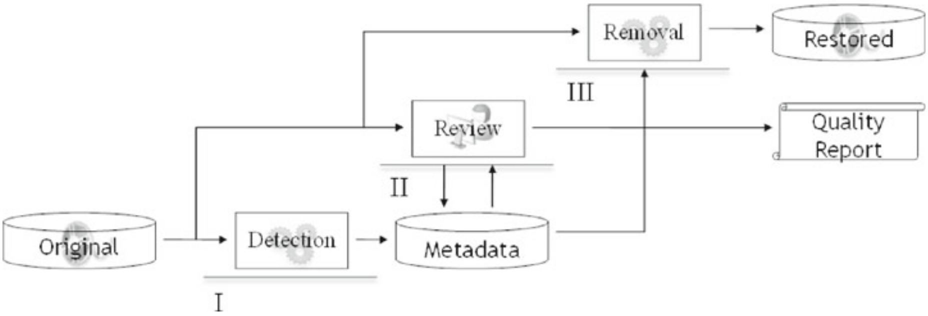


Fig. 5 Three step film restoration workflow with interactive review

usually overlap with content object pdf that should not be detected. This means that the rate of object misdetection and undetection cannot be reduced to zero. Detection therefore requires quality control.

The second step, the review follows detection and requires user interaction for checking metadata quality and improving it if necessary. After metadata has been approved a quality report for the original file is generated. Misdetected or undetected objects can be quickly gathered this way. The review application also allows a quick preview on the results that the succeeding removal step will produce. In the final step, scratch and dirt object removal is carried out automatically based on the generated metadata to build the restored output. Therefore, results are completely predictable and the original film file is only modified at locations where objects have been detected and metadata has been approved for removal.

Thus, the CONTENTUS research created a solution for film restoration that is optimized for automatic processing of large content archives with efficient quality control and reporting capabilities. In the future additional algorithms will be integrated into this metadata driven restoration concept to eliminate effects as, e.g. noise, flicker, and color degradation that affect the experience of viewing film.

4.2.3 Semantic analysis

In the proposed semantic video analysis method, video sequences are temporally segmented into shots and sub-shots [68]. The median picture of each sub-shot is selected as key frame to represent the entire corresponding shot in the subsequent processing steps. The key pictures are collected and annotated using a bag-of-features based approach. Over 90 categories as indoor/outdoor, male/female can be determined. This approach has demonstrated its effectiveness and efficiency for image categorization and classification. The key frame annotations are assigned to corresponding shots and sub-shots. The latter are then grouped into scenes for which the shot annotations apply. Hence a coarse-to-fine navigation through the content based on semantic labels is enabled.

4.3 Audio

The audio processing within CONTENTUS includes segmentation, speaker and speech recognition and musical analysis. The CONTENTUS system handles the audio part of audio/visual recordings like news broadcasts and movies, as well as recordings of music and speech from digitized tapes and records.

4.3.1 Quality analysis

Analog audio material differs from printed material as it is stored on a huge variety of carriers and its digitization offers more degrees of freedom. Many experts still prefer “human touch” over automated transfer [78]. Analog tape recorders could sport varying tape head (azimuth) angles, their heads could be set to any unknown magnetic bias or suffer from irregular playback speed (flutter) to name a few. Digitization of records can be equally challenging, since equalization curves were not standardized until 1954 and the selection of the right needle and pickup counterweight for playback still is dependent on subjective perception and conservation condition of the record. While evaluating commercial products and research activities regarding audio quality measurement during the project planning phase, we found that a fully automated audio quality analysis was out of reach for our project as it could not cope with the multitude of quality impairments that could occur while digitizing the various analog media.

Accordingly, audio quality was ensured by manual control and by selecting only service providers offering a fully IASA [40] compliant workflow.

4.3.2 Segmentation

In order to enable a search in audio documents, their contents must be automatically analyzed to extract the required metadata. A first step is to identify the structure of the audio documents, segmenting them into homogeneous segments and classifying them into categories. This allows for segment-wise browsing of the files and for content analysis targeted specifically to the category. For example automatic speech recognition obviously is only useful if the audio contains speech.

The segmentation and classification into *Speech*, *Silence*, and *Music* is done by assessing the heterogeneity of the audio spectrum of the file: A hypothetical segment border is shifted through the audio document and for each position it is assessed whether the spectral heterogeneity of the material up to this point is large or small. If it is large, a segment border is inserted, and the rest of the file is processed in the same manner, starting from the new boundary. The approach is similar to the one described in [83] together with improvements shown in [19].

After segmentation, the type of audio content of each segment is estimated with probabilistic models for various categories: Segments containing speech and music are identified and further processing is applied to those segments. For speech segments, it is determined whether they stem from a telephone line or are recorded directly by analyzing the available spectral bandwidth in the source material, as this makes a considerable difference for automatic speech recognition systems. Also, the gender of the speaker can be identified. The categorization module uses Gaussian Mixture Models (GMMs) of Mel-frequency Cepstral Coefficients (MFCCs), so the models are generative representations of the spectral content of the audio signal, capturing acoustic properties of the documents.

For higher-level information, namely the words spoken in the document, automatic speech recognition is done for the segments containing speech.

4.3.3 Automatic speech recognition (ASR)

The speech recognition system used is an extension of the one described in [9, 75]. It produces a German word and a syllable transcript of the spoken content of the audio documents. It uses acoustic triphone models trained on broadcast news data, pronunciation dictionaries with 200,000 words and 10,000 syllables, and 3-gram word and 4-gram syllable language models trained from newswire texts. Both transcriptions can be used to enable full text search in audio, where the syllable transcript brings the advantage that even words not in the word dictionary (such as uncommon names) can be found by searching for their syllable representation. The ASR also yields timestamps for the words and syllables. Thus, it is possible for the search engine user to directly jump to the position of a found keyword within an audio document, without the need of reviewing the whole file.

It is also possible to semi-automatically adapt the ASR models to better fit the word usage and grammar of a specific domain, such as *politics* or *music theory*. This increases the quality of the transcription and search results for documents from the subject. Also, the models can be adapted to better fit a particular speaker prominent in the given data. If the speaker is correctly identified by the speaker recognition module (see next Section 4.3.4) or if it is known from metadata that he/she is present in the audio file, using the adapted model gives better ASR performance.

4.3.4 Speaker recognition

In order to be able to search not only for the words spoken in an audio document, but for a specific speaker, the audio analysis process encompasses automatic speaker recognition. This requires that a model is built from speech samples for each desired speaker before the analysis takes place. During analysis the speakers can then be recognized via their models and a speaker name can be put to the corresponding audio segments.

The basic method used is the one described in [71], where speakers are modeled by Gaussian Mixture Models (GMMs) of Mel-frequency Cepstral Coefficients (MFCCs), capturing their voice's spectral characteristics. During analysis, these models output a probability of the speaker being present in a given audio segment, and the segment is assigned the name of the speaker with the highest probability. An important feature of the approach is that each speaker's probability is normalized with the probability from a "background model", which represents the average human speaker. This makes the resulting scores interpretable as something like "this speaker is ten times more likely than the average person".

This voice-based method can be extended if further information and training material (in the order of magnitude from 10 minutes to some hours) for the speaker is available. After producing a word transcript for the extended training material with ASR, the speakers' favorite words, often reflecting their topics of expertise, can be identified. This is done by finding the words that carry the most speaker information, words that are more frequent in the speakers' material than in other documents [8]. Adding this information to the voice-based method improves speaker identification performance.

Another source of information about the speakers is their idiosyncratic pronunciation, which can be captured by using the phonemes in the syllable transcript of the speakers' documents. For this, the syllables are split into phonemes, and speaker specific phoneme sequences are identified [10]. This can capture, for example, a speaker's habit to omit some phonemes at the end of words or the use of a different vowel set stemming from the speaker's dialect. Incorporating this pronunciation information further improves the speaker recognition module's performance.

