

**Verwendung von
Qualitäts-Metadaten zur
verbesserten Wissensauffindung
und Testimplementierung im
xFIND System**

Diplomarbeit
an der
Technischen Universität Graz

vorgelegt von

Johann Weitzer

Institut für Informationsverarbeitung
und Computergestützte neue Medien
Technische Universität Graz
A-8010 Graz

© Copyright 2000, Johann Weitzer
Begutachter: o.Univ.-Prof. Dr. Dr. h.c. Hermann Maurer
Betreuer: Dipl.-Ing. Christian Gütl

31. März 2000

Ich versichere hiermit, diese Arbeit selbständig verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt und mich auch sonst keiner unerlaubten Hilfsmittel bedient zu haben.

Kurzfassung

Das exponentielle Wachstum der Inhalte im Internet – insbesondere im WWW – und dessen Unstrukturiertheit führen dazu, daß die Zuverlässigkeit der vorhandenen Informationen zunehmend sinkt. Daher gewinnen solche Suchdienste immer mehr an Bedeutung, welche in der Lage sind, dem Suchergebnis auch Informationen über die Qualität der gefundenen Dokumente hinzuzufügen. Die Kombination von Qualitätsmetadaten und beschreibenden Metadaten in Verbindung mit Suchmaschinen stellt einen vielversprechenden Ansatz in Richtung verbesserter Wissensauffindung dar. Auf diese Weise erhalten Benutzer nicht erst im Suchergebnis Informationen über die Qualität der gefundenen Ressourcen, sondern können schon bei der Festlegung der Suchkriterien ihre individuellen Bedürfnisse nach inhaltlicher Qualität angeben.

Es ist daher ein naheliegendes Ziel, Dokumente im Netz nicht nur bezüglich ihres Inhaltes zu beschreiben, sondern auch in Bezug auf ihre Qualität. In der vorliegenden Arbeit werden Systeme untersucht, welche die Ressourcen im Netz zu unterschiedlichen Zwecken mit Metadaten beschreiben. Solche Metadaten dienen dazu, auf brauchbare Inhalte hinzuweisen, bzw. von ungenügenden Informationsquellen abzuhalten, vor allem aber dazu, die Suche nach relevanten und guten Dokumenten zu ermöglichen.

Das Ziel der vorliegenden Arbeit ist es, ein System zu schaffen, welches die Bewertung von Online-Ressourcen hinsichtlich der Qualität ermöglicht und die verknüpfte Suche von Qualitätsmetadaten und Dokumentinhalten erlaubt. Zu diesem Zweck wurde ein System von Attributen zur Beschreibung der Qualität, das xFIND Quality Metadata Scheme xQMS, definiert, und eine Testimplementierung entwickelt. Autoren und Webserver-Betreiber können ihre Ressourcen an diesem System voranmelden, und eine Bewertung erstellen. Der Administrator kann diese anpassen und im System aufnehmen. Fachexperten können dann die Ressourcen bzw. Bewertung überprüfen und ihrerseits bewerten. Eine entsprechende Schnittstelle zum xFIND Suchsystem ermöglicht es, nach xQMS Qualitäts Metadaten zu suchen.

Die vorliegende Arbeit zeigt somit einen Weg auf, wie die Wissensauffindung angesichts der Informationsmenge im Netz sinnvoll verbessert werden kann.

Abstract

The exponential growth of content on the internet – especially on the WWW – and its non-existent structure lead to a decline in reliability of the existing information. Search services which are able to add some extra information about the quality of the found documents become ever more important. The combination of quality-metadata and metadata describing resources in conjunction with a search-engine are a good approach for a better knowledge discovery. This way, users not only get information about the quality of the search-result, but they can express their needs for quality when they enter their search criteria.

Thus, the wish to not only describe the content of resources on the net, but also to define their quality seems obvious. In the present work some systems will be analysed which use metadata to describe resources for different purposes, e.g. to recommend good quality, or to steer away from useless information, and generally to assist in the search for relevant information of high quality.

It is the goal of this work to develop a system which enables the rating of online resources with regard to their quality, and which allows the combined search for quality metadata and document content. For this purpose a system of attributes describing the quality, the xFIND Quality Metadata Scheme xQMS, was defined, and a prototype developed. Authors and information service providers will thus be enabled to register their resources in the system and to prepare a rating. The administrator will be able to change the rating and include it into the system. Experts will then inspect the resource and the rating, and give their own ratings. An interface to the xFIND search system makes it possible to search for the quality metadata .

This work demonstrates how knowledge discovery can be improved in the view of the huge amount of information on the net.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Aufbau der Arbeit	3
1.2	Zusammenfassung	4
I	Untersuchungsbereich	5
2	Qualität im Internet	6
2.1	Qualität, Datenqualität	7
2.2	Kriterien inhaltlicher Qualität von Online Ressourcen	9
2.3	Bewertungssysteme	13
2.3.1	Upstream Filtering	13
2.3.2	Downstream Filtering	13
2.4	Zusammenfassung	15
3	Metadaten	17
3.1	Definition von Metadaten	18
3.2	Technische Umsetzung	19
3.2.1	HTML META Tag	20
3.2.2	PICS	21
3.2.3	RDF	24
3.3	Metadaten Schemata	26
3.3.1	MARC	26

<i>INHALTSVERZEICHNIS</i>	ii
3.3.2 SOIF/RDM	27
3.3.3 Dublin Core	27
3.3.4 LOM, IMS	30
3.4 Zusammenfassung	30
4 Blockieren durch Einsatz von Metadaten	32
4.1 Blocking Systeme	34
4.1.1 CYBERsitter von Solid Oak	34
4.1.2 Cyber Patrol von Microsystems Inc.	35
4.1.3 RSACi	35
4.1.4 INCORE	36
4.2 Zusammenfassung	36
5 Empfehlungs- und Annotationssysteme	39
5.1 Empfehlungssysteme	40
5.1.1 Content-based recommendation	40
5.1.2 Collaborative recommendation	41
5.1.3 Soziale Folgen	42
5.2 Annotationssysteme	43
5.2.1 Architektur eines Annotationssystems unter Verwendung von Metadaten	44
5.2.2 Das Interface des Annotationssystems von Hyperwave	45
5.3 Zusammenfassung	46
6 Suchen durch Einsatz von Metadaten	49
6.1 Index Suchmaschinen	50
6.2 Metasucher	51
6.3 Katalogdienste	51
6.3.1 Subject Catalogues	52

6.3.2	Annotated Directories	52
6.3.3	Annotated Directories with Ratings or Reviews	53
6.3.4	Subject Directories with Ratings	53
6.3.5	Subject Guides	53
6.3.6	Information Gateways	53
6.4	xFIND	54
6.5	Auswahl eines Systems zur verbesserten Wissensauffindung	56
6.5.1	Auswahl geeigneter Attribute zur Beschreibung und Bewertung von Ressourcen im Netz	59
6.5.2	Wer soll Bewertungen erstellen?	64
6.6	Zusammenfassung	66

II Gestaltungsbereich **68**

7 xFIND Quality Metadata Scheme
xQMS v1.0 **69**

7.1	Beschreibung der Attribute	72
7.1.1	Creator	72
7.1.2	Identifier	72
7.1.3	Scope of description	73
7.1.4	Signatur	73
7.1.5	Type	74
7.1.6	Title	75
7.1.7	Version	76
7.1.8	Subject	77
7.1.9	Description	77
7.1.10	Date	78
7.1.11	Quotation hints	79
7.1.12	Language	79

7.1.13	Audience	80
7.1.14	Authority	80
7.1.15	Accuracy	81
7.1.16	Rating	83
7.2	Verwendung der xQMS Attribute	85
7.3	Zusammenfassung	85
8	Testimplementierung QMRatingSystem	89
8.1	Das QMRatingSystem aus der Sicht der unterschiedlichen Benutzergruppen	90
8.1.1	Anmeldung	90
8.1.2	Der Benutzer-Modus	91
8.1.3	Der Administrator-Modus	92
8.1.4	Der Experten-Modus	96
8.2	Aufbau des Programms	97
8.2.1	Verzeichnis der Benutzerdaten	99
8.2.2	Verzeichnis für Voranmeldungen	100
8.2.3	Verzeichnis der Bewertungen	102
8.2.4	Schnittstelle zum xFIND Suchsystem	102
8.2.5	Struktur der Java Applikation	104
8.2.6	QMStart	104
8.2.7	QMServer	104
8.2.8	QMApplication	106
8.2.9	QMWebForm	107
8.2.10	QMWebOutput	107
8.2.11	QMFileHandling	108
8.2.12	Userinteraktion	108
8.3	Zusammenfassung	108
9	Ausblick	110

10 Zusammenfassung	112
III Anhang	114
A Qualitätskriterien	115
A.1 Kriterienkatalog nach [Mitretek97]	115
B Metadaten	117
B.1 PICS Optionen	117
B.2 Liste von Typen in Dublin Core	118
B.3 Kompatibilität von Dublin Core zu LOM	120
C Blockingsoftware	121
C.1 CYBERSitter	121
C.2 Die Listen von CyberPatrol	122
D Verbesserte Wissensauffindung im Internet	124
D.1 Themenklassifikation des Dewey Decimal Code	124
Glossar	125
Abbildungsverzeichnis	130
Tabellenverzeichnis	133
Listingsverzeichnis	135
Literaturverzeichnis	136

Kapitel 1

Einleitung

Das Internet hat seine Wurzeln als Netzwerk in der Wissenschaft und Technik, und noch immer stellen technische und wissenschaftliche Dokumente von Universitäten, Forschungs- und Lehreinrichtungen einen großen Teil der im World Wide Web¹ verfügbaren Dokumente. Dass jederman sehr leicht jegliche Art von Information publizieren kann, bedeutet jedoch ein Sinken der Zuverlässigkeit der vorhandenen Informationen, was die Nützlichkeit des Internet für Fachleute im wissenschaftlichen Bereich erheblich beeinträchtigt. [Palme98]

In [Ciolek96] wird die These aufgestellt, daß sich das Web wie eine sich selbst-organisierende und selbst-verbessernde Einheit verhält. Dies deshalb, weil die Online-Autoren und Herausgeber ständig voneinander lernen, und sich die gesamte Qualität ihrer vernetzten Aktivitäten langsam aber sicher bessert. Wird das Netz jedoch überschwemmt mit Dokumenten schlechter Qualität, könnte das diesen Prozess behindern oder ihm sogar entgegenwirken. Die Entwicklung des Mediums Buch hat ungefähr 400 Jahre gedauert, um die heutige Qualität von Inhalt und Aussehen zu erreichen. Deshalb sei es erlaubt anzunehmen, dass in einem Bruchteil dieser Zeit - vielleicht innerhalb von 10-15 Jahren - die strukturellen und organisatorischen Probleme des Internet verschwunden sein werden.

Dagegen spricht aber das steile, exponentielle Wachstum des Internet, welches die Gesetze von Moore und Rutkowski zu beweisen scheint: Das Gesetz von Moore besagt, dass sich elektronische Technologien durchschnittlich alle zwei Jahre wesentlich ändern. Rutkowski² ist der Ansicht, dass sich diese Zeitspanne im hoch-dynamischen Umfeld des Internet in Monaten messen lässt. [Ciolek96]

¹Das WWW ist der am häufigsten benutzte Dienst im Internet, und hat wohl die rasche Verbreitung des Netzes der Netze erst ermöglicht.

