Metadaten aus der Cloud: Technologien und Anwendungen der CONTENTUS-Diensteplattform zur Medienerschließung

Metadata from the Cloud: Technologies and Applications of the CONTENTUS service platform for digital content enrichment

Dr. Michael Eble und Dr. Stefan Paal
Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloss Birlinghoven – 53754 Sankt Augustin – Deutschland
Tel.: 0049 22 41 / 14 34 06 (Eble) und 0049 22 41 / 14 34 38 (Paal)
E-Mail: Michael.Eble@iais.fraunhofer.de, Stefan.Paal@iais.fraunhofer.de
Internet: http://www.metadaten-aus-der-cloud.de

Zusammenfassung:
Das Erzeugen von reichhaltigen Metadaten gehört zu einer wesentlichen Tätigkeit von Kultureinrichtungen, um die eigenen Bestände zu nutzen und zu vernetzen. Dazu werden manuelle und automatische Verfahren der Medienerschließung eingesetzt. Entsprechende Software-Lösungen können zunehmend mehr auch via Cloud Computing als Software-as-a-Service genutzt werden. Hard- und Software-Infrastrukturen müssen so nicht lokal vorgehalten werden, stattdessen kommen Metadaten aus der Cloud. Dadurch können finanzielle, organisatorische und technische Hürden sinken, um Bestände zu erschließen.
Ein Beispiel für eine solche Lösung ist die CONTENTUS-Diensteplattform, die eine Reihe von Verfahren zur strukturellen und inhaltlichen Erschließung und Anreicherung bündelt. Aufbereitete Digitalisate und Metadaten können in Formaten wie PDF/A, METS, MODS, ALTO u. a. exportiert werden. Nutzer der Plattform übernehmen diese Daten dann direkt in elektronische Lesesäle und andere Anwendungen.

Abstract:
Creating rich Metadata is a core activity of cultural organizations to better exploit and link up their content. Therefore manual as well as automatic methods are used to open up and enrich inventories. Corresponding software solutions can increasingly be used via cloud computing as software-as-a-service. Thus, hardware and software infrastructure have not to be held locally, instead, metadata are delivered from the cloud. In this way, financial, organizational and technical hurdles of projects may decrease.
An example of such a solution is the CONTENTUS services platform that combines a number of methods for structural and content exploitation as well as enrichment. Processed images and metadata can be exported in formats such as PDF/A, METS, MODS, ALTO etc. Users of the platform integrate this data directly into electronic reading rooms and further applications.

1. Einleitung

Das Aufbereiten und Erschließen von Medienbeständen durch Bibliotheken, Archive und andere Kultur- und Medienorganisationen ist eine wesentliche Voraussetzung, um Inhalte verwenden und neuen Nutzergruppen zugänglich machen zu können. Daher gehört es zu einer wesentlichen Tätigkeit dieser Organisationen, reichhaltige Metadaten zu erzeugen und zu pflegen. Dazu können heute manuelle und automatische Verfahren zur Medienerschließung unmittelbar miteinander kombiniert und Ergebnisse der Verfahren zur gegenseitigen Ergänzung verschränkt werden. Zur effizienten automatischen Erschließung von großen Volumina digitaler Audio-/Video- sowie Bild-/Dokumentenbestände sind dabei leistungsfähige Hard- und Softwarearchitekturen erforderlich. Diese sind in der Regel mit hohen Investitionskosten verbunden.

Demgegenüber versprechen Lösungen aus dem Bereich des Cloud Computings nun, Software zur Unterstützung der automatischen und manuellen Medienerschließung so bereitzustellen, dass sie eine Verwendung bei weitgehend variablen Betriebskosten ermöglichen. Damit besteht die Option, softwaregestützte Erschließungsverfahren in Projekten „on demand" zu nutzen. Über Diensteplattformen wird eine solche Software-as-a-Service (SaaS) auf eine Weise verfügbar, die nicht den lokalen Betrieb von Hardware- und Software-Infrastrukturen für Erschließungsverfahren erfordert, gleichzeitig aber eine weitgehend hausinterne Projektdurchführung ermöglicht. Über Programmierschnittstellen werden Algorithmen und Daten hierbei in eigene Anwendungen integriert. Dadurch können die finanziellen, organisatorischen und technischen Hürden sinken, um automatische Verfahren der Medienerschließung zu nutzen. Ein Beispiel dafür ist die CONTENTUS-Diensteplattform, die im Rahmen des Forschungsprogramms THESEUS entwickelt wurde und eine Reihe von Verfahren zur Medienerschließung sowohl in Form von einzelnen Diensten als auch in orchestrierten Workflows bereitstellt.
Ziel des vorliegenden Beitrags ist es, Rahmenbedingungen von Medienerschließung und Cloud Computing sowie verschiedene Technologien und Anwendungsfälle der CONTENTUS-Diensteplattform vorzustellen. Der Fokus liegt auf Projekten zur Erschließung von großvolumigen Medienbeständen in Medien- und Kulturorganisationen.