An Evaluation of the CONTENTUS ASR and Speaker Recognition Algorithms on a broadcast dataset can be found in [9]. We evaluated several Speaker Recognition algorithms against each other in [8, 10]. An evaluation of the algorithms on the CONTENTUS historical speech recordings has yet to be performed—this will be done in the remaining project lifetime once reference transcripts of the material are available through manual annotation.

4.3.5 Musical features

The consumption of music has been changed dramatically by its availability in digital form. While access to music is increasingly simple, it has become more difficult to manage the growing music collections. Archives and libraries are facing similar problems. Even though there is big potential in using or selling archive audio material, the content often lacks complete description, categorization and semantic links.

This challenge is addressed by the automatic music analysis performed within the CONTENTUS framework. First the audio signal is analyzed and basic features are extracted. We also employed low-, mid- and high-level audio features from the MPEG-7 standard to which the project partner *mufin* has contributed. Other, more recent developments include rhythm analysis.

As a next step, state-of-the-art machine learning technology is employed to extract semantic musical attributes. We used classifiers based on K-nearest Neighbor, Gaussian Mixture Models and Support Vector Machines. Musical attributes include mood descriptions such as *happy*, *sad*, *calm* or *aggressive* but also other descriptive tags such as *synthetic*, *acoustic*, *presence of electronic beats*, *distorted guitars*, etc.

This information can then be used to narrow down search results, to offer browsing capabilities or to contextually describe content. By combining these attributes with information from other sources such as editorial metadata the user can, for instance, search for "aggressive rock songs from the 1970s with a track-length of more than 8 min".

The technology also allows the user to find similar music, e.g., for playlist generation or to support an editor who needs an alternative to a piece of music he is not allowed to use in a certain context. Similar music is found by combining all available information about the music in the catalog using a music ontology. In addition to the signal-based similarity engine other data sources such as editorial data or user tags can be employed which makes the recommendation system a hybrid of multiple approaches.

As the recommendation system features a module based on digital signal-processing algorithms it can generate musical similarity measures for all songs within a music catalog. Due to the fact that it makes use of mathematical analysis of the audio signals, it is completely deterministic and can work independent of any "human factor" such as cultural background, listening habits, etc. In contrast to other technologies such as collaborative filtering [34, 81], the technology can provide recommendations for any track, even if there are no tags or social data. Recommendations are not limited by genre boundaries, target groups or biased by popularity. In case that genre boundaries or the influence of popularity is desired, this can be addressed by combining the results with other data sources such as social data.

Finally, the technology *audioid* enables the identification of unknown audio material by comparing a fingerprint excerpt to a reference fingerprint database. As of January 2011, mufin operates a content database of close to 10 million tracks acquired through cooperations with major and independent music labels. An automated ingestion system handles the analysis of new musical content and the update of similarity relations. Via a web service this database powers commercial applications such as the mufin player [61] or MAGIX MP3 deluxe as well as further third-party web portals, PC and mobile applications.

5 Content analysis and semantic linking: media independent processing

Once the media assets have undergone their respective media specific analysis steps, all information extracted is available as a textual representation (e.g. speech transcript, OCR transcript, audio/visual genre). The *media independent* processing steps of named entity recognition, indexing and ontology linking described in this section allow for making links between existing catalog data and media assets and are essential building blocks for the semantic search and retrieval engine of CONTENTUS.

5.1 Named entity recognition and disambiguation

A basic component in the semantic processing of textual information with the goal of enhanced information extraction is the detection of named entities. In the context of the project we restrict our work to the recognition of names of persons, locations, and organizations in German natural language texts. In named entity disambiguation we identify these entities by matching them to entries in an ontology, in our case Wikipedia. Many libraries create and maintain collections of index terms and synonyms for persons, locations, topics, organizations that all have unique identifiers for disambiguating, the so called *authority files* [69]. Since the German Wikipedia already contains community maintained links to the authority files of the German National Library, this enables us to connect catalog entries for persons, organizations and works to the media assets automatically.

The goal of Named Entity Recognition is to assign to every token of a text a label indicating whether or not the token is part of a named entity, such as a person, a location, or an organization. Our method is to make use of Conditional Random Fields, a supervised learning algorithm to label structured data [50]. In a supervised method labeled training data is needed. We use the well-established BIO-notation to label the data [70]. Table 3 gives an example sentence with its labeling. A 'B' indicates the beginning of a named entity, an 'I' indicates the continuation ("Inside") of a named entity, and an 'O' indicates other tokens not part of any named entity.

Table 3 Labeling of words in a sentence after applying Named Entity Recognition

Angela	Merkel	(	CDU	)	met	Obama	in	Berlin
B_PER	I_PER	O	B_ORG	O	O	B_PER	O	B_LOC

To train the learning algorithm we model every word w_i as a vector x_i. The vector is composed of real valued features representing the properties of the word and its neighborhood. Features can be syntactical (such as capitalization, the occurrence of certain prefixes or suffixes or special characters, etc.), structural (such as the part-of-speech tag) or based on external resources (such as the occurrence in some dictionary). Conditional Random Fields also allow to take into account properties of neighboring words, which allows us to represent the context of words explicitly. Thus, we represent each word w_i as a pair (x_i, y_i) where x_i is the feature vector and y_i the label (in BIO-notation as seen in Table 3).

We train the model using the annotated German dataset of the German news agency DPA [29].

After words have been identified to represent entities the second goal is to identify them. For example, the name “Michael Jackson” refers to a well-known American pop-star, but also to an English journalist who authored several books about beer and whiskey. Named Entity Recognition only gives us the information that “Michael Jackson” is a person.

In order to identify the person in question we relate the context of the occurrence of the person’s name to some background information we have about candidate persons. We use Wikipedia to provide us with this background information. Hence, the first step is to identify candidate persons and the second step is to select the most likely candidate.

To identify candidate persons we rely first and foremost on the complete name of the persons. Furthermore, we employ heuristics to deduce the complete name if only parts of it is given in the current occurrence, i.e., we look for previous occurrences of a possibly complete name in the same paragraph or document. Last, we employ some simple rules to identify very specific people, e.g., by their function or profession, e.g. “Joseph Ratzinger” and “Pope Benedict XVI.”.

To select the most likely candidate we perform two steps. First, we compute the cosine similarity of the current document to each of the Wikipedia entries of the candidates. We perform length normalization and stop word removal. Second, we combine this score with a prior score for each candidate. The prior score is determined by counting the incoming links within Wikipedia for each candidate.

5.2 Data linking

Data linking in CONTENTUS comprises terminological knowledge in shape of an OWL ontology and assertional knowledge found in a number of sources as described in the following.

5.2.1 Ontology

To facilitate integrated semantic search and consumption of multi-media corpora and background knowledge, CONTENTUS provides its metadata in a knowledge base employing an OWL ontology. The CONTENTUS ontology (CONTO) primarily describes and links works, persons, corporations, locations, and topics. For multi-media corpora it is important that CONTO differentiates types of media like e.g. text, image, video, audio. Because works are often sought with regard to persons, relations between persons are modeled in detail.

The ontology CONTO is developed in cooperation with other Semantic Web related projects. Its current modeling and re-used vocabularies are shown in Fig. 6.

Currently employed vocabularies are the model of the LinkedData project [52] of the DNB as well as bibliographic (RDA [72], SKOS [76], dc [24]), geographic (GeoWGS84 [91], FAO geopolitical ontology [30]), and other vocabularies (FoaF [32]), employed dataset URIs primarily refer to DNB data records.

As available vocabularies for person-to-person relations like FoaF, RELATIONSHIP [73] contained too few types of relations for our existing bibliographical dataset (person authority file, PND) or were not thoroughly structured, we developed a customized model for personal relations. CONTO defines about 50 hierarchical, gender-neutral person relations describing mainly friendship, kinship, and occupation connections. Cooperating projects include e.g. ALEXANDRIA, the DNB LinkedData project [3, 36], and the DNB GND project [11]. CONTO is still to be considered as work in progress and will be published in the upcoming months once a more stable intermediate state is reached.