²Rutkowski, Anthony-Michael: Today's Cooperative Competitive Standards Environment for Open Information and Telecommunication Networks and the Internet Standards-Making Model. <http://info.isoc.org/papers/standards/amr-on-standards.html>

Das enorme Wachstum des Internet lässt sich anhand der Anzahl von im Netz enthaltenen Hosts zeigen, welche zweimal pro Jahr durch das Internet Software Consortium³ ermittelt wird. Die Anzahl betrug im Jänner 2000 rund 72,4 Millionen. Sieben Jahre zuvor lag die Zahl noch bei einem Bruchteil davon, nämlich bei rund 1 Million (vgl. Abbildung 1.1). [ISC2000]

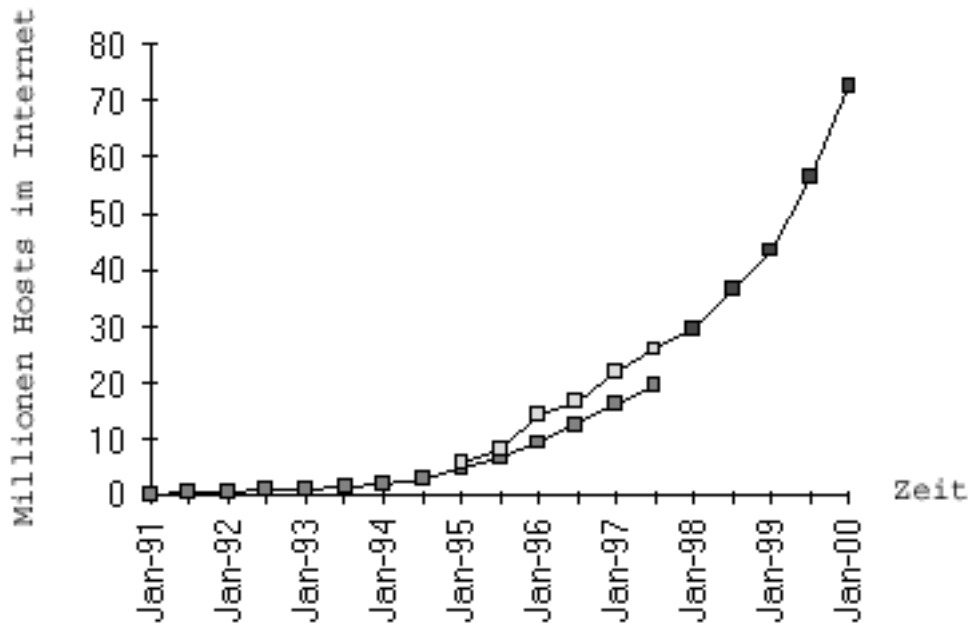


Abbildung 1.1: Die Anzahl der Hosts im Internet von 1991 bis 2000. Bis Jänner 1998 hat man diese Anzahl mit einer anderen Methode berechnet (unterer Zweig der Kurve), weshalb man bis zu diesem Datum interpolieren musste (oberer Zweig). [ISC2000]

In [AC96] wird Clifford Lynch⁴ zitiert, der behauptet, dass elektronisch verfügbare Informations-Ressourcen zu erfolgreich und bequem seien: Die Benutzer werden geneigt sein zu glauben, dass diese die Gesamtheit der verfügbaren Information bilden. Was man im Internet nicht finden kann, mag ganz einfach gar nicht existieren.

³<http://www.isc.org/ds/>

⁴Lynch, Clifford: Rethinking the Integrity of the Scholarly Record in the Networked Information Age. In: Educom Review, März/April 1994. http://www.educom.edu/educom.review/↔review.94/mar.apr/lynch_article bzw. <http://www.educom.edu/web/pubs/review/reviewArticles/29238.html>

Auch nach der Einschätzung des Philosophen Walther Zimmerli führt die Verbreitung des Internet zu einem gravierenden Wissensverlust der Menschen. Da immer mehr Daten im Internet zu finden seien, brauche der Einzelne immer weniger Informationen abrufbar im Kopf zu haben. Das Problem sei weiters, dass man benötigtes Wissen in den Datenmengen nicht mehr finde. Selbst Wissenschaftler, Wirtschaftsexperten, Politiker und Journalisten hätten bereits den Überblick über die ständig anschwellenden Datenmengen in ihren Fachgebieten verloren. Erst Suchmaschinen und Internet-Dienste für einzelne Branchen und Wissenschaften machen das Internet für viele Anwender nutzbar. [Heise99]

Die EU unterstützt das Projekt SELECT, in dessen Rahmen Suchmöglichkeiten entwickelt werden, die es den Menschen erlauben sollen, wertvollere und interessantere Dokumente zu finden, und dabei Datenmüll und Redundanzen zu vermeiden. Im Projektbericht wird der wirtschaftliche Nutzen angesprochen: Die Bevölkerung der EU soll schätzungsweise schon im Jahr 1999 mehr als tausend Millionen Stunden pro Jahr mit dem Lesen von meist nicht relevanten Dokumenten im Internet beschäftigt sein. Wenn die Auswahl relevanter Dokumente um nur 10 Prozent effizienter gestaltet würde, könnte man tausende Millionen Euro sparen. [Palme98]

1.1 Aufbau der Arbeit

In den Kapiteln 2 bis 6 des folgenden Untersuchungsbereichs werden die theoretischen Grundlagen zur Thematik der verbesserten Wissensauffindung im Internet untersucht, und im darauffolgenden Gestaltungsbereich wird ein praktisches System entwickelt, welches auf diesen Grundlagen basiert.

Im nachfolgenden Kapitel 2 wird der Begriff „Qualität“ im Zusammenhang mit Online Ressourcen näher beleuchtet. Es wird darauf eingegangen, welche Attribute einer Ressource für Konsumenten wichtig sind, und mit welchen Kriterien sich die Qualität von Dokumenten im Internet beschreiben lässt.

In Kapitel 3 wird erklärt, wie solche Kriterien in Form von Metadaten computer-verarbeitbar gespeichert werden können. Es werden allgemein Metadaten beschrieben, sowie deren technische Implementierung im Internet diskutiert. Weiters werden Metadaten Schemata behandelt, die heute in Verwendung sind, um Ressourcen im Internet zu charakterisieren.

Kapitel 4 erläutert, wie Metadaten dazu eingesetzt werden können, im Speziellen Kinder von schädlichen Inhalten im Internet fernzuhalten, ohne dabei Methoden der Zensur verwenden zu müssen. Es werden auch Systeme vorgestellt, die sich bereits im Internet im Einsatz befinden.

In Kapitel 5 wird anschließend erklärt, wie sich Empfehlungen mit Hilfe von Metadaten umsetzen lassen. Dabei muß zwischen Empfehlungs- und Annotations-systemen unterschieden werden, wobei konkrete Beispiele derartiger Systeme gezeigt werden.

Kapitel 6 behandelt den Einsatz von Metadaten in Verbindung mit Suchmaschinen. Zuerst werden verschiedene Arten von Suchsystemen vorgestellt, wie zum Beispiel Indexsuchmaschinen, Metasucher und Katalogdienste. Im Anschluss wird diskutiert, mit Hilfe welcher Attribute und Systeme sich die Suche nach relevanten Inhalten verbessern lässt.

Der folgende Untersuchungsbereich baut auf den vorangestellten Untersuchungen auf. In Kapitel 7 wird ein Metadaten Schema gezeigt, mit dem sich Inhalte sowohl beschreiben als auch bewerten lassen. Dieses spezielle Metadaten Schema, xQMS genannt, soll in Verbindung mit einem verteilten Suchsystem, xFIND, die Suche nach relevanten Inhalten verbessern. Zu diesem Zweck wurde eine Testimplementierung entwickelt, QMRatingSystem genannt, mit deren Hilfe Ressourcen von den Autoren, Administratoren von Suchsystemen und Fachexperten bewertet werden können. Das System verwaltet die xQMS Qualitätsmetadaten, und stellt sie dem xFIND Suchsystem zur Verfügung. Dieser Prototyp wird in Kapitel 8 beschrieben. Im Ausblick in Kapitel 9 wird gezeigt, wie das System in Zukunft ausgebaut werden könnte. Zuletzt folgt in Kapitel 10 eine Zusammenfassung, sowie der Anhang und die verschiedenen Verzeichnisse.

1.2 Zusammenfassung

Es wird deutlich, wie dringend notwendig die verbesserte Suche nach qualitativ hochwertigem Inhalt ist. In der herkömmlichen, wissenschaftlichen Kommunikation gibt es eine Reihe von Instrumenten zur Wahrung der Qualität von Informationen. Beispiele dafür sind wissenschaftliche-kontrollierte Journale und Konferenzen, sowie im akademischen Bereich die Dissertationen [Palme98]. Im Internet müssen neue Wege gesucht werden, qualitativ hochwertigen Inhalt von anderen Informationen zu trennen. Eine teilweise Überführung des Systems der Bewertung durch Experten gelingt durch die Verwendung von Metadaten, welche – wie in den folgenden Kapiteln gezeigt wird – sich zum Blockieren von schädlichen Inhalten, für Empfehlungen und zum Suchen relevanter Ressourcen eignen.

Im folgenden Kapitel wird untersucht, wodurch sich inhaltliche Qualität von Ressourcen im Netzwerk auszeichnet, und mit welchen Attributen sie sich beschreiben lässt. Es wird weiters untersucht, von wem Bewertungen auf welche Weise durchgeführt werden können, und wie es den einzelnen Benutzern möglich wird, die für sie interessanten Informationen aus dem Netz zu filtern.

Teil I

Untersuchungsbereich

Kapitel 2

Qualität im Internet

Bevor man über Systeme nachdenken kann, welche dazu dienen sollen, beispielsweise bessere Suchergebnisse im Internet zu liefern, muss man sich darüber im Klaren sein, was „Qualität“ in diesem Zusammenhang bedeutet. Nachdem Informationen im Internet keiner Qualitätskontrolle unterzogen werden, und tatsächlich jeder publizieren kann, könnten Online Informationen von minderer Qualität, die unkritisch aufgenommen werden, mitunter schlimme Folgen haben. In [ED99] wird dabei auf den medizinischen Sektor eingegangen, bei dem es besonders gefährlich sein kann, auf unkontrollierte Informationen im Internet zu vertrauen.

Es gibt viele verschiedene Möglichkeiten, die Qualität von Ressourcen zu beschreiben. Es wird jedoch nicht ausreichen, Webpages ein bis fünf Sterne zu verleihen, oder sie als „cool“ bzw. „langweilig“ zu bezeichnen. Man muß die Qualität hinsichtlich mehrerer Attribute beschreiben, wobei es schwierig ist, diese auszuwählen. Die Kriterien sollten die inhaltliche Qualität beschreiben und auf so viele verschiedene Ressourcen wie möglich anwendbar sein.

Verschiedene Organisationen haben Vorschläge für Kriterien-Kataloge erstellt, und z.T. durch Umfragen oder Experten bestätigen lassen. In den folgenden Abschnitten werden einige dieser Vorschläge diskutiert. Sie werden auch als Grundlage für das xFIND Quality Metadata Set, xQMS (vgl. Kapitel 7) dienen. Es sei in diesem Zusammenhang auch auf Kapitel 3.3 hingewiesen.

Im nachfolgenden Abschnitt 2.1 werden die Kriterien zur Beschreibung und Bewertung von Ressourcen diskutiert. Weiters muss noch überprüft werden, wer die Arbeit der Rezension und Kritik übernehmen soll. Es ist auch zu entscheiden, wo diese Metadaten gespeichert und wie sie abrufbar gemacht werden sollen. Diese Thematik wird in Abschnitt 2.3 diskutiert.

