> CONTENTUS – NEXT GENERATION MULTIMEDIA LIBRARY

NICOLAS FLORES-HERR/STEFAN EICKELER/JAN NANDZIK/STEFAN PAAL/
IULIU KONYA/HARALD SACK

CONTENTUS ist ein Anwendungsszenario des Forschungsprogramms THESEUS, welches durch das Bundesministerium für Wirtschaft und Technologie gefördert wird. Die Deutsche Nationalbibliothek (DNB) leitet das Projekt. Weitere Partner sind die Deutsche Thomson OHG (DTO, ein Tochterunternehmen von Technicolor), das Institut für Rundfunktechnik (IRT), das Fraunhofer Institut für Nachrichtentechnik - Heinrich-Hertz-Institut (HHI), das Fraunhofer Institut für Intelligente Analyse- und Informationssysteme (IAIS), das Hasso-Plattner-Institut (HPI) und die mufin GmbH.

ZIELE VON CONTENTUS

In rund 30.000 Bibliotheken, Museen und Archiven Deutschlands lagern Millionen von Büchern, Bildern, Tonbändern und Filmen. Das Forschungs- und Entwicklungsprojekt CONTENTUS schafft technische Lösungen und Konzepte, wie dieses kulturelle Erbe einer möglichst großen Zahl von Menschen zugänglich gemacht werden kann. Konkret unterstützt das Projekt Kultureinrichtungen und Informationsanbieter darin, einen internetbasierten und medienübergreifenden Zugriff auf Wissens- und Kulturgüter anzubieten.

Multimediale Sammlungen von Bibliotheken, Medienarchiven und Sendeanstalten werden durch die von THESEUS und CONTENTUS geschaffene internetbasierte Wissensinfrastruktur zum Teil zu einer neuen Informations- und Interaktionskultur: der Next-Generation Multimedia Library (siehe Abbildung 1).

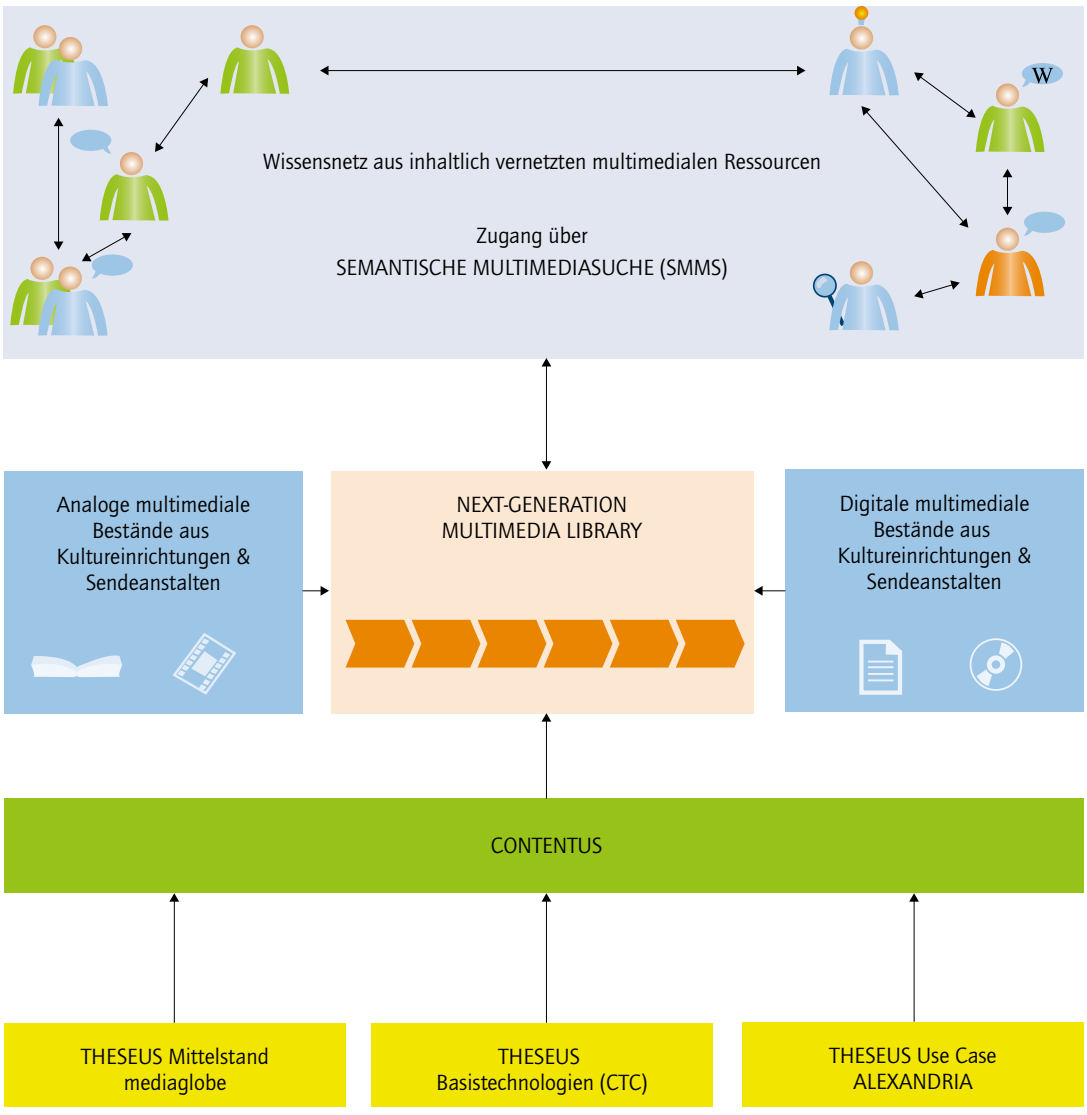
Im Zentrum steht der semi-automatische Aufbau eines Wissensnetzes: ein Netz aus Fakten, die den Bestand beschreiben und seine inneren Zusammenhänge sichtbar machen – Fakten, die sowohl von einem Experten oder interessierten Nutzer als auch, rein maschinell, durch einen Algorithmus zur Inhaltsanalyse von Medien hervorgebracht werden können.

Eine semantische Multimediasuche (SMMS) ermöglicht dem Nutzer einen intuitiven Zugang zu und eine medienübergreifende Suche in diesem Wissensnetz (siehe Abbildung 1, grauer Kasten). Des Weiteren wird es Anbietern von Inhalten (blau) ermöglicht, effizient die Qualität ihrer digitalisierten multimedialen Inhalte automatisch zu bestimmen und diese an die Nutzeranforderungen anzupassen. Mit Hilfe der Technologien in CONTENTUS kann ein effizientes Qualitätsmanagement für digitalisierte multimediale

Inhalte realisiert werden, was zum Beispiel bei einer Einspeisung von Ressourcen (mit hohem Durchsatz und Datenvolumen) in das Wissensnetz nutzbringend eingesetzt werden kann.

Grundlage für die Next-Generation Multimedia Library ist eine kohärente Sammlung von Methodologien, Werkzeugen und Verfahren für Inhalteanbieter mit großen multimedialen Beständen. Zur Erreichung dieses Ziels kooperiert CONTENTUS (grün) mit verschiedenen Vorhaben aus THESEUS: dem Anwendungsszenario ALEXANDRIA (gelb), dem Core Technology Cluster (gelb) und MediaGlobe (gelb), einem KMU-Projekt.

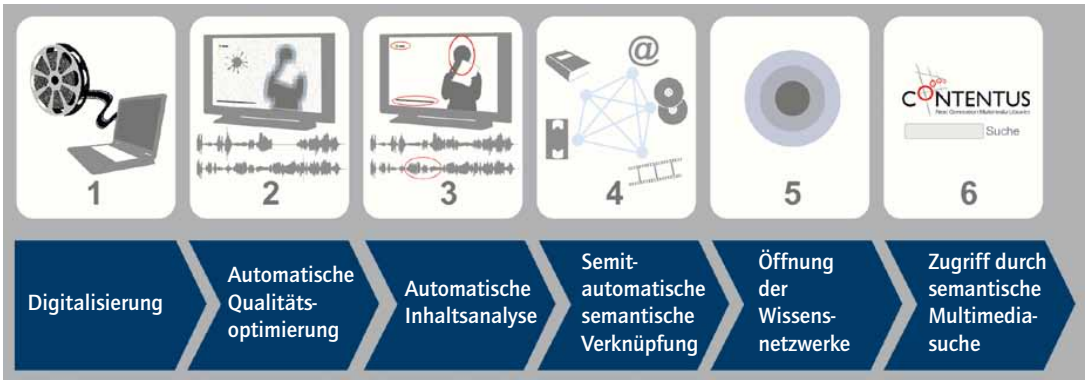
Abbildung 1: CONTENTUS als Technologielieferant für die Next-Generation Multimedia Library, einer internetbasierten Wissensinfrastruktur.



