Open Citation Project
– OpCit –

Hans-Georg Eßer
Matr. Nr. 184 383

Betreuer: F. Gatzemeier

im Wintersemester 2001/2002

Inhaltsverzeichnis

1 Einleitung

Ein wichtiger Aspekt wissenschaftlicher Arbeit ist die Publikation der eigenen Ergebnisse in Fachzeitschriften, die zur Veröffentlichung eingesandte Arbeiten zunächst von (in der Regel anonymen) Gutachtern (engl.: referees) überprüfen lassen, bevor dem Abdruck zugestimmt wird. Durch diesen Prozess wird sichergestellt, dass alle veröffentlichten Arbeiten dem Anspruch des Journals genügen.

Elementare Voraussetzung für die Akzeptanz jeder Publikation ist, dass alle aus anderen Büchern, Artikeln oder sonstigen Quellen übernommenen Teile (wortwörtliche Zitate oder sinngemäße Wiedergaben) kenntlich gemacht werden, so dass sie auffindbar sind. Üblicherweise geschieht dies über ein Literaturverzeichnis im Anhang.

Für den Leser eines Artikels oder Buches kann das Nachverfolgen der Quellenangaben mühsam sein. Insbesondere die meist kurzen Journal-Artikel setzen häufig gute Kenntnisse der angegebenen Literatur voraus, so dass zum Verständnis des Textes die übrigen Werke beschafft werden müssen.

Der lange Zeit übliche Weg, benötigte Artikel zunächst in der eigenen Universitätsbibliothek zu suchen und eventuell per Fernausleihe von einer anderen Bibliothek anzufordern, ist sehr mühsam und zeitaufwendig.

Das Internet zeigt eine Alternative: Das Konzept der Hyperlinks erlaubt die schnelle Navigation von einer Web-Seite zur nächsten. Ähnlich wie Literaturangaben verweisen auch Links auf Informationen, die an anderer Stelle verfügbar sind[1] Ein Link enthält eine genaue Internet-Adresse (URL, Uniform Resource Locator), die von jedem Web-Browser interpretiert werden kann. Der Wunsch, eine ähnlichen Funktionalität auch für zitierte Quellen in einer wissenschaftlichen Veröffentlichung zu bieten, war der Ausgangspunkt für verschiedene Projekte.

1.1 Überblick

Dieser Text stellt das *Open Citation Project* vor, dessen Ziel es ist, Literaturangaben in wissenschaftlichen Veröffentlichungen automatisch die entsprechenden Dokumente zuzuordnen und ihren Download zu ermöglichen, sofern die referenzierten Dokumente online verfügbar sind. Seit viele Journals dazu übergegangen sind, ihre Artikel nicht nur in gedruckter Form sondern auch im Internet anzubieten, ist eine solche automatische Verknüpfung der Dokumente interessant, da sie die aufwendige Suche nach den Dokumenten erspart.

[1] HTML kennt auch interne Links auf andere Dokumentpositionen; dabei kann über `<a href="#ref">` auf einen Abschnitt verwiesen werden, der durch `<a name="ref">` eingeleitet wird.

Die bereits entwickelte Software [HCJ+00] bietet dabei die Möglichkeit, einen Artikel als PDF-Datei zu betrachten und Literaturangaben im laufenden Text zu aktivieren: Sofern Informationen zu dieser Referenz gespeichert sind, wird dadurch ein Web-Browser-Fenster geöffnet, in dem verschiedene Speicherorte des referenzierten Textes erscheinen. Die hier verwendeten PDF-Dateien sind gegenüber den Originalen also um Web-Links erweitert, welche die Literaturangaben auflösen; ansonsten sind sie mit den Originaltexten identisch.

Ein alternativer Ansatz [BL01a] eignet sich für Journals, die ihre Texte im HTML-Format ins Netz stellen; die hier entwickelte Software analysiert diese Dokumente. Aus der Analyse wird dann eine neue HTML-Datei erzeugt, die wie beim obigen Ansatz beim Aktivieren einer Referenz eine Liste von Fundstellen des referenzierten Dokuments anzeigt.

Eine Gemeinsamkeit beider Ansätze ist, dass die Analyse der Dokumente und Generierung der um Links ergänzten Dateien vollautomatisch erfolgt, so dass sich ganze Archive in kurzer Zeit bearbeiten lassen.

Älter ist das ResearchIndex-Projekt [LGB99]. Beim hier verfolgten Ansatz werden auch Suchmaschinen und Postings in News Groups als Quellen für den Aufbau einer Datenbank verwendet.

1.2 Aufbau der Arbeit

Abschnitt 2 gibt eine Einführung in den Begriff der Literaturangaben, ihre Einsatzgebiete und unterschiedlichen Notationen. Anhand der Begriffe „Work“ und „Item“ wird zwischen einem abstrakten Dokument und Dateien, in denen dieses Dokument gespeichert ist, unterschieden.

In Abschnitt 3 werden Metadaten (insbesondere Dublin Core Metadata) und digitale Identifier sowie XML Namespaces behandelt. Dieser Abschnitt kann übersprungen werden, wenn die Begriffe bereits hinreichend bekannt sind.

Das Hauptthema der Arbeit, das Open Citation Project, wird in Abschnitt 4 vorgestellt. Beginnend mit einer kurzen Beschreibung der online verfügbaren Testseite, auf der nach verlinkten Dokumenten gesucht werden kann, wird der Prozess des Reference Linking, d. h. der Erzeugung einer Datei mit im Browser oder PDF-Reader aktivierbaren Links aus einer unverlinkten Datei beschrieben.

Die Arbeit schließt mit einer Übersicht verwandter Projekte (Abschnitt 5) und einer kurzen Zusammenfassung.

1.3 Notationen

Dieser Text folgt der neuen deutschen Rechtschreibung und Silbentrennung, wobei einige neue Schreibweisen nicht angewendet werden. Dazu gehören u. a. „nummerieren“, „aufwändig“ und „selbstständig“.

Viele Beispiele, die in diesem Text verwendet werden, sind englischsprachig, da sie aus Quellen übernommen wurden, die ebenfalls in dieser Sprache vorliegen. Englisch ist die dominierende Sprache im Bereich der naturwissenschaftlichen und technischen Veröffentlichungen.

2 Literaturangaben

In diesem Abschnitt beschäftigen wir uns mit einigen grundlegenden Überlegungen zu Zitaten und Literaturangaben in wissenschaftlichen Arbeiten.

2.1 Funktion und Sinn

Eine Literaturangabe ist ein Querverweis auf eine andere Arbeit. Ihre Verwendung kann verschiedene Gründe haben; wir gehen im Folgenden davon aus, dass sich in Dokument D ein Verweis auf Dokument V befindet.

Wir unterscheiden vier Kategorien, die wir als „Quelle“, „Grundlagen“, „Querverweis“ und „Sonstiges“ charakterisieren:

Quelle: Wenn der Autor von D ganze Sätze oder Absätze identisch aus V übernommen hat, dann ist dies in Arbeit D als Zitat kenntlich zu machen; anderenfalls würde der Eindruck erweckt, es handelte sich hier um originäre Ergebnisse des Autors von D. Der aus V übernommene Bereich wird dann deutlich als Zitat ausgezeichnet (etwa durch das Setzen von Anführungszeichen) und ein Verweis auf V eingefügt. Es spielt dabei keine Rolle, ob die Formulierungen aus V wortwörtlich übernommen wurden. Im naturwissenschaftlichen Bereich ist nicht nur eine Umformulierung, sondern auch die Änderung der Notation üblich, um etwa einen mathematischen Satz der eigenen Notation anzupassen.

Grundlagen: Wenn der Autor Konzepte und Anregungen aus V übernommen hat oder V einfach zur Grundlagenliteratur gehört, kann dies durch einen Verweis auf V angegeben werden. Dies ist auch dann üblich, wenn kein Inhalt aus V verwendet wurde.

Querverweis: Auch der Verweis auf Arbeiten, die sich mit demselben oder angrenzenden Themen beschäftigen („Related Work“) ist üblich. Dabei kann

es sich um Veröffentlichungen handeln, die einen ähnlichen Ansatz verfolgen, aber auch um solche, die vollständig anders vorgehen. In beiden Fällen helfen diese Angaben der Abgrenzung der aktuellen Arbeit von bisherigen Forschungsergebnissen.

Sonstiges: Es gibt verschiedene andere Gründe für ein Zitat, darunter das Vermitteln oder Anrufen von Autorität, die Dokumentation von Nähe oder Dankbarkeit gegenüber einem Lehrer.

Der erste dieser drei Punkte findet sich in vielen Prüfungsordnungen wieder – so heißt es etwa in § 19 (6) der Aachener DPO Informatik [RWT97]:

> Bei der Abgabe der Diplomarbeit hat die Kandidatin oder der Kandidat schriftlich zu versichern, dass sie oder er die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht hat.