2.1 Qualität, Datenqualität

Bevor der Begriff der inhaltlichen Qualität von Ressourcen im Netz beschrieben wird, sollte die Bedeutungen des allgemeineren Begriffs "Qualität" erläutert werden: Laut Brockhaus bedeutet das aus dem Lateinischen stammende Wort allgemein Beschaffenheit oder Eigenschaft. In der Philosophie werden damit sinnliche Seiten der Wahrnehmung (Farbe, Geruch, Härte etc.) bezeichnet. In der Wirtschaft bedeutet es die Beschaffenheit einer Ware nach ihren Unterscheidungsmerkmalen gegenüber anderen Waren. „Der Begriff Qualität wird einmal objektiv auf messbare Eigenschaften, wie die Reinheit chemischer Erzeugnisse [...], angewendet. Er bringt zum anderen die Abstufung des Eignungswertes gleichartiger Güter für die Befriedigung bestimmter Bedürfnisse zum Ausdruck und ist insoweit subjektiv bestimmt.“ [Brockhaus56]

In [WS96] wird „*data quality (DQ)*“ definiert als Daten, die tauglich sind, vom Benutzer verwendet werden zu können¹. Nach dieser Studie lässt sich die Datenqualität in mehreren Dimensionen beschreiben, wobei sich eine Dimension wieder aus mehreren Attributen zusammensetzt. Zur Feststellung, welche Attribute sinnvoll zur Beschreibung von DQ sind, wurden in einem mehrstufigen Prozess zuerst alle möglichen Attribute ausgemacht, bewertet und zu Kategorien zusammengefasst. Die resultierenden vier Kategorien sind die Innere DQ (*intrinsic DQ*), die Kontextuelle DQ (*contextual DQ*), die Darstellungs DQ (*representational DQ*) und die Zugangs DQ (*accessibility DQ*) (vgl. Abbildung 2.1).

Die Innere DQ umfasst nicht nur Genauigkeit und Objektivität, also die vorurteilsfreie, richtige Wiedergabe von Fakten, die für Informationstheoretiker schon lange offensichtliche Attribute hoher DQ darstellen, sondern auch Glaubwürdigkeit und Ansehen, welche den guten Ruf und die Vertrauenswürdigkeit einer Quelle bezeichnen. [WS96]

Die Studie hat ergeben, dass die Qualität der Daten immer in dem gegebenen Kontext einer Aufgabe zu beurteilen ist. Als Attribute im Rahmen der Kontextuellen DQ sind vor allem die Zeitabhängigkeit und Vollständigkeit zu nennen. Weil Aufgaben und deren Kontext sich mit der Zeit und den Daten-Konsumenten ändern, ist es eine große Herausforderung, gute kontextuelle DQ zu erlangen [WS96]. Der Kontext ist aber auch sehr stark vom Leserkreis abhängig, welcher sich beispielsweise durch das Alter, die Religion oder den Kulturkreis definiert. In diesem Zusammenhang sind Bewertungen durch verschiedene Institutionen vorstellbar.

Als Beispiel für Kontextuelle DQ wird in [WS96] eine Problematik genannt, die die US Navy während der Operation *Desert Storm* im Golfkrieg erkannte: Die

¹ „*data that are fit for use by data consumers*“ [WS96]

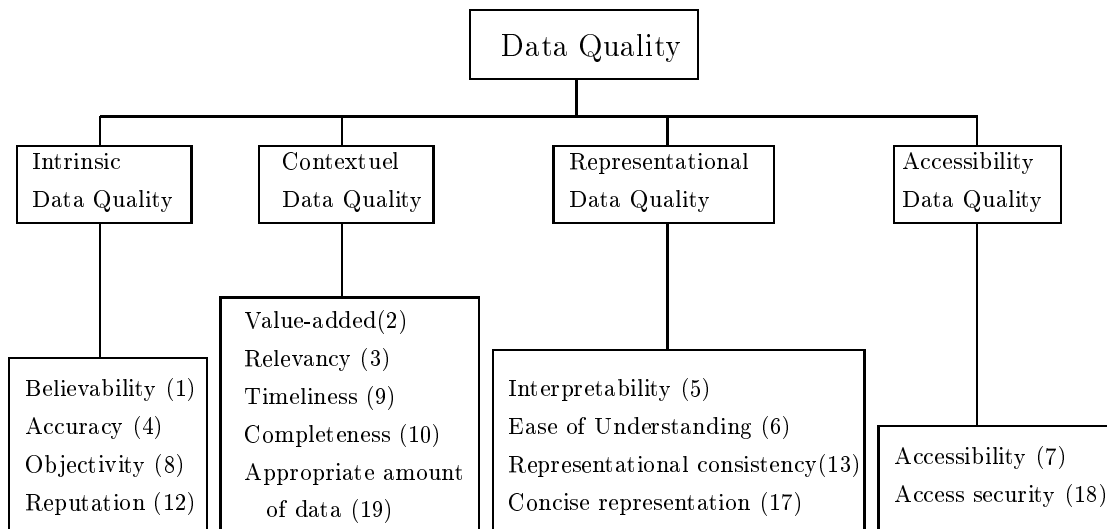


Abbildung 2.1: In [WS96] werden vier Arten der Datenqualität unterschieden. Jede dieser Qualitäten wird durch eine Reihe von Eigenschaften repräsentiert.

Karten und Satellitenfotos waren nicht für alle Stellen (z.B. für Piloten, Angriffsplaner oder Schadensbewerter) gleich brauchbar. Deswegen erstellte man einen Prototypen, bei dem die kontextuellen Dimensionen für jede Aufgabe parametrisiert werden können. Man gibt beispielsweise an, welche Satellitenaufnahmen man in Bezug auf den Ort, die Aktualität, Auflösung, Zieltypen etc. auf seiner Karte verzeichnet haben möchte.

Die Darstellungs DQ inkludiert Aspekte des Formats der Daten (Prägnanz und Konsistenz) und der Bedeutung (Interpretierbarkeit und Verständlichkeit). Bei der Zugangs DQ geht es schließlich darum, wie leicht man an gewünschte Daten herankommt. Auch diesem Aspekt wird von den Konsumenten ein großer Wert beigemessen. [WS96]

Bezogen auf das World Wide Web kann man behaupten, dass die Darstellungs- und Zugangs DQ an Wichtigkeit gegenüber der Inneren- und er Kontextuellen DQ verlieren. Dokumente im Internet liegen nahezu immer in Varianten von HTML oder PDF vor, und werden alle über Browser angezeigt und navigierbar gemacht. Zwar gibt es Unterschiede im Design, die aber meist keine große Bedeutung haben. Hingegen wird es immer wichtiger sein, genau zu wissen, in welchem Kontext Inhalte präsentiert werden.

In [KW95] werden im Besonderen zwei Ansätze benutzt, um „Informationsqualität“ (*Information Quality*) zu definieren: Der produkt- und der wert-basierte Ansatz. Der wert-basierte Ansatz beleuchtet den Vergleich von der Verwendung von Information mit deren Einfachheit der Bedienung, Zeitersparnissen und Ko-

stensenkungen. Der produkt-basierte Ansatz definiert Informationsqualität als Übereinstimmung mit Erfordernissen. Das bedeutet, dass jede Abweichung von den Bedürfnissen, die ein Benutzer erkannt hat, einen Verlust von Informationsqualität impliziert.

Nachdem nun die Begriffe inhaltlichen Qualität, Informations- und Datenqualität weitestgehend erläutert und eingeschränkt wurden, muß im nächsten Abschnitt untersucht werden, mit welchen Eigenschaften sich diese beschreiben lassen, und wie sie gemessen werden können.

2.2 Kriterien inhaltlicher Qualität von Online Ressourcen

Einen Ansatz, die Eigenschaften inhaltlicher Qualität von Ressourcen im Netz zusammenzustellen, stammt vom *Department of Instructional Technology* der Universität von Georgia. Dort hat man versucht, die Qualitätsaspekte von Online Ressourcen herauszufiltern, und hat erst einmal einen Katalog mit allen in Frage kommenden Kriterien aufgestellt. Dazu wurden Bibliotheken, Experten und Rezensenten wissenschaftlicher Journale befragt. In weiteren Schritten wurde die so erlangte, große Anzahl der Kriterien durch ein Auswahlverfahren und Befragungen von 509 auf 11 Kategorien zusammengefasst. In Tabelle 2.1 sind diese aufgezählt, und es ist vermerkt, wieviele der Befragten die jeweiligen Kriterien eher der Qualität der Information oder der Präsentation (innerhalb einer Website) zusprechen. [OWB97]

Kriterium	Site	Information
Site Zugang und Nützlichkeit	18	2
Ressourcen Identifikation und Dokumentation	4	13
Identifikation des Autors	3	9
Ansehen des Autors	-	5
Informations-Struktur und -Design	13	19
Relevanz und Umfang des Inhalts	-	6
Gültigkeit des Inhalts	-	9
Genauigkeit und Gleichgewicht der Information	-	8
Navigation innerhalb des Dokuments	12	8
Qualität der Links	10	12
Ästhetische Aspekte	13	6

Tabelle 2.1: Qualitätskriterien aus der Studie von [OWB97]

Abschließend wurde noch ein Fragenkatalog entwickelt, der auf die Kriterien der inhaltlichen Qualität abzielt. Damit sollte es möglich werden, die Eigenschaften möglichst objektiv zu bestimmen. Beispielsweise, „Gibt es offensichtliche Fehler oder irreführende Auslassungen?“ oder „Wie lautet der Name des Autors?“ [OWB97].

Aus dieser Studie ist erkennbar, dass ästhetische Aspekte, Navigierbarkeit und Zugang – also die Darstellungs- und Zugangs DQ – eher Attribute der Site sind, und nicht der Information. Die Kriterien Ansehen des Autors, Relevanz, Umfang und Gültigkeit des Inhalts, sowie Genauigkeit und Gleichgewicht der Information beschreiben die inhaltliche Qualität. Mit dem Fragenkatalog wird versucht, die jeweiligen Kriterien zu „messen“. Manche der Fragen lassen sich jedoch nicht objektiv beantworten (z.B. „Ist das Design so komplex, dass es vom Inhalt ablenkt?“) oder machen es schwierig, auf eine Eigenschaft des Dokuments rückzuschliessen: Mit Ja/Nein-Fragen wie „Gibt es offensichtliche Fehler oder irreführende Auslassungen?“ kann man sicherlich keinen Wert z.B. für die Genauigkeit der Information bestimmen. In diesem Fall ist es leichter, man ordnet einem Attribut einen diskreten Wertebereich zu, und verknüpft jeden dieser möglichen Werte mit Eigenschaften für das Attribut. Beispielsweise kann dem Wert 0 für das Attribut *Ansehen* die Eigenschaft „unbekannt“ zugewiesen werden, bis zum Wert 9 für „höchstes Ansehen“. Damit wird es leichter, Ressourcen zu bewerten und untereinander vergleichbar zu machen.

Das OMNI Konsortium hat untersucht, wie sich die Beurteilung herkömmlicher Printmedien auf die elektronischen Medien umsetzen lässt. Einige Eigenschaften haben die beiden Medientypen gemeinsam, andere Attribute sind jedoch spezifisch für elektronische Medien. Die Kriterien zur Beschreibung der elektronischen Medien sind in Tabelle 2.2 angeführt.

Das erste Attribut von Tabelle 2.2 lautet Ansehen (*Authority*) und bezeichnet den Ruf und die Autorität des Erstellers einer Ressource. Es wird in der Arbeit diskutiert, wie schwierig es ist, diese Daten herauszufinden. Jeder kann im Internet vorgeben, jemand anderer, mit bestimmten Fähigkeiten, zu sein. [SC94]

In einzelnen Fachgebieten ließen sich genaue Angaben über die Reputation oder die Ausbildung von Autoren treffen. Beispielsweise auf dem Gebiet der Medizin im amerikanischen Raum: Dort existiert eine Datenbank der American Medical Association (AMA), in der die meisten Ärzte mit ihren akademischen Ausbildungen gespeichert sind. Da sich solche genauen Angaben über das Ansehen nur in Ausnahmefällen machen lassen, wird selbst bei dem speziellen Kriterienkatalog für medizinische Ressourcen der Mitretek Systems ein anderer Weg beschritten: Es wird nicht das Ansehen des Autors, sondern die Glaubwürdigkeit der Ressource (nach Tabelle 2.3) bewertet (in Anhang A.1 ist der gesamte Kriterienkatalog angeführt). [Mitretek97]

	Kriterium	Englischer Begriff
1	Ansehen	Authority
2	Herkunft	Genealogy
3	Umfang und Behandlung	Scope and Treatment
3.1	Zweck	Purpose
3.2	Umfang	Coverage
3.3	Aktualität von Überarbeitungen	Currency and methods of revision
3.4	Genauigkeit	Accuracy
3.5	Objektivität	Objectivity
3.6	Publikum	Audience
4	Format	Format
5	Anordnung	Arrangement
6	Technische Überlegungen	Technical Considerations
7	Preis und Verfügbarkeit	Price and Availability
8	User Support	User Support

Tabelle 2.2: Kriterien zur Beschreibung elektronischer Ressourcen, aufgestellt durch das OMNI Konsortium [SC94]

Stichhaltigkeit des Beweises	worauf zu achten ist
++++ (bester Beweis)	zufällige, kontrollierte Stichproben
+++	nicht zufällige, kontrollierte Stichproben
++	gut geplante Stichproben
+	Meinungen respektierter Persönlichkeiten
Kein Beweis	falsche Darstellung, Betrug

Tabelle 2.3: In der Arbeit von [Mitretek97] werden Dokumente nach ihrem „Ansehen“ bewertet. Es wird beispielsweise darauf geachtet, ob im Dokument persönliche Meinungen ausgedrückt werden, oder Aussagen mit Beweisen unterlegt werden (beispielsweise durch Stichproben).