TECHNOLOGIEN

Zur Realisierung der Vision der Next-Generation Multimedia Library werden im Rahmen von CONTENTUS Technologien und Konzepte zur effizienten Verarbeitung multimedialer Inhalte entwickelt. Diese dienen dazu, multimediale Bestände aufzubereiten, zu erschließen und schließlich zugänglich zu machen. Ziel ist es, eine umfassende Lösung für Bibliotheken und Medienarchive anzubieten, die alle notwendigen Verarbeitungsschritte beinhaltet. Zusammen ergibt sich eine Prozesskette, wie sie in Abbildung 2 dargestellt ist.

Abbildung 2: CONTENTUS-Prozesskette zur Verarbeitung analoger und digitaler multimedialer Inhalte. CONTENTUS-Technologien decken alle essenziellen Verarbeitungsschritte vom analogen Medium bis hin zur semantischen Multimediasuche ab.



Die jeweiligen zur Realisierung der Prozesskette benötigten Technologien sind auf Bedürfnisse von Bibliotheken, Medienarchive und Rundfunkanstalten abgestimmt. Dies wird durch die Deutsche Nationalbibliothek und das Institut für Rundfunktechnik sichergestellt

Neben dem für analoge Medien notwendigen ersten Schritt der Digitalisierung sind folgende Schritte wichtig:

- **Automatische Qualitätsoptimierung:** Mittels Hochdurchsatzverfahren ist es möglich, in kurzer Zeit große Mengen analogen Materials zu digitalisieren. Je schneller der Prozess ist, desto höher ist aber auch die Wahrscheinlichkeit, dass einzelne digitale Abbilder fehlerbehaftet sind – Druckseiten verknicken, Filme verkratzen oder verschmutzen, Videobänder werden beschädigt, Tonaufnahmen übersteuern, oder einzelne Seiten und Titel fehlen ganz. Darüber hinaus kann der Archivbestand aus Lagerungs- und Veralterungsgründen schon beschädigt vorliegen. Auch wenn die Digitalisierung ohne Fehler verläuft, kann es sein, dass

die Originale bereits mit Qualitätsproblemen behaftet sind (z. B. Flecken und Risse auf Buchseiten, Kratzer und Staub bei audiovisuellen Aufnahmen), die sich auf die digitalisierten Inhalte übertragen. Deren manuelle Prüfung und Aufbereitung ist einer der größten Kostenfaktoren beim Aufbau von digitalen Archiven. CONTENTUS widmet sich dieser Problematik durch die Entwicklung von Verfahren zur automatischen Erfassung und Optimierung der Qualität und zur Erzeugung von Präsentationsformaten. Für Bibliotheken und Sendeanstalten soll die Qualität großer Mengen digitalisierter Inhalte effizient erfassbar werden, so dass im Fehlerfall eine Neudigitalisierung stattfinden kann, bevor der analoge Träger weiterem alterungsbedingten Verfall anheimfällt. Weiterhin können zu jeder Zeit digitalisierte multimediale Inhalte in schlechter Qualität mit neuen Verfahren gezielt erneut optimiert werden. Da die Güte der automatisch extrahierten Metadaten – insbesondere der deskriptiven Metadaten, welche für die Suche relevant sind – oft stark von der Bild-, Ton- bzw. Videoqualität abhängt, ist ein effizientes Qualitätsmanagement von digitalisiertem Material von großer Wichtigkeit für den späteren Zugriff auf die Inhalte.

- **Automatische Inhaltsanalyse:** Die Nutzererwartungen hinsichtlich der Durchsuchbarkeit sind durch das digitale Abbild eines analogen Mediums, wie es aus der Analog/Digital-Wandlung bzw. aus der Qualitätsoptimierung kommt, in der Regel noch nicht erfüllt. Ein Abbild einer Buchseite, eine Videodatei eines Filmes oder eine Audiodatei eines Tonbandes sind, ohne ihren Inhalt zu kennen, für den Nutzer praktisch wertlos, insbesondere wenn es sich um große Mengen von Inhalten handelt. Darum werden automatische Verfahren zur inhaltlichen Beschreibung der Medien auf die unbearbeiteten „Rohinhalte" angewendet, um ihre Auffindbarkeit deutlich zu erhöhen. Zum Einsatz kommen u. a. Verfahren zur Strukturierung, zur Objekt- und Texterkennung, Personen- und Stichworterkennung, Sprecher- und Spracherkennung sowie zur Ermittlung musikalischer Eigenschaften. Zudem können sich kulturelle Institutionen, durch solche automatische Verfahren entlastet, auf ihre Kernkompetenz konzentrieren: die intellektuelle Erschließung von Medien. Sprache, Musik, Videos und andere Dokumente werden mit Hilfe von Technologien aus CONTENTUS maschinenlesbar – dies bildet die Grundlage dafür, dass man die Inhalte später automatisch kategorisieren sowie gezielt nach Werken mit bestimmten Eigenschaften suchen kann (siehe Abbildung 2 Schritt 3).
- **Semiautomatische semantische Verknüpfung:** Zunächst werden, basierend auf den zuvor extrahierten Metadaten, inhaltlich verwandte Medien identifiziert und miteinander verknüpft. Zusätzlich können die Inhalte auf diese Weise auch mit externen Informationsquellen – z. B. den bereits vorhanden Metadaten kultureller Einrichtungen oder kollaborativ gepflegten Datenbeständen aus der Wikipe-