2. Medienerschließung und Cloud Computing

Hinter der Digitalisierung und Erschließung von Medienbeständen stehen je nach Organisation unterschiedliche Ziele, so z. B. die Bewahrung von Kultur- und Mediengütern, die Distribution von digitalen Inhalten oder der Vertrieb von digitalen Produkten. In den meisten Fällen werden dazu u. a. identifizierende, beschreibende, struktur- und relationsbezogene sowie administrative und technische Metadaten benötigt. Das Erzeugen von reichhaltigen Metadaten ist damit nicht nur eine wesentliche, sondern auch eine ständig wiederkehrende Tätigkeit.
Um Metadaten aus der Cloud zu beziehen, bieten inzwischen verschiedene Anbieter ihre SaaS-Lösungen für die Erschließung von Dokument- sowie Audio- und Videobeständen an (Eble & Kirch 2012a und b): Abbyy bietet mit der OCR Cloud SDK eine Plattform zur Erschließung von Dokumenten an, Nuance hat zu diesem Zweck den OmniPage Cloud Service gestartet. Zur Erschließung von Audio- und Videobeständen stehen Cloud-Dienste von SpeakerText und 3Play Media zur Verfügung. Eine semantische Anreicherung von textuell vorliegenden Daten ist mit Thomson Calais und Orchestr8 AlcheMyAPI möglich, während mit der Kooba Technology Platform ein Cloud-Angebot genutzt werden kann, um Bilddaten anzureichern.
Bei den derzeitigen Angeboten liegt der Fokus vielfach auf englischsprachigen Inhalten; die Anbieter sind in mehreren Fällen nicht in Deutschland, sondern in den USA ansässig. Zudem ist eine integrierte Medienerschließung, also die Verarbeitung von Dokument-, Audio- und Videobeständen mit einer einzigen Plattform, derzeit noch wenig verbreitet, so dass hierzu die Angebote unterschiedlicher Anbieter miteinander kombiniert werden müssen. Eine Ausnahme stellt die Ramp MediaCloud dar, die bereits heute mehrere Medientypen in einer einzigen Umgebung erschliessen kann und die Metadaten anschließend für die Verwendung in eigenen Anwendungen (Content-Management-Systeme, Archivsysteme etc.) und Produkten bereitstellt.
Verschiedene Untersuchungen der vergangenen Jahre verweisen auf eine Reihe möglicher Vorteile von Cloud Computing als Gründe für die Nutzung solcher Dienste (Holtkamp 2010, PwC 2011 und Weiner 2011): Verringerte Einstiegshürden bei gleichzeitiger Konzentration auf Kernkompetenzen, flexible Einsatzmöglichkeiten und schnelle Verfügbarkeit zusätzlicher Ressourcen für Projekte, Kosteneinsparung und -kontrolle sowie eine positive Veränderung der Kostenstrukturen von Investitions- zu Betriebskosten. Demgegenüber stehen jedoch auch Risiken bzw. Gründe gegen die Nutzung, die nicht zu unterschätzen sind (aaO): Auf der technischen Ebene zählt dazu eine derzeit noch eingeschränkte Leistungsfähigkeit bei gleichzeitig fehlender Interoperabilität und schwieriger Integration sowie unzureichende bzw. unklare Sicherheitsmechanismen. Auf der kaufmännischen bzw. fachlichen Ebene wird der Markt als zu unübersichtlich bewertet und das Fehlen spezifischer Angebote für bestimmte Branchen sowie fehlende Best Practices hemmt die Nutzung.

3. Technologien der CONTENTUS-Diensteplattform

Im skizzierten Kontext ist auch die CONTENTUS-Diensteplattform verortet. Das Angebot wurde im Rahmen des vom Bundesministerium für Wirtschaft und Technologie (BMWi) geförderten Forschungsprogramms THESEUS entwickelt und ermöglicht eine cloud-basierte und integrierte Medienerschließung. Sie ist als verteiltes und skalierbares System nach dem SOA-Prinzip konzipiert und ermöglicht die Integration verschiedener Verarbeitungsdienste sowie eine adaptive Einbindung leistungsfähiger Rechnerressourcen verschiedener Anbieter.