Das zweite Kriterium der Arbeit von [SC94] bezeichnet die Herkunft (*Genealogy*), also den Prozess von der Erstellung einer Ressource über deren Modifikationen bis zum gegenwärtigen, aktuellen Status. Weiters kann mit diesem Kriterium z.B. die Herkunft von Bildern beschrieben werden: sind es fotografische Aufnahmen oder künstliche Animationen, etc. Auf diese Weise lassen sich wieder Rückschlüsse auf die Vertrauenswürdigkeit und Glaubwürdigkeit einer Ressource ziehen.

Es finden sich jedoch keine Angaben darüber, welche Werte man dem Attribut zuweisen kann. Ein objektiver und zugleich leicht zu realisierender Weg ist es, das Datum der Erstellung und jenes der letzten Änderung anzugeben. Daran lässt sich erkennen, wie aktuell oder gepflegt eine Ressource ist, und welches von zwei

Dokumenten zuerst publik gemacht wurde. Der Nachteil dieser Methode ist, dass damit keine Angaben über die Art der Erstellung gemacht werden können, was aber in den meisten Fällen ohnehin schwierig bis unmöglich festzustellen ist.

Der Umfang und die Behandlung (*Scope and Treatment*) sind sehr viel schwieriger als in Printmedien zu beschreiben, weil elektronischen Informationen in den meisten Fällen der Kontext fehlt, weil sie informeller sind, und weil sie oft keine Quellenangaben enthalten. Der Benutzer sollte über den Zweck, den Umfang, die Aktualität, die Genauigkeit, Objektivität und das Zielpublikum informiert werden. [SC94]

Autoren oder Kritiker könnten Angaben über das Zielpublikum hinsichtlich des Alters und des Vorwissens auf dem zu behandelnden Gebiet machen, genauso wie dies im Schulwesen üblich ist. Auch dort erfolgt die Einteilung üblicherweise nach Alter, Wissensstand oder Interessensgruppen. Weiters kann man dem Benutzer Hinweise über die Genauigkeit und den Umfang einer Ressource machen, indem man Angaben über die Breite und Tiefe der gebotenen Information bereitstellt. Die übrigen Attribute der Aufstellung beschreiben das äußere Erscheinungsbild, den Preis und die Verfügbarkeit, sowie den User Support. Diese Eigenschaften beziehen sich nur indirekt auf die inhaltliche Qualität.

In der Arbeit von [Skov98] wird wiederum eine andere Einteilung der Qualitätskriterien getroffen (8 Hauptkriterien): Eignung zum Zweck (*Fitness for Purpose*) beschreibt, wie gut sich eine Ressource eignet, den angekündigten Zweck zu erfüllen. Die Kriterien Inhalt (*Content*), Ansehen und Glaubwürdigkeit (*Authority/Credibility*) entsprechen im Wesentlichen den bereits gemachten Aussagen. Aktualität (*Timeliness/Concurrency*) bezeichnet, ob eine Ressource am neuesten Stand ist und gehalten wird, und Navigierbarkeit (*Navigation*) sowie Einfachheit des Zugangs (*Ease of Access*) charakterisieren den Umgang mit der Ressource. Das Kriterium Design/Style (*Design/Style*) beschreibt das Aussehen, und Leistung (*Performance*) kennzeichnet schließlich die Fähigkeit der schnellen Datenverarbeitung und Interaktion einer Ressource.

Zusammenfassend kann man angeben, dass die verschiedenen Autoren Attribute unterschiedlichen Qualitätskategorien zuordnen. Einmal werden Angaben über den Autor zur inhaltlichen Qualität gezählt, ein anderes Mal zu einer eigenen Kategorie. Wichtig ist es jedoch festzustellen, dass bei jeder der genannten Untersuchungen im Grunde dieselben Attribute vorkommen. Das xFIND Quality Metadata Set, xQMS, welches an späterer Stelle definiert wird (vgl. Kapitel 7), baut auf diesen Untersuchungen auf. Man wird auch in Zukunft keine scharfe Grenze zwischen der inhaltlichen Qualität und anderen Qualitätserfordernissen ziehen können.

2.3 Bewertungssysteme

Bisher wurde untersucht, welche Kriterien die inhaltlicher Qualität von Internet Ressourcen am ehesten charakterisieren können. Man muß sich aber auch darüber im Klaren sein, von welcher Person diese Bewertungen auf welche Art durchführen werden sollen. Wenn die Qualitätskontrolle zum Zeitpunkt der Erstellung entweder nicht möglich oder wünschenswert ist, kann sie dezentralisiert werden. Sie besteht dann aus der Selektion von Produkten, die den Qualitätserfordernissen des Konsumenten entsprechen. [ED98]

Man trennt bei dieser Methode das Bewerten einer Ressource – das sog. Rating – vom Auswählen – dem Filterprozess. Dieses Auswählen kann ein *Upstream*- oder *Downstream*-Filtern sein, was in den folgenden Abschnitten erläutert wird. Beide Methoden sind Formen der verteilten Qualitätskontrolle (*distributed quality management*). [ED98] [Legenstein99]

2.3.1 Upstream Filtering

Beim *Upstream Filtering* erfolgt die Selektion durch einen Dritten (also weder durch den Autor einer Ressource, noch durch den Benutzer), so wie dies z.B. bei den heute üblichen Linksammlungen von Kritikern oder manchen Katalogdiensten (vgl. 6.3) üblich ist (vgl. Abbildung 2.2). Hierbei setzen ein paar Kritiker die Qualitätskriterien fest, führen gleichzeitig die Bewertungen durch. Das Verfahren hat jedoch einige wesentliche Nachteile. [ED98]

Das Internet wächst und verändert sich zu rasch, um von ein paar Bewertungs-Agenturen beobachtet werden zu können. Eine mögliche Lösung wäre, dass sich mehr und mehr Agenturen auf jeweils einem sehr speziellen Gebiet etablieren, und so den Überblick behalten könnten. Ein weiterer Nachteil ist die fragwürdige Gültigkeit und Zuverlässigkeit der Bewertungs-Instrumente: Es ist nicht anzunehmen, dass alle Bewertungs-Agenturen dieselben Maßstäbe an Qualität anwenden werden. Ein großer Nachteil des *Upstream Filtering* ist, dass die Bewertungen den Kontext und die Bedürfnisse des Benutzers nicht berücksichtigen können. Beispielsweise könnte ein von einem Arzt als hervorragend bewertertes Dokument eines Spezialisten für einen Patienten wertlos sein. In Abschnitt 6.3 werden Katalogdienste beschrieben, welche im Wesentlichen nach dem Prinzip des Upstream Filtering funktionieren.

2.3.2 Downstream Filtering

Ein Ansatz, der gewisse Nachteile des Upstream Filtering umgeht, ist das *Downstream-Filtering*. Hierbei werden Qualitätskriterien durch einen Dritten

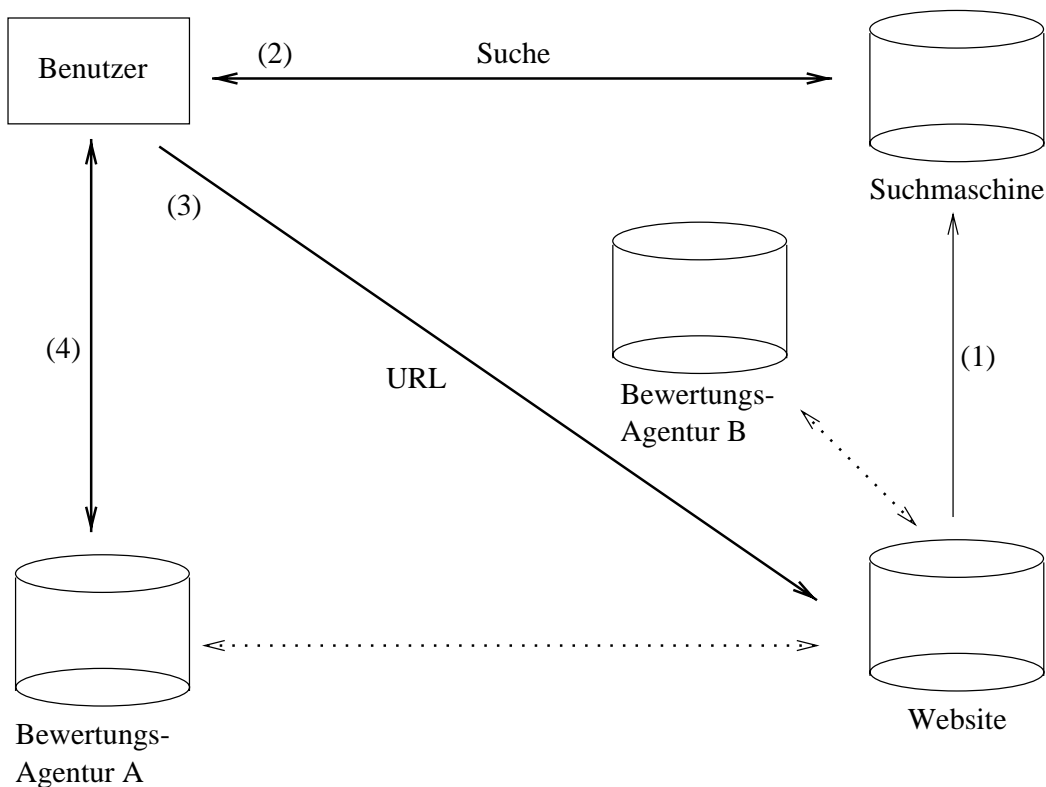


Abbildung 2.2: Normalerweise finden Benutzer Websites mit herkömmlichen Suchmaschinen (1,2,3). Alternativ könnten sie den Katalog einer Bewertungs-Agentur A verwenden (4). Bewertungs-Agenturen treffen aufgrund festgelegter Qualitätskriterien eine Vorauswahl für ihre Benutzer. Der Nachteil dabei ist, dass bereits die Wahl einer Bewertungs-Agentur für das Suchergebnis ausschlaggebend ist. Möglicherweise kennt der Benutzer die Agentur B nicht, obwohl diese für ihn passendere Ergebnisse liefern würde.[ED99]

festgelegt, und in ein computer-lesbares Vokabular übersetzt. Die Filterung erfolgt dann zumindest teilweise beim Benutzer (vgl. Abbildung 2.3). [ED98]

Eine Voraussetzung für diesen Ansatz ist, dass Informationen im Internet mit Metadaten in einer standardisierten Form „etikettiert“ sind (*label*), um es einer Software zu erlauben, Informationen zu überprüfen und zu suchen, die für einen individuellen Benutzer passen. Diese, den Inhalt beschreibenden Metadaten, können vom Autor selbst innerhalb der Information bereitgestellt werden oder – mit Einschränkungen – automatisch generiert werden. Die Software des Benutzers kann aber auch eine Anfrage an eine dritte Stelle richten, ob dort zusätzliche beschreibende und/oder bewertende Informationen (Metadaten) über die Site vorhanden sind. Software Produkte (Browser) können vom Benutzer angewiesen werden, jede Information, die nicht den persönlichen Qualitätserforder-