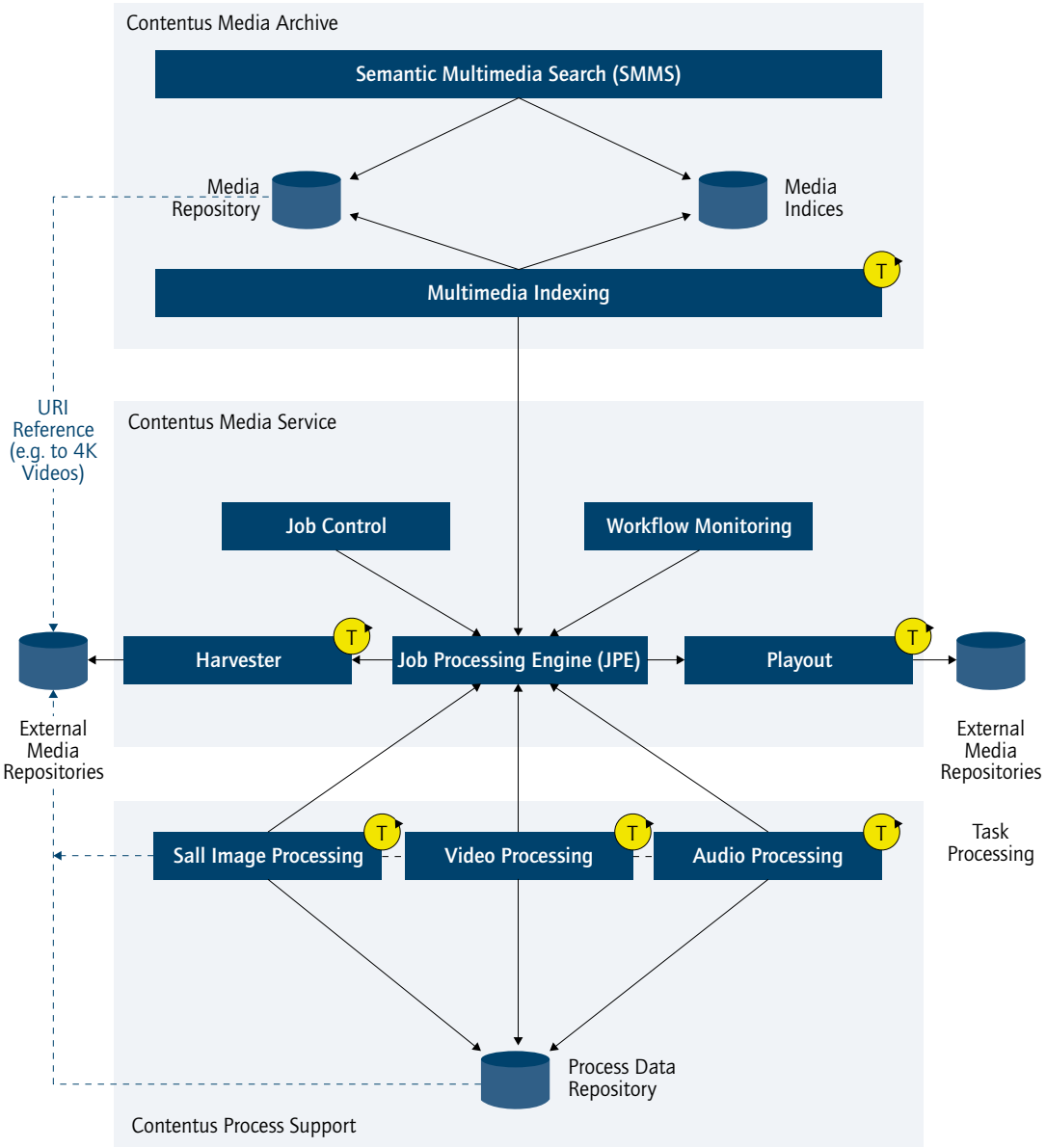
dia – in Verbindung gebracht werden. Dabei werden die gewonnenen Fakten semantisch vernetzt und mit den externen Informationen so in Beziehung gesetzt, dass sie von Computern interpretiert und weiterverarbeitet werden können. Mit Hilfe einer formalen Sprache (z. B. dem W3C-Standard Resource Description Framework (RDF)) kann beispielsweise modelliert werden, dass Person A mit Person B verwandt ist, Musikstück 1 ähnlich zu Musikstück 2 ist oder es kann eine direkte Verbindung eines im Video gefundenen bekannten Politikers mit allen Beiträgen aus dem News-Bereich hergestellt werden. Die Verwendung einer derart modellierten Wissensbasis eröffnet neue Nutzungsszenarien wie maschinelles Schlussfolgern („Reasoning"), so dass neben einer klassischen Volltextsuche auch komplexere Anfragen durchführbar sind: z. B. „Finde langsame Gitarrenstücke von einer Band, in der Eric Clapton einmal Mitglied war". Darüber hinaus eröffnen standardisierte Technologien und Konzepte, insbesondere aus dem „Semantic Web Stack" des W3Cs, Kultureinrichtungen und Informationsanbietern neue Möglichkeiten für Interoperabilität und Datenaustausch.

- **Öffnung der Wissensnetze und Community Building:** Durch die Veröffentlichung von Metadaten als Linked Open Data und die Einbindung sozialer Wissensnetzwerke können Inhalte und Metadaten mit einem großen Kreis von Benutzern geteilt werden. Linked Open Data und soziale Netzwerke werden in Zukunft sowohl die Sichtbarkeit unseres digitalisierten kulturellen Erbes im Internet erhöhen als auch eine tragende Rolle bei der Ausweitung der Kooperationsradien von Kultur- und Gedächtniseinrichtungen spielen. Bei der Vernetzung mit anderen Institutionen stehen neben weiteren Kultureinrichtungen zusätzlich kommerzielle Inhalteanbieter (z. B. Verlage, Video-Portale, Tonträgerhersteller, Buchhandel) im Fokus. Im Rahmen der Kooperation der THESEUS-Projekte CONTENTUS und ALEXANDRIA werden in Zukunft Konzepte und Technologien für die Einbeziehung von Netzwerken aus Nutzern erarbeitet.
- **Semantische Multimediasuche:** Ein wichtiger Bestandteil der Forschungs- und Entwicklungsarbeiten in CONTENTUS ist die semantische Multimediasuche. Hierbei sollen, durch semantische Technologien gestützt, Endnutzer Zugang zu den Inhalten aus dem Wissensnetz (Abbildung 1) erhalten. Weiterhin sollen sie in die Lage versetzt werden, zufällige und unerwartete Inhalte zu finden (Serendipity) und explorativ in multimedialen Sammlungen zu stöbern. Auf diese Weise eröffnet CONTENTUS neue Möglichkeiten zur Wissensnavigation und -recherche.

CONTENTUS: DIENSTBASIERTER WORKFLOW ZUR MEDIENVERABEITUNG

Für die automatisierte Erschließung von Multimediadaten wurde eine Diensteplattform konzipiert und realisiert, welche die verschiedenen CONTENTUS-Prozessierungskomponenten in die medienspezifischen Workflows einbindet. Die Diensteplattform besteht dazu aus einer Job Processing Engine (JPE), die die Workflow-Steuerung übernimmt und aus sogen. Task Prozessoren (TP), die jeweils einen Schritt ausführen (siehe Abbildung 3).

Abbildung 3: Diensteplattform zur automatisierten Erschließung von Multimediainhalten

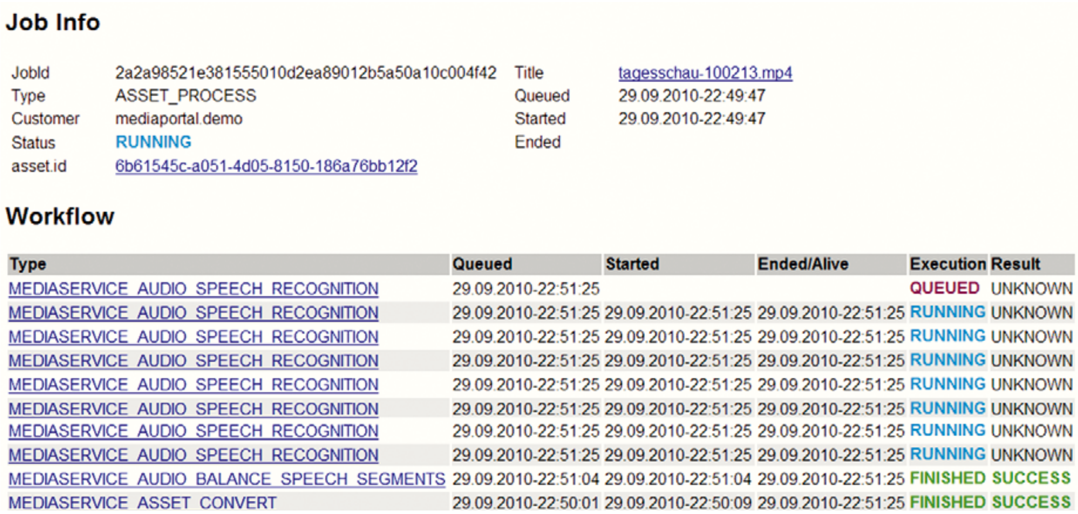


Ein besonderes Merkmal ist die beliebige Verteilung und Skalierung der TP, z. B. in einer Cloud, zur Verarbeitung von großvolumigen Medienbeständen. Damit konnte im Projekt vor allem die aufwendige Audioanalyse um ein Vielfaches beschleunigt werden. Dazu gehört auch die Möglichkeit zur Nutzung spezifischer TP „vor Ort“, zum Beispiel einer Kultureinrichtung. Dieser Betriebsmodus bietet bei der zur Verarbeitung großer Multimediaobjekte Vorteile, weil hierdurch die Verarbeitungszeit zu verkürzt wird (z. B. bei der Segmentierung und Konvertierung von HD-Videos). Weiterhin können TP auch dynamisch zur Laufzeit hinzugefügt und entfernt werden, so dass gerade in Cloud-Umgebungen eine dynamische Anpassung der Verarbeitungsleistung und -merkmale vorgenommen werden kann.

