Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung Seminar

Sven Schliesing

Metadaten

Fakultät Technik und Informatik
Studiendepartment Informatik

Sven Schliesing

Thema der Ausarbeitung
Seminar

Metadaten im Bereich von Dokumentenverwaltungssystemen

Stichworte

Metadaten

Kurzzusammenfassung

Metadaten erlauben es, bekannte Dokumente mit beschreibenden Daten auszustatten. Es lassen sich Gemeinsamkeiten erkennen und Verknüpfungen zwischen den Dokumenten dokumentieren und verwerten.
Besonders im Bereich von Dokumentensystemen (digitale Bibliotheken) ist dies wichtig. Im Folgenden wird näher auf bestehende Techniken eingegangen und das "'OAIS"'-Modell sowie der "'METS"'-Standard vorgestellt

Inhaltsverzeichnis

1 Einleitung

Metadaten sind weit mehr als nur “Daten über Daten“. (Abbildung: 1) 1997 wandte Tim Berners-Lee den Begriff der Metadaten auf das Web an: “Metadata is machine understandable information about web resources or other things“.Berners-Lee (1997)

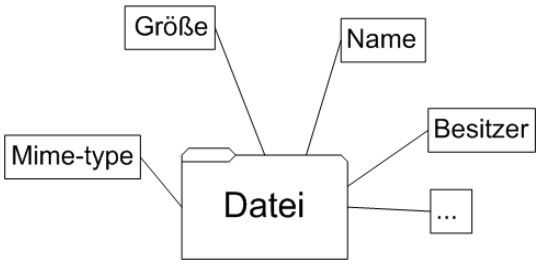

Abbildung 1: einfache Sicht der Dinge

Metadaten sind also nicht einfach nur Daten sondern vielmehr ein Hilfsmittel um mit bestehenden Daten besser umgehen zu können, sie zu verwalten und zu interpretieren.
Die Technik, Dinge auf einer höheren Ebene (der Metaebene) zu betrachten findet sich nicht nur in der Informatik sondern z.B. auch in der Psychologie. Hier begeben sich die Gesprächspartner zum Zwecke der Gesprächsanalyse auf eine höhere Ebene um Zusammenhänge in ihrere Kommunikation zu erkennen.brockhaus (2001)

Für die Anwendung in der Informatik gibt es eine vielzahl von Möglichkeiten. Angefangen bei der Auswertung oder, noch simpler, der Verwaltung von Dateien auf einem Speichermedium bis hin zur komplexen Interpretation von, sich in einer Suchmaschine befindenden, Dokumenten. Letzteres wird intensiv bei Knowledge-Management-Systemen sowie Bibliotheksverwaltungssystemen praktiziert um das Auffinden von entsprechenden Dokumenten einfach und effektiv zu gestalten.

In dieser Ausarbeitung zum Vortrag “Metadaten“ soll nun genauer darauf eingegangen werden in wie weit Metadaten sinnvolle Hilfen seien können und in wie verlässlich und sinnvoll implizite und explizite Metadaten sind.

Im Hauptteil dieser Ausarbeitung wird auf die verwendeten Technologien eingegangen. Es wird ein Überblick über „MARC21“, „DCMES“ sowie deren Kombination gegeben. Außerdem erfolgt ein kurzer Exkurs in das Thema „Semantic Web“. Bevor dieser Teil mit einer Betrachtung des Themas im Bezug auf die Master Thesis endet, wird das Modell „OAIS“ umfassend vorgestellt und auf den „METS“ Standard eingegangen um eine Grundlage zu schaffen.

2 Technologien

Die Technologien rund um die Metadaten wurden schon früh von wichtigen Institutionen vorangetrieben. So entstand 1968 die erste Version von “MARC 21“, dem “Machine-readable cataloging“.
In der jüngeren Zeit gab es diverse Neuentwicklungen. Hervorzuheben sei hier das “Dublin Core Metadata Element Set“ der “Dublin Core“-Initiative, welches im Jahre 1995 verabschiedet wurde. Zusätzlich gib es Techniken die noch einen Schritt weiter gehen. Die Idee des Semantik Web nutzt, hauptsächlich explizite, Metadaten um Strukturen in Dokumenten zu erkennen und diese sinnvoll mit anderen Dokumenten in Verbindung zu bringen.