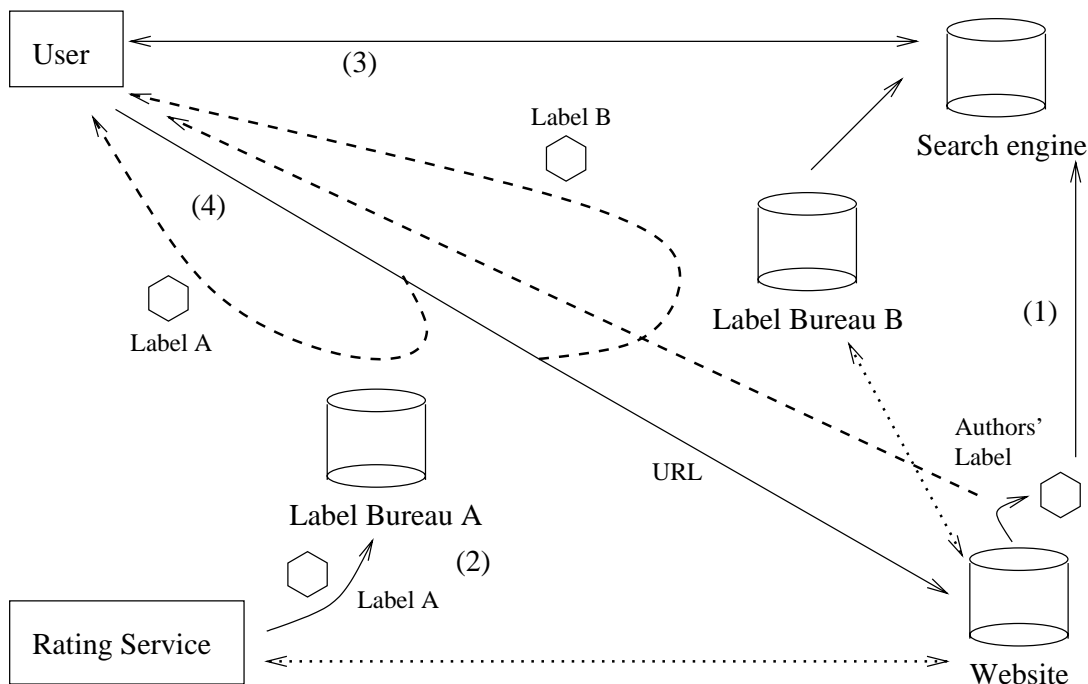


Abbildung 2.3: Autoren können zusätzlich zum Inhalt ihrer Website standardisierte, beschreibende Labels ihrer Sites bereitstellen, die von einer Suchmaschine indiziert werden können (1,3). Zusätzlich beschreiben und/oder bewerten *Rating Services* die Site, und speichern ihre Labels auf separaten Servern, „Label Büros“ genannt (2). Der Benutzer abonniert einen oder mehrere Bewertungs-Dienste, denen er vertraut, und bekommt automatisch deren Labels zu einer Site, die er abrufen (4). [ED99]

nissen oder Interessen entspricht, auszufiltern. Da beide Typen von Metadaten (die des Autors, und jene einer dritten Organisation) auch von Suchmaschinen indiziert werden können, hilft dieser Ansatz auch bei der Suche nach für den Benutzer passenden und relevanten Informationen. [ED98]

2.4 Zusammenfassung

In diesem Kapitel wurde gezeigt, was Qualität ist, und wodurch sich Datenqualität auszeichnet. Im Anschluss daran wurde erläutert, welche Eigenschaften die inhaltliche Qualität von Internet Ressourcen kennzeichnen. Hier bestehen große Unterschiede zu herkömmlichen Printmedien, bei denen Dokumente sich nicht so leicht verändern lassen. Bei der Aufzählung der Attribute wurde auch gezeigt, wie man durch Zuweisung von Werten zu Eigenschaften Angaben über die Attribute in einer computer-interpretierbaren Form speichern kann.

Es wurde weiters beschrieben, wie Bewertungen von Online-Ressourcen durchgeführt werden können, wobei im speziellen die Methode des *Downstream Filtering* interessant erscheint, weil dabei die individuellen Bedürfnisse der Benutzer hinsichtlich Qualität berücksichtigt werden können. Die Benutzer sind nicht von einigen wenigen Bewertungs Agenturen abhängig, die nach ihren eigenen Maßstäben Beurteilungen von Sites durchführen, sondern können selbst Qualitätsmaßstäbe vorgeben.

Die genannten Eigenschaften – auch Attribute genannt – sind sogenannte Metadaten, und sie müssen für das *Downstream Filtering* standardisiert und computer-lesbar bzw. -verarbeitbar gespeichert werden. Dazu wurden für das Internet spezielle Techniken entwickelt, die im folgenden Kapitel 3 beschrieben werden. Derart erstellte Metainformationen erlauben es dem Benutzer, unerwünschte Inhalte von sich fern zu halten (vgl. Kapitel 4), Empfehlungen mit anderen Benutzern auszutauschen (vgl. Kapitel 5) und in Verbindung mit Suchmaschinen Inhalte zu suchen (vgl. Kapitel 6).

Kapitel 3

Metadaten

Die Inhalte des World Wide Web, des meist benutzten Internetdienstes, sind von Menschen gestaltet, und dazu gedacht, von anderen Menschen aufgenommen zu werden. Und obwohl alles im Web auch maschinen-*lesbar* ist, ist es nicht unmittelbar maschinen-*verstehbar*. Deshalb ist es schwierig, Vorgänge des Filterns und Suchens zu automatisieren. Maschinen – im speziellen Suchmaschinen – können nicht ohne spezielle Vorkehrungen automatisch spezielle Informationen wie z.B. den Autor, das Publikationsdatum, das Thema etc. aus einem natürlich sprachigen Dokument, welches dem Computer in Form von Daten vorliegt, extrahieren. Solche Informationen über die Daten des Dokuments werden als Metadaten bezeichnet. [Lassila98]

In [SJ95] werden Metadaten bzw. Metainformationen als beschreibende Daten eines Dokuments bezeichnet. Sie können etwas über die Situation, die Form oder den Kontext eines Dokuments aussagen. Metadaten werden üblicherweise von einer Menge beschreibender Attribute repräsentiert, durch die eine Datenmenge oder eine andere Art von Information beschrieben wird. [AT99]

Oft ist es nicht leicht, Daten von Metadaten zu unterscheiden, denn dieselben Metadaten können in einem anderen Kontext Daten sein. Beispielsweise beinhaltet ein Bibliothekskatalog *Metadaten*, weil er die *Daten* von Büchern und Publikationen beschreibt. Um den Bibliothekskatalog zu bedienen, bedarf es meist einer speziellen Software. Für diese Software stellt der Katalog wiederum *Daten* dar. [Lassila98]

Zur Klärung des Begriffes Metadaten folgt deshalb im nächsten Abschnitt eine Definition und eine Aufzählung von Anwendungsmöglichkeiten. Daran anschließend werden Möglichkeiten der technischen Implementierung im World Wide Web besprochen, wozu das HTML META Tag, PICS und RDF gehören. Schließlich werden noch im Einsatz befindliche Metadaten Schemata besprochen. Das

sind Vereinbarungen darüber, welche Metadaten über Ressourcen auf welche Art gespeichert werden sollen.

3.1 Definition von Metadaten

Die Definition der Metadaten als „Daten *über* Daten“ bedarf noch einiger zusätzlicher Bemerkungen [Daniel97]:

- Es gibt eigentlich keine eindeutige Unterscheidung zwischen Daten und Metadaten. Wir können eine solche Entscheidung nur in Bezug auf eine ganz gewisse „*über*“-Relation treffen. Im obigen Beispiel (vgl. Kapitel 3) ist die Beziehung zwischen dem Bibliothekskatalog und den Daten der Büchern eine Relation, die andere besteht zwischen dem Bibliothekskatalog und der Software.
- Es gibt nicht nur eine „*über*“-Relation. Eine Relation könnte beispielsweise Informationen über die Qualität von Daten enthalten, andere Metadaten könnten etwas über deren Format aussagen, oder über das Layout.
- Ressourcen können ohne Rücksicht auf ihren Ort in Beziehung gesetzt werden. Die vernetzte Informationsarchitektur macht es möglich, dass Daten in einem Speicher andere Daten in einem anderen Speicher beschreiben.
- Die Rechenkapazitäten in vernetzten Architekturen machen es weiters möglich, auch über aktive oder dynamische Relationen von Datenmengen nachzudenken. Metadaten brauchen physisch gar nicht ständig zu existieren, sondern sie könnten automatisch abgeleitet werden.

In [RM96] werden folgende Verwendungsmöglichkeiten für Metadaten aufgezählt:

1. Metadaten werden eingesetzt, um Dokumente im Internet zu beschreiben und zu bewerten. Verbunden mit Indexern und Suchmaschinen, die diese Metadaten interpretieren können, erleichtert sich dadurch die Suche nach qualitativen Inhalten (Vgl. Kapitel 6 'Suchen durch Einsatz von Metadaten'). Zugleich wird es damit möglich, den Zugang von Kindern zu Seiten mit schädlichem Inhalt zu blockieren (Vgl. Kapitel 4 'Blockieren durch Einsatz von Metadaten').
2. Empfehlungssysteme erlauben es jedem, Empfehlungen beizusteuern, und diese zu verwenden, um andere auf interessante Inhalte hinzuweisen. Diese Beratung kann auch auf Personen mit ähnlichem Geschmack abgestimmt werden. Dieser drückt sich in der Einschätzung von Dokumenten aus, die

- beide Partner getroffen haben (Vgl. Kapitel 5 'Empfehlungs- und Annotationssysteme')
3. Online-Journale könnten alle Einreichungen veröffentlichen, aber Rezensionen in Form von Metadaten anhängen, die jeder Leser als Empfehlung oder Ablehnung interpretieren kann.
 4. Metadaten zur Beschreibung des geistigen Eigentums können entwickelt werden, um Benutzer auf den Urheber oder die erlaubte Art der Verwendung einer Ressource hinzuweisen.
 5. Benutzer können mittels Metadaten darauf hingewiesen werden, welche Art von Informationen während der Interaktion mit einer Website gesammelt werden.
 6. Desweiteren sind Wortregister zur Beschreibung des Ansehens von Firmen oder Usenet Autoren denkbar. Autoren mit einem schlechten Ruf können dann einfach ausgeblendet werden.

3.2 Technische Umsetzung

Im Alltag findet man Metadaten beispielsweise als Buch- und Filmkritiken, als Post-it® Notes oder in Empfehlungen. Die elektronische Entsprechung existiert ebenso in mehreren Ausprägungen. Die einfachste und zugleich älteste Form ist das HTML Meta Tag. Weiters wurde vom W3C, dem World Wide Web Consortium, die Platform for Internet Content Selection - kurz PICS - ins Leben gerufen, ursprünglich um den Zutritt zu verbotenem oder schädlichen Inhalt zu unterbinden. Und schließlich gibt es noch das Resource Description Framework, RDF, zur Erstellung komplexerer Metadaten Strukturen. Zusätzlich gibt es drei verschiedene Varianten in der Verfügbarmachung von Metadaten [AT99] [Fielding94]:

1. Die Metainformation wird extern vom Dokument gespeichert. Sie kann unabhängig abgerufen werden.
2. Die Metainformation und das Dokument sind in einem Container verpackt, welcher die Informationen bereitstellt, wenn sie benötigt werden.
3. Die Metainformation ist im Dokument selbst enthalten.

Die zur Zeit am häufigsten benutzte Form zur Repräsentation von Metadaten ist das HTML META Tag, welches im nachfolgenden Abschnitt beschrieben wird. Es folgen dann die Erläuterungen zu den erwähnten Systemen PICS und RDF.