Über eine Webservice-Schnittstelle kann die Diensteplattform und die CONTENTUS Prozesskomponenten auch in anderen Anwendungssystemen, wie z. B. die Deutsche Digitale Bibliothek (DDB)[1], eingebunden werden. Externe Medienquellen können dazu eingelesen (Harvester), erschlossen und die Verarbeitungsergebnisse (Playout) anschließend wieder ausgespielt werden. Das System ist hierzu mehrmandantenfähig, so dass die gleiche Infrastruktur verschiedene Anwendungsszenarien bedienen kann.

Für die Kontrolle der aktuellen Verarbeitung durch einen Operator steht ein Webinterface (Workflow Monitoring) zur Verfügung, welches die einzelnen Jobs und die damit verbundenen Tasks visualisiert. In Abbildung 4 ist der Verarbeitungsfortschritt einer Prozessierung von audiovisuellen Inhalten zu sehen, in der beispielhaft die Audiotranskription durch die Workflowsteuerung adaptiv auf mehrere Task Prozessoren verteilt wurde. Der Ausführungszustand eines einzelnen Jobs aber auch des kompletten Systems kann darüber jederzeit leicht abgerufen werden.

Abbildung 4: Workflow Monitoring der automatisierten Prozessierung von digitalen audiovisuellen Inhalten



Job Info

JobId	2a2a98521e381555010d2ea89012b5a50a10c004f42	Title	tagesschau-100213.mp4
Type	ASSET_PROCESS	Queued	29.09.2010-22:49:47
Customer	mediaportal.demo	Started	29.09.2010-22:49:47
Status	RUNNING	Ended	
asset.id	6b61545c-a051-4d05-8150-186a76bb12f2		

Workflow

Type	Queued	Started	Ended/Alive	Execution	Result
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25			QUEUED	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_SPEECH_RECOGNITION	29.09.2010-22:51:25	29.09.2010-22:51:25	29.09.2010-22:51:25	RUNNING	UNKNOWN
MEDIASERVICE_AUDIO_BALANCE_SPEECH_SEGMENTS	29.09.2010-22:51:04	29.09.2010-22:51:04	29.09.2010-22:51:25	FINISHED	SUCCESS
MEDIASERVICE_ASSET_CONVERT	29.09.2010-22:50:01	29.09.2010-22:50:09	29.09.2010-22:51:25	FINISHED	SUCCESS

1 http://www.deutsche-digitale-bibliothek.de/

Neben der Überwachung der eigentlichen Verarbeitung stehen weitere Administrationsansichten bereit, die eine Auswertung und Begutachtung der Verarbeitungsergebnisse im Workflow erlauben.

Die wiederkehrenden Arbeitsschritte zur inhaltlichen Erschließung werden modularisiert. Aus ihnen können dadurch spezifische Verarbeitungsketten für neue Medientypen erstellt werden, welche automatisiert und parallelisiert ablaufen, um eine unbeaufsichtigte Massenverarbeitung zu ermöglichen.

Innerhalb von CONTENTUS wurden in Abhängigkeit des Medientyps maßgeschneiderte Workflows definiert und realisiert: Printdokumentverarbeitung (z. B. Zeitungen, Bücher und Presseclippings), Fotos und audiovisuelle Daten. Die Ergebnisse dieses medientypabhängigen Workflows werden anschließend in einem gemeinsamen Workflow semantisch angereichert und verlinkt.

Im Folgenden werden anhand des Beispiels der Medientypen Print und Audio/Video die Prozessierung von Inhalten detailliert dargestellt.

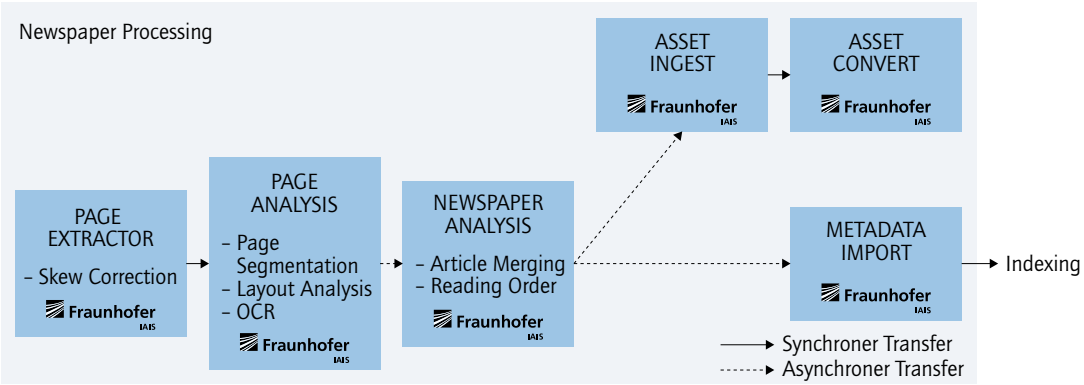
PRINTDOKUMENTENVERARBEITUNG

Speziell für die Verarbeitung von digitalisierten Printdokumenten wurden für das CONTENTUS-Anwendungsszenario Workflows konzipiert und Medien mit der Diensteplattform verarbeitet. In Abschnitt 0 wird beispielhaft die Verarbeitung von Büchern und Zeitungen vorgestellt und detailliert beschrieben (siehe Abbildung 5 und Abbildung 12).

ZEITUNGEN

Der Workflow zur Zeitungsverarbeitung hat die meisten Arbeitsschritte. Der erste Schritt ist die Freistellung der Seiten mit Korrektur von Verzerrungen. Die folgenden Schritte sind die geometrische und die logische Layoutanalyse. Erst danach werden die Digitalisate und die Ergebnisse ins Repository eingelesen.

Abbildung 5: Workflow für die Verarbeitung digitalisierter Zeitungen



GEOMETRISCHE LAYOUTANALYSE

Die geometrische (oder physikalische) Layoutanalyse unterscheidet zwischen den verschiedenen Bereichen einer gedruckten Seite. Es werden Text, Bilder, Grafiken, sowie trennende horizontale und vertikale Linien und trennende Bereiche identifiziert. Im Rahmen von CONTENTUS wurde ein Seitensegmentierer entwickelt, der insbesondere bei schwierigen Layouts, wie es bei Büchern, Zeitungen und Zeitschriften vorkommt, gute Ergebnisse liefert.

Es wurden verschiedene Verfahren kombiniert und optimiert:

1) **Vorverarbeitung.** Es wird eine Binarisierung des Zeitungsscans durchgeführt.
2) **Black Separator Detection.** Die Qualität der horizontalen und vertikalen Separierungslinien werden durch Morphologische Operatoren verbessert und anschließend die „Directional Single-Connected Chain" (DSCC) extrahiert. Die resultierenden DSCC werden anhand ihrer Größe gefiltert. Folgendes Bild zeigt die DSCC und die Filterschritte:

Abbildung 6: Darstellung der DSCC und Filterschritte.

3) **White Separator Detection.** Die maximalen leeren Rechtecke werden detektiert. Hierbei müssen einige Randbedingungen eingehalten werden. So muss zum Beispiel die Höhe eines trennenden Bereiches deutlich größer als die Buchstabengröße sein. Die horizontalen weißen Separatoren sind grün dargestellt:

Abbildung 7: White Separator Detection. Die horizontalen weißen Separatoren sind grün dargestellt.

Warum jedes Mal das Rad neu erfinden? Wenn Behörden in unterschiedlichen Bundesländern einheitliche Technologien einsetzen, spart das Entwicklungskosten. Die fachlichen Anwendungen müssen nur einmal analysiert und in IT-Konzepten umgesetzt werden. Nur für welches Modell sollte man sich entscheiden? Bei der Suche nach einer geeigneten KOOPERATION kann das Fraunhofer E-Government-Zentrum behilflich sein. Die Experten treten als Mediatoren auf. Diese Vernetzung macht allein schon deshalb Sinn, weil sich viele Politikfelder nicht lokal eingrenzen lassen. Auch in den Verwaltungen lässt sich der digitale Fortschritt nicht mehr zurückdrehen: Die Informationsgesellschaft ist ohne ein intelligentes E-Government als entscheidendes »Betriebssystem« undenkbar geworden.