When involving diverse cataloging or annotation sources, the knowledge base receives metadata of quite different quality and reliability. Therefore, data provenance, confidence, and temporal validity is stored in CONTO by meta properties using the n-ary approach [27], representing attributes with concepts rather than with properties.

The facts explicitly stated in the ontology can be padded with implicit facts given by hierarchy, symmetry, inversity, transitivity, and specific connections between properties (e.g. grandparents, aunt/uncle). Reasoning will be employed for this, as in many authority file entries only one direction of relation is explicitly stated (A is father of B) while the corresponding relation (B is child of A) has to be derived.

5.2.2 Data integration

One of the challenges in the CONTENTUS project is the need for integration of information and metadata from a variety of different sources. Typically, such data can comprise products of digitization efforts, born-digital documents, but also external knowledge contributed by user communities and more. Even if the quality of the

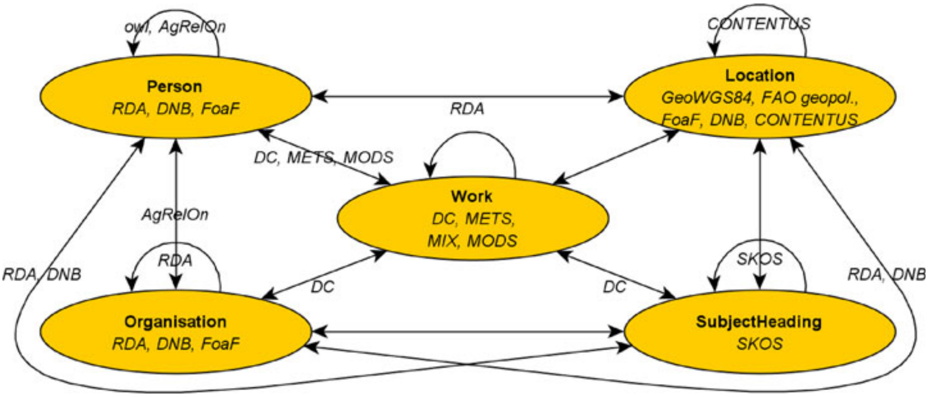


Fig. 6 Vocabularies used for concepts and relations in CONTO Ontology

external metadata is sometimes (but not necessarily) lower than ideal, it can still complement existing data. For example, the catalog of the German Music Archive, the German national archive hosting a central collection of sheet music and sound recordings and serving as the center of bibliographic information related to music, does not list the individual songs or tracks of a recording. When linking the catalog data to a music database, the user has access to more detailed information, such as the track lists.

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

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

The archives' authority and catalog information serves as a reference—the mapping from Wikipedia URL to the authority file is maintained manually by volunteers. This mapping is already used in the German Wikipedia and the catalog system of the German National Library. Audio tracks were enriched with corresponding MusicBrainz metadata using the Picard tagging application making use of audio fingerprints as an identifier as the MusicBrainz information was openly available and free of royalty issues.

For the mapping of extracted but not yet disambiguated locations to Wikipedia and authority file data a combination of heuristics and similarity metrics is used. 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 [36].

According to Linked Open Data principles it is recommended to use dereferenceable and persistent URIs as identifiers for entities. URIs allow us to tie together different sources of information regarding entities of interest (persons, places, organizations, etc.) and integrate them into the CONTENTUS knowledge base.

5.3 Metadata

Metadata from all data sources within CONTENTUS need to be represented and transported from the first import stages to the search interface. We managed to narrow down the number of different metadata formats needed for the import to the following two:

- MPEG-7 for audio/visual media (recordings from public broadcasters, digitized film material)
- METS/MODS for printed media (newspapers/books), audio/music (CDs, shellac and vinyl records) and still images (photographs).

The integration of broadcast archives in CONTENTUS is realized by a mapping of the broadcast specific metadata formats. Manually generated data sets can be integrated over web service interfaces to the CONTENTUS system. The architecture of this mapping service is also designed to export from CONTENTUS to other systems. The core of the transformation service is a central exchange data model called BMF (Broadcast Metadata Exchange Format) [17]. The BMF data model covers the information requirements of a wide variety of domains in television as production, planning, distribution and archiving.

5.3.1 MPEG-7

To exchange AV metadata within CONTENTUS, the selected data model for audio/visual metadata is MPEG-7 [41, 54]. Compared to other data models, MPEG-7 has advantages regarding its flexibility because it is based on general and widely applicable concepts. While many efforts have been made to extend the standardized MPEG-7 schemes for semantic retrieval, such as [1, 12, 18, 46, 84], many of these focus on expressing domain knowledge as ontologies. The generality and flexibility of MPEG-7, however, caused practical problems during the development of the CONTENTUS-system. On the one hand, the flexible definition of MPEG-7 elements causes ambiguities—semantically identical data can be represented in multiple ways—which become problematic when mapping one data model into another [21]. On the other hand, MPEG-7 was not specifically designed to describe (broadcast) content for the CONTENTUS archive, which means that certain features are missing. To adapt MPEG-7 to the broadcast specific requirements e.g. of including spatio-temporal information such as subtitles from existing BMF metadata, CONTENTUS extended the MPEG-7 data model for use by the AV modules in the project.

The CONTENTUS MPEG-7-based data model was designed to include all metadata generated by the automatic content analysis modules (cf. Section 4) and to fulfill the semantic constraints that are set by the field of application, in our case the broadcast domain. For example, to address certain elements of a media item (e.g. a video) the extended data model allows the description of whole media items as well as of small spatial, temporal or spatio-temporal fragments of the media item (e.g. one *scene* or *shot*). For the representation of AV features like image regions or object descriptions, the definition of the structure should be explicit and still be flexible enough.

To fulfill these requirements several new data types had to be created to guarantee unambiguousness: For example, as MPEG-7 provides only one option for a title description, a specific "TitleType" has been created for every title that is used in the broadcast domain, e.g. main title, broadcast title, working title. Another example is the "SpokenContentType" which has been added to the MPEG-7 data model to display the results of the Speech Recognition module (e.g. the Speech-to-Text transcript or the speaker's name). Furthermore, the elements "SubtitleSegment" and "SubtitleSegmentType" were added to the type "AudioVisualSegmentTemporalDecomposition" to incorporate readily available subtitle information that could then be displayed in the CONTENTUS semantic media viewer component of the search user interface. Since CONTENTUS uses PND-URIs for disambiguating entities and linking to external data sources, we extended the MPEG-7 model by adding "URI" to the existing "PersonType". Several further types have been introduced to describe quality attributes such as video quality, video defects or perceptual attributes. Some quality-related MPEG-7 extensions proposed by the CONTENTUS project are currently being standardized by the European Broadcasting Union (EBU) as an MPEG Profile [58].

The extended MPEG-7 data model is used for the exchange of metadata between the AV processing modules, thus guaranteeing a consistent, end-to-end pass-through of metadata throughout the CONTENTUS AV processing chain.

5.3.2 METS

The Metadata Encoding and Transmission Scheme METS [59] is used as the XML based metadata import format for printed and audio media into the CONTENTUS media repository. METS is a container format that stores structural information (e.g. references to all single digitized pages or audio files) and allows to be extended using other XML schemes. For printed media, bibliographical data from the catalogs of the German National Library was stored in METS using the MODS [60] extension schema. Within the MODS section, persons such as authors and publishers were referenced using a PND [69] authority file URI wherever applicable.

We also used the METS container format for preserving a set of basic information about the digitization process itself. For still image files the NISO MIX [64] extension schema was utilized to store descriptive data about scanner hardware, data encoding and compression used, an MD5 checksum and information about the different digitization service providers. These informations allow for tracking of digitization defects in case of flawed hardware or workflow.