Der Ausgangspunkt der Forschungs- und Entwicklungsarbeiten war, dass Kultur- und Medienorganisationen vor mehreren typischen Problemen stehen, die sich vom Zerfall von Archivgut über unzureichende Erschließung und fehlende Sinnzusammenhänge bis zu unklaren Verwertungsperspektiven erstrecken. Wesentlich war dabei, dass Entwicklungen auf die Bedarfe von Kultureinrichtungen wie Bibliotheken und Archive zugeschnitten sind, indem sie auch den Umgang mit schwierigen Digitalisaten von Zeitungen und Büchern adressieren. Ziel des Projekts war es daher, Lösungen mit einem möglichst hohen Automatisierungsgrad für die einzelnen Prozessschritte und Medientypen zu entwickeln (siehe Abbildung 1).

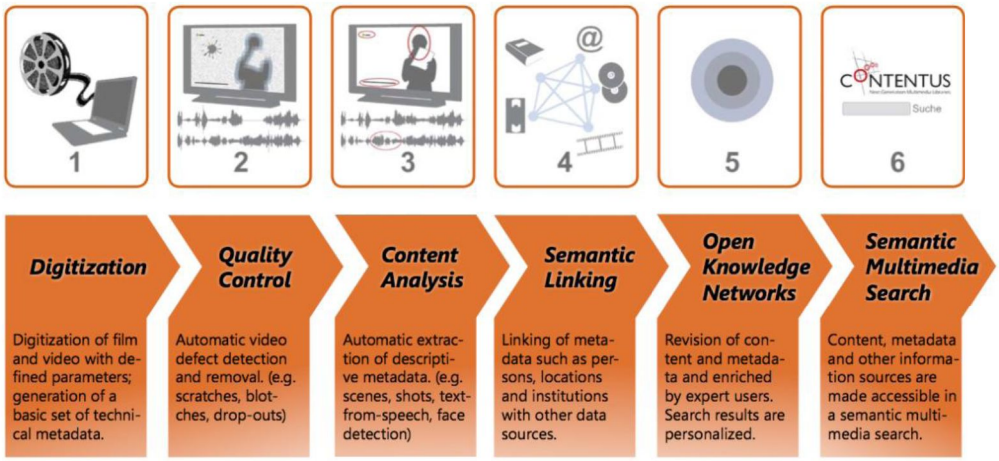

Abbildung 1: Die CONTENTUS-Prozesskette der Medienerschließung

Die Plattform bündelt eine Reihe von Technologien, die sich an tradierten Prozessen der Medienerschließung orientieren (Abbildung 2) und im Folgenden für Dokumentbestände illustriert werden:

- Im ersten Schritt findet mittels automatischer Verfahren ein **Aufbereiten der Mediendaten** statt, wodurch die Digitalisate von gescannten Zeitungs- oder Buchseiten geschärft und entzerrt werden. Zudem werden Alterungseffekte entfernt und durch eine Rotationskorrektur die Digitalisate angemessen ausgerichtet.
- Im zweiten Schritt wird ein **Erkennen von Strukturen** durchgeführt: Die Buch- oder Zeitungsseiten werden in ihre logischen Einheiten unterteilt (Seitensegmentierung) und zusammengehörende Zeitungsartikel werden identifiziert (Artikelsegmentierung), so dass XML-Beschreibungen der logischen und physischen Struktur von Dokumenten vorliegen. Zudem wird die jeweilige Titelseite von Zeitungen oder Büchern automatisch erkannt und extrahiert, so dass sie zur Darstellung einer Vorschau o. ä. zur Verfügung steht.
- Im dritten Schritt wird ein **Erschließen und Anreichern von Inhalten** umgesetzt, so dass mittels einer Texterkennung (OCR) Bild- in maschinenlesbare Textdaten umgewandelt werden und die pixelgenaue Position von Texten, Grafiken, Bildern, Tabellen etc. sowie logische Bezeichner für jedes Element eines Digitalisats vorliegen. Aus den maschinenlesbaren Texten können anschließend z. B. benannte Entitäten wie Personen, Organisationen und Orte extrahiert, identifiziert und mit Datensätzen wie Biografien oder Dossiers verlinkt werden.
- Schließlich kann ein vollständiger **Export der Digitalisate und Metadaten** in Formaten wie TIFF, JPG, PDF/A, METS, MODS, ALTO u. a. erfolgen, so dass Kultureinrichtungen die erzeugten Daten umfassend in eigenen Anwendungen wie elektronischen Lesesälen u. ä. nutzen können.