Bilder: MEV

Umweltschutz ist Ländersache: *Da sich Luft und Gewässer naturgemäß nicht nach Ländergrenzen richten, müssen einheitliche IT-Werkzeuge zur Kontrolle der Wasserqualität geschaffen werden.*

KOOPERATION

Gewässerschutz kennt keine Ländergrenzen

Flexibles Software-System informiert über die Wasserqualität von Gewässern

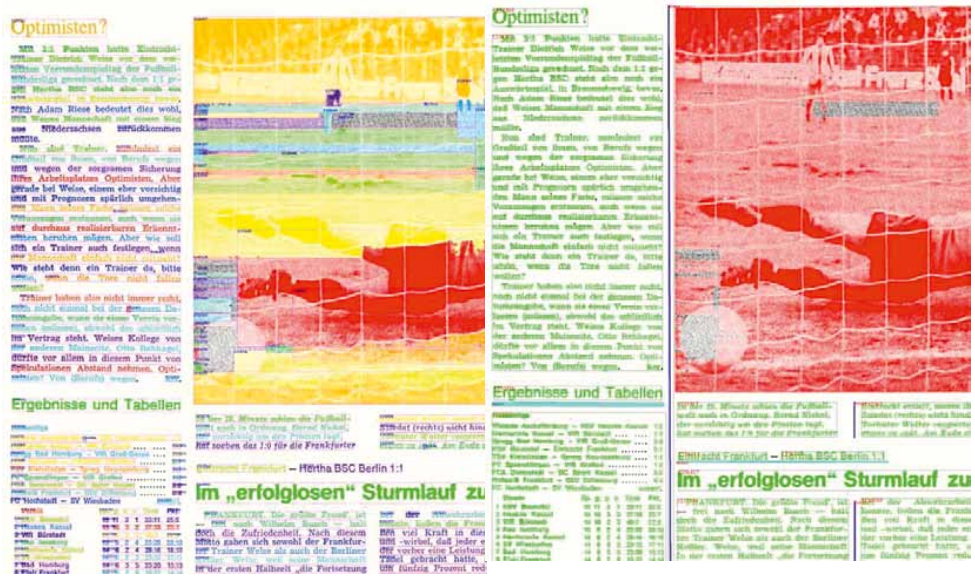
Karlsruhe, Fraunhofer IITB – Neun europäische Länder durchquert der Rhein bevor er in die Nordsee mündet. Offensichtlich keine einfachen Bedingungen für ein integriertes Wassermanagement, wie es die Europäische Wasserrahmenrichtlinie fordert. Bis 2015 müssen alle europäischen Gewässer zumindest einen »guten Zustand« erreicht haben. Die Kontrolle der Wasserqualität innerhalb der natürlichen Flussgrenzen anstatt wie bisher innerhalb von Verwaltungseinheiten bringt eine Herausforderung für die Informationstechnologie der betroffenen Behörden. Damit die gesammelten Daten aus der Umweltüberwachung länderübergreifend nutzbar sind, müssen einheitliche IT-Systeme und Standards geschaffen werden. Forscher des Fraunhofer IITB haben hierfür den Software-Baukasten »WaterFrame®« entwickelt. Auch Mitarbeiter ohne besondere IT-Kenntnisse können so ohne Probleme neue Daten importieren oder exportieren. Benutzerdefinierte Objekte wie Parameterlisten oder Gruppierungen von Messstellen gestatten dem Sachbearbeiter die Anwendung auf seine Bedürfnisse zuzuschneiden. Benötigt er beispielsweise Daten über die Konzentration des Pflanzenschutzmittels Atrazin im Grundwasser – allerdings nur in der Nähe von Flüssen und nur für die Jahre 1985 bis 1990 – so gibt er diese Kriterien einfach in die Maske von »WaterFrame®« ein. Als Suchergebnis erhält er Tabellen, Diagramme oder Karten, die nach Experten-Vorgaben erstellt wurden. »WaterFrame« vernetzt bereits die Bundesländer Baden-Württemberg, Bayern und Thüringen.

Marktreife: 2006
Innovationsgrad:
Mehr Informationen: Webkey • 20644

4) **Page Segmentation.** Hier wird ein hybrider Ansatz verwendet, der aus einem „Bottom-Up" Prozess, der von Top-Down Informationen über die Spalten gesteuert wird, besteht. Die Informationen über die Spalten wurden durch dynamische Programmierung aus den Listen der Separatoren erzeugt. Die Unterscheidung zwischen Text und Nicht-Text wird anhand der statistischen Eigenschaften von Text entschieden.

Folgendes Bild zeigt die vereinfachten Zeilen eines Bildes vor und nach der Filterung:

Abbildung 8: Page Segmentation. Die Bilder zeigen die vereinfachten Zeilen eines Bildes vor (links) und nach (rechts) der Filterung.

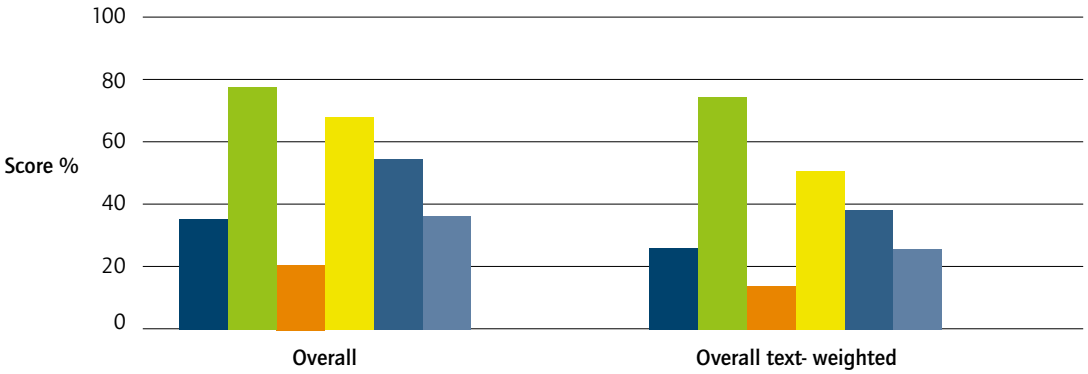


5) **Text Line and Region Extraction.** Die Erkennung der exakten Zeilen ist eine Verfeinerung der Textbereichserkennung des vorhergehenden Schritts. Textregionen mit ähnlichen Eigenschaften werden anhand von Merkmalen wie der Strichstärke und der x-Höhe bestimmt.

Abbildung 9: Ergebnis der Seitensegmentierung auf einer Seite der Jahreschronik (rot: Foto, grün: Text, blau: weiße Seperatoren, türkis: horizontale Seperatoren, dunkelblau: vertikale Seperatoren)

Das Seitensegmentierungsverfahren hat an dem „Page Segmentation Contest" der „International Conference on Document Analysis and Recognition" teilgenommen. Hier hat es mit deutlichem Abstand – noch vor Google, welches unter „OCRopus" firmiert – den ersten Platz belegt (siehe Abbildung 10).

Abbildung 10: CONTENTUS-Technologie siegt im Wettbewerb. PRImA Segmentierungsmaß auf dem ICDAR09 Page Segmentation Competition Datensatz.



DICE	35,31	25,97
Fraunhofer	77,55	73,74
REGIM-ENIS	20,81	13,94
Tesseract	68,15	50,30
FineReader	54,52	37,90
OCRopus	36,19	25,59

LOGISCHE LAYOUTANALYSE

Die logische Layoutanalyse segmentiert eine Dokumentenseite in die einzelnen Artikel. Diese Zerlegung ist besonders für Zeitschriften und Zeitungen wichtig, da Artikel zu unterschiedlichen Themen auf einer Seite vorhanden sind. Es wäre für eine weitergehende semantische Analyse sehr störend, wenn diese Artikel zu einem Seitentext kombiniert würden. Bei Büchern sind ähnliche Probleme vorhanden. In der Regel ist hier aber die Anzahl von Artikeln oder Kapiteln geringer.