For the audio data similar digitization metadata had to be kept. Since we did not find a suitable corresponding schema for the documentation of audio digitization we created an own schema containing detailed information about the digitization hardware and every step of the subsequent software processing chain. Once this schema has undergone an internal reconciliation within the German Music Archive in order to ensure its suitability for upcoming digitization projects, we plan to release it to the public.

6 Semantic multimedia search

The CONTENTUS search engine interface demonstrates how semantic technologies can be used to facilitate a better and livelier search experience within large media collections. The main challenge of the user interface is to overcome the heterogeneity of multimedia and semantic metadata and make it accessible in a consistent, clear, and meaningful way.

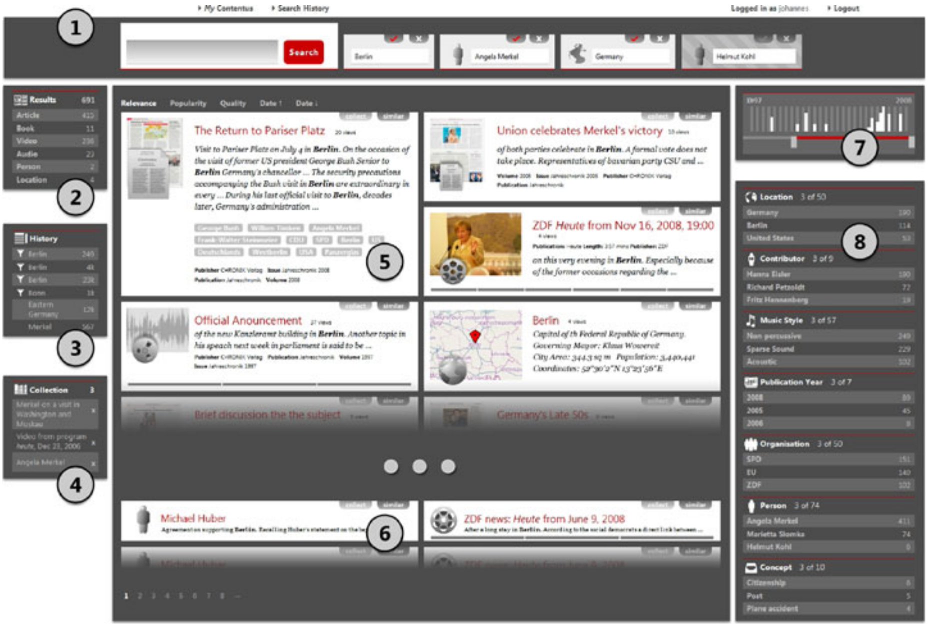
The objective of the CONTENTUS user interface is to combine multi-modal and semantic search by improving search results and navigation while preserving usability. Therefore, the interface not only has to provide information found in documents, but also information about documents and resources, its comprising entities, and their in-between relationships. For example, a person could arise as an author of a document, but also as a subject that a document describes.

6.1 User interface

The layout of the CONTENTUS user interface as shown in Fig. 7 is arranged in a Search Area (1) and a Result Area. The Search Area contains the Search Field and the Search Path. The Result Area below is divided into three columns. The left column (2, 3, 4) hosts the functions to organize queries and filter by media type, the right column (7, 8) the functions to filter of the result set by content. The main column (5, 6) in the middle shows the current search results.

6.2 Search results

The multimodal search interface shows different media and entity types (2) within one result page. They comprise books, newspaper-articles, videos, images, audio, locations, and persons and all have an individual *entity page*—for example, search results for locations reveal a geo-map, and an image preview is presented in video image results. The relevance ranking of the search result list is visually emphasized by the size of the single result items (5, 6). Every search result is featured with a title and specific information depending on its media or entity type like production companies, broadcast dates or publisher. Results based on transcripts (e.g. OCR, speech) are presented with text snippets and semantic entities extracted from the documents content (as highlighted in the first result item). The interface makes use of Ajax by adding and changing elements on the screen to minimize the separation between entering a search term and browsing its results.



The Return to Pariser Platz

*Visit to Pariser Platz on July 4 in **Berlin**. On the occasion of the visit of former US president George Bush Senior to **Berlin** Germany's chancellor ... The security precautions accompanying the Bush visit in **Berlin** are extraordinary in every ... During his last official visit to **Berlin**, decades later, Germany's administration ...*

Union celebrates Merkel's victory

*of both parties celebrate in **Berlin**. A formal vote does not take place. Representatives of bavarian party CSU and ...*

ZDF *Heute* from Nov 16, 2008, 19:00

*on this very evening in **Berlin**. Especially because of the former occasions regarding the ...*

Official Anouncement

*of the new Kanzleramt building in **Berlin**. Another topic in his speach next week in parliament is said to be ...*

Berlin

Capital of th Federal Republic of Germany.
Governing Mayor: Klaus Wowereit
City Area: 344.3 sq m Population: 3,440,441
Coordinates: 52°30'2"N 13°23'56"E

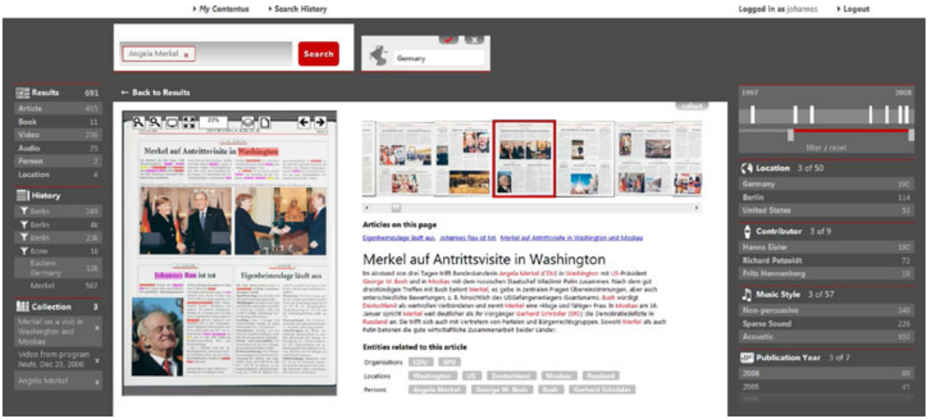
Fig. 7 Overview of the CONTENTUS user interface

Dates and timestamps are represented by a timeline slider (7), the result frequency per year is shown as a histogram over the timeline. In case of persons suitable timestamps can be the date of birth, the date of death, or any other event related to that person. For documents, the timeline in the CONTENTUS user interface shows publication dates. Users can select any desired time frame with the timeline sliders and filter the results accordingly.

The list of facets (8) is located below the timeline. The facets used by CONTENTUS are extracted entities grounded to Wikipedia articles and concepts assigned to the documents by means of automatic classification. The interface distinguishes between individuals in an active and a passive relation to the results—contributors (e.g. authors, directors, musicians or composers) and persons (entities found within the text transcripts). Selected facets are collected as *breadcrumbs* for a maximum of user transparency. These breadcrumbs can be individually deselected or deleted to change the result set instantly. Previous searches for the user are stored in a history (3) and can thus be easily recalled.

The user can store and collect relevant search result items in his own collection (4). A preview of this collection is always shown on the lower end of the left column. This collection of media items is used to generate a user profile based on classification of the media assets. In the final version of the system the user interest profile will be utilized to personalize the search results.

By clicking on any item within the search results, the media or entity detail page displays all relevant information about this item. An example for an article in a yearly chronicle book is shown in Fig. 8. The article viewer (left) displays the original scanned document page. All detected semantic entities within the page are highlighted and clickable as an additional search facet or a starting point of a new search. The color key encodes the entities' class (e.g. red for locations, purple for persons and so on). Below, the article text extracted by OCR is provided and all related entities extracted from the text are shown. By clicking on an entity, its detail page appears (not depicted) and provides further information. On the entity detail page, relationships to other resources and entities are exposed in a relationship graph that can be navigated.