Der erste Punkt stellt damit eine Minimalanforderung dar: Alle derartigen unmittelbaren Zitate müssen kenntlich gemacht werden, anderenfalls handelt es sich um Plagiat.

Die Funktion eines „autoritären Zitats“ beschreibt [Sch85] wie folgt:

> Das richtige Zitat signalisiert nicht nur Belesenheit, sondern macht Ansprüche geltend, die nicht ohne weiteres entwertet werden können, es sei denn, der Gesprächspartner halte sich für eine noch größere Autorität. Kurzum: Das Zitat einer Autorität ist ein autoritäres Zitat.

2.2 Form und Verwendung

Im Folgenden geht es um die syntaktische Form, in der Zitate verwendet werden, und um die Fragen, welche Texte „zitierfähig“ sind und welche Arbeiten in eine Bibliographie aufgenommen werden sollten.

2.2.1 Zitatform

Für die Art und Weise, in der diese Verweise angegeben werden, gibt es verschiedene Standards. In den meisten naturwissenschaftlichen Publikationen ist es üblich, in den Text an geeigneter Stelle einen Hinweis in eckigen Klammern einzufügen; der darin enthaltene Text kann eine rein numerische oder alphanumerische Angabe sein. Diese Ergänzung wird im Folgenden als *Anker* (engl.: anchor) bezeichnet. Am Ende des Textes gibt es dann einen Abschnitt, in dem diese Angaben aufgelistet werden. Bei sehr kurzen Arbeiten werden die Literaturangaben häufig schlicht

durchnumeriert; im Text haben die Anker dann die Form [1], [2], [3] etc. Diese Notation wird z. B. vom Journal of Computer and System Sciences für Artikel vorgeschrieben. [Jou00]

Bei längeren Texten oder Texten mit vielen Literaturangaben ist eine alphanumerische Bezeichnung übersichtlicher: Diese setzt sich z. B. aus Initialen oder den ersten Buchstaben der Autoren/des Autors und der Jahreszahl des Erscheinungsdatums zusammen, etwa [Ber99] für eine Arbeit von Donna Bergmark von 1999. Die meisten wissenschaftlichen Journals besitzen eigene Stil-Hinweise, die beim Erstellen dieser Angaben zu befolgen sind.

Texte aus dem geisteswissenschaftlichen Umfeld folgen in der Regel einem anderen System: Hier ist es üblich, am Ende des Zitats eine Fußnote einzufügen, die einen Verweis auf das Literaturverzeichnis enthält. Eine solche Fußnote hat dann z. B. die gleiche Gestalt wie die Fußnote zu diesem Satz.[2] Weitere Referenzen zum gleichen Dokument, die sich auf der selben Seite wie die erste befinden, werden dann oft mit „Ebd." (ebenda) oder „op.cit." (opus citatum) zitiert.[3] Diese Notation erschwert automatische Analyseprozesse. Da sich alle dieser Ausarbeitung zugrundeliegenden Texte nur mit naturwissenschaftlichen/technischen Artikeln beschäftigen, gehen wir auf abweichende Zitierweisen nicht ein und setzen die Verwendung von Ankern voraus.

In den Literaturangaben tauchen alle Anker wieder auf und bieten hier möglichst vollständige Hinweise zum referenzierten Text, also Name und Autor des Textes, Erscheinungsjahr (und -monat), bei regelmäßig erscheinenden Publikationen auch Jahrgangsnummer und Ausgabennummer, ferner Name und Ort des Verlags. Sind Dokumente auch online verfügbar, wird gelegentlich zusätzlich die Web-Adresse angegeben, unter der diese zu finden sind.

Ein typischer Eintrag hat die Form:

[BL01] Donna Bergmark and Carl Lagoze. An architecture for automatic reference linking. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 2001. To appear. Preprint available at http://www.cs.cornell.edu/cdlrg/Reference%20Linking/tr1842.ps.

Am Anfang stehen die Autorennamen, es folgt der Name des Artikels, dann der Konferenzband mit Jahresangabe. Der Hinweis „To appear" weist darauf hin, dass dieser Artikel noch nicht erschienen ist, es gibt ihn jedoch bereits als Pre-Print, und die Web-Adresse, unter der eine PostScript-Datei des Pre-Prints bezogen werden kann, schließt diese Literaturangabe ab.

Dank der Angaben ist es dem Leser möglich, sich die Literatur zu besorgen, die dem aktuellen Text zugrundeliegt bzw. mit ihm in Zusammenhang steht. Wo im

[2] Vgl. Eßer, Seminararbeit, 2001, S. 5-7
[3] Ebd., S. 5-7

Artikel etwa eine mathematische Aussage nur zitiert wird, finden sich im referenzierten Werk ein Beweis oder ausführlichere Erklärungen.

Einige Arbeiten bieten kommentierte Literaturverzeichnisse: Neben den reinen Referenzangaben finden sich zu einigen Einträgen noch zusätliche Hinweise, etwa eine kurze Inhaltsbeschreibung oder eine Bemerkung des Autors, inwieweit die jeweilige Arbeit für weiterführende Lektüre geeignet ist.

2.2.2 Unterschiede in der Verwendung

Die Analyse der Techniken und Stile bei der Verwendung von Zitaten ist ein eigenes Forschungsgebiet – dort betrachtet man die unterschiedliche Behandlung von Zitaten in den verschiedenen wissenschaftlichen Disziplinen. So findet sich etwa in den „Hinweisen für die Anfertigung wissenschaftlicher (Haus-) Arbeiten“ eines wirtschaftswissenschaftlichen Lehrstuhls [Wol00] folgender Hinweis:

> Als zitierfähig gelten „alle Quellen und Sekundärmaterialien, die in irgendeiner Form . . . veröffentlicht worden sind . . .“ Unveröffentlichte Literatur darf nur mit Genehmigung des jeweiligen Verfassers zitiert werden.
>
> Nicht zitierfähig sind demnach Seminar- und Diplomarbeiten oder Vorlesungsskripte. Werden diese ausnahmsweise doch zitiert, so ist der Urheber der Quelle, ggf. mit Adresse für weitere Informationen, in einer Fußnote und nicht im Literaturverzeichnis anzugeben.
>
> Nicht zitierwürdig sind genau genommen alle nichtfachlichen Veröffentlichungen, insbesondere jedoch sog. Publikumszeitschriften (Amica, Freundin, Bild der Frau oder Max) und vergleichbare Publikationen. Aus gegebenem, insbes. aktuellem Anlaß können im Ausnahmefall Zeitschriften wie „Der Spiegel“, „Focus“, „Die Zeit“ oder ähnliche zitiert werden. Dasselbe gilt für die Tagespresse.

In der Informatik ist es durchaus üblich, Diplomarbeiten oder Vorlesungsskripte zu zitieren, und die entsprechenden Referenzen erscheinen unter den übrigen Angaben am Ende des Textes; Fußnoten werden für Literaturangaben üblicherweise nicht verwendet.

Auch bei der Frage, welche Referenzen in das Literaturverzeichnis aufgenommen werden sollen, besteht keine Einigkeit: Klar ist, dass jeder zitierte Text in der Bibliographie einen Eintrag erhalten muss; die Umkehrung gilt nicht automatisch. Oft wird eine Arbeit in die Auflistung aufgenommen, weil sie zum Erstellen des Textes verwendet wurde, auch wenn kein Ausschnitt dieser Arbeit zitiert wird. Das Textsatzsystem LaTeX kennt hierfür den `\nocite`-Befehl. Auch in kommentierten Literaturverzeichnissen finden sich oft solche ergänzenden Einträge.

2.3 Artikelgraph

Betrachtet man die Menge aller Artikel und die Literaturreferenzen zwischen ihnen, so lässt sich aus ihnen ein gerichteter Graph erstellen: Die Knoten des Graphs sind die Artikel, und eine gerichtete Kante von D nach V existiert, wenn D einen Verweis auf V enthält.

Eine offensichtlich ähnliche Struktur bildet die Menge aller im Internet verfügbaren Web-Seiten – ein Link einer Seite D auf eine Seite V entspricht dem Literaturhinweis, und es lässt sich auf gleiche Weise ein Graph bilden. Es fallen aber einige Unterschiede auf:

- Zuordnungen sind eindeutig, da ein (korrekter) Link einen eindeutigen Speicherort angibt. Bei Literaturangaben ist eine Suche notwendig, und dafür erforderliche Angaben können fehlen.
- Anders als wissenschaftliche Artikel sind Web-Seiten veränderlich: Ihr Inhalt kann zu unbestimmten Zeitpunkten geändert werden, oder eine Seite kann völlig verschwinden und einen ehemals korrekten Link damit ungültig machen.[4]
- Kreise im Graph sind bei Web-Seiten viel häufiger, da Seiten zum gleichen Thema häufig untereinander verlinkt sind; ein typisches Beispiel sind Web-Ringe, bei denen die Mitglieder im Kreis aufeinander verweisen.