3.2.1 HTML META Tag

Das META Tag spezifiziert eine Eigenschaft (z.B. „Autor“) und weist ihr einen Wert zu („Dave Raggett“). Listing 3.1 zeigt eine HTML Anweisung.

```
<META name="Author" content="Dave Raggett">
```

Listing 3.1: Einfache HTML Anweisung

Die Spezifikation definiert keine Menge gültiger Eigenschaften. Die Bedeutung einer Eigenschaft und die Menge gültiger Werte für die Eigenschaft wird in einem Referenz-Lexikon festgehalten, dem sog. *Profil*. Beispielsweise könnte ein Profil, welches entworfen wurde um Suchmaschinen beim Indizieren zu unterstützen, die Eigenschaften „author“, „copyright“, „keywords“ etc. enthalten. Um die Profile, auch Schemata genannt, geht es in Kapitel 3.3. Neben den Attributen *name* (zur Definition der Eigenschaft) und *content* (für den entsprechenden Wert) gibt es das Attribut *lang* zur Beschreibung der Sprache. Eine Suchmaschine kann so beispielsweise nach Schlüsselwörtern in verschiedenen Sprachen suchen (Vgl. Listing 3.2).

```
<META name="keywords" lang="en-us" content="vacation, sunshine">  
<META name="keywords" lang="en" content="holiday, sunshine">  
<META name="keywords" lang="fr" content="vacances, soleil">
```

Listing 3.2: Demonstration des Parameter lang

Statt des Attributes *name* existiert weiters das Attribut *http-equiv*, welches eine besondere Bedeutung hat wenn das Dokument via HTTP geladen wird, wie das Beispiel in Listing 3.3 demonstriert. Der Browser-Cache kann mit der Eigenschaft „Expires“ feststellen, wann eine neue Version von einem Dokument heruntergeladen werden soll.

```
<META http-equiv="Expires" content="Tue, 20 Aug 1999 14:32:00 GMT">
```

Listing 3.3: Beispiel für http-equiv

Die Metadaten müssen nicht wie in den bisherigen Beispielen im selben Dokument wie die Daten, welche durch sie beschrieben werden, liegen. Man verwendet in diesem Fall statt des META-Tags das LINK-Tag. [W3C Meta]

Die Verwendung des META-Tag hat aber einen großen Nachteil: Es lassen sich zwar Metainformationen in einem Dokument einbetten, aber es ist damit nicht möglich, einem Client – beispielsweise dem Browser eines Benutzers – mitzuteilen, was die Metainformationen beschreiben, und wie der Client damit umzugehen

hat. Deshalb hat das World Wide Web Consortium weitere Entwicklungen vorangetrieben, und die beiden Metadaten Frameworks PICS und RDF entwickelt, welche in den nächsten beiden Abschnitten beschrieben werden.

3.2.2 PICS

Wie im vorigen Abschnitt erläutert, hat die Verwendung des HTML META-Tag gewisse Nachteile, die von PICS größtenteils beseitigt werden sollten. PICS¹ steht für Platform for Internet Content Selection, und ist eine Konvention zur Einführung sog. *Labels* für das Internet. PICS Label beschreiben Inhalte in ein oder mehr Dimensionen unter Verwendung zweck-gebundener Wortschätze, und sie erlauben einer Selektions Software über den Zutritt zu fragwürdigen Inhalten zu entscheiden. [Armstrong97]

Die bekannteste und zugleich ursprüngliche Verwendung von PICS liegt in der Kontrolle des Internetzugangs zu schädlichem und verbotenem Inhalt (vgl. Kapitel 4). Der lokale Browser kann z.B. angewiesen werden, den Zutritt zu Inhalten zu unterbinden (blocking), wenn darin Darstellungen von Gewalt einen zuvor von Eltern oder Lehrern bestimmten Grad (level) überschreiten. Der Besitzer einer Ressource oder ein Bewertungs-Service zeigt auf einer gegenseitig akzeptierten Skala den Pegel von Gewalt, Obszönitäten etc. an, und die Selektionssoftware blendet Seiten aus, deren Label die voreingestellten Levels überschreiten. Bei Sites ohne Label ist zu entscheiden, ob sie generell eingeblendet oder alle ausgeblendet werden sollen. Wobei es wichtig ist festzuhalten, dass PICS selbst nur eine Methode bzw. Infrastruktur (*Framework*) zur Verfügung stellt, und kein Bewertungs-Service (vgl. Kapitel 2.3) oder aktives System der Zensur oder Auswahl ist. Es ist wertfrei. [Armstrong97]

Die Mechanismen um Zutritt zu beschränken bzw. zu erreichen sind jedoch zwei Seiten derselben Münze: Anforderungen, den Zugang zu allen Seiten, welche mit einem speziellen Thema zu tun haben, zu unterbinden sind sehr ähnlich mit Anforderungen, Seiten mit demselben Inhalt zu finden. [Armstrong97]

Ein Label besteht aus den drei Teilen *Service identifier*, *Label options* und einem *Rating*. Der Service identifier gibt die URL des Bewertungs-Service an, welche diesen eindeutig identifiziert. Die Label options halten zusätzliche Eigenschaften der zu bewertenden Ressource sowie der Bewertung selbst fest, wie z.B. das Datum der Kritik. Die Bewertung (Rating) selbst ist eine Menge von Attribut-Wert Paaren, die eine Ressource in ein oder mehr Dimensionen beschreiben. Ein oder mehr Labels können zusammen als Liste transportiert werden, wobei die allgemeine Form wie in Listing 3.4 aussieht. [W3C PICS]

¹<http://www.w3.org/PICS/>

```
(PICS-1.1
  <service url> [option...]
    labels [option...] ratings (<category> <value> ...)
      [option...] ratings (<category> <value> ...)
      ...
  <service url> [option...]
    labels [option...] ratings (<category> <value> ...)
      [option...] ratings (<category> <value> ...)
...)
```

Listing 3.4: Allgemeine Form von PICS Labels

Es sind zwei Arten von Labels zu unterscheiden: Ein *specific* Label gilt für ein einzelnes Dokument, und trifft nicht auf Dokumente zu, auf die evt. verwiesen wird. Ein *generic* Label wird durch die **generic** Option gekennzeichnet, und gilt für jedes Dokument, dessen URL mit dem durch die **for** Option vorgegebenen String beginnt. Eine Liste mit den wichtigsten Optionen findet sich im Anhang B.1. Das Beispiel in Listing 3.5 zeigt ein spezielles Label. [W3C PICS]

```
(PICS-1.1
  "http://www.gcf.org/v2.5" by "John Doe"
    labels on "1994.11.05T08:15-0500"
      until "1995.12.31T23:59-0000"
      for "http://w3.org/PICS/Overview.html"
      ratings (suds 0.5 density 0 color/hue 1)
      for "http://w3.org/PICS/Underview.html"
      by "Jane Doe"
      ratings (subject 2 density 1 color/hue 1) )
```

Listing 3.5: Beispiel für ein spezielles Label

Ein PICS Label kann auf drei verschiedene Arten übertragen werden:

1. Labels können in ein HTML Dokument unter Verwendung des META Tags (vgl. Abschnitt 3.2.1) eingebettet werden. Der dazugehörige Befehl sieht wie in Listing 3.6 aus.

```
<META http-equiv="PICS-Label" content='label-list'>
```

Listing 3.6: Einbettung von Labels in HTML

Ein auf diese Weise eingebettetes Label kann die oben erwähnte **for** Option weglassen. Es ist für das Dokument gültig, egal durch welche URL es

aufgerufen werden kann. Handelt es sich um ein *generic* label, das in ein Dokument integriert ist, welches via einer „home“-URL (z.B. ein URL Pfad der mit einem Slash endet) aufgerufen werden kann, dann gilt das Label für alle URLs, welche die „home“-URL als Prefix besitzen. [W3C PICS]

Wenn ein Client am Dokument „http://www.docs.com/foo/bar/bat.html“ interessiert ist, kann er zuerst überprüfen ob ein spezielles Label integriert ist. Wenn nicht kann er das Dokument „http://www.docs.com/foo/bar/“ anfordern. Der Server wird dann das home-Dokument zurücksenden, z.B. foo/bar/index.html oder foo/bar/home.htm, abhängig vom Server. Wenn dieses Dokument ein *generic* Label enthält, kann der Client es als gültig auch für bat.html ansehen. Wenn der Client auch dort kein Label findet, kann er eine Stufe höher in der Hierarchie nach labels suchen. [W3C PICS]

2. Viele Protokolle, wie z.B. Email, HTTP und Usenet benutzen einen US-ASCII Header, wie in RFC-822² beschrieben. Für die Benutzung in solchen Protokollen kann man einen neuen Header definieren, PICS-Label, dessen Syntax lautet: [W3C PICS]

PICS-Label: <labellist>

Listing 3.7: Neuer Header PICS-Label

3. PICS Label können auch unabhängig von dem Dokument, welches sie beschreiben, gespeichert werden. Sie können beispielsweise auf dem Server eines sogenannten Label-Büros abgelegt werden. Wenn ein Konsument an den Beschreibungen eines speziellen Label-Büros interessiert ist, so kann dieser Einstellungen in seinem Browser treffen, damit jedesmal, wenn er eine beliebige URL abrufen möchte, gleichzeitig eine Anfrage an den Server des Label-Büros gestellt wird. Besitzt der entsprechende Server ein Label, welches das Dokument an der URL beschreibt, so wird es an den Konsumenten geschickt (in der Form eines Dokuments vom Typ „application/pics-label“). Der Browser, oder jede andere Client-Software, können dann angewiesen werden, das ursprünglich angeforderte Dokument entweder anzuzeigen oder zu blockieren, wenn das Label vorher festgelegte Kriterien nicht erfüllt (vgl. Kapitel 4). Die dritte Möglichkeit ist, dass dem Konsumenten zuerst die Beschreibung der Ressource gezeigt wird, und er dann entscheiden kann, ob er sie sehen will oder nicht.[W3C PICS]

In [ED99] wird ein Prototyp eines auf PICS basierenden Bewertungsvokabular für medizinische Informationen (med-PICS) erläutert, welcher sowohl *beschreibende* (*descriptive*) als auch *bewertende* (*evaluative*) Kategorien enthält, die vom Webautor und unabhängigen Bewertungsagenturen (wie z.B. medizinischen Vereinen) verwendet werden können.

²<http://puma.germany.net/internic/rfc/rfc822.txt>

Wie bereits einangs erwähnt, erlaubt PICS den Attributen als Wertzuweisungen nur Zahlen innerhalb eines vorzugebenden Bereichs. Man kann z.B. festlegen, dass das Attribut *Note* die Werte 1 für *Sehr gut* bis 5 für *Nicht genügend* annehmen darf. Es gibt daher auch in med-PICS keine der im Bibliothekswesen üblichen, beschreibenden Attribute wie *Autor* oder *Titel*. [ED99]

Aufgrund der erwähnten Einschränkungen, die für PICS gelten, wurde *PICS - Next Generation (PICS-NG)*³ entwickelt. Dieses sollte die Anforderungen von PICS 1.1 unterstützen, aber zusätzlich Attribute mit textlichen Werten zulassen, wie diese z.B. in Dublin Core (vgl. Abschnitt 3.3.3) verwendet werden. [Wood97] Im folgenden Abschnitt wird eine andere Technik zur Implementierung von Metadaten im World Wide Web erläutert, die es ebenso erlaubt, Metadaten zwischen verschiedenen Clients, also verschiedenen Applikationen, auszutauschen. Bei dieser Technik, RDF genannt, ist man nicht auf die oben erwähnten speziellen Wertzuweisungen beschränkt.