3 MARC 21

“MARC 21“ ist der älteste und am meisten genutzte Standard im Bereichn der Metadaten-Erfassung. Er ist standardisiert unter dem US-nationalen Standard Z39.2 sowie unter ISO 2709.
Das Haupteinsatzgebiet sind große Bibliotheken mit Millionen von Dokumenten wie z.B. die “Library of Congress“[1]. Ziel war es, ein Format zu entwickeln was allen Bedürfnissen einer Bibliothek gerecht wird. Das Format beschreibt nicht nur eine Elementmenge für die Daten definiert, sondern auch deren Semantik festlegt. Zur weiteren Definition der Feldinhalte werden weitere Standards wie die “Anglo-American Cataloguing Rules“ (AACR2) oder “International Standard Bibliographic Description“ (ISBD) verwendet. “MARC 21“ ist also ein Format, dass nicht nur Metadaten speichert sondern auch die Definition der einzelnen Felder vorgibt sowie einen Informationsaustausch ermöglicht.
Dies ist, unter anderem, ein Grund für die Verwendung von “MARC 21“. Kataloge können über Rechnergrenzen hinweg gemeinsam verwendet werden und verhindern so die Erzeugung von Duplikaten. Durch die Verknüpfung von verschiedenen MARC-Dokumenten wird die Möglichkeit geschaffen, semantische Verknüpfungen zwischen beschriebenen Dokumenten herzustellen. Hierdurch kann auf schon vorhandene Dokumente zurückgegriffen werden und es entstehen zusätzliche Verbindungen die einer verbesserten Qualifizierung eines Dokuments zu Gute kommen.
1991 wurde begonnen die Möglichkeiten von “MARC 21“ und “AACR2“ weiter für die Verwendung im Internet zu verbessern. Das “MARC Advisory Committee“ betrachte Vorschläge die eine Erweiterung des MARC-Standards nach sich zogen um z.B. den Bibliothekseintrag mit dem eigentlichen Dokument zu verknüpfen oder das Dokument an sich näher zu beschreiben.
Kurzer Auszug aus einem Beispieldokument:

```
040 ## $a DLC
       $c DLC
       $d DLC
050 00 $a GV943.25
       $b .B74 1990
```

[1] http://www.loc.gov/

```
082 00 $a 796.334/2
       $2 20
100 1# $a Brenner, Richard J.,
       $d 1941-
245 10 $a Make the team.
       $p Soccer :
       $b a heads up guide to super soccer! /
       $c Richard J. Brenner.
```

3.1 DCMES

Das “Dublin Core Metadata Element Set“ beinhaltet in seiner Grundversion 15 Elemente in ihrer Definion und Semantik. Es beinhaltet allerdings keine Definition der speziellen Syntax, so dass z.B. ein Datum in jeder beliebigen Form angegeben werden kann:
“In includes some suggested encodings for specific applications, but it does not mandante on any particular syntax.“(Wayne Jones, 2002, S. 42)
Auch die Syntax des kompletten Dokumentes ist nicht festgelegt, da lediglich die Menge der Elemente definiert ist.
Die ursprüngliche Menge der DCMES beinhaltete folgende Elemente: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type.
Als konzeptionelles Framework wurde 1996 während eines Dublin Core-Workshops das konzeptionelle “Warwick Framework“ entwickelt. Es ermöglicht die Koexistenz verschiedenster Metadaten und wird als Vorgänger des “Resource Description Framework“ (RDF) angesehen.
Aufgrund seiner Einfachheit und vielseitigen Einsetzungs-Möglichkeiten findet das DCMES breite Unterstützung z.B. von der “Ëuropean Commission for Standardization“ (CEN) sowie von der “Internet Engineering Task Force“ (IETF). Durch die Bildung der “Dublin Core Metadata Initiative“ und dem “Dublin Core Advisory Committee“ wurde die Formalisierung weiter vorangetrieben.