Im Zusammenspiel mit den genannten Technologien können manuelle Schritte umgesetzt werden: Das kann erstens ein **Trainieren von Algorithmen** sein – z. B., um Dokumente, Entitäten oder andere Daten für einen Trainingskorpus zu klassifizieren, der dann als Referenzpunkt für maschinelle Lernverfahren dient. Zweitens kann es sich um ein **Optimieren von Metadaten** wie etwa das Korrigieren von Fehlern handeln, die bei der Ausführung von automatischen Verfahren entstanden sind. Drittens kann es ein **Ergänzen von Metadaten** sein, wenn manuell Informationen zu Personen oder Links zu biografischen Informationen und Standorten hinzugeführt werden.
Die Notwendigkeit und der Nutzen einer solchen Verschränkung von manuellen und automatischen Verfahren stehen damit in Zusammenhang, dass einerseits Erschließungsprojekte für große Medienbestände durch ausschließlich manuelle Arbeit vielfach nicht auf wirtschaftliche Weise umgesetzt werden können und andererseits automatische Verfahren trainiert werden müssen und nicht immer vollständig zufriedenstellende Ergebnisse liefern können. Wesentlich ist daher, dass beide Formen der Medienerschließung genutzt und ergänzt werden können.

Abbildung 2: Arbeitsweise und Leistungen der CONTENTUS-Diensteplattform

Die beschriebenen Verarbeitungsschritte der CONTENTUS-Diensteplattform kommen nicht nur für die Dokumentenerschließung zum Einsatz, sondern in ähnlicher Weise auch für die Erschließung von Audio- und Videobeständen in Archiven und Mediatheken. Dabei handelt es sich hinsichtlich des Erkennens von Strukturen z. B. um Audio- bzw. Videosegmentierung oder um Sprechergruppierung. Hinsichtlich des Erschließens von Inhalten zählt dazu die Sprach- und Sprechererkennung sowie die Identifikation von Gesichtern. Auch hier können sich automatische und manuelle Verfahren ergänzen.

4. Anwendungen der CONTENTUS-Diensteplattform

Derzeit verwenden unterschiedliche Kultureinrichtungen und Digitalisierungsdienstleister die Plattform, um Digitalisate aufzubereiten und zu erschließen. Dabei stehen ihnen die Verfahren sowohl über eine grafische Benutzeroberfläche als auch über Web-Service-Schnittstellen zur Verfügung: Im ersten Fall erfolgt die Verwendung der Dienste vollständig im Web-Browser, im zweiten Fall nutzen die Anwender die Dienste in eigenen Systemen. Beispiele für solche Systeme sind Archiv- und Annotationsumgebungen. Zwei Anwendungsbeispiele dafür sind Projekte zur Erschließung von Zeitungs- und Zeitschriftenarchiven:

- Über die Diensteplattform wurden alle seit 1977 erschienenen Ausgaben der **Zeitschrift EMMA** erschlossen und sind nun vollständig auf der Webseite des Verlages durchsuchbar. Die insgesamt 25.348 Zeitschriftenseiten wurden dabei vollständig mittels automatischer Texterkennung verarbeitet. Im Auftrag des Bonner Unternehmens ImageWare hat das Fraunhofer IAIS gemeinsam mit dem Hochschulbibliothekszentrum NRW den gesamten Bestand der Zeitschrift als neues Produkt „EMMAdigital“ zugänglich gemacht.
- In ähnlicher Weise wurden mit der Diensteplattform insgesamt 1,8 Millionen Zeitungsseiten aus dem Bestand der **Regionalzeitung Donaukurier** erschlossen. Dadurch stehen Lesern und der Redaktion alle 25.000 Ausgaben seit 1945 vollständig erschlossen und durchsuchbar zur Verfügung. Dabei sind aus den Digitalisaten auch eigene Archiv-Produkte entstanden, die im Lesermarkt vertrieben werden.

Wie auch in vorherigen Erschließungsprojekten wie z. B. mit der Neuen Zürcher Zeitung (NZZ) orientierte sich das Vorgehen an der beschriebenen CONTENTUS-Prozesskette, die für den jeweiligen Anwendungsfall angepasst wurde (Abbildung 3).