3.2.3 RDF

Gewissermaßen als Nachfolger von PICS veröffentlichte das World Wide Web Consortium das sogenannte Resource Description Framework⁴ als einen Standard für Web-Metadaten. Es erlaubt das Austauschen von Metadaten zwischen verschiedenen Anwendungen. Das Design wurde von verschiedenen Quellen beeinflusst, allen voran von der Webgemeinschaft selbst, im speziellen durch die HTML Meta Tags und die Platform for Internet Content Selection (vgl. Abschnitte 3.2.1 und 3.2.2). Andere Einflüsse und Konzepte stammen aus dem Bibliothekswesen, den strukturierten Sprachen (in der Form von SGML⁵ und XML⁶), der objekt-orientierten Programmiersprachen und Datenbanken sowie aus der Wissens-Repräsentation. Jede Art von Ressource, die mit einem Uniform Resource Identifier (URI) benannt werden kann, kann mit RDF beschrieben werden. Das ist die Klasse der Web Bezeichner, zu der auch die bekannten URLs (Uniform Resource Locator) zählen. [Lassila98]

RDF Daten bestehen aus Knoten und Attribut-Wert Paaren. Die Knoten können jede Art Web Ressource sein, inklusive anderer Instanzen von Metadaten. Attribute sind benannte Eigenschaften der Knoten, und deren Werte sind entweder atomar (Text Strings, Datum, Zahl) oder andere Knoten. Das RDF Modell kann in einer objekt-orientierten Sichtweise auch als gerichteter Graph aufgefasst werden, wobei die Attribute die Kanten zwischen den Knoten bezeichnen. Die

³<http://www.w3.org/PICS/NG>

⁴<http://www.w3.org/RDF/>

⁵<http://www.oasis-open.org/cover/>

⁶<http://www.w3.org/XML/>

Entwickler wählten XML, um Instanzen dieses Modells in Dateien speichern zu können. [Lassila98]

Im folgenden Beispiel wird die Terminologie von Dublin Core verwendet, einem Metadaten Schema zur Bildung digitaler Bibliothekskataloge (vgl. Abschnitt 3.3.3). Das Metadaten Fragment in Listing 3.8 beschreibt eine Web-Ressource mit einer gegebenen URL, und besagt, dass „John Smith“ mit der Email Adresse „smith@some.org“ der Autor ist. Die Web-Page ist ein Knoten mit einer Eigenschaft, genannt „DC:Creator“. Deren Wert ist wiederum eine Struktur bestehend aus zwei weiteren Eigenschaften, dem Namen und der Email Adresse. Abbildung 3.1 stellt den Zusammenhang als Graphen dar. [Lassila98] Wie PICS enthält auch RDF keine vordefinierten Vokabel für die Erstellung von Metadaten.

```
<?xml:namespace ns="http://www.w3.org/TR/WD-rdf-syntax"prefix="RDF"?>
<?xml:namespace ns="http://purl.org/metadata/dublin_core"prefix="DC"?>
<?xml:namespace ns="http://some.org/schemata/people"prefix="P"?>

<RDF:RDF>
  <RDF:Description about="http://www.some.org/smith">
    <DC:Creator>
      <RDF:Description>
        <P:Name>John Smith</P:Name>
        <P:Email>mailto:smith@some.org</P:Email>
      </RDF:Description>
    </DC:Creator>
  </RDF:Description>
</RDF:RDF>
```

Listing 3.8: Beispiel für RDF Metadaten [Lassila98]

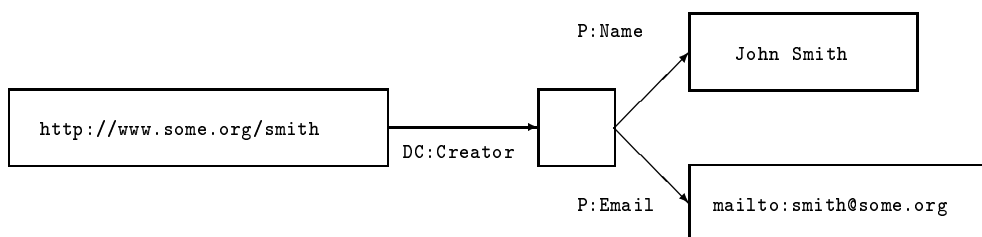


Abbildung 3.1: Darstellung des Beispiel 3.8 als Graph [Lassila98]

Das fehlende Bindeglied zwischen der technischen Implementierung und der tatsächlichen Verwendung von Metadaten sind Vereinbarungen darüber, *welche*

Informationen (Kriterien) über Daten man mit *welchen Vokabeln* und dazugehörigen Eigenschaften beschreibt. Solche Vereinbarungen, Schemata genannt, werden in den nächsten Abschnitten vorgestellt.

3.3 Metadaten Schemata

Wie im vorigen Abschnitt beschrieben, stellt insbesondere RDF unabhängigen Gemeinschaften⁷ eine Basis zur Verfügung, auf welcher diese Metadaten Attribute (vocabulary) nach ihren speziellen Bedürfnissen kreieren können. Damit auch andere mit solcherart erstellten Metadaten umgehen und sie verstehen können, muss zuerst die Bedeutung der Ausdrücke detailliert formuliert werden. Diese Beschreibung der Wortschätze heißt in RDF Schema. Ein Schema definiert die Bedeutung, die Eigenschaften und die Beziehungen einer Menge von Eigenschaften untereinander. Dazu zählt auch die Einschränkung mancher Attribute auf zulässige Werte, und die Vererbung von Attributen durch andere Schemata. RDF macht es jedem Dokument, welches Metadaten enthält, möglich, klarzustellen welcher Wortschatz verwendet wird, indem jedem Wortschatz eine Web-Adresse zugewiesen wird. [W3C Activity]

Neben den in den folgenden Abschnitten beschriebenen Schemata MARC und SOIF ist Dublin Core (vgl. Abschnitt 3.3.3) eines der bekanntesten Metadaten Schemata, welches aus der Bibliotheksgemeinschaft kommt. Der Name stammt daher, dass das erste Treffen in Dublin, Ohio, USA, stattgefunden hat. Ein weiteres bekanntes Schema heißt LOM (vgl. Abschnitt 3.3.4), welches sich die Beschreibung von Lernstoff mit Metadaten zum Ziel setzt.

3.3.1 MARC

MARC⁸ kann als der Wegbereiter der elektronischen Metadaten Formate bezeichnet werden, nachdem es in den sechziger Jahren für die elektronische Integration von weitverbreiteten Bibliotheks Ressourcen entworfen wurde. MARC ist eigentlich nicht ein einziges Format, sondern eine Familie von Formaten, die alle eine ähnliche Struktur besitzen. Wegen unterschiedlicher lokaler Anforderungen, wie z.B. zweisprachige Beschreibungen oder unterschiedliche Zahlenformate, haben sich nationale Formate entwickelt, und auch innerhalb dieser haben sich Varianten ausgebildet⁹. [Heery96]

⁷communities

⁸*machine readable catalogue*

⁹Die nationalen Varianten US-MARC, UK-MARC und Canadian-MARC sollten bis 1. Jänner 1999 vereinheitlicht worden sein. [Heery96]

Das Format ist in der Benutzung nicht sehr bequem, und erfordert fortgeschrittene Kenntnisse in der Katalogisierung und das Festhalten an einer komplizierten Syntax. Für Bibliotheken, die alte Systeme betreiben, ist MARC aber vielleicht der beste Weg, neue elektronische Ressourcen mit den alten Katalogen zu integrieren. MARC wird nämlich durch Dublin Core mit den „Warwick Framework“¹⁰ Erweiterungen unterstützt, wodurch sich die Vorteile mehrerer Metadaten Formate nutzen lassen. [AT99]

3.3.2 SOIF/RDM

SOIF (Summary Object Interchange Format), welches im Web-Indexing Shareware Paket „Harvest“ der Universität von Colorado enthalten ist, kann man nicht im eigentlichen Sinn als Metadaten Schema bezeichnen. Es ist ein Format bzw. *Framework*, welches die Ermittlung der Metadaten aus diversen Textformaten und deren Übertragung regelt. Da in der Standardkonfiguration aber eine bestimmte Menge von Attributen enthalten ist, kann man in diesem Fall auch von einem Metadaten Schema sprechen.

Abhängig von der Konfiguration der Harvest Software kann eine große Anzahl an Tags¹¹ unterstützt werden. In der erwähnten Standardkonfiguration (vgl. Tabelle 3.1) werden grundlegende Informationen über den Autor, das Format, Stichwörter und das Thema unterstützt. Dabei ist der Wertebereich jeder Eigenschaft nicht eingeschränkt oder standardisiert. SOIF ist mit den wichtigsten Dokumentformaten wie SGML, HTML, PostScript oder RTF kompatibel. Das Format hat einen bedeutenden Fortschritt erlangt, nachdem Netscape erklärt hat, ihre RDM¹² Such-Syntax arbeitet direkt mit SOIF Objekten zusammen, um die Suche zu verbessern. SOIF/RDM ist neben Dublin Core wahrscheinlich eines der populärsten Formate, und profitiert von der Verbindung zu Harvest und Netscape. [AT99]

3.3.3 Dublin Core

Dublin Core¹³ ist ein Metadaten Schema zur Beschreibung elektronischer Ressourcen. Ursprünglich gedacht für die Autor-generierte Beschreibung von Web Ressourcen, hat es bald auch die Aufmerksamkeit von Museen, Bibliotheken

¹⁰<http://www.lub.lu.se/tk/warwick.html>

¹¹Ein *Tag* bezeichnet ein Element innerhalb eines Dokuments, welches dadurch eine spezielle Bedeutung erhält. Beispielsweise kann die Überschrift mit einem Tag als solche gekennzeichnet werden.

¹²Resource Description Messaging

¹³<http://purl.org/dc>

Bezeichnung	Kurzbeschreibung
Abstract	Kurze Inhaltsangabe über das Objekt
Author	Autor(en) des Objekts
Description	Kurze Beschreibung des Objekts
File-Size	Datei Größe in Bytes
Full-Text	Gesamter Inhalt des Objekts
Gatherer-Host	Host, auf dem Gatherer lief, der SOIF-Objekt erstellte
Gatherer-Name	Name dieses Gatherers
Gatherer-Port	Portnummer des Gatherers
Gatherer-Version	Versions Nummer des Gatherers
Keywords	Schlüsselwörter im Objekt
Last-Modification	Zeit der letzten Modifikation des Objekts
MD5	MD5 16-byte Checksumme des Objekts
Refresh-Rate	Zeit nach Update-Time, um SOIF-Objekt zu aktualisieren
Time-to-Live	Zeit nach Update-Time, ab der SOIF-Objekt ungültig
Title	Titel des Objekts
Type	Objekttyp
Update-Time	Zeit, zu der SOIF-Objekt zuletzt aktualisiert wurde
URL-References	Die URLs von Links in HTML Objekten

Tabelle 3.1: Standardattribute bei der Verwendung von SOIF [SOIF96]

und Regierungsinstitutionen auf sich gezogen. Das zentrale Merkmal von Dublin Core ist ein inter-disziplinäres und internationales Übereinkommen über eine Kernmenge von Elementen, welche in Tabelle 3.2 aufgelistet sind. Dabei ist der erlaubte Wertbereich für jede Eigenschaft nicht eingeschränkt. Es existieren aber Vorschläge, wie man beispielsweise das Datum in einer standardisierten Form angeben kann. [DublinCore99]

Dublin Core weist folgende Merkmale auf, welche es als einen prominenten Kandidaten unter den Metadaten Schemata auszeichnet: Dublin Core ist gedacht für die Verwendung durch Spezialisten auf dem Gebiet der Ressourcen Beschreibung und die Ersteller von Ressourcen. Es ist unabhängig von der zu beschreibenden Disziplin einsetzbar, und international (in derzeit 20 Ländern) anerkannt. Desweiteren ist Dublin Core eine ökonomische Alternative zu umständlichen Beschreibungsmodellen, wie z.B. die volle MARC-Katalogisierung der Bibliotheken. [DublinCore99]

Doch gerade der Anspruch auf die universelle Einsetzbarkeit, der große Einfluss aus dem Bibliothekswesen, und die internationale Übereinkunft schränken die Verwendung von Dublin Core bei der ganz speziellen Bewertung der inhaltlichen Qualität von Ressourcen ein. Dublin Core dient der völlig objektiven Beschreibung von Ressourcen, und bietet keine Möglichkeit, diese auch zu bewerten

Element	Beschreibung
Title	Benennung einer Ressource
Creator	Primär verantwortliche Einheit für Inhalt
Subject	Thema des Inhalts
Description	Beschreibung des Inhalts
Publisher	Herausgeber der Ressource
Contributor	Beitragende
Date	Üblicherweise das Datum der Erstellung
Type	Art der Ressource (vgl. Anhang B.2)
Format	Digitale oder physikalische Form der Ressource
Identifier	Eindeutige Referenz, Bsp. ISBN Nr, URL
Source	Quelle der Ressource
Language	Sprache des Inhalts
Relation	Referenz zu verwandter Ressource
Coverage	Ausmaß des Inhalts
Rights	Rechte über die Ressource

Tabelle 3.2: Dublin Core Metadata Element Set, Version 1.1 [DublinCore99]

oder zu kritisieren. Diese Tätigkeit wird auch nicht von Bibliothekaren durchgeführt, sondern von Kritikern, Prüfern und Publizisten.