3.2 MARC 21 + DCMES

„MARC 21“ und „DCMES“’ wurden parallel entwickelt. Beide verfolgen die gleichen Ziele: effiziente Suche und Archivierung. Eine Kombination aus beiden ermöglicht die Verwendung von flexiblen Elementenmengen in einem ausgereiften Framework: DCMES im MARC-Framework:

```
<?xml version=’1.0’ ?>
<dc xmlns="http://purl.org/dc/elements/1.1/">
  <title>Arithmetic / </title>
  <creator>Sandburg, Carl, 1878-1967.</creator>
  <creator>Rand, Ted, ill.</creator>
```

```
  <type/>
  <publisher>San Diego :Harcourt Brace Jovanovich,</publisher>
  <date>c1993.</date>
  <language>eng</language>
  <description>A poem about numbers and their characteristics.
  Features anamorphic, or distorted, drawings which can be
  restored to normal by viewing from a particular angle or
  by viewing the image's reflection in the provided Mylar
  cone.</description>
  <description>One Mylar sheet included in pocket.</description>

  <subject>Arithmetic</subject>
  <subject>Children's poetry, American.</subject>
  <subject>Arithmetic</subject>
  <subject>American poetry.</subject>
  <subject>Visual perception.</subject>
</dc>
```

3.3 Semantic Web

Wie schon erwähnt wurden mit „MARC 21“ und „DCMES“ Schritte in Richtung Web-Auszeichnungen gemacht. Die Verknüpfung von Ressourcen durch Auszeichnungen auf der Meta-Ebene wird im Bereich des „Semantic Web“ weiter verfeinert:
„For the Semantic Web, semantic indicates the the meaning of data on the Web can be discovered-not just by people, but also by computers“ Passin (2004)
Es wird also die Möglichkeit geschaffen, durch die Auszeichnung von Dokumenten durch weitere Daten auf der Metaebene, das Web weiter zu strukturieren und Informationen intelligenter zu verwalten. Auch Zusammenhänge lassen sich so effizienter erkennen und gewinnen. Die Auszeichnung muss hierbei nicht zwingend manuell erfolgen. Es gibt u.A. die Möglichkeit implizite Daten aus dem Dokument oder seiner Umgebung zu gewinnen. Aus der Umgebung erfährt man den Besitzer, die Dateiendung sowie die Größe. Das Dokument selbst bietet die Möglichkeit Informationen aus der Überschrift, dem Abstract oder sogar dem eigentlichen Inhalt/Text des Dokuments zu extrahieren:

> „To find information, a Semantic Web approach would expect to go beyond keyword and alphabetical indexes to let users search by concepts and categories.“

Passin (2004)
Diese Konzepte und Kategorien lassen sich, mit den entsprechenden Techniken aus dem Dokument selber gewinnen. Der Titel, sofern er extrahiert werden kann, bietet hierbei den ersten Ansatzpunkt.

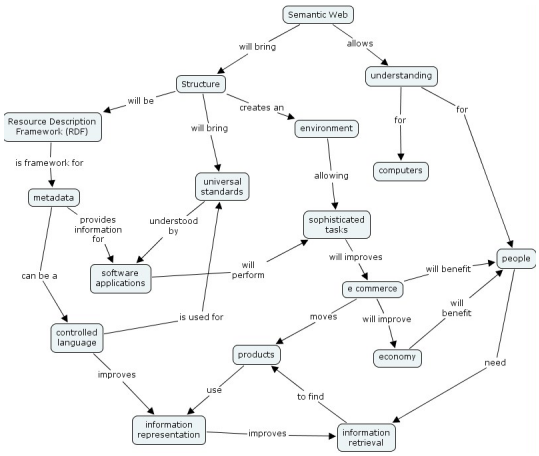

Abbildung 2: Semantic web

4 Mehrwert

Den Mehrwert den man sich durch Metadaten erhofft ist nicht nur die Suche in einer Menge von Datensätzen. Auch wenn die ersten Ansätze mit „MARC 21“ unter Anderem aus diesem Grund gestartet wurden. Ein ganz wichtiger Punkt ist die Strukturierung der Dokumente, die Festlegung der Beziehungen. Wie schon erwähnt ermöglichen die Techniken wie „MARC 21“ und auch, etwas allgemeiner gefasst, das semantic web diese Verknüpfungen. Mit ihnen lassen sich Hierarchien und Netze von Dokumenten erkennen und ausnutzen. Es können so weitere Erkenntnisse gewonnen werden, die z.B. für eine Auswertung/Statistik oder die Anreicherung/Analyse genutzt werden können.