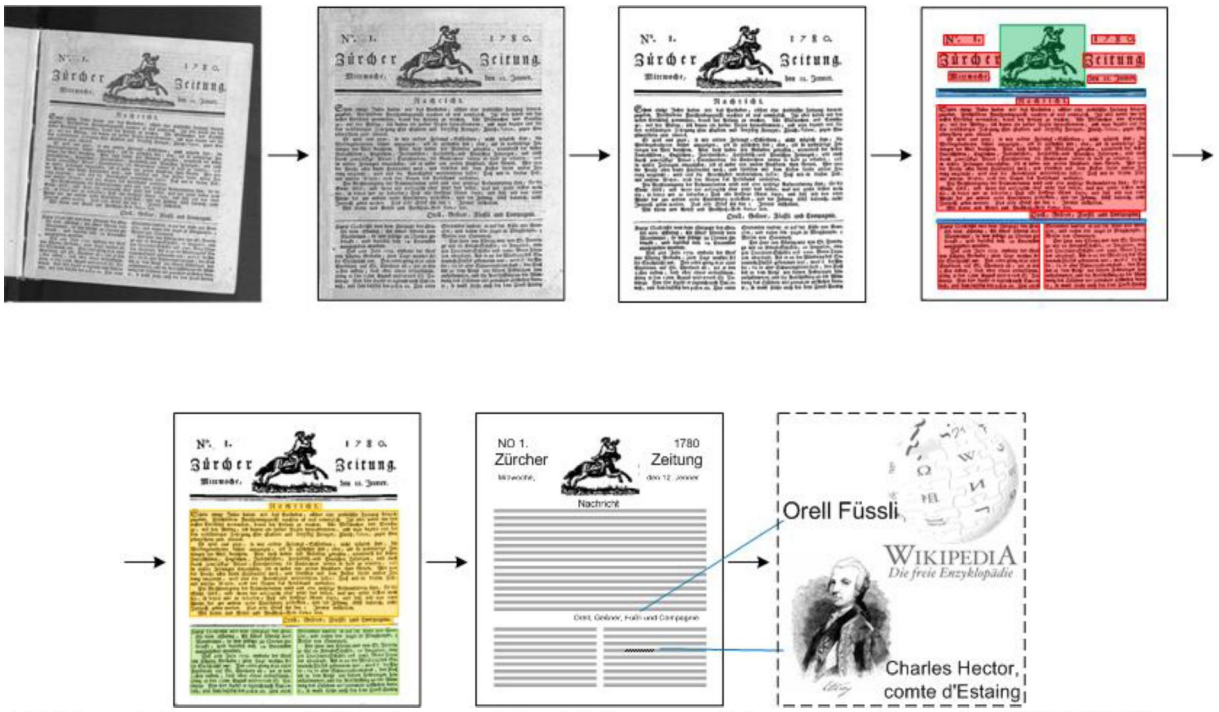

Abbildung 3: Prozess zur Aufbereitung und Erschließung von Zeitungsdigitalisaten in CONTENTUS

Wesentlich war in allen Fällen, dass die reine Erschließung der Archive nur der erste Schritt war. Hinzu kamen Fragen zur Darstellung und Nutzung der Digitalisate. Entscheidend war dabei, dass eine konventionelle Bereitstellung etwa via Web-Browser für die Medienanbieter nicht in Frage kam, da die wertvollen Archivgüter auf zu einfache Weise kopierbar wären. Daher wurden die Digitalisate und ihre Metadaten in den Elektronischen Lesesaal MyBib eL® übernommen: Dieser Lesesaal ist eine in Kooperation mit der Firma ImageWare Components GmbH entwickelte Präsentationsplattform für Digitalisate und eröffnet die Möglichkeit, unter Einhaltung des Urheberrechts eine Vielfalt an Dokumenten sicher digital bereitzustellen. Dadurch schützt der

Lesesaal sowohl Rechteinhaber als auch Nutzer vor Datenmissbrauch. Unterschiedliche Authentifizierungs- und Autorisierungsmöglichkeiten erlauben es den Benutzern, rechtssicher auf (Lehr)Bücher und Zeitschriften zuzugreifen (Paal & Eickeler 2011; Paal 2012).