2.4 Items und Works

Beim Vergleich von Literaturverweisen und Web-Links ist die Eindeutigkeit von Links besonders interessant. Wir wollen im Folgenden einen Artikel von seinen (möglicherweise vielen) Speicherorten begrifflich trennen.

Als *Work* bezeichnen wir eine wissenschaftliche Veröffentlichung. Das kann ein Artikel, ein Buch, ein Kapitel eines Buches oder ein sonstiges Dokument sein. Referenzen beziehen sich in der Regel abstrakt auf das Work. Davon unterschieden wird ein *Item*,[5] welches eine von möglicherweise mehreren Repräsentationen

[4] Auch Artikel können Veränderungen unterliegen; so kann die endgültig in einem Journal erschienene Fassung marginal von einem Pre-Print abweichen, obwohl beide Dokumente als „im wesentlichen“ identisch betrachtet werden. Durch exakte Literaturangaben kann dieses Problem umgangen werden.

[5] Da sich „Work“ und „Item“ nicht vernünftig ins Deutsche übersetzen lassen, werden in dieser Arbeit die englischen Begriffe verwendet, wie sie in [BL01a] definiert wurden. „Work“ lässt sich als „Werk“ übersetzen, zu „Item“ passt am ehesten der Begriff „Instanz“. Da letzterer im Englischen aber „instance“ genannt wird und dies eine Abwandlung der ursprünglichen Notation darstellt, wurde auf eine Übersetzung verzichtet.

des Works ist. Neben unterschiedlichen Speicherorten kann auch das Format abweichen (PDF, PostScript, verlinktes PDF). Für unsere Zwecke sollen dies Web-Adressen (http oder ftp, etwa http://www.uni-x.de/pub/tk2012.pdf) sein (allgemeiner kann man auch beispielsweise in einer Bibliothek gesammelte Konferenzbände dazu zählen; wir beschränken uns auf online verfügbare Dateien).

Zu einem Work gehören somit 0 oder mehr Items [BL01a]. Während sich jedes dieser Items eindeutig benennen lässt (Web-Adressen sind eindeutige *Identifier*, mehr dazu in Abschnitt 3.2), gilt dies für das Work nicht. Ein Work ohne Item ist nach dieser Begriffsbildung eine Veröffentlichung, die nicht online zugänglich ist.

Wir betrachten die neu eingeführten Begriffe an einem Beispiel: Der Artikel „Integrating, navigating and analyzing eprint archives through Open Citation Linking (the OpCit Project)“ von Stevan Harnad und Leslie Carr, erschienen in Current Science Online (Vol. 79, No. 5) ist ein Work. Im Internet finden sich zwei Items, also digitale Repräsentationen dieses Works, eine PDF- und eine HTML-Version:

- *http://tejas.serc.iisc.ernet.in/~currsci/sep102000/629.pdf*
- *http://www.cogsci.soton.ac.uk/~harnad/Papers/Harnad/harnad00.citation.htm*

Beide Dateien sind zum gleichen Work gehörende Items. Wir legen hier fest, dass ein Item über seine URL identifiziert wird. Ist eine Datei durch symbolische Links auf dem Web-Server über verschiedene URLs erreichbar, so betrachten wir diese als separate Items, da die Dateistrukturen auf dem Web-Server von außen nicht transparent sind.

2.4.1 Varianten eines Works

Varianten eines Dokumentes wie verschiedene Auflagen eines Buches oder Pre-Print und endgültige Version eines Artikels sind zu unterscheiden. Im strengen Sinne handelt es sich um unterschiedliche Works. Lässt man die strenge Abgrenzung fallen, ergibt sich der Vorteil, besseren Zugriff auf Dokumente zu haben: So wird dann etwa die online verfügbare PostScript-Version eines Pre-Prints in der Liste der zu einem Work gehörenden Items angezeigt, obwohl das Work selbst die endgültige Fassung ist.

Im Verlagswesen wird zwischen identischen und überarbeiteten Neuauflagen unterschieden: So behält ein Buch beim unveränderten Nachdruck seine ISBN, während bei wesentlicher Überarbeitung eine neue ISBN vergeben wird: Unter [LLC] finden sich die entsprechenden Hinweise

> When a title is reprinted, regardless of price change, the original ISBN must be maintained. [...]
>
> Revised editions require a new ISBN.

Dies legt nahe, dass zwei Auflagen eines Buches mit gleicher ISBN als identisches Work (und dementsprechend alle Online-Versionen als gemeinsam zu diesem Work gehörende Items) angesehen werden, während Auflagen mit unterschiedlichen ISBN auch als Works unterschieden werden.

3 Technische Grundlagen

Es folgen nun einige weitere Grundlagen zu digitalen Identifiers (Handle System, DOI – Digital Object Identifiers) und Metadaten (Dublin-Core-Standard).

3.1 Metadaten

Metadaten sind Informationen, die in einem Dokument enthalten sind, aber im Grunde nicht Teil des Dokumentinhalts sind, sondern lediglich seiner Beschreibung dienen. In HTML-Seiten finden sich etwa im Kopf oft `meta`-Tags der Form

```
<head>
  <meta name="generator" content="HTML Tidy, see www.w3.org">
  <meta name="htdig-noindex">
  ...
</head>
```

Im Folgenden geben wir HTML-Code in XHTML-Syntax [W3C00] an.

Metadaten werden vom Browser nicht angezeigt und in der Regel auch nicht ausgewertet, dennoch stellen sie mehr als etwa Kommentare in einem Programm-Quelltext dar: Wenn sie bestimmten Syntaxregeln und Standards folgen, können sie von verarbeitenden Programmen ausgewertet werden.

Ohne eine Auflistung möglicher `meta`-Tags[6] geben zu wollen, ist hier die Beobachtung interessant, dass viele Tags dem Aufbau

```
<meta name="Name" content="Inhalt" />
<meta name="Name" />
```

folgen; auf diese Weise werden Web-Browsern oder Tools, die sich automatisch durch das Web bewegen (etwa Suchmaschinen), Informationen übergeben, die beim Anzeigen einer Web-Seite ignoriert werden – bestimmte Programme werten diese aber aus. So lässt sich zum Beispiel dem Indizierer ht://dig (*http://www.htdig.org/*)

[6] Laut HTML-4.01-Spezifikation [W3C99a] (Abschnitt 7.4.4: „Meta data“) sind die Standard-Meta-Tags aus der RDF-Spezifikation (Resource Description Framework, [W3C99b]) zu übernehmen.

durch das Vorhandensein eines Meta-Tags mit dem Namen *htdig-noindex* mitteilen, dass er zwar Links auf der aktuellen Seite weiterverfolgen, die Seite selbst aber nicht indizieren soll.

Auch in E-Mails werden Metadaten verwendet; der Header einer E-Mail enthält Zeilen der Form

```
From: <Autor>
To: <Empfänger>
Cc: <Durchschlag-Empfänger>
Subject: <Thema der E-Mail>
Reply-To: <Rückantwort-Adresse>
```

Diese sind in RFC 821 [Pos82] und RFC 822 [Cro82] definiert. Sogenannte „Extension Fields“ erlauben zusätzliche Angaben, die vom RFC nicht berücksichtigt wurden: Diese werden durch „X–“ eingeleitet. Ein Beispiel ist die Angabe des zur Erstellung verwendeten E-Mail-Programms in der Form

```
X-Mailer: Mozilla 4.77 [en] (X11; U; Linux 2.4.3-20mdksmp i6
86)
```

3.1.1 Dublin Core Metadata

Dokumente im Netz bieten im Allgemeinen recht wenig automatisch erfassbare Informationen über den Inhalt: Außer einer Volltextindizierung lässt sich höchstens der Inhalt des `<title>`-Elements zur Beschreibung einer Seite heranziehen. Durch die Verwendung zusätzlicher Meta-Tags können Zusatzinformationen (etwa über den Autor, das Erstellungsdatum etc.) integriert werden. Die Dublin Core Metadata Initiative [Hil01], die aus einem Workshop in Dublin, Ohio (1995) hervorgegangen ist, hat einen Standard für die Benennung solcher Meta-Tags geschaffen, um hier für Einheitlichkeit zu sorgen und ein Metadaten-Vokabular zu entwickeln, das komplexere Inhalte der Metadaten erlaubt, die von Automaten analysiert werden können. Eine HTML-Datei könnte dann etwa die folgenden Tags enthalten:

```
<meta name    = "DC.Title"
      content = "Seminar: Open Citation Project" />
<meta name    = "DC.Subject"
      content = "Open Citation Project, Seminarvortrag" />
<meta name    = "DC.Date.Created"
      content = "2001-09-19" />
<meta name    = "DC.Format.Medium"
      scheme  = "IMT"
      content = "text/html" />
<meta name    = "DC.Creator"
      content = "Eßer, Hans-Georg" />
<meta name    = "DC.Identifier"
      content = "http://esser-books.com/hg/inf/opcit-sem.htm
l" />
```

Die Verwendung der Dublin-Core-Metadaten ist aber nicht auf HTML beschränkt; auch in XML-Dateien, die zunehmend für die Speicherung aller möglichen Arten von Daten verwendet werden, können Dublin-Core-Daten in Form von XML-Elementen eingefügt werden. Der obige Block könnte in XML zum Beispiel wie folgt aussehen:

```
<DC:Title>Seminar: Open Citation Project</DC:Title>
<DC:Subject>Open Citation Project, Seminarvortrag</DC:Subjec
t>
<DC:Date.Created>2001-09-19"</DC:Date.Created>
<DC:Format.Medium>text/html</DC:Format.Medium>
<DC:Creator>Eßer, Hans-Georg</DC:Creator>
<DC:Identifier>http://esser-books.com/hg/inf/opcit-sem.html<
/DC:Identifier>
```

Die Bedeutung des allen Elementen vorangehenden „DC:“ erläutern wir im Abschnitt 3.3 über XML Namespaces.