5 Risiken

Im Bereich der Metadaten, und hier besonders im Bereich der Web-Dokumente, gibt es viele Risiken die eine sinnvolle Verwendung der festgelegten/extrahierten Daten unmöglich machen können:

- Zu viele Dokumente: Alleine Google indiziert über 8 Milliarden Dokumente.
- Viele Dokumente existieren nur für einen **temporären Zweck**. Referenziert man diese nun aus anderen Dokumenten bzw aus Metadaten dieser Dokumente heraus, entstehen Verweise die „ins Leere zeigen“.
- Eine Kontrolle über die **Richtigkeit der angegebenen Daten** ist nahezu unmöglich. Menschen lügen zu ihren Vorteil verdrehen zumindest Wahrheiten. Ein Verlass auf

gegebene Informationen kann demnach im Zweifelsfall dazu führen, dass eine aufgebaute Hierarchie komplett nutzlos ist.

- **Unvollständigkeiten** sowie falsche Angaben sind nicht immer eine Boshaftigkeit des Benutzers. Oftmals sind Faulheit und Unwissenheit der Grund dafür, dass Angaben falsch oder unzureichen sind oder schlichtweg fehlen.

- **Subjektive Bewertung** bzw. Benennung erschweren weiterhin die korrekte Einordnung von Dokumenten:

  > „There's more than one way to describe something.“, Cory Doctorow

  Einen Sachverhalt mit verschiedenen Wörtern zu beschreiben ist dabei noch nicht mal das größte Problem. Werden zwei verschiedene Sachverhalte mit den, in diesem Fall korrekten aber doppeldeutigen, gleichen Wörtern beschrieben, so werden Hierarchien und Verknüpfungen aufgebaut, die nicht dem entsprechen was man erwarten würde.

Doctorow (2001)

6 Thesis Outline

6.1 OAIS

> „Digital Information lasts forever-or for five years, whichever comes first.“, Jeff Rothenberg

Zur Langzeitarchivierung von Dokumenten wurde das OAIS „Open Archival Information System“ [2] von „Consultive Committe for Space Data Systems“ entwickelt und unter ISO 14721 im Jahre 2002 standardisiert.
Es beschreibt einen Standard mit dem entsprechenden zugehörigen Referenz-Modell. Im Standard sowie im Modell sind keinerlei Datentypen oder -strukturen festgelegt. Lediglich eine Skizzierung der Systemarchitektur (konzeptionell) sowie die Abläufe im System und Zuständigkeiten der einzelnen Teile wurden bestimmt.

Das Handling der Daten erfolgt, wie in Abbildung 3 veranschaulicht, in sogenannten Informationsobjekten. Diese Packages stellen einen Verbund von Daten und Transportinformationen dar.

- SIP: Submission Information Package
  Alles was das System von aussen annimmt wird in diesen Paketen gespeichert und an das entsprechende Modul „Ingest“ weitergereicht.

[2] http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html

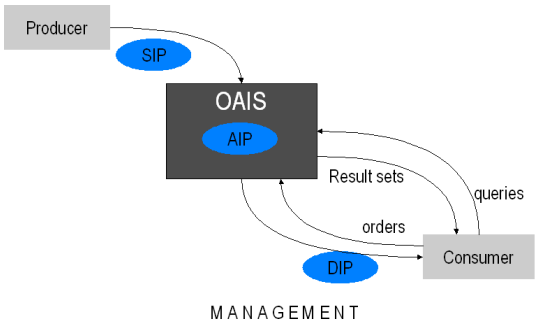

Abbildung 3: OAIS

- AIP: Archive Information Package
  Die vom Modul „Ingest“ entgegengenommenen Daten wurden durch Beschreibungsinformationen angereichert und in Form eines AIP an die interne Datenverwaltung übergeben.

- DIP: Dissemination Information Package
  Für die Ausschüttung der entsprechenden Daten bei Abfrage werden die DIP verwendet, die vom Access-Modul an den Consumer geschickt werden.

Das System der *IP hat den recht simplen Grund, dass man eine konsistente Archivierung für alle Arten von Dokumenten benötigt. Auch wenn sich die Struktur der Daten in den Jahren ändert, so soll zu jedem Zeitpunkt eine eindeutige Rekonstruktion möglich sein.Brübach (2003)

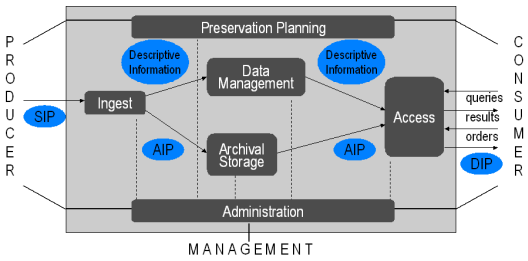

Abbildung 4: OAIS

Im Modell des OAIS gibt es, wie Abbildung 4 zeigt, 3 Akteure:

- Der **Producer** ist ein Synonym für jegliche Datenliefernde Quelle. Er erzeugt SIPs die im „Ingest“ übernommen werden. Hier werden nicht nur beschreibende Informationen

(Metadaten) erzeugt sondern auch die übernommenen Daten auf Konsistenz und Gültigkeit geprüft sowie die notwendigen Anpassungsschritte an der interne Datenformat vorgenommen.