Articles on this page

Eigenheimzulage läuft aus. Johannes Rau ist tot. Merkel auf Antrittsvisite in Washington und Moskau

Merkel auf Antrittsvisite in Washington

Im Abstand von drei Tagen trifft Bundeskanzlerin Angela Merkel (CDU) in Washington mit US-Präsident George W. Bush und in Moskau mit dem russischen Staatschef Wladimir Putin zusammen. Nach dem gut dreistündigen Treffen mit Bush betont Merkel, es gebe in zentralen Fragen Übereinstimmungen, aber auch unterschiedliche Bewertungen, z. B. hinsichtlich des USGefangenenlagers Guantanamo. Bush würdigt Deutschland als wertvollen Verbündeten und nennt Merkel eine »kluge und fähige« Frau. In Moskau am 16. Januar spricht Merkel weit deutlicher als ihr Vorgänger Gerhard Schröder (SPD) die Demokratiedefizite in Russland an. Sie trifft sich auch mit Vertretern von Parteien und Bürgerrechtsgruppen. Sowohl Merkel als auch Putin betonen die gute wirtschaftliche Zusammenarbeit beider Länder.

Entities related to this article

Organisations	CDU SPD
Locations	Washington US Deutschland Moskau Russland
Persons	Angela Merkel George W. Bush Bush Gerhard Schröder

Fig. 8 Book detail page

6.3 The CONTENTUS index

The central building block of our semantic search engine is an index that is populated with textual information within and about the documents from the repository. The main purpose of the index is to retrieve suitable document identifiers given a user query. With the document identifier the complete document can then be fetched from the repository and be presented to the user.

The first part of the index is the traditional full-text index. Here, the document texts are indexed so that queries containing words from these documents will return the correct documents. Second, the index represents metadata provided through human annotation, such as the date of creation for each document. Metadata can be used to filter and refine search results, for example by restricting the time period for documents.

Most importantly, the index also stores the facets obtained by the automatic document annotation process: Named and disambiguated entities as well as the automatically detected topics. Facets again allow the user to filter and refine search results. A typical usage scenario would be to first enter a text query, such as “Merkel”, then to refine the result by the uniquely identified person “Angela Merkel” and then to further filter the result by specific user interests, such as the person “Nicolas Sarkozy” or the location “Washington, DC”. The index further returns the count of any potential filter facets. Hence, the user can draw conclusions about which facets are closely related to one another and which are not.

The index is implemented using the Apache Lucene/Solr package. Figure 9 gives an insight into the Solr schema of the index. Some of the metadata is explicitly modeled, such as the publication date or the quality of the document. Furthermore, some items, such as audio and video documents are organized in segments, which needs to be taken into account by the index. The automatically obtained document facets are all stored in one field. The value of the facet encodes the type of the facet, i.e., whether the facet refers to a person, a location, an organization, or the classification.

```
<fields>

  <field name="uuid" type="string" indexed="true" stored="true" required="true"/>
  <field name="uri" type="string" indexed="true" stored="true" required="true"/>
  <field name="pubDate" type="tdate" indexed="true" stored="true" sortMissingLast="true"/>
  <field name="quality" type="tfloat" indexed="true" stored="true" sortMissingLast="true"/>
  <field name="text" type="text" indexed="true" stored="false" multiValued="true" .../>
  <field name="itemTypeUri" type="string" indexed="true" stored="true" required="true"/>
  <field name="maxSegNr" type="pint" indexed="false" stored="true" required="true"/>

  <dynamicField name="stored_*" type="text" indexed="false" stored="true"/>
  <dynamicField name="info_*" type="text" indexed="false" stored="true"/>
  <dynamicField name="segment_*" type="text" indexed="false" stored="true"/>
  <dynamicField name="popularity_*" type="tlong" indexed="true" stored="true" .../>
  <dynamicField name="facets_*" type="facet" stored="false" multiValued="true"/>
  <dynamicField name="ignored_*" type="ignored" multiValued="true"/>

</fields>
```

Fig. 9 The Solr schema for the CONTENTUS index

The index is filled and queried via a Web service interface. For querying a query string and a list of facets is given and a list of document identifiers is returned. For each resulting document details can be requested via a separate call for the detail page in the user interface. Access to the repository is only needed for the detail page; the index duplicates the information necessary for the result overview page. There is also a feedback loop: Document popularity can be increased from the outside via the Web service interface.

During the development process we found that using two different indices for storing full-text and ontology information was impractical since merging the result sets while maintaining a useful ranking could not be achieved. Therefore we decided to do a two-step approach: First, textual transcriptions containing named entities (persons, organizations, locations) are enriched with ontology information like alternate names, relatives and organizational members. The whole augmented transcript is then indexed so that the document can be retrieved even when searching only for alternate names that were not originally present in the transcript. User generated knowledge is treated similarly—first, the knowledge editing interface returns user edits along with the user's role (like *expert*, *librarian*, *normal user*) which is then fed into the index. That way, the user interface can distinguish between knowledge relevant only for single users (edits by "normal users") and knowledge relevant for broader user groups (edits by "experts"). Depending on the user currently logged in, the result set can be personalized/different even for the same search query entered. We will evaluate the user acceptance of this feature in the remaining project lifetime.

7 Conclusion and outlook

In this paper we have described the challenges that are faced by multimedia archives and have presented solutions in shape of a modular process chain. The deterioration of analog media can be countervailed by digitization. The deficient quality of analog as well as digitized content needs to be ameliorated with media specific restoration techniques. The knowledge acquisition bottleneck needs support by automatic information extraction approaches. The vast amount of heterogeneous metadata needs to be consolidated through data linking. The content can be accessed through a semantic multimedia search interface.

Regarding the algorithms developed in the project, several stand out: The award-winning page segmentation providing a significant robustness advantage over the state of the art, the audio/visual restoration that is an industry's first to allow for unattended film and video restoration and our extraction of musical features and similarity that already has shown its scalability over millions of songs and is deployed in a commercial product.

In addition, the CONTENTUS *service platform* provides a scalable infrastructure capable of processing the output of mass digitization projects and large multimedia collections. We have shown that it is possible to fully automate the process from ingest to access for printed, audio and audio/visual media without any need for operator interaction.

The machine-based generation of metadata from multimedia objects, as described in this paper, is not aimed at substituting existing catalog metadata traditionally used in cultural heritage organizations—rather, CONTENTUS aims at integrating different metadata sources. We have shown that it is possible to automatically merge automatically generated metadata, catalog metadata and external resources in a data model using W3C standard Semantic Web technologies. All these metadata sources are successfully integrated in a single search user interface using a single search index.

To sum up, the combination of tools and workflows developed in CONTENTUS have the potential to greatly lower the entry barrier for building accessible digital collections of cultural heritage organizations.

At present the project focuses on developing its final software demonstrator. It will integrate personalization techniques into the user interface and will allow for user feedback and editing of machine-generated knowledge, in addition to an uploading functionality for user-generated content.

CONTENTUS research technologies have been used to process test content, however, large-scale testing in production environments has not yet been conducted. Parallel to commercial exploitation activities, research technologies from CONTENTUS are therefore being transferred to the German Digital Library project [25], to be evaluated under real-world conditions in early 2012. Our interim performance investigations show that there still is the need and potential for optimizations of the data transfer overhead within the processing cluster.

The comparison of the CONTENTUS semantic multimedia search with other search engines proves to be difficult. This is due to incomparable media and metadata corpora used by other products or projects as well as CONTENTUS' unique search paradigm that combines automatically generated metadata, catalog metadata and external resources. A quantification of the search engine performance using precision and recall is difficult as well as the relevancy and expected ranking of the results is highly subjective. Many questions arise, such as "Which result type has the highest relevancy?" "Which result is the most relevant whenever more than one entity is selected within the user interface?" Therefore, the final phase of the project will focus on user tests to evaluate the search experience. These tests will also investigate how to best exploit and improve the provenance-dependent and personalized relevance ranking of our index in order to measure and increase user satisfaction with respect to search results. In addition, specific corpora and ground truth will have to be generated to compare CONTENTUS search results with technologies of related projects and systems.