3.2 Object Identifiers

Ziel von Object Identifiers ist es, Objekte absolut zu benennen, so dass eine eindeutige Zuordnung eines Objektes zu seinem Identifier möglich ist. Bei Web-Seiten ist dies der Fall: So ist etwa die Startseite der RWTH Aachen über die URL http://www.rwth-aachen.de/index.html eindeutig zu bestimmen.

Veröffentlichungen (Works) eindeutige Identifier zuzuordnen, ist schon schwieriger: Verschiedene Verlage, Hochschulen oder sonstige Dokumentanbieter verwalten eine Unzahl von Dokumenten, und es ist sicher zu stellen, dass kein Identifier doppelt vergeben wird. Wir stellen im Folgenden zwei Projekte vor, die eindeutige Identifier-Vergabe garantieren: das *Handle System* und die *International Digital Object Identifier (DOI) Foundation*. Da DOI auf dem Handle System basiert, beschreiben wir dieses zuerst.

3.2.1 Handle System

> „The Handle System is a general-purpose global name service that allows secured name resolution and administration over the public Internet.“ [SL01]

Das Handle System der amerikanischen Non-Profit-Organisation „Corporation for National Research Initiatives“ (*http://www.cnri.reston.va.us/*) bezeichnet Identifier als „Handles“ und sichert ihnen u. a. die folgenden Eigenschaften zu [SL01]:

Einmaligkeit Jedes Handle ist global einmalig innerhalb des Handle Systems.

Beständigkeit Da Handles nicht automatisch aus Daten über das Objekt abgeleitet, sondern vom System vergeben werden, kann ein Objekt auch bei Veränderung äußerer Umstände (wie Speicherort) dem Handle zugeordnet bleiben.

Mehrfache Instanzen Ein Handle kann auf mehrere Instanzen des Objekts verweisen, also in der Notation von oben auf mehrere Items eines Works.

Delegation Handle Services können die Verwaltung eines eingeschränkten, lokalen Namensraumes übernehmen. Damit fallen Administration und Namensauflösung für diesen Teilraum an die lokale Handle Authority. Diese Delegation ist auch mehrstufig möglich.

Ein Handle besteht immer aus zwei, durch „/“ getrennten, Teilen: Der erste identifiziert die *Handle Naming Authority*, der zweite den lokalen Handle-Namen. Bei den lokalen Namen muss nun keine Rücksicht auf Übereinstimmung mit Bezeichnern von anderen Naming Authorities genommen werden, da die vollständigen Handles sich schon durch den ersten Teil unterscheiden.

Ein Beispiel ist „10.1045/january99-bearman“, wobei 10.1045 der Code für das D-Lib Magazine ist. Der Punkt zwischen 10 und 1045 trennt einzelne Hierarchiestufen der Handle Naming Authorities ab.

3.2.2 Digital Object Identifier

Digital Object Identifiers [Int01] bauen auf dem Handle-System auf. Neben der Vergabe und Gewährleistung von Eindeutigkeit und Beständigkeit der DOI bietet das DOI-System den Service, Informationen über DOI zur Verfügung zu stellen und eventuell Zugang zu einer digitalen Repräsentation (im Falle eines Artikels etwa zu einem Item) zu verschaffen. Dazu werden zu jedem Objekt mindestens die folgenden Metadaten erfasst:

Identifier Ein aus einem anderen System stammender Identifier (z. B. ISBN)

Titel Name des Objekts

Modus Typ des Objekts: etwa eine schriftliche Arbeit oder eine Aufführung

Primäragent In der Regel der erstgenannte Ersteller des Objekts

Agentenrolle Funktion, die der Primäragent bei der Erstellung hatte, z. B. Autor.

Um die Speicherung zusätzlicher Informationen zu ermöglichen, gibt es die sogenannten DOI Application Profiles (DOI-AP). Diese dienen der späteren Nutzung in Programmen, als Beispiel gibt die DOI-Homepage Digital-Rights-Management-

(DRM-) Systeme an. Jedes Objekt wird mindestens einem dieser Profile zugeordnet; so lassen sich beliebige Arten von Objekten verwalten.

Die Zuordnung zu einer digitalen Repräsentation (im DOI-System Auflösung genannt), übernimmt das Handle-System (Abschnitt 3.2.1). Dabei können einem DOI auch mehrere URLs, E-Mail-Adressen etc. zugeordnet werden (Abbildung 1).

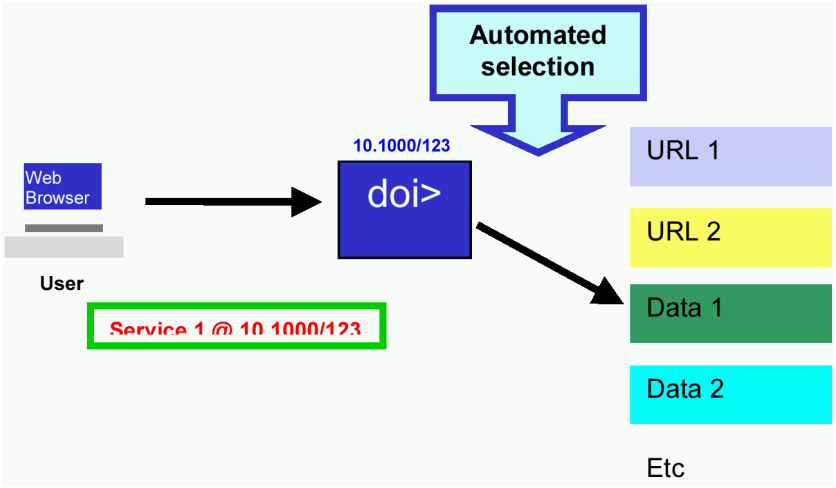

Abbildung 1: Zuordnung verschiedener Internet-Ressourcen über das DOI-System (Quelle: [Int01])

DOI sind Handles; Sie beginnen immer mit „10.“ – die allgemeine Form ist also „10.*/*“. Das Beispiel im Abschnitt zum Handle-System („10.1045/january99-bearman“) ist also ein DOI. Die Nummer zwischen „10.“ und „/“ ist die Registriernummer der Organisation, die diesen DOI registriert hat.

3.3 XML Namespaces

XML, die Extensible Markup Language [BPSMM00], ist – anders als etwa HTML oder LaTeX – eine Meta-Markup-Sprache, die eine Definition eigener Markup-Sprachen erlaubt; eine in XML definierte Sprache wird *XML-Anwendung* genannt.[7] Anwendungen benutzen in einer DTD erklärte Element- und Attributnamen.

Sollen nun in einem Dokument zwei XML-Anwendungen verwendet werden, so könnte es zu Namenskonflikten zwischen doppelt definierten Elementen oder Attributen kommen – aus diesem Grund wurden in XML *Namespaces* [BHL99] definiert: Jede Anwendung wird an einen separaten Namensraum gebunden, so dass Zuordnungen stets eindeutig sind.

[7] Für einen ersten Einstieg in XML empfiehlt sich die Lektüre eines Buches wie „Inside XML“ [Hol00].