- Der **Consumer** kann zur Erlangung von DIPs Anfragen an die „Access“-Komponente stellen. Es werden Anfragen entgegengenommen, Zugriffsrechte geprüft und die entsprechenden AIPs erzeugt.

- Die Aufgaben des **Management** enthalten „ strategic planning, defining the scope of the OAIS's archived collection, and articulating the preservation 'guarantee' associated with items entrusted to the archive. “Lavoie (2004) Desweiteren beinhaltet das Management die Regeln für die Arbeitsvorgänge im Archivsystem.

Desweiteren existieren 6 funktionale Entities:

- Ingest: Accepts Submission Information Packages (SIPs) from Producers, prepare contents for storage and management.

- Archival Storage: Storage, maintenance and retrieval of Archival Information Packages

- Data Management: Populating, maintaining, and accessing both descriptive information and internal archive administrative data.

- Access: Supports consumers in determining the existence, description, location and availability of information; allows consumers to request and receive information products

- Administration: Manages the overall operation of the archive system

- Preservation Planning: Monitors the environment of the OAIS and provides recommendations to ensure that the information stored in the OAIS remain accessible to the Designated User Community over the long term

Ullmann (2004)

6.2 OAIS und Metadaten

Die Verwaltung eines digitalen Archives erfordert ebenfalls die Verwaltung der dazugehörigen Metadaten. Die notwendigen Daten für eine erfolgreiche Verwaltung und Benutzung der digitalen Dokumente sind weitaus komplexer als die von gedruckten Dokumenten.
Während eine Bibliothek die beschreibenden Daten ihrer Werke sammeln kann wird sich kein Buch in seine Bestandteile auflösen oder Informationen verlorengehen. Es wird auch niemand Probleme haben ein Buch zu lesen weil er seine beschreibenden Daten nicht kennt.

In digitalen Bibliotheken ist dies anders. Ohne entsprechende Metadaten die die Struktur eines Dokuments beschreiben wäre es unmöglich die einzelnen Teile zu assoziieren oder zu verstehen wie die digitalisierten Teil zu interpretieren sind. Es wäre sogar z.B. bei graphischen Daten unmöglich diese anzuzeigen wenn man nicht weiss in welchen Graphikformat diese vorliegen.

Das „Making of America II“-Projekt[3] (MOA2) ist dieses Problem angegangen mit dem Ziel ein Format für beschreibende, administrative sowie strukturelle Metadaten für text- und grafik-basierte Dokumente zu entwickeln. Auf diesen Ergebnissen baute METS ein XML-Metadaten-Format für den Austausch wie auch die Verwaltung auf. Es kann also, im Bezug auf OAIS, in für alle Paketarten (SIP, AIP, DIP) verwendet werden.metstutorial (2005)

- Submission Information Package (SIP)
  - METS as transfer syntax
- Dissemination Information Package (DIP)
  - METS as transfer syntax
  - METS as input to display applications
- Archival Information Package (AIP)
  - METS stored internally in an archive

Proffitt (2003)
METS in Verbindung mit Dublin Core-Elementen

```
<dmdSec ID="dmd002">
<mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core Metadata">
  <xmlData>
    <dc:title>Alice’s Adventures in Wonderland</dc:title>
    <dc:creator>Lewis Carroll</dc:creator>
    <dc:date>between 1872 and 1890</dc:date>
    <dc:publisher>McCloughlin Brothers</dc:publisher>
    <dc:type>text</dc:type>
  </xmlData>
</mdWrap>
</dmdSec>
```