Acknowledgements We would like to kindly thank Andreas Hess for initial input, Jan Hannemann for coordination and supporting of the review process and Klaus Bossert for review. The project CONTENTUS was funded by means of the German Federal Ministry of Economy and Technology under the promotional reference "01MQ07003".

Our project partners are:

- German National Library (Project lead, content provision and semantic technologies),
- Deutsche Thomson OHG (Processing of audio/visual media),
- Fraunhofer Institute for Intelligent Analysis and Information Systems (Text and speech processing, workflow engine),
- Fraunhofer Heinrich Hertz Institute (Audio/visual and still image processing),
- Hasso Plattner Institut (since 2009) (User interface, personalization),
- Institut für Rundfunktechnik GmbH (Requirements specification, content provision),
- Moresophy GmbH (until 2009) (User interface and semantic technologies), and
- Mufin GmbH (Audio processing and classification).

References

1. Agius HW, Angelides MC (2009) From mpeg-7 user interaction tools to hanging basket models: bridging the gap. Multimedia Tools Appl 41(3):375–406
2. ALEXANDRIA—a collaborative knowledge engine, a use case of the Theseus research program. http://alexandria.wefind.de. Accessed 1 Dec 2011
3. Altenhöner R, Hannemann J, Kett J (2010) Linked data aus und für bibliotheken: rückgratstärkung im Semantic Web. In: Proc of 1. DGI-Konferenz Semantic Web und linked data—elemente zukünftiger informationsstrukturen, pp 67–75
4. Amato G, Debole F, Peters CPS (2008) The multimatch prototype: multilingual/multimedia search for cultural heritage objects. In: Proc of the 12th European conf on digital libraries. Aarhus, Denmark
5. Antonacopoulos A, Pletschacher S, Bridson D, Papadopoulos C (2009) Page segmentation competition. In: Proc of 10th int conf on document analysis and recognition (ICDAR), pp 1370–1374
6. Avrithis Y, Kompatsiaris Y, Staab S, O'Connor N (eds) (2006) Semantic multimedia: first international conference on semantic and digital media technologies. In: SAMT 2006, Athens, Greece, 6–8 December 2006, Proceedings, lecture notes in computer science, vol 4306. Springer, Berlin. doi:10.1007/11930334
7. Bartolini I, Patella M, Romani C (2010) Shiatsu: semantic-based hierarchical automatic tagging of videos by segmentation using cuts. In: Proc of the 3rd int'l workshop on automated information extraction in media production, AIEMPro '10. ACM, New York, pp 57–62
8. Baum D (2009) Topic-based speaker recognition for German parliamentary speeches. In: Proc of IEEE automatic speech recognition and understanding workshop (ASRU '09). Merano, Italy
9. Baum D, Schneider D, Bardeli R, Schwenninger J, Samlowski B, Winkler T, Köhler J (2010) DiSCo—a German evaluation corpus for challenging problems in the broadcast domain. In: Proc of the 7th int'l conf on language resources and evaluation (LREC'10)
10. Baum D, Schneider D, Mertens T, Köhler J (2010) Constrained subword units for speaker recognition. In: Proc of the speaker and language recognition workshop odyssey
11. Behrens-Neumann R, Pfeifer B (2011) Die Gemeinsame Normdatei—ein Kooperationsprojekt. Dialog mit Bibliotheken
12. Benitez AB, Zhong D, Chang SF (2007) Enabling MPEG-7 structural and semantic descriptions in retrieval applications. J Am Soc Inf Sci Technol 58:1377–1380
13. Berners-Lee T, Hendler J, Lassila O (2001) The semantic Web. Sci Am 284(5):34–43
14. Blinkx: a video search engine. http://www.blinkx.com. Accessed 1 Dec 2011
15. Breuel T (2002) Two algorithms for geometric layout analysis. In: Proc of workshop on document analysis systems, vol 3697, pp 188–199
16. Breuel TM (2003) High performance document layout analysis
17. Broadcast metadata exchange format—specification. http://www.irt.de/en/activities/production/bmf.html. Accessed 1 Dec 2011
18. Celma O, Dasiopoulou S, Hausenblas M, Little S, Tsinaraki C, Troncy R (2007) MPEG-7 and the semantic Web. W3C Incubator Group Editors
19. Cheng S, Wang H, Fu H (2010) BIC-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. IEEE Trans Audio Speech Lang Process 18(1):141–157
20. Contentus: a use case of the Theseus research program. http://www.contentus-projekt.de. Accessed 1 Dec 2011
21. Corda U (2008) Multimedia semantics—from MPEG-7 metadata to semantic Web ontologies
22. Dasiopoulou S, Tzouvaras V, Kompatsiaris I, Strintzis MG (2009) Capturing MPEG-7 semantics. In: Sicilia MA, Lytras MD (eds) Metadata and semantics. Springer, New York, pp 113–122
23. Dasiopoulou S, Giannakidou E, Litos G, Malasioti P, Kompatsiaris Y (2011) Knowledge-driven multimedia information extraction and ontology evolution. Chap A survey of semantic image and video annotation tools. Springer, Berlin, pp 196–239
24. DC (Dublin Core): metadata element set, version 1.1. http://purl.org/dc/elements/1.1/. Accessed 1 Dec 2011
25. DDB: the German Digital Library project, a portal for culture and science. http://www.deutsche-digitale-bibliothek.de. Accessed 1 Dec 2011
26. Debald S, Nejdl W, Nucci FS, Paiu R, Plu M (2006) Pharos—platform for search of audiovisual resources across online spaces. In: Proc of the 1st int'l conf on semantic and digital media technologies (SAMT2006). Athens, Greece