Wir geben dazu ein Beispiel: RDF-Dateien (Resource Description Framework) werden häufig verwendet, um Inhalte von News-Tickern und ähnlichem auf beliebigen Web-Seiten verfügbar zu machen. Ein XML-Dokument, welches nur die RDF-Elemente verwenden möchte, könnte dann den folgenden Aufbau haben:

```
<RDF xlmns="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <Description>Ein Beispiel</Description>
  <Bag>
    <li resource="http://www.w3.org/" />
    <li resource="http://www.rwth-aachen.de/" />
  </Bag>
</RDF>
```

Soll nun eine zweite XML-Anwendung im gleichen Dokument verwendet werden, werden Name Spaces explizit benannt und den Elementen vorangestellt:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <rdf:Description about="http://www.rwth-aachen.de/">
    <dc:Format>HTML</dc:Format>
    <dc:Language>de</dc:Language>
    <dc:Type>Homepage</dc:Type>
  </rdf:Description>
</rdf:RDF>
```

XML-Dateien dieser Form werden vom Open Citation Project zum Verwalten der Artikeldaten verwendet.

4 Open Citation Project

Im Original-Proposal des Open Citation Project [GHL+99] wird das folgende Ziel beschrieben:

> It is easy to say what would be the ideal online resource for scholars and scientists: all papers in all fields, systematically interconnected, effortlessly accessible and rationally navigable from any researcher's desk worldwide.

Diese Aussage enthält zwei getrennte Ziele: Der Wunsch, alle wissenschaftlichen Veröffentlichung über das Internet für jeden zugänglich zu machen, dürfte noch eine Weile an dem Wunsch nach kommerzieller Vermarktung der Journals scheitern. Das zweite Ziel hingegen, die Verknüpfung der (frei verfügbaren) Artikel, kann durch die vom OpCit-Projekt entwickelte Software schon teilweise gelöst werden.

4.1 Vorführung der Implementation

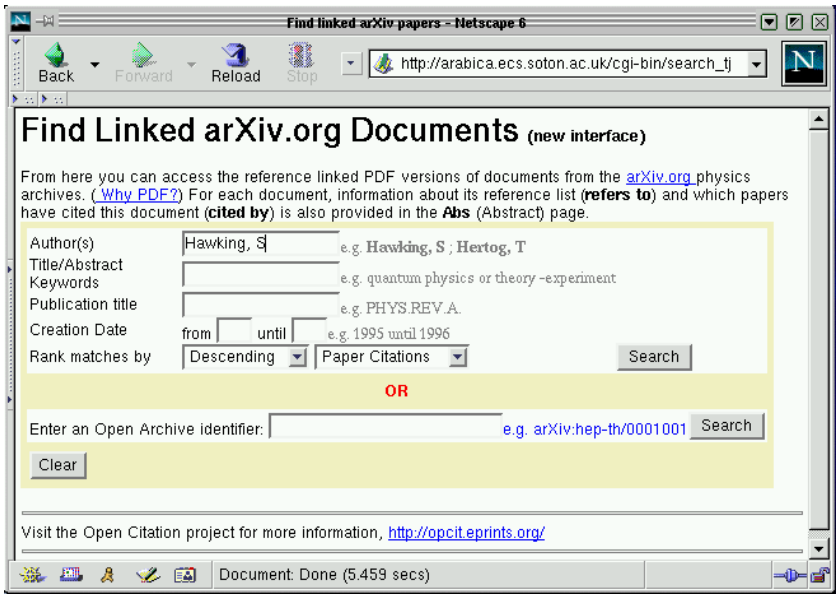

Abbildung 2: Der Test beginnt mit der Suche eines Dokuments. Hier lassen sich wahlweise verschiedene Schlüsselworte oder ein „Open Archive Identifier“ eingeben.

In diesem Abschnitt betrachten wir die „Citation and Reference Linking Demo“, *http://arabica.ecs.soton.ac.uk/*. Für die Demo wurden die über 150.000 Dokumente in den Physik-Archiven, die unter *http://arXiv.org* zur Verfügung stehen, um Links ergänzt. Die Demo-Seite erlaubt einen Einstieg über die Suche nach einer Veröffentlichung. Aus den Suchergebnissen kann eine PDF-Datei ausgewählt werden, und nach dem Download wird diese im Acrobat Reader bzw. (falls ein entsprechendes Plugin installiert ist) im Browser angezeigt. Die PDF-Dokumente sind dabei über URLs der Form

```
http://arabica.ecs.soton.ac.uk/cgi-bin/lpdf?id=oai:arXiv-hep
-th/9909205
```

erreichbar.

Im Folgenden zeigen wir anhand eines Beispiels die Verwendung dieser Demo.

Zunächst wird mit einem beliebigen Dokument gestartet. Dazu suchen wir, wie auf der Startseite als Beispiel empfohlen, nach Veröffentlichungen von Stephen Hawking und geben im *Author*-Feld „Hawking, S“ ein (Abbildung 2).

Es erscheint dann eine Liste passender Veröffentlichungen, aus der ein Dokument ausgewählt werden kann. Nach Auswahl eines Links wird das PDF-Dokument geladen und im Browser oder einem externen PDF-Viewer angezeigt.

Die folgenden zwei Screenshots wurden aus [Hit00] entnommen, da die Demo zum Zeitpunkt der Fertigstellung dieser Arbeit nicht funktionierte: Die Suchseite war zwar noch online, konnte aber keine Suchergebnisse finden. Abbildung 3 zeigt einen Ausschnitt des Literaturverzeichnisses: Alle Einträge verfügen über einen Link, der zum referenzierten Dokument führt. Eine Sonderrolle spielt dabei der fünfte, grün hervorgehobene Eintrag: Dieser führt zu einer HTML-Seite, die eine Auswahl verschiedener Texte erlaubt (Abstract, verlinktes PDF des OpCit-Projekts, Original-Text; Abbildung 4).

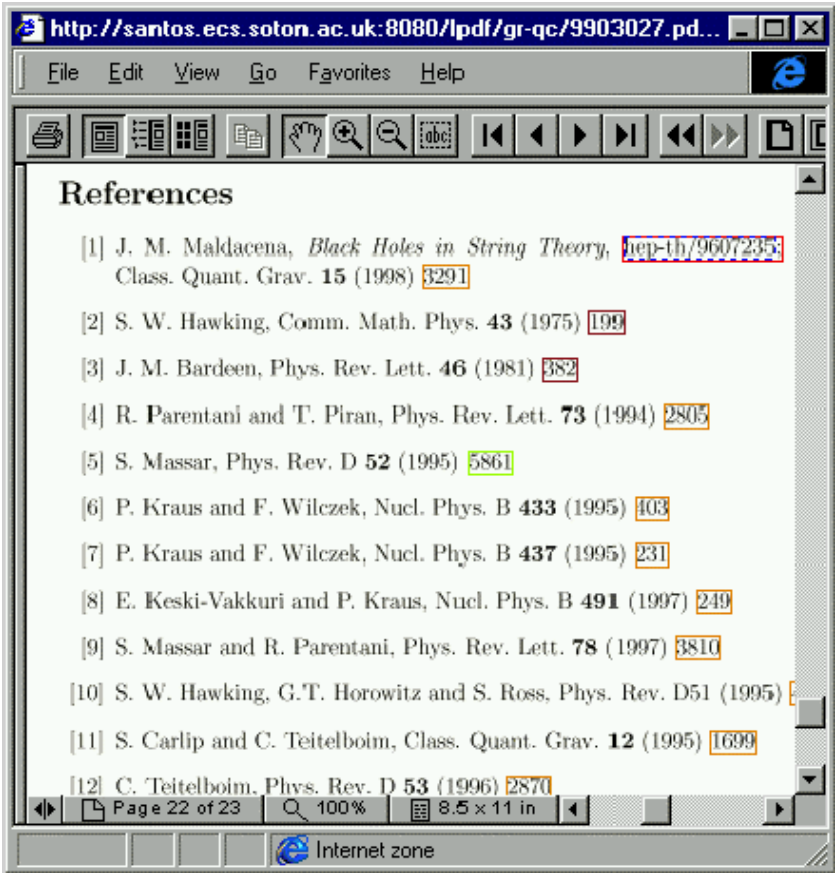

Abbildung 3: Die Literaturangaben dieses Dokuments sind verlinkt. (Quelle: [Hit00])

4.2 Reference Linking

Reference Linking ist der Prozess, der ein online verfügbares Dokument so bearbeitet, dass Literaturverweise (also Angaben zu Works) in Links auf zugehörige Items umgewandelt werden (Abbildung 5).

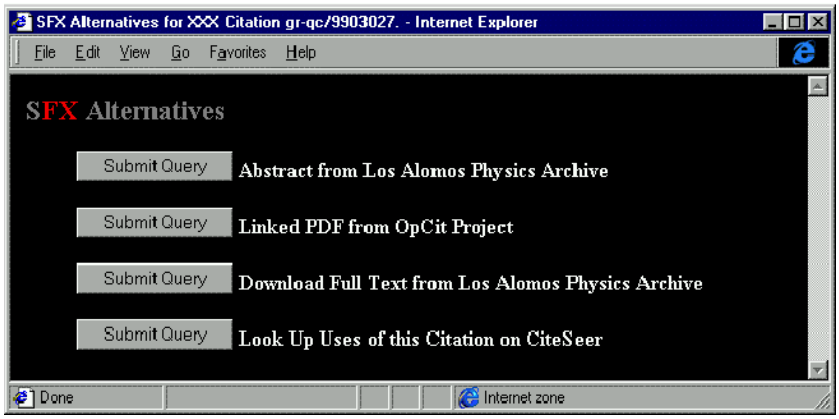

Abbildung 4: Hier sind zum Work mehrere Items verfügbar, darunter der Original-Text und eine um Links ergänzte PDF-Version. Das ebenfalls wählbare Abstract ist im eigentlichen Sinn *kein* Item. (Quelle: [Hit00])

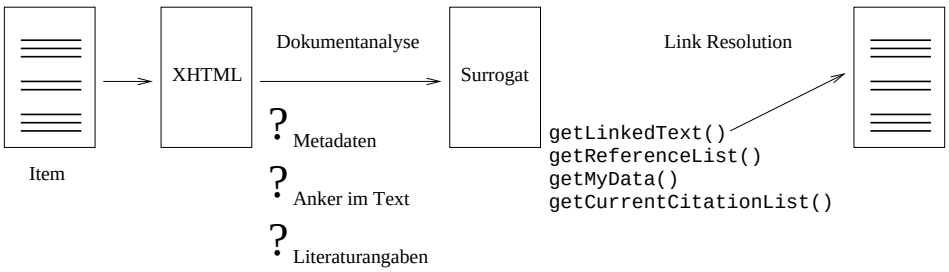

Abbildung 5: Die Schritte des Reference Linking: Durch Extraktion von Metadaten, Ankern im Text und Literaturangaben wird ein „Surrogat“ erzeugt. Dies ist ein Objekt, das neben dem Original-Dokument die gewonnenen Daten speichert und auf Wunsch ein verlinktes Dokument erzeugt.

Zwei Varianten von Verweisen auf online verfügbare Literatur sind besonders häufig anzutreffen:

- Eine Literaturangabe kann die konkrete URL eines Dokuments enthalten. In diesem Fall ist die Umwandlung in einen Link beinahe trivial. Allerdings wird so nur das Dokument in einer speziellen Form referenziert, es handelt sich also um ein Item. Die ebenfalls um Links erweiterte Fassung (so sie existiert) kann sich an anderer Stelle befinden.
- Eine Literaturangabe gibt Name des Journals, Nummer (Ausgabe), Autor und Titel an. Hier hilft die Suche in einem online verfügbaren Index des Journals weiter.

4.3 Dokumentanalyse

Bei der automatischen Bearbeitung einer Sammlung online verfügbarer Dokumente wird jedes Dokument in verschiedenen Schritten nach Informationen durchsucht. Neben den Referenzen eines Dokuments muss auch festgestellt werden, um welches Dokument es sich gerade handelt. Diesen Prozess beschreibt der folgende Abschnitt.

4.3.1 Extraktion der Metadaten

Metadaten eines Dokumentes sind die zentralen Informationen über dieses Dokument, die gespeichert werden: Autor, Titel und Jahr der Veröffentlichung. [Ber00] beschreibt einen Algorithmus, der ein XHTML-Dokument nach diesen Daten durchsucht:

```
set title1 = value of <title> element if there is one
Scan for any of the following:
  <H1>text</H1>, <H2>text</H2>, <font size="+3">text</font>,
  <font size="+2">text</font>, <font size="5">text</font>
set title2 = "text"
if title2 is shorter than title1,
  then scan for subtitle and append to title2
```

Dieser Algorithmus greift auf Layout-Informationen im XHTML-Quelltext zurück; dabei wird davon ausgegangen, dass ein Teil des Dokumenttitels im `<title>`-Tag enthalten ist, während ein weiterer Teil durch `<h1>`, `<h2>` oder eine ähnliche Auszeichnung im Body des Dokuments hervorgehoben wurde.

Komplizierter gestaltet sich die Suche nach dem Autor bzw. den Autoren. Hier gibt [Ber00] die folgenden vier Regeln, um den Abschnitt mit den Autoreninformationen zu finden:

1. Suche den Autor in der ersten Zeile nach dem Titel.
2. Suche Text innerhalb von `<p>...</p>`, `<center>...` oder `<strong>...`
3. Einzelne Autorennamen sind durch Tags (z. B. `<br>`) oder Kommata getrennt.
4. Der Autorenabschnitt wird durch die erste Überschrift (etwa `<h3>`) beendet.

Das Veröffentlichungsjahr ergibt sich manchmal aus der URL des untersuchten Dokuments; im Text selbst ist es meist nicht enthalten. Falls bereits Dokumente untersucht wurden, die auf diesen Text verweisen, kann das Jahr eventuell aus diesen Literaturangaben entnommen werden.

4.3.2 Extraktion der Anker aus dem Body

Als „Body“ des Dokuments definieren wir den gesamten Bereich des Textes zwischen dem Titel/Autor-Block und dem Beginn der Literaturangaben.

Innerhalb des Bodys wird der Text nach Verweisen auf die Literaturangaben (Ankern) durchsucht; diese haben im naturwissenschaftlichen Umfeld in der Regel die Form [1], [2,4], [3–5] etc. oder (Müller, 1999). Komplizierter wird es, wenn indirekte Verweise der Form „Wie Müller bereits 1999 gezeigt hat …“ verwendet werden. Auch für die Analyse solcher Formulierungen gibt [Ber00] einen Algorithmus.

Nachdem der Typ der Indizierung eines Dokuments anhand des ersten Verweises erkannt wurde, wird für den Rest des Dokuments angenommen, dass diese Konvention eingehalten wird. Bei der Analyse der D-Lib-Artikel wurde festgestellt, dass die Beachtung von sechs Konventionen ausreicht:

[1]: Numerische Aufzählung der Literaturangaben, typischerweise in der Reihenfolge des ersten Auftretens im Text

(Müller, 1999): Name des Autors und Jahr der Veröffentlichung, durch Komma getrennt in runden Klammern

(Maier, 1998, Müller, 1999): Variante mit mehreren Verweisen in einer Klammerung

[Maier, 1998, Müller, 1999]: Variante mit eckigen Klammern

[Mül1999]: Autorenkürzel (erste drei Buchstaben eines Einzelautors oder Anfangsbuchstaben der Nachnamen mehrerer Autoren) plus Jahr der Veröffentlichung in eckigen Klammern

{**Mül1999**}**:** Variante in geschweiften Klammern

In späteren Schritten werden die gefundenen Verweise durch Links auf den entsprechenden Eintrag im Literaturverzeichnis ersetzt.

4.3.3 Extraktion der Literaturangaben

Die meisten Dokumente besitzen am Ende des Textes separate Abschnitte mit Titeln der Form „References“, „Bibliography“ oder „Notes and References“.[8]

Der Literaturabschnitt wird zunächst in einzelne Literaturangaben zerlegt. Von jeder wird dann das Anker-Tag (z. B. „[1]“) entfernt, und der Rest wird von der `deciter`-Routine aus dem Distributed-Link-Software-Paket (DLS) [CRHH95] ausgewertet.

4.3.4 Dokument-Identifier

Als Dokument-Identifier für das gerade untersuchte Dokument wird ein String aus den folgenden Bestandteilen zusammengesetzt: Nachname des Autors, Jahr der Veröffentlichung und die ersten 20 Zeichen des Titels (in Kleinbuchstaben) [Ber00]; alternativ wird ein Handle-System-Identifier der Form „10.1045/january-99-bearman“ (siehe Abschnitt 3.2.1) verwendet – in [BL01a] sogar beide gleichzeitig in zwei separaten `dc:identifier`-Elementen.

4.4 Link Resolution

Bei der „Link Resolution“, also Link-Auflösung, werden Literaturangaben, zu denen entsprechende Online-Dokumente gefunden wurden, in echte Links umgewandelt, die anklickbar sind. Es findet also eine Zuordnung von Works und Items statt.

4.4.1 Statische Link Resolution

Bei der statischen Link Resolution, wie sie z. B. von ResearchIndex (siehe Abschnitt 5) angewandt wird, werden beim Analyseprozess die bekannten Items mit den Literaturangaben verknüpft. Diese Verknüpfungen werden in einer Datenbank abgelegt und können beim Abruf eines Dokumentes zum Erzeugen einer HTML- oder PDF-Datei mit direkten Links auf die Referenzen verwendet werden. Diese Vorgehensweise hat den Vorteil, dass jedes neue Item nur einmal analysiert werden muss.

[8] Alle diesem Seminarvortrag zugrundeliegenden Arbeiten beziehen sich ausschließlich auf englischsprachige Veröffentlichungen; siehe Einleitung.

4.4.2 Dynamische Link Resolution

Bei der dynamischen Link Resolution findet dieser Abgleich erst zum Zeitpunkt der Anfrage statt, so dass mit der Zeit mehr und mehr Dokument-Orte gefunden werden können; zwischenzeitlich nicht mehr gültige Orte können auch weg fallen. Hier ist auch kontextsensitive Link Resolution möglich, so dass etwa bei mehreren Speicherorten der FTP- oder Web-Server gewählt wird, der die geringste Entfernung zum Anwender hat.

4.4.3 Surrogate statt Datenbank

Anstelle einer umfassenden Datenbank, die sämtliche Informationen über alle analysierten Dateien speichert, verwendet das OpCit-Projekt sogenannte Surrogate (dt.: Stellvertreter).

> A surrogate is a digital object that encapsulates reference linking information relating to one single item on the Web. [BAL00]

Ein Surrogat ist ein Objekt (im Sinne objektorientierter Programmierung), in dem alle Daten aus dem Analyseprozess abgelegt werden. Sind zu einem Work mehrere Items analysiert worden, so führt das auch zu mehreren Surrogaten. Über ein API können die Surrogate angesprochen werden; sie liefern dann XML-Code mit den folgenden Informationen zurück:

- `getLinkedText`: Inhalt des Items (Text, PDF), ergänzt um Reference-Linking-Daten
- `getReferenceList`: Liste der Literaturangaben in diesem Item
- `getMyData`: Metadaten des Items
- `getCurrentCitationList`: Die Liste der Works, die dieses Item zitieren (soweit es dem Surrogat bisher bekannt ist)

Ein Ausschnitt einer beispielhaften Ausgabe der Funktion `getReferenceList` für ein Dokument mit 17 Literaturangaben sieht folgendermaßen aus:

```
<api:reference_list length="17"
  xmlns:api="http://www.cs.cornell.edu/cdlrg/..."
  xmlns:dc="http://purl.org/DC">
...
<api:reference ord="2">
<dc:title>
Smart Objects, Dump Archives: A User-Centric, Layered
```

```
Digital Library Framework
</dc:title>
<dc:date>1999-03-01</dc:date>
<dc:identifier>10.1045/march99-maly</dc:identifier>
<dc:creator>K Maly</dc:creator>
<api:displayID>
http://www.dlib.org/dlib/march99-maly/03maly.html
</api:displayID>
<api:literal tag="2.">
Maly K, "Smart Objects, Dumb Archives: A User-Centric,
Layered Digital Library Framework" in D-Lib Magazine,
March 1999,
&lt;http://www.dlib.org/dlib/march99-maly/03maly.html&gt;.
</api:literal>
<api:context_list>
<api:context>
The need for standards to support the interoperation of
digital library systems has been reported on before in
D-Lib[1],[2] as have efforts to discover common ground in
related standard processes(Dublin Core and INDECS[3]).
</api:context>
</api:context_list>
</api:reference>
...
</api:reference_list>
```

Im `api:reference_list`-Element wird zunächst über das `length`-Attribut angegeben, wieviele Referenzen im Folgenden aufgelistet werden (hier 17).

Das `api:reference`-Element beschreibt eine der Referenzen; die vollständige Ausgabe enthält also 17 `api:reference`-Elemente innerhalb des `api:reference_list`-Elementes. Die üblichen bibliographischen Informationen sind als Dublin-Core-Elemente dargestellt (siehe Abschnitt 3.1.1), und diese sind von den Zusatzinformationen durch Verwendung zweier Namespaces (`dc` und `api`) getrennt. Das `api:literal`-Element enthält die exakte Schreibweise der Literaturangabe, wie sie im Item gefunden wurde; im `tag`-Attribut wird separat der Anker dieser Literaturangabe aufgeführt – in diesem Fall „2.“ und nicht „[2]“.

Im `context` wird der vollständige Satz gespeichert, in dem diese Referenz zitiert wurde. In [BL01a] wird das gleiche Beispiel angeführt, hier ist die Zeile

```
<api:context>
```

allerdings durch

```
<api:context ord="9" anchor="[2]" normalization="[2]">
```

ersetzt: Die Zusatzinformationen heben nochmals die Schreibweise des Ankers im Text (und eine hier identische, normalisierte Form) hervor. Gibt es mehrere Zitate

dieser Arbeit, so finden sich im `context_list`-Element mehrere `context`-Elemente.

Das `dc:identifier`-Element enthält hier übrigens einen Handle, wie in Abschnitt 3.2.1 beschrieben.

Die PDF-Dateien, die in der Demo-Vorführung (Abschnitt 4.1) besprochen wurden, werden aus den Ergebnissen der `getLinkedText`-Funktion generiert.

4.5 Aufbereitung der Darstellung

Sind alle Daten vorhanden, ist das Generieren eines um Links erweiterten Artikels eine einfache Aufgabe, solange der Originaltext im HTML-Format vorliegt: `getLinkedText` ersetzt in der ursprünglichen Datei Ausschnitte der Form