[3] http://sunsite.berkeley.edu/MOA2/

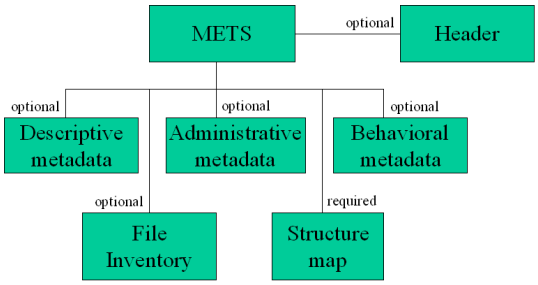

Abbildung 5: Mets-Schema

6.3 Anwendung in der Master-Thesis

Für die Master-Thesis habe ich mir vorgenommen die Einsatzmöglichkeit von METS in Verbindung mit OAIS zu evaluieren. Dadurch dass es fertige Toolkits für METS wie z.B. das METS Java Toolkit [4] gibt wird der Einstieg erleichtert und es sollte sich schnell eine solide Grundlage für eine weitere Forschung bilden lassen. Ein anderes Problem könnte eine freie Implementation des OAIS-Modells sein, so dass eine Implementierung dieses Modells ebenfalls Gegenstand der Thesis wird.
Desweiteren möchte ich untersuchen in wie weit sich Erkenntnisse und Techniken aus der Psychologie („Theorie der Metaebene“) in der Welt der Informatik einsetzen lassen können. Die Kombination beider Welten könnte entscheidende Impulse geben so dass die Verarbeitung, Speicherung und Handhabung der Metadaten in der Informatik entscheidende Vorteile erfahren könnte.

[4] `http://hul.harvard.edu/mets/`

7 Zusammenfassung

In dieser Ausarbeitung wurden die Notwendigkeit aber auch die Risiken von Metadaten betrachtet. Große System die mit digitalen Dokumenten arbeiten können diese nur effizient verwalten wenn entsprechende Beschreibungsdaten über sie vorliegen. Es gibt sogar Fälle in denen ohne diese „Daten über Daten“ die Interpretation oder sogar bloßes Anzeigen nicht möglich ist.
Große Systeme wie das OAIS-Modell nutzen die Metadaten-Technik so extensiv, dass das Metadaten-System entsprechend umfangreich und ausgereift sein muss.
In der Welt der Bilbiotheks- und Dokumentenverwaltungssysteme haben sich die Techniken „MARC 21“ (mit DCMES) sowie „METS“ (in Verbindung mit OAIS) etabliert.

Mit Hilfe dieser Techniken werden aber nicht nur die Beschreibungsdaten verwaltet sondern diese auch analysiert, angereichert und vorhandene Dokumente durch sie verknüpft. Dieser Ansatz, der auch im Semantik Web benutzt wird ermöglicht komplexe Recherchen durch „vernetzte“ Dokumente.

Literatur

[brockhaus 2001] *Der Brockhaus - Psychologie*. 2001

[metstutorial 2005] : *METS: An Overview & Tutorial*. Mai 2005. – URL `http://www.loc.gov/standards/mets/METSOverview.v2.html`

[Berners-Lee 1997] BERNERS-LEE, Tim: *Web architecture: Metadata*. Januar 1997. – URL `http://www.w3.org/DesignIssues/Metadata.html`

[Brübach 2003] BRÜBACH, Nils: *OAIS–Das "Open Archival Information System": Ein Referenzmodell zur Organisation und Abwicklung der Archivierung digitaler Unterlagen*. August 2003. – URL `http://www.sachsen.de/de/bf/verwaltung/archivverwaltung/pdf/pdf_onlinepublikationen/pp_bruebach.pdf`

[Doctorow 2001] DOCTOROW, Cory: *Metacrap: Putting the torch to seven straw-men of the meta-utopia*. August 2001. – URL `http://www.well.com/~doctorow/metacrap.htm`

[Lavoie 2004] LAVOIE, Brian F.: *The Open Archival Information System Reference Model: Introductory Guide*. Januar 2004. – URL `http://www.dpconline.org/docs/lavoie_OAIS.pdf`

[Passin 2004] PASSIN, Thomas B.: *Explorer's guide to the semantic web*. Manning, 2004

[Proffitt 2003] PROFFITT, Merrilee: *News from the Digital Library*. April 2003. – URL `http://dc-mrg.english.ucsb.edu/conference%202003/documents/merrilee_proffitt.ppt`

[Ullmann 2004] ULLMANN, Richard: *ISO 14721:2003 - OAIS; A Reference Model for an Open Archival Information System*. Juli 2004. – URL `http://apan.net/meetings/cairns2004/presentation/escience-ullman.ppt`

[Wayne Jones 2002] WAYNE JONES, Josephine C.: *Cataloging the Web*. 2002