27. Defining N-ary relations on the semantic Web, W3C working group note 12 April 2006. http://www.w3.org/TR/swbp-n-aryRelations/. Accessed 1 Dec 2011
28. Ding H, Sølvberg IT (2005) Semantic data integration framework in peer-to-peer based digital libraries. JDIM 3(2):71–75
29. dpa (deutsche presse-agentur gmbh). http://www.dpa.de. Accessed 1 Dec 2011
30. FAO (Food and Agriculture Organization of the United Nations): geopolitical ontology. http://aims.fao.org/aos/geopolitical.owl. Accessed 1 Dec 2011
31. Ferzli R, Karam LJ (2009) A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). IEEE Trans Image Process 18(4):717–728
32. FOAF (Friend Of A Friend): vocabulary specification. http://xmlns.com/foaf/spec/. Accessed 1 Dec 2011
33. Gatos B, Danatsas D, Pratikakis I, Perantonis SJ (2005) Automatic table detection in document images. In: Proc of 3rd int conf on advances in pattern recognition (ICAPR), LNCS 3686, pp 609–618
34. Goldberg D, Nichols D, Oki BM, Terry D (1992) Using collaborative filtering to weave an information tapestry. Commun ACM 35:61–70
35. Guha R, McCool R, Miller E (2003) Using the semantic Web: semantic search. In: WWW '05 proceedings of the 14th international conference on world wide Web, pp 700–709. doi:10.1145/775152.775250
36. Hannemann J, Kett J (2010) Linked data for libraries. In: Proc of the world library and information congress of the Int'l Federation of Library Associations and Institutions (IFLA)
37. Hinze A, Buchanan G, Bainbridge D, Witten IH (2009) Semantics in Greenstone. In: Kruk SR, McDaniel B (eds) Semantic digital libraries. Springer, New York, pp 163–176. doi:10.1007/978-3-540-85434-0_12
38. Hobson P, Kompatsiaris Y (2006) Advances in semantic multimedia analysis for personalised content access. In: ISCAS. IEEE, Piscataway
39. Huiskes MJ, Lew MS (2008) The Mir Flickr retrieval evaluation. In: Proceeding of the 1st ACM int'l conf on multimedia information retrieval, MIR '08. ACM, New York, pp 39–43
40. IASA (International Association of Sound and Audiovisual Archives) TC 04: guidelines on the production and preservation of digital audio objects. http://www.iasa-web.org/audio-preservation-tc04. Accessed 1 Dec 2011
41. Information technology—multimedia content description interface—part 1: systems. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34228. Accessed 1 Dec 2011
42. Informedia-i: integrated speech, image and language understanding for creation and exploration of digital video libraries. http://www.informedia.cs.cmu.edu/dli1/index.html. Accessed 1 Dec 2011
43. Informedia-ii digital video library: auto summarization and visualization across multiple video documents and libraries. http://www.informedia.cs.cmu.edu/dli2/index.html. Accessed 1 Dec 2011
44. Jain A, Yu B (1998) Document representation and its application to page decomposition. IEEE Trans Pattern Anal Mach Intell 20(3):294–308
45. Kaprykowsky H, Ndjiki-Nya P (2009) Restoration of digitized videos: efficient drop-out detection and removal. In: Proc of IEEE int'l conf on image processing (ICIP '09)
46. Kim Hg, Moreau N, Sikora T (2005) MPEG-7 audio and retrieval. Communication
47. Koeppel M, Doshkov D, Ndjiki-Nya P (2009) Fully automatic inpainting method for complex image content. In: Proc of int'l workshop on image analysis for multimedia interactive services (WIAMS'09)
48. Kompatsiaris Y, Hobson P (2008) Semantic multimedia and ontologies: theory and applications, 1st edn.
49. Konya I, Seibert C, Eickeler S, Glahn S (2009) Constant-time locally optimal adaptive binarization. In: Proc of 10th int'l conf document analysis and recognition. IEEE, Piscataway, pp 738–742
50. Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. Williamstown, MA, USA, pp 282–289
51. Lindbloom B (1994) Delta E (CIE 1994). In: Delta E (CIE 1994)
52. Linked data service of the German National Library. http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm. Accessed 1 Dec 2011
53. Liu M, Konya I, Nandzik J, Flores-Herr N, Eickeler S, Ndjiki-Nya P (2011) A new quality assessment and improvement system for print media. EURASIP (Special issue on image and video quality improvement techniques for emerging applications), submitted

54. Manjunath BS (2002) Introduction to MPEG-7, multimedia content description interface. Wiley, New York
55. MEDIAGLOBE—the digital archive, a SME project of the Theseus research program. http://www.projekt-mediaglobe.de/. Accessed 1 Dec 2011
56. Mediamill—semantic video search engine. http://www.science.uva.nl/research/mediamill/index.php. Accessed 1 Dec 2011
57. Mesh—multimedia semantic syndication for enhanced news services. http://www.mesh-ip.eu. Accessed 1 Dec 2011
58. Messina A, Sutter RD, Bailer W, Sano M, Evain JP, Ndjiki-Nya P, Schroeter B (2010) MPEG-7 audiovisual description profile (avdp). Report MPEG2010/M17744, MPEG (ISO/IEC JTC1/SC29/WG11)
59. METS—metadata encoding and transmission standard specification. http://www.loc.gov/standards/mets. Accessed 1 Dec 2011
60. MODS—metadata object description schema specification. http://www.loc.gov/standards/mods. Accessed 1 Dec 2011
61. Mufin player: a recommendation based music player. http://player.mufin.com/en. Accessed 1 Dec 2011
62. Müller S, Bühler J, Weitbruch S, Thebault C, Doser I, Neisse O (2009) Scratch detection supported by coherency analysis of motion vector fields. In: ICIP'09, pp 89–92
63. Nandzik J, Heß A, Hannemann J, Flores-Herr N, Bossert K (2010) Contentus—towards semantic multi-media libraries. In: Proc of 76th IFLA general conf and assembly (2010)
64. NISO metadata for images in XML (NISO MIX) schema. http://www.loc.gov/standards/mix. Accessed 1 Dec 2011
65. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
66. Petasis G, Karkaletsis V, Krithara A, Paliouras G, Spyropoulos C (2009) Semi-automated ontoloy learning: the Boemie approach. In: Proceedings of the 1st ESWC workshop on inductive reasoning and machine learning. Heraklion, Greece
67. Petersohn C (2004) Fraunhofer HHI at TRECVID 2004: shot boundary detection system. In: Proc TREC video retrieval evaluation workshop
68. Petersohn C (2009) Temporal video structuring for preservation and annotation of video content. In: Proc of IEEE int'l conf on image processing (ICIP '09)
69. PND, name authority file of the German National Library. http://www.d-nb.de/eng/standardisierung/normdateien/pnd.htm. Accessed 1 Dec 2011
70. Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. Boulder, CO, USA, pp 147–155
71. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted Gaussian mixture models. Digit Signal Process 10(1–3):19–41
72. RDA (Resource Description and Access): vocabularies. http://metadataregistry.org/rdabrowse.htm. Accessed 1 Dec 2011
73. RELATIONSHIP: a vocabulary for describing relationships between people. http://vocab.org/relationship/.html. Accessed 1 Dec 2011
74. Rushes—European research project on multimedia search and retrieval of rushes data. http://www.rushes-project.eu. Accessed 1 Dec 2011
75. Schneider D, Schon J, Eickeler S (2008) Towards large scale vocabulary independent spoken term detection: advances in the Fraunhofer IAIS audiomining system. In: Köhler J, Larson M, Jong de F, Kraaij W, Ordelman R (eds) Proc of the ACM SIGIR workshop "searching spontaneous conversational speech". Singapore
76. Skos (simple knowledge organization system): reference. http://www.w3.org/2004/02/skos/. Accessed 1 Dec 2011
77. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and trecvid. In: Proc of the 8th ACM int'l workshop on multimedia information retrieval, MIR '06. ACM, New York, pp 321–330
78. Smith K (2006) Capturing analog sound for digital preservation: report of a roundtable discussion of best practices for transferring analog discs and tapes
79. Snoek CGM, Worring M (2009) Concept-based video retrieval. Found Trends Inf Retr 4(2):215–322
80. Snoek CGM, Smeulders AWM (2010) Visual-concept search solved? IEEE Computer 43(6):76–78