```
... it was said [5] that ...
```

durch

```
... it was said
<api:reflink ord="5" author="last-name-of-first-author"
title="title of this work" year="1999">
<api:url>"http://www.some.org/..."</api:url>[5]</api:reflink>
that ...
```

und das Ergebnis kann via XSLT in eine HTML-Datei gewandelt werden, in der der Link aktivierbar ist. [BL01a]

Im Falle von PDF-Dokumenten ist dies schwieriger: In [BL01b] findet sich im Ausblick der Hinweis, dass mit Tools von Adobe experimentiert wird, um PDF-Dateien unmittelbar zu bearbeiten.

5 Verwandte Projekte

In diesem Abschnitt geben wir eine kurze Auflistung ähnlicher Projekte.

Distributed Link Service

DLS [CRHH95, CHH98] wird seit 1995 entwickelt. In der ursprünglichen Form musste der Benutzer des Service mit einer speziellen Client-Software auf den DLS-Server zugreifen; die aktuelle DLS-Version ist als Web-Proxy implementiert. Ein angefordertes Dokument wird vom Proxy analysiert und um Links ergänzt, die in der internen Link-Datenbank gefunden wurden.

DLS wertet dabei nicht nur Literaturverweise aus; auch Informationen zu Personen, Begriffen (Lexikon-Einträge) sind verfügbar; ferner kennt DLS sogennante „Konzept-Links“, die aus einer Analyse des Gesamtdokuments erstellt werden – DLS versucht hier also, das Thema des Dokuments zu erkennen.

Durch Bereitstellung eines zusätzlichen Kontrollfensters (ebenfalls über den Proxy realisiert) erlaubt DLS die Anpassung des Proxy; unter anderem lässt sich einstellen, ob Links im Text oder in Form eines bibliograhischen Anhangs dargestellt werden sollen.

ResearchIndex

Das ehemals Citeseer genannte Projekt [LGB99] basiert auf „Autonomous Citation Indexing“: Das System durchsucht ähnlich wie eine Suchmaschine das Internet und analysiert alle gefundenen Dokumente; zusätzlich werden Newsgroups und Mailinglisten durchsucht sowie Datenbanken von Verlagen ausgewertet. Durch Analyse der Literaturverzeichnisse in den gefundenen Dokumenten wird eine Datenbank aufgebaut.

Die Dokumente selbst werden nicht verlinkt, aber Benutzer können sich in der Datenbank von Dokument zu (zitiertem) Dokument bewegen; auch eine Rückwärtssuche ist möglich, also die Navigation zu Dokumenten, die das aktuelle Dokument zitieren.

SFX

SFX ist die Kurzform für den Arbeitstitel Special Effects [VdSH99a, VdSH99b, VdSH99c]. Das Projekt der Universität Gent versucht in erster Linie, verschiedene (kommerzielle) Bibliothekssysteme zu verknüpfen, wie sie in einer großen Universitätsbibliothek eingesetzt werden.

Hier kommt dynamisches Linking zum Einsatz, und die Links in einem von SFX verlinkten Dokument müssen nicht notwendig direkt auf ein Zieldokument verweisen, sondern können etwa eine Suchmaschinenanfrage starten oder einen lokalen Index abfragen. Ein Beispiel ist ein geeigneter Aufruf der Büchersuche des Online-Buchhändlers Amazon.

SFX wird auch vermarktet: URL: *http://www.sfxit.com/*.

S-Link-S

Das Scholarly Link Specification Framework nimmt sich des Problems an, dass große Verlage zwar bereits mit Linking-Tools arbeiten, diese jedoch nur die eigene Datenbank bearbeiten und nicht mit den Formaten anderer Verlage kompatibel

sind. Hier bietet S-Link-S eine XML-basierte Syntaxdefinition für den Austausch von Linking-Informationen.

URLs werden aus Standard-Bibliographie-Informationen erzeugt.

Zu S-Link-S ist kein wissenschaftlicher Beitrag verfügbar, Informationen finden sich auf der Projekt-Homepage: *http://www.openly.com/SLinkS/*.

CrossRef

Auch CrossRef [ALR+00] zielt auf die Verknüpfung der Daten mehrerer Verlage; zur Zeit wird mit 96 Verlagen zusammengearbeitet, die gemeinsam 5.604 wissenschaftliche Zeitschriften veröffentlichen. Für das Verlinken werden dabei DOIs 3.2.2 eingesetzt. Die DOIs werden als Metadaten von den Verlagen in die Texte integriert.

Der Austausch der Daten läuft über CrossRefs Metadata Database (MDDB): Alle teilnehmenden Verlage erzeugen Metadaten ihrer eigenen Texte gemäß einer von CrossRef definierten XML DTD und laden diese Informationen auf den MDDB-Server; für die Integration fremder Links lassen sich dann Datenbankanfragen an den Server stellen.

6 Zusammenfassung

Reference Linking vereinfacht den wissenschaftlichen Alltag, indem die sonst zeitaufwendige Suche nach zitierter Literatur aus einem gerade bearbeiteten Artikel wegfällt: Im Falle online verfügbarer Dokumente reicht ein Mausklick, um zum zitierten Text zu gelangen. Die Verknüpfungen zwischen den Texten müssen dabei nicht manuell angelegt werden, sondern ein automatisierter Prozess kann in kurzer Zeit komplette Online-Archive verarbeiten.

Automatisches Reference Linking funktioniert: Einige der hier zitierten Arbeiten bieten Statistiken über die Erfolgshäufigkeit: [BL01a] unterscheidet hier zwischen „Item accuracy“ und „Reference accuracy“. Erstere misst das korrekte Parsen der Dokument-Metadaten (Autor, Titel etc.) und Zitat-Kontexte, letztere die Korrektheit erkannter Referenzen. Die dabei erreichten Quoten wurden von den Autoren positiv bewertet: Bei einem Test mit 66 Artikeln hatten 68% perfekte (also 100-prozentige) Item accuracy und 57% perfekte Reference accuracy. (Ein Durchschnittswert wurde nicht angegeben.) Diese Zahlen enthalten noch keine Informationen über das Auffinden der zu den erkannten Referenzen gehörenden Items.

Wenn die noch in der Entwicklung befindliche Software ausgereift ist, wird dies den Zugriff auf wissenschaftliche Artikel erleichtern.

Literatur

[ALR+00] Helen Atkins, Catherine Lyons, Howard Ratner, Carol Risher, Chris Shillum, David Sidman, and Andrew Stevens. Reference Linking with DOIs. *D-Lib Magazine*, 6(2), Februar 2000.
http://www.dlib.org/dlib/february00/02risher.html. 5

[BAL00] Donna Bergmark, William Arms, and Carl Lagoze. An Architecture for Reference Linking. Technical report, Cornell University, 2000.
http://www.cs.cornell.edu/bergmark/ReferenceLinkingArchitecture.ps. 4.4.3

[Ber00] Donna Bergmark. Automatic Extraction of Reference Linking Information from Online Documents. Technical report, Cornell University, 2000.
http://www.cs.cornell.edu/cdlrg/Reference%20Linking/extraction.pdf. 4.3.1, 4.3.2, 4.3.4

[BHL99] Tim Bray, Dave Hollander, and Andrew Layman. *Namespaces in XML*, 1999.
http://www.w3.org/TR/1999/REC-xml-names-19990114/. 7

[BL01a] Donna Bergmark and Carl Lagoze. An Architecture for Automatic Reference Linking. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, 2001. LNCS 2163 [CS01], S. 115–126; auch als Cornell University Technical Report, [BL01b]. 1.1, 5, 4.3.4, 4.4.3, 4.5, 6

[BL01b] Donna Bergmark and Carl Lagoze. An Architecture for Automatic Reference Linking. Technical report tr2001-1842, Cornell University, 2001.
http://www.cs.cornell.edu/cdlrg/Reference%20Linking/tr1842.ps. 4.5, 6

[BPSMM00] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, and Eve Maler. Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, Oktober 2000.
http://www.w3.org/TR/2000/REC-xml-20001006. 3.3

[CHH98] Les A. Carr, Wendy Hall, and Steve M. Hitchcock. Link services or agent services? In *Ninth Conference on Hypertext*, Pittsburgh, 1998.
http://www.bib.ecs.soton.ac.uk/data/1438/pdf/carr1998a.pdf. 5

[CRHH95] Leslie Carr, David De Roure, Wendy Hall, and Gary Hill. The Distributed Link Service: A Tool for Publishers, Authors and Readers. In

Fourth International World Wide Web Conference, Boston, Massachusetts, USA, December 11–14, 1995.
http://www.w3.org/Conferences/WWW4/Papers/178/. 8, 5

[Cro82] David H. Crocker. *RFC 822: Standard for the Format of ARPA Internet Text Messages*, 1982.
http://www.faqs.org/rfcs/rfc822.html. 6

[CS01] P. Constantopoulos and I.T. Solvberg, editors. *Research and Advanced Technology for Digital Libraries – 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4–9, 2001. Proceedings*, volume 2163 of *Lecture Notes in Computer Science*. Springer, 2001. 6

[GHL+99] Paul Ginsparg, Joe Halpern, Carl Lagoze, Stevan Harnad, Wendy Hall, and Les Carr. Integrating and Navigating eprint Archives through Citation-Linking: The Open Citation (OpCit) Linking Project, 1999.
http://www.cogsci.soton.ac.uk/~harnad/citation.html. 4

[GST64] Eugene Garfield, Irving H. Sher, and Richard J. Torpie. *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information Inc., Philadelphia, USA, 1964.
http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf.

[HCJ+00] Steve Hitchcock, Les Carr, Zhuoan Jiao, Donna Bergmark, Wendy Hall, Carl Lagoze, and Stevan Harnad. Developing services for open eprint archives: globalisation, integration and the impact of links. In *Proceedings of the 5th ACM Conference on Digital Libraries*, 2000.
http://opcit.eprints.org/dl00/htdl00.pdf. 1.1

[Hil01] Diane Hillmann. *Using Dublin Core*, 2001. DCMI Recommendation,
http://dublincore.org/documents/2001/04/12/usageguide/. 3.1.1

[Hit00] Steve Hitchcock. The Open Citation Project, First Year Report to JISC. Technical report, 2000.
http://opcit.eprints.org/y1report/y1report-final.pdf. 4.1, 3, 4

[Hol00] Steven Holtzner. *Inside XML*. New Riders, 2000. 7

[Int01] International DOI Foundation. *The DOI Handbook*, 2001.
http://www.doi.org/handbook_2000/. 3.2.2, 1

[Jou00] Journal of Computer and System Sciences. *Information for Authors*, 2000.
http://www.academicpress.com/www/journal/ss/ssifa.htm. 2.2.1

[LGB99] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

http://www.neci.nj.nec.com/homepages/lawrence/papers/aci-computer98/aci-computer99.html; *.../aci-computer99.pdf*. 1.1, 5

[LLC] R.R. Bowker LLC. Major Issues in the Implementation of the ISBN. *http://www.isbn.org/standards/home/isbn/us/major.asp*. 2.4.1

[Pos82] Jonathan B. Postel. *RFC 821: Simple Mail Transfer Protocol*, 1982. *http://www.faqs.org/rfcs/rfc821.html*. 6

[RWT97] RWTH Aachen. Diplomprüfungsordnung für den Studiengang Informatik der Rheinisch–Westfälischen Technischen Hochschule Aachen, März 1997. *http://www-i1.informatik.rwth-aachen.de/stube/dpo-1997.html*. 2.1

[Sch85] Peter Schlobinski. Durchs wilde Germanistan. *wecker. Zeitschrift am FB Germanistik, Universität Hannover*, 10, 1985. *http://www.fbls.uni-hannover.de/sdls/schlobi/splitter/germanistan.htm*. 2.1

[SL01] Sam X. Sun and Laurence Lannom. *Handle System Overview*. Corporation for National Research Initiatives, 2001. *http://www.handle.net/overview-current.html*, *http://www.handle.net/overview-current.pdf*. 3.2.1

[VdSH99a] Herbert Van de Sompel and Patrick Hochstenbach. Reference Linking in a Hybrid Library Environment, Part 1: Frameworks for Linking. *D-Lib Magazine*, 5(4), April 1999. *http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html*. 5

[VdSH99b] Herbert Van de Sompel and Patrick Hochstenbach. Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution. *D-Lib Magazine*, 5(4), April 1999. *http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html*. 5

[VdSH99c] Herbert Van de Sompel and Patrick Hochstenbach. Reference Linking in a Hybrid Library Environment, Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment. *D-Lib Magazine*, 5(10), Oktober 1999. *http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html*. 5

[VL01] Herbert Van de Sompel and Carl Lagoze. *The Open Archives Initiative Protocol for Metadata Harvesting*, 2001. *http://www.openarchives.org/OAI/openarchivesprotocol.htm*.

[W3C99a] W3C Consortium. HTML 4.01 Specification, Dezember 1999. W3C Recommendation,
http://www.w3.org/TR/REC-html40/. 6

[W3C99b] W3C Consortium. Resource Description Framework (RDF) Model and Syntax Specification, Februar 1999. W3C Recommendation,
http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/. 6

[W3C00] W3C Consortium. XHTML 1.0: The Extensible HyperText Markup Language, Januar 2000. W3C Recommendation,
http://www.w3.org/TR/xhtml1/. 3.1

[Wol00] Birgitta Wolff. Hinweise für die Anfertigung wissenschaftlicher (Haus-) Arbeiten, Lehrstuhl Internationales Management, Univ. Magdeburg, 2000.
http://www.ww.uni-magdeburg.de/bwl2/pruefungen/FormhinweiseIM.pdf. 2.2.2