Ein drittes Anwendungsbeispiel ist die Erschließung der Zeitung **Neues Deutschland** in einer Kooperation zwischen der Staatsbibliothek Berlin, dem Archivdienstleister ArchivInForm und dem Fraunhofer-Institut IAIS. Hier wurden zunächst 2,5 Millionen Artikel aus 15.000 Ausgaben der historischen Zeitung strukturell und inhaltlich über die CONTENTUS-Diensteplattform erschlossen. Dazu wurde einerseits die grafische Benutzeroberfläche (Abbildung 4) verwendet und andererseits die programmatischen Schnittstellen innerhalb einer Anwendung zur manuellen Erschließung. Auf diese Weise nutzte der Archivdienstleister die Ergebnisse der automatischen Erschließung und steigerte deren Qualität durch manuelle Annotationen, um den zukünftigen Nutzern des Archivs von Neues Deutschland ein hochwertiges Erlebnis bieten zu können.

Abbildung 4: Beispiel der grafischen Benutzeroberfläche der CONTENTUS-Diensteplattform

Ein viertes Anwendungsbeispiel ist die Verbindung zwischen der CONTENTUS-Diensteplattform und dem Kernsystem der Deutschen Digitalen Bibliothek. Hierdurch haben die beteiligten Kultureinrichtungen und ihre Digitalisierungsdienstleister die Möglichkeit, die im Rahmen von CONTENTUS entwickelte Plattform zu nutzen, um Digitalisate aufzubereiten sowie strukturell und inhaltlich zu erschließen und dann direkt in die Datenbanken der Deutschen Digitalen Bibliothek zu ingestieren. Zudem tragen sie in der derzeit laufenden Evaluierungsphase mit ihrem Feedback zur Verbesserung der Plattform bei.

5. Fazit und weitere Entwicklungen

Abschließend lässt sich festhalten, dass die Medienerschließung via Cloud Computing derzeit in weiten Teilen noch am Anfang steht. Es zeichnet sich jedoch bereits deutlich ab, dass die Verschränkung von manuellen und automatischen Verfahren zu einem wesentlichen Bestandteil von Plattformen zur integrierten Medienerschließung wird, um die Vorteile beider Zugänge zu nutzen. Dabei wird die Erschließung von Dokument-, Audio- und Videoarchiven auch dadurch wichtiger, dass Mehrfachverwertung und Reichweitensteigerung für Mediengüter u. a. im Social Web wichtige Bausteine von Digital-Strategien sind und Medieninhalte entsprechend anschluss- und zitierfähig sowie auffind- und verlinkbar sein müssen (Eble & Schwenninger 2012). Metadaten aus der Cloud können dabei unterstützen.

6. Quellen

Eble, Michael & Kirch, Sebastian (2012): Metadaten aus der Cloud: Erschließung von Medienarchiven u□ber Diensteplattformen und Software as a Service In: Info 7 - Medien | Archive | Information, 2012/2.

Eble, Michael & Kirch, Sebastian (2012): Metadaten aus der Cloud: Technologien und Anwendungsfälle der Medienerschließung mittels Software as a Service. In: Proceedings zur WissKom 2012 als 6. Konferenz der Zentralbibliothek Forschungszentrum Jülich.

Eble, Michael & Schwenninger, Jochen (2012): Ziele, Strategien, Leistungswerte: Wie Medienarchive und Kultureinrichtungen das Social Web nutzen. In: Info 7 - Medien | Archive | Information, 2012/2.

Holtkamp, Bernd (2010): Cloud Computing für den Mittelstand am Beispiel der Logistikbranche. Dortmund: Fraunhofer ISST.

Paal, Stefan (2012): Metadaten aus der Cloud: Von der CONTENTUS-Diensteplattform in den elektronischen Lesesaal. Vortrag auf dem 12. Oracle Bibliotheken Summit „Zukunft der Bibliotheken. Engere Vernetzung von Inhalten und Dienstleistungen“ am 20. Juni 2012 in Bonn.

Paal, Stefan & Eickeler, Stefan (2011): Automatisierung vom Scan bis zum elektronischen Lesesaal. In: Information - Wissenschaft & Praxis (IWP) 62 (2011) 8, S. 351-354.

PwC – PricewaterhouseCoopers(2011): Cloud Computing im Mittelstand. Erfahrungen, Nutzen und Herausforderungen.

Weiner, Nico (2011): Der Einsatz von Cloud Computing in KMUs. Ein Beispiel aus dem Bereich der Medienagenturen. Vortrag vom 07.07.2011 am THESEUS Innovationszentrum, Berlin.