81. Su X, Khoshgoftaar TM (2009) A survey of collaborative filtering techniques. Adv Artif Intell 2009:1–19
82. Theseus: a research program. http://www.theseus-programm.de. Accessed 1 Dec 2011
83. Tritschler A, Gopinath RA (1999) Improved speaker segmentation and segments clustering using the Bayesian information criterion. In: Proc of 6th European conf on speech communication and technology (EUROSPEECH'99). Budapest, Hungary, pp 679–682
84. Tsinaraki C, Christodoulakis S (2007) An MPEG-7 query language and a user preference model that allow semantic retrieval and filtering of multimedia content. Multimedia Syst 13(2):131–153
85. Ulges A, Schulze C, Keysers D, Breuel TM (2008) A system that learns to tag videos by watching Youtube. In: Proc of the 6th int'l conf on computer vision systems (ICVS'08). Springer, Berlin, pp 415–424
86. Verge—hybrid interactive video retrieval system. http://mklab.iti.gr/verge/. Accessed 1 Dec 2011
87. Vidi video: improving the accessibility of video. http://www.vidivideo.info. Accessed 1 Dec 2011
88. Vitalas (video and image indexing and retrieval in the large scale): a European fp6 research project. http://vitalas.ercim.org. Accessed 1 Dec 2011
89. Waitelonis J, Osterhoff JP, Sack H (2011) More than the sum of its parts: Contentus—a semantic multimodal search user interface. In: Proc of workshop on visual interfaces to the social and semantic Web (VISSW), co-located with ACM IUI 2011, 13 February 2011, Palo Alto, US, CEUR workshop proceedings, vol 694
90. Waitelonis J, Sack H (2010) Exploratory semantic video search with Yovisto. In: Proc of the 4th IEEE ICSC. Pittsburgh, PA, USA
91. Wgs84 geo positioning: an rdf vocabulary. http://www.w3.org/2003/01/geo/wgs84_pos. Accessed 1 Dec 2011
92. Witten IH, Bainbridge D, Nichols DM (2009) How to build a digital library, 2nd edn. Morgan Kaufmann, San Francisco
93. Worring M, Schreiber G (2007) Semantic image and video indexing in broad domains. IEEE Trans Multimedia 9(5):909–911
94. WS-BPEL: Web services business process execution language (specification). http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html. Accessed 1 Dec 2011
95. WS-RF: Web services resource framework (primer). http://docs.oasis-open.org/wsrf/wsrf-primer-1.2-primer-cd-02.pdf. Accessed 1 Dec 2011
96. Wu L, Hua XS, Yu N, Ma WY, Li S (2008) Flickr distance. In: Proceeding of the 16th ACM int'l conf on multimedia, MM '08. ACM, New York, pp 31–40
97. Yan R, Hauptmann AG (2007) A review of text and image retrieval approaches for broadcast news video. Inf Retr 10:445–484
98. Zheng Y, Liu C, Ding X, Pan S (2001) Form frame line detection with directional single-connected chain. In: Proc of int conf on document analysis and recognition (ICDAR). IEEE Computer Society, Los Alamitos, pp 699–703

Jan Nandzik is the co-founder of the research consultancy firm Acosta Consult (Frankfurt, Germany) and works as a principal consultant with a focus on mass digitisation, large-scale databases and

Semantic Web technologies. His scientific career started in the field of food chemistry at the Johann Wolfgang Goethe University in Frankfurt/Germany, where he was researching on enantioselective chromatography and received a diploma in 1998. Working in the industry research department of Goldwell (Darmstadt), a cosmetics company, he started his professional career and received his government certification as a food chemist after working for the Chemical Investigation Office (Kassel). Parallel to studying computer science he worked for the IT department of the Commerzbank until 2008, where he was responsible for process automation and business analysis as a project coordinator in several large-scale business intelligence data warehousing projects. As a freelancing consultant for the German National Library between 2006 and 2007 Jan Nandzik was involved in the multimedia Semantic Web project CONTENTUS, part of the government funded THESEUS research programme. Since 2008 he is member of the project steering group of CONTENTUS.

Berenike Litz is a researcher at the German National Library. Her main research interests are in the areas of natural language processing and artificial intelligence. She holds a PhD in Computer Science from Bremen University and a Master in Computational Linguistics from the University of Heidelberg.

Nicolas Flores-Herr is a partner at the German consulting firm Acosta Consult. He has a focus on mass digitisation and digitisation quality management, automated analysis as well as semantic retrieval of multimedia content. He worked as a scientist and as a project manager in the field of cultural heritage organisations (German National Library), in the industry sector (Sanofi Aventis) and in basic research (Max Planck Institute, University of Sydney). In the last years he worked for the project CONTENTUS which is part of the German research programme THESEUS (funded by the Federal Ministry of Economics and Technology) and for the German Digital Library (Deutsche Digitale Bibliothek).

Iuliu Konya holds two M.Sc. degrees—in Intelligent Systems from the Babes-Bolyai University, Romania in 2004 and in Media Informatics (summa cum laude) from the RWTH Aachen University, Germany in 2006. At present he works as a researcher at Fraunhofer IAIS, Germany and is a Ph.D. candidate at the University of Bonn on the topic robust document image understanding. His research interests lie in the areas of multimedia analysis and machine learning.

Doris Baum studied computer science in Nürnberg, Edinburgh, and Vienna and obtained a Bachelor's and a Master's degree (Dipl.-Ing.) in 2004 and 2007, respectively. She was involved in the Contentus project as a PhD student in the field of audio analysis, specialising in automatic speaker recognition. Her research interests include machine learning, information retrieval, and automatic audio analysis, particularly for music and speech.

André Bergholz is a senior research engineer and project manager at Fraunhofer IAIS. He coordinates the text mining activities in the CONTENTUS project. His areas of specialization include text analytics, data mining, and data management. André Bergholz holds a PhD in Computer Science from Humboldt-University Berlin. Prior to joining Fraunhofer he was working at Xerox Research Centre Europe and Stanford University.

Dirk Schönfuß has studied business informatics at the University of Technology in Mittweida. Before building up and leading research and development at mufin he had worked as a consultant and freelance software developer for many years.

Christian Fey studied Television Technology and Electronic Media at the University of Applied Sciences FH Wiesbaden. He received his Dipl.-Ing. Degree in 2007 and started working at the IRT shortly after graduating. The topic of his diploma thesis was the “Development of a software module for automatic analysis of mxf-files”. He is active at the IRT in the project CONTENTUS as software developer with experience in C++, C#, Java and XML coding. Furthermore, he is experienced in software development for video file formats metadata modeling.

Johannes Osterhoff is not a media but an interface artist. Since the rounded buttons of Windows were replaced by cornered ones during its redesign in 1995, he keeps a wary eye on the more and more baroque graphical user interfaces of contemporary pop culture media and reflects on this topic as scientist, artist and designer. Osterhoff lives in Berlin, researches in the field of semantic information retrieval interfaces at Hasso-Plattner-Institute in Potsdam and lectures Interface Design at Berliner Technische Kunsthochschule and at Merz Akademie in Stuttgart.

Jörg Waitelonis is Research Assistant at the Hasso Plattner-Institute for IT-Systems Engineering (HPI) at the University of Potsdam. After graduating in computer science at the Friedrich-SchillerUniversity in Jena in 2006 he developed the video search engine yovisto.com. Yovisto started as an ESF/BMWi (European Social Fund/German Government) funded project with the objective to develop a video search engine for academic lecture recordings and was relocated from FriedrichSchiller-University Jena to Hasso Plattner-Institute to become a research platform for semantic multimedia technologies. Furthermore, Jörg worked on the multimedia processing system REPLAY at Swiss Federal Institute of Technology Zürich.

Harald Sack is Senior Researcher at the Hasso Plattner-Institute for IT-Systems Engineering (HPI) at the University of Potsdam. After graduating in computer science at the University of the Federal Forces Munich Campus in 1990, he worked as systems/network engineer and project manager in the signal intelligence corps of the German federal forces from 1990–1997. In 1997 he became an associated member of the graduate program 'mathematical optimization' at the University of Trier and graduated with a PhD thesis on formal verification in 2002. From 2002–2008 he did research and teaching as a postdoc at the Friedrich-Schiller-University in Jena and since 2007 he has a visiting position at the HPI, where he is head of a research group on semantic technologies. His areas of research include multimedia retrieval, semantic web, knowledge representations and semantic enabled retrieval. Since 2008 he is also general secretary of the German IPv6 council.

Ralf Köhler has been a Principal Scientist, Distinguished R&D Engineer and fellowship member at Deutsche Thomson OHG, Technicolor R&I since 2000. Since 2009 he leads an activity in R&I developing content enhancement software technologies for Technicolor broadcast and cinematography applications. He is a system architect and software expert for high-end multimedia solutions. 1995–2000 he was with SICAN GmbH as system architect and project lead of customer oriented product development. He received his Dipl.-Ing. in Electrical Engineering and Computer Science 1995 at University of Hannover, Germany.

Patrick Ndjiki-Nya received the Dipl.-Ing. title (corr. to M.S degree) from the Technische Universität Berlin in 1997. In 2008 he also completed his doctorate at the Technische Universität Berlin. He is currently a group manager at Fraunhofer Heinrich Hertz Institute, where he has been working since 1998.