Schon auf dem OCLC/NCSA¹⁴ Metadaten Workshop im Jahr 1995, wo man sich auf das *Dublin Metadata Core Element Set* einigte, fasste man den Entschluss zwei spezielle Einschränkungen im Design zu akzeptieren: Erstens sollte es die Aufgabe der Elementmenge sein, „dokument-ähnliche Objekte¹⁵“ zu beschreiben, obwohl dieser Terminus nicht genauer spezifiziert wurde. Zweitens einigten sich die Teilnehmer des Workshop darauf, dass nur Elemente zur Ressourcenentdeckung aufgenommen werden sollen, und nicht zum Suchen und Auffinden. [Heery96]

Im nächsten Abschnitt wird ein anderes Metadaten Schema erläutert, welches ebenso einen speziellen Typus von Ressourcen beschreiben soll, nämlich die sogenannte Lernobjekte. Darunter fällt alles, was im weitesten Sinn mit Lehren und Lernen zu tun hat.

¹⁴National Centre for Supercomputer Applications

¹⁵DLO *document like objects*

3.3.4 LOM, IMS

LOM¹⁶ steht für „Learning Object Metadata“, und stammt vom IEEE Learning Technology Standards Committee (LTSC). IMS¹⁷ variiert LOM in der Hinsicht, dass nicht ein generelles Schema für Lernobjekte definiert wird, sondern fünf unterschiedliche, zur Beschreibung der verschiedenen Typen von Ressourcen zum Lernen.

Lernobjekte sind in diesem Kontext als digitale oder nicht digitale Einheiten definiert, die wiederverwendet oder referenziert werden können, während eines technik-unterstützten Lernens, beispielsweise im *Computer Based Training* (CBT), in *Distance Learning Systems* (DLS) und *Collaborative Learning Environments*. Beispiele für Lernobjekte sind multimedialer Inhalt, Anweisungen, Lernziele, Software Tools, etc. Zu den wichtigsten Anwendungsgebieten zählen folgende Punkte: [IMS99]

1. Den Lernenden und Lehrenden das Suchen, Erwerben und Gebrauchen von Lernobjekten zu ermöglichen.
2. Das Austauschen von Lernobjekten über verschiedene Arten des technik-unterstützten Lernens zu erleichtern.
3. Computer Agenten das automatische und dynamische Zusammenstellen von persönlichen Lektionen zu ermöglichen.

Die Elemente in LOM und IMS sind hierarchisch und in Gruppen angeordnet. Sie besitzen verschiedene, festgelegte und standardisierte Datentypen, und sie sind obligatorisch oder fakultativ auszufüllen. Es besteht eine gewisse Kompatibilität zwischen LOM/IMS und Dublin Core (vgl. Anhang B.3) [LOM98]

LOM und IMS charakterisieren eine ganz spezielle Art von Ressourcen (nämlich Lernobjekte) nach objektiven Kriterien. Es ist wie bei DC nicht möglich, in den Metadaten Bewertungen oder Kritik über die Ressource anzubringen. Es ist außerdem fraglich, wieviele Menschen sich die Zeit nehmen, eine Ressource mit bis zu 80 Feldern zu charakterisieren.

3.4 Zusammenfassung

Bisher fehlte dem Web jener Teil, welcher Informationen über Informationen enthält - katalogisierende und beschreibende Information, strukturiert in einer

¹⁶<http://ltsc.ieee.org/doc/wg12/LOMdoc2.1.html>

¹⁷<http://www.imsproject.org>

Form, die es Computern erlaubt, Webseiten auf geeignete Art zu suchen und zu indizieren. Die Technik dazu wurde vom W3C, dem World Wide Web Consortium, entwickelt: RDF, welches ein umfassendes Gerüst bereitstellt, um jedwede Art von Metadaten zu implementieren, und PICS, das vorwiegend zum Ausblenden von schädlichem Inhalt geeignet ist. Die Entwicklung von PICS ist abgeschlossen, RDF befindet sich noch in Arbeit. [W3C Activity]

Das fehlende Bindeglied zwischen der technischen Implementierung und der tatsächlichen Verwendung von Metadaten sind sogenannte Schemata. Das sind Vereinbarungen darüber, *welche* Informationen (Kriterien) über Daten man mit *welchen Vokabeln* und dazugehörigen Eigenschaften beschreibt. Eines der bekanntesten und verbreitetsten Schemata ist Dublin Core, welches der Beschreibung von Dokumenten dient, sie aber nicht bewerten kann.

Im folgenden Kapitel 4 wird erläutert, wie man mit Hilfe der vorgestellten Techniken (Verwendung von PICS Labels) im Besonderen Kinder vor gefährlichen Inhalten aus dem Internet schützen kann. Danach wird in Kapitel 5 aufgezeigt, wie man Metadaten dazu einsetzen kann, ähnlich Denkende auf besonders interessante Dokumente hinzuweisen bzw. von schlechten Inhalten abzuraten.

Schließlich behandelt Kapitel 6 die Einsatzmöglichkeiten von Metadaten in Kombination mit Suchmaschinen, zur besseren und schnelleren Auffindung von interessanten Informationen im World Wide Web. Zu diesem Zweck benötigt man ein Metadaten Schema, welches Ressourcen nicht nur inhaltlich beschreibt (wie z.B. Dublin Core), sondern auch deren Qualität bewertet.

Aus den in Abschnitt 3.2.2 gemachten Überlegungen ergibt sich, dass durch die Verwendung von Attributen, welche einen eingeschränkten Wertevorrat besitzen, eine automatische Auswertung der Metadaten durch den Computer erfolgen kann. Wenn man nun die inhaltliche Qualität mit Attributen beschreiben könnte, deren Wertevorrat eingeschränkt ist, wäre dadurch in weiterer Folge auch die Suche nach Qualität möglich. Bei der Entwicklung eines Qualitäts-Metadaten Schemas ist daher zu beachten, dass so viele Attribute wie möglich in einer computerinterpretierbaren Form definiert werden. Im Gestaltungsbereich der Arbeit wird ein solches Schema entwickelt, welches zu einem grossen Teil auf den in diesem Kapitel gemachten Untersuchungen beruht.

Kapitel 4

Blockieren durch Einsatz von Metadaten

Es wird in diesem Kapitel gezeigt, wie Benutzer einen Browser unter Einsatz von Metadaten anweisen können, gewisse Inhalte nicht anzuzeigen (zu blockieren). Beispielsweise könnten Eltern daran interessiert sein, dass ihre Kinder beim „surfen“ im Web keine gewaltverherrlichenden Seiten besuchen können.

Mit dem explosiven Wachstum des Internet ist ein Problem verbunden, welches im Grunde genommen alle Medien betrifft, die für verschiedenste Zuhörerschaften gedacht sind: Nicht alle Materialien sind für jedes Publikum gleich gut geeignet. Jede Gesellschaft reagiert unterschiedlich auf die Charakteristiken der Medien. In den meisten Ländern gibt es jedoch mehr Einschränkungen für leicht zugängliche Medien wie Rundfunk und Fernsehen als für gedruckte Medien. Die Verteilung und Verbreitung gedruckter Medien kann man nämlich wesentlich leichter kontrollieren und beschränken, als beispielsweise Publikationen im Internet. Und obwohl der Zugang zum Internet beinahe so leicht möglich ist wie zum Fernsehen, gibt es keine oder nur geringe gesetzliche Beschränkungen¹. Die unterschiedlichen Gesetze zur Beschränkung von Übertragungen werden für die einen immer zu restriktiv sein, während sie für andere noch zuwenig restriktiv sind. [RM96]

Im Internet ist es jedoch nicht unbedingt nötig, die Verbreitung von Inhalten zu zensurieren. Denn es gibt Techniken, die es dem Benutzer erlauben, den Zugang zu möglicherweise schädlichem Inhalten selbst zu kontrollieren. [Resnick97]

¹Allerdings gibt es Länder auf der Welt, in denen der Zugang zum Internet stark eingeschränkt oder verboten ist. Aber auch Firmen oder Universitäten, die ihren Angestellten bzw. Studenten einen kostenfreien Zugang zum Internet ermöglichen, können dessen Gebrauch limitieren.

Beispielsweise kann der Zugang zu fragwürdigen Usenet News durch die Software leicht unterbunden werden, weil jede Newsgroup einen Namen besitzt, der das Thema beschreibt (wie z.B. rec.music.folk, soc.culture.peru oder alt.tv.x-files). Eltern könnten also ihren Kindern den Zugang zu jenen Newsgroups blockieren, deren Name anzeigt, dass sie sexuell freizügiges Material enthalten (wie z.B. alt.sex.stories). [Weinberg97]

Den Zugang zu diesen Inhalten im Web zu unterbinden ist jedoch viel schwieriger. Blocking Software der ersten Generation basiert auf Listen mit URLs von Webseiten, die z.B. für Kinder nicht geeignet sind. Diese Listen werden mit zwei Methoden erstellt: einerseits arbeiten sich Bewerter per Hand durch das Web und verfolgen Links auf Seiten mit fragwürdigem Inhalt, andererseits sucht eine Software nach Seiten, die verbotene Wörter wie „sex“ oder „xxx“ in der URL enthalten. [Weinberg97]

Mit Hilfe von PICS Labels (vgl. Kap. 3.2.2) können wesentlich bessere Blocking Software Systeme entwickelt werden: Anstatt ein Dokument nur als „für Erwachsene“ oder als „sicher für Kinder“ einzustufen, erlaubt PICS den Bewertungs-Agenturen die Klassifikation in mehreren Dimensionen, wie z.B. den Grad an Gewaltdarstellungen, Gewaltausdrücken, Pornographie oder Obszönitäten. Entlang jeder Dimension gibt es eine Menge zulässiger Werte, beispielsweise eine Skala von 1 bis 10 für den zunehmenden Grad an Darstellungen von Gewalt. Eltern wird es damit z.B. möglich gemacht, Seiten, die mit einem Gewaltlevel von über 3 und einem Level von über 8 für Darstellungen von Sex bewertet wurden, für ihre Kinder auszublenden. [Weinberg97]

Eine einzelne Seite oder ein einzelnes Dokument kann mehrere Labels besitzen, die von verschiedenen Organisationen bereitgestellt werden. Diese Organisationen können das Vokabular für ihre Labels frei wählen (vgl. Abbildung 4.1). Dadurch ist es möglich, Ressourcen angepasst an unterschiedliche Kulturkreise, Religionen oder persönliche Einstellungen zu bewerten. Im Folgenden sind einige ausgewählte Beispiele² von Blockingsoftware näher erläutert. [RM96]

²Weitere nennenswerte Systeme sind:
Net Nanny, <http://www.netnanny.com>
Safe Surf, <http://www.safesurf.com> und
Surf Watch, <http://www.surfwatch.com>