Die Artikelseparierung arbeitet durch die Kombination von mehreren Verfahren:

1) Alle Blockpaare der Seitensegmentierung werden betrachtet und diesen Paaren werden Kantengewichte zugeordnet. Die Kantengewichte ergeben sich aus der geometrischen Anordnung, Schriftähnlichkeit und der Anzahl der Separatoren zwischen den Blöcken.
2) Der „Minimum Spanning Tree" der Blöcke mit den Kantengewichten wird berechnet.
3) Eine vorläufige Artikelsegmentierung wird über die Textgrößen vorgenommen. Hierbei wird davon ausgegangen, dass die Artikelüberschrift größer als der Artikeltext ist.

4) Über die Lesereihenfolge können die Blöcke eines Artikels richtig sortiert werden.
5) Mit Hilfe von Nachbearbeitungsregeln werden verbleibende Artikelblöcke ohne Überschrift bestehenden Artikeln zugeordnet.

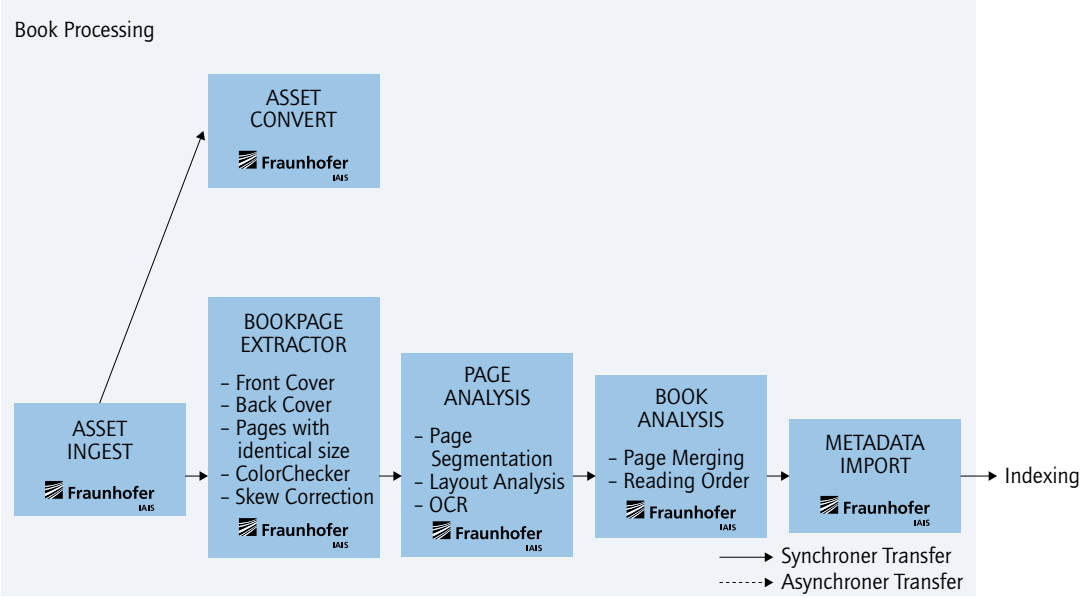
Abbildung 11: Verarbeitungsschritte bei der Artikelsegmentierung: Initialer Graph, Minim Spanning Tree, vorläufige Artikelsegmentierung und Lesereihenfolge, fusionierte Artikel

BÜCHER

Die Richtlinien zur Buchdigitalisierung sind für die exakte Verarbeitung zu ungenau. Informationen über die Reihenfolge der Inhalte werden bei der Digitalisierung nicht aufgenommen und müssen daher automatisch bestimmt werden.

Eine weitere wichtige Randbedingung für die Seitenextraktion bei Büchern ist die exakt gleiche Größe aller Seiten für den Page-Flip Effekt bei der Buchdarstellung. Die optimale Seitengröße wird anhand von Samples innerhalb des Buches bestimmt. Anschließend wird die Seitengröße auf alle Seiten des Buches bei der Seitenextraktion angewendet.

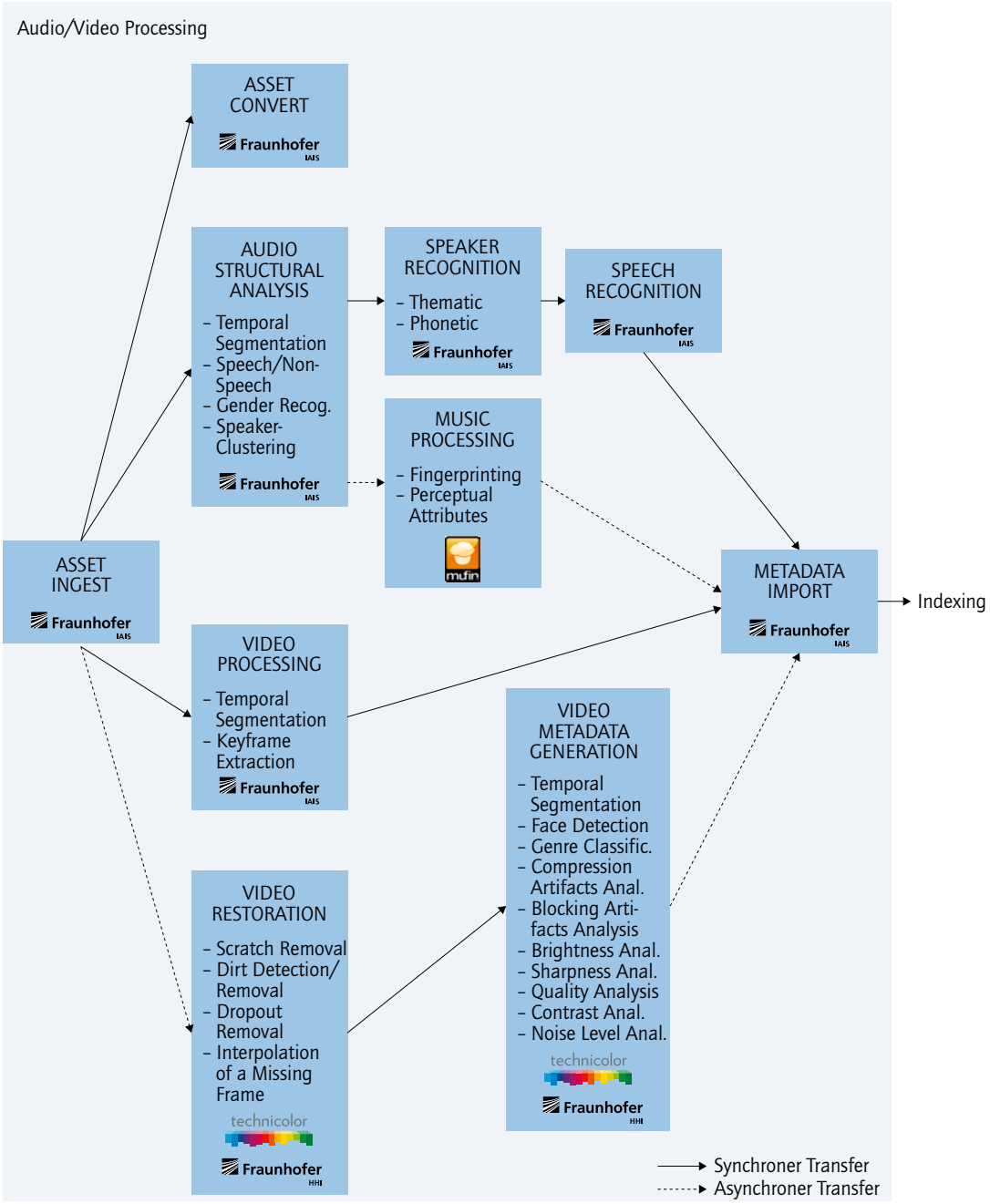
Abbildung 12: Workflow für die Verarbeitung von digitalisierten Büchern



VERARBEITUNG VON AUDIO/VIDEO MATERIAL

Der Workflow für die Verarbeitung von A/V kann Videosequenzen mit Ton sowie Videosequenzen ohne Ton und reine Audioaufnahmen verarbeiten (siehe Abbildung 13). Hier sind die Verarbeitungs-Module vom Fraunhofer IAIS synchron in den Workflow eingebunden. Für die Verarbeitung von großen Mengen von Videosequenzen gibt es ein vereinfachtes Videoverarbeitungsmodul von IAIS.

Abbildung 13 Workflow für die Verarbeitung von audiovisuellen Inhalten



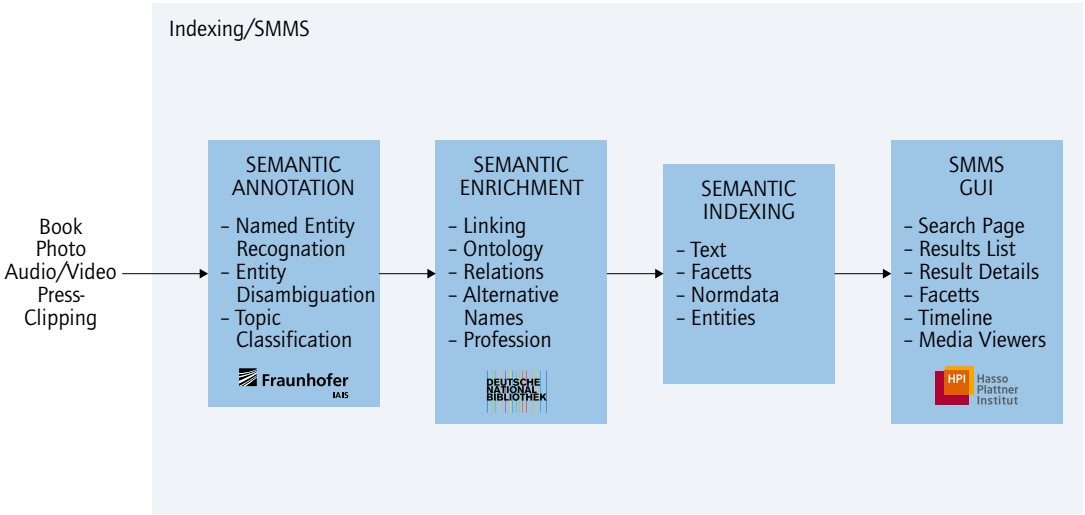
Für die manuelle Qualitätskontrolle von automatisch erzeugten Metadaten wurde ein Webinterface geschaffen, welches diese übersichtlich präsentiert (siehe Abbildung 14)

Abbildung 14: Qualitätskontrolle der automatisierten AV-Daten-Prozessierung

WORKFLOW FÜR SEMANTISCHE ANREICHERUNG UND SEMANTISCHE SUCHE

Die generierten Metadaten der vorherigen Verarbeitungsketten werden semantisch angereichert und in der semantischen Multimediasuche verwendet (siehe Abbildung 15).

Abbildung 15: Workflow für die Semantische Anreicherung und die Semantische Suche



Dabei werden die automatisch generierten semantischen Annotationen durch Information aus Ontologien angereichert. Diese integrierte semantische Information wird indexiert; der dabei entstandene Index geht in seiner Funktionalität weit über einen herkömmlichen Volltextindex hinaus. So können z. B. Volltexte über Personen auch über Eingabe ihres Berufes gefunden werden oder Personen-Entitätenseiten über Eingabe des Geburtsortes oder –jahres. Es wird zwischen verschieden Rollen für Personen in Dokumenten unterschieden, wie z. B. „Von-Person" und „Über-Person". Die GUI der SMMS greift über eine Web-Service Schnittstelle auf den Index zu. Zu einer Anfrage wird neben den Suchtreffern auch eine Liste von Facetten zurückgeliefert. Diese Facetten erlauben das weitere Einschränken des Suchergebnisses. Facetten werden z. B. aus den erkannten Entitäten oder Klassen gebildet, aber auch aus den Metadaten aus der Ontologie.

SEMANTISCHE MULTIMEDIASUCHE

Die semantische Multimediasuche basiert auf dem semantischen Index, der aus automatisch und manuell erstellten Metadaten generiert wird. Die Erstellung des Indexes ist als separater Dienst in die im vorigen Abschnitt beschriebene Diensteplattform integriert.

Die grafischer Nutzeroberfläche für SMMS des aktuellen CONTENTUS-Prototyps (siehe auch Abbildung 16) besteht aus folgenden Elementen

1. Suchschlitz mit Anzeige des Suchpfads
2. Suchhistorie und darüber die Funktionen zur Sortierung der Ergebnisse
3. Eine kontextabhängige grafische Zeitleistendarstellung für das Publikationsdatum der dargestellten Suchergebnisse und darunter Suchfacetten
4. Suchergebnisse (Printmedien, AV-Medien, Audioaufnahmen, Orte usw.)

Zusätzlich zur Darstellung der Suchergebnisse kommen noch folgende zu visualisierende Kontexte hinzu:

- Darstellung der ausgewählten Ressource durch einen „Semantic Media Viewer"
- Darstellung der Detailinformation für eine ausgewählte Ressource aus den Suchergebnissen
- Detaildarstellungen von Entitäten inklusive einer grafischen Repräsentation ihres Beziehungsgeflechts mit anderen Entitäten.
- Darstellung der Metadaten einer ausgewählten Ressource mit der Möglichkeit der Nutzerinteraktion bezüglich persönlicher Metadaten, Bestätigung, Ablehnung oder Korrektur

- Darstellung eines „Medien-Warenkorbs" mit dessen Hilfe der Benutzer ausgewählte Ressourcen zum persönlichen Gebrauch zwischenspeichern, aggregieren und – auf Metadatenebene – weiterbearbeiten kann
- Grafische Benutzeroberfläche für die Benutzer- und Benutzergruppenverwaltung sowie zur Administration und Verwaltung der Nutzerinteressenprofile („User Interest Profiles"); diese werden im Rahmen des Beitrags nicht gezeigt.

Sofern sich der Nutzer für ein Suchergebnis entschieden hat, kann er entweder per Semantic Media Viewer auf den multimedialen Inhalt – zur Zeit Audio-, AV- und Printmedien (inkl. Fotos) – zugreifen (siehe Abbildung 17 und Abbildung 18) oder Details zu einer der darin vorkommenden Entitäten, also zum Beispiel der Person „Ezer Weizmann", ansehen (siehe Abbildung 19).

Abbildung 16: Die grafische Benutzeroberfläche für Suchergebnisse der CONTENTUS-SMMS

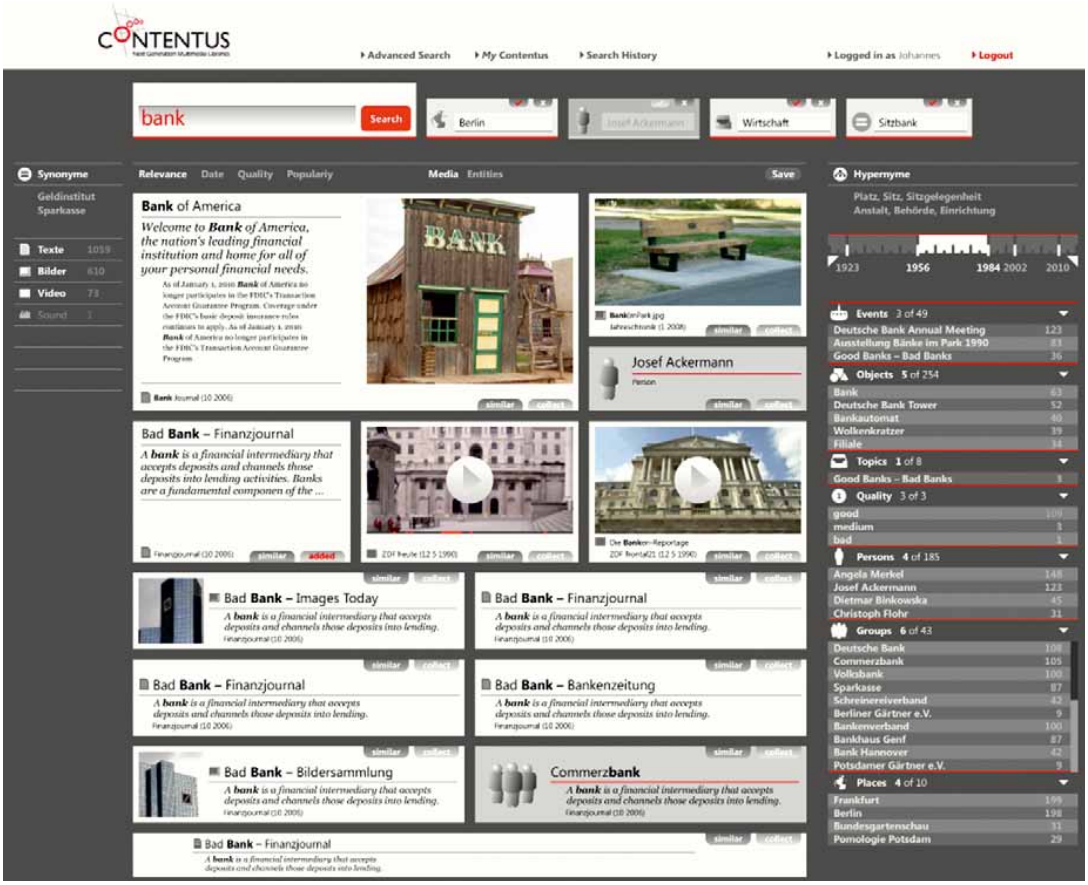


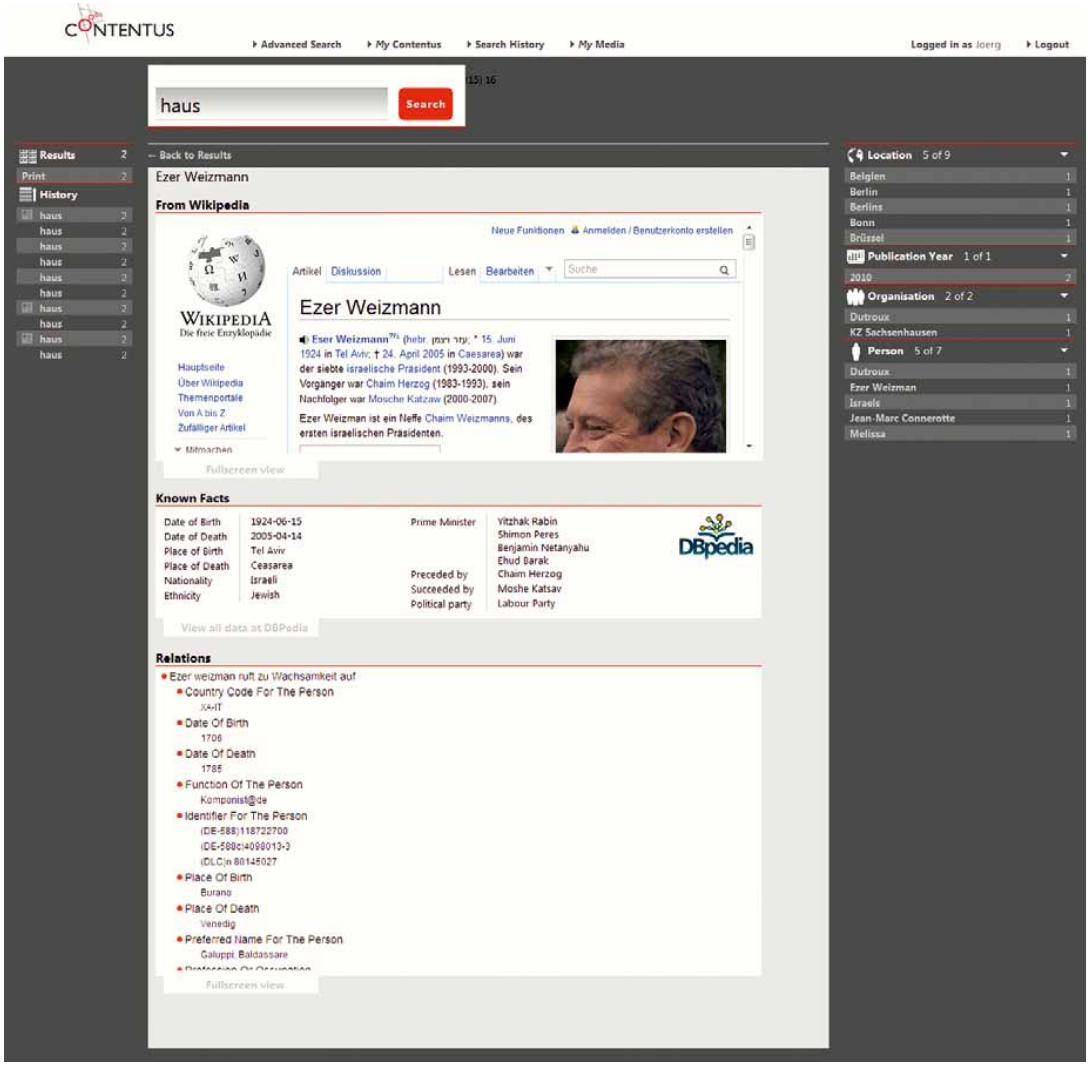
Abbildung 17: Der Semantic Book Viewer aus CONTENTUS mit digitalisiertem Dokument (links), Vorschau (rechts oben), OCR-erfasstem Text (rechts mitte) und erkannten Entitäten wie Organisationen (rechts unten).

Der Semantic Book Viewer zeigt jeweils eine digitalisierte Seite eines Dokuments an und erlaubt mit der Maus das intuitive Zoomen und Blättern in der jeweiligen Publikation.

Abbildung 18: Der Semantic Book Viewer zeigt ein digitisalisiertes Dokument einschließlich der hervorgehobenen Entitäten, welche automatisch aus dem Text extrahiert wurden.

Über ein GUI-Element lässt sich die Anzeige der erkannten Entitäten layoutgetreu ein- und ausschalten, so dass erkannte Personen, Orte und Institutionen leicht im Text identifiziert werden können. Beim Darüberfahren mit der Maus wird eine Infobox angezeigt und durch Klicken darauf weitere Informationen und Relationen zu der gewählten Entität angezeigt.

Abbildung 19: Anzeige von Entitätendetails: die CONTENTUS-SMMS-GUI zeigt Entitätenrelationen aus der CONTENTUS-Wissensbasis (unter „Relations") kombiniert mit Informationen aus der Wikipedia und der DBPedia („Known Facts").



ZUSAMMENFASSUNG UND AUSBLICK

Mit Hilfe von CONTENTUS-Technologien sollen Nutzer in Zukunft einfacher im digitalisierten Kulturerbe recherchieren und navigieren – dies ist die Vision der Next-Generation Multimedia Library. Die Grundlage hierfür ist ein Workflow zur automatischen inhaltlichen und semantischen Erschließung von multimedialen Inhalten. Im Rahmen dieses Prozesses werden Metadaten zu den Multimediaobjekten erzeugt, Datensätze semantisch vernetzt und zu einer Wissensbasis zusammengefügt. Im Rahmen der semantischen Vernetzung können Medieninhalte ebenfalls mit externen Wissensquellen vernetzt werden. Für den CONTENTUS-Demonstrator wurden DBpedia und Wikipedia mit Elementen aus der CONTENTUS-Wissensbasis verknüpft.

Der ebenfalls in CONTENTUS entwickelte Demonstrator zur semantischen Multimediasuche ermöglicht bereits heute Nutzern über eine grafische Oberfläche eine prototypische Nutzung dieser Wissensbasis.

In der nächsten Ausbaustufe sollen Nutzer auf einfache Weise Ergebnisse und Erkenntnisse mit anderen Nutzern und Anbietern teilen und Inhalte selbstständig durch Verknüpfungen anreichern.

Durch die Nutzung von CONTENTUS-Technologien können Anbieter von Inhalten die Kosten zur Verarbeitung ihrer multimedialen Sammlung und zur Realisierung einer semantischen Multimediasuche deutlich senken. Weiterhin werden die Inhalte Teil eines wachsenden Wissensnetzes aus Nutzern und Experten; damit erhöht sich die Sichtbarkeit der Inhalte.

So können durch die in CONTENTUS geschaffenen Lösungen, neben großen Informationsanbietern und Inhalteinhabern, auch kleinere Bibliotheken und Archive profitieren indem sie Teil der internetbasierten Wissensinfrastruktur werden und ihre multimedialen Sammlungen einem breiteren Nutzerkreis zugänglich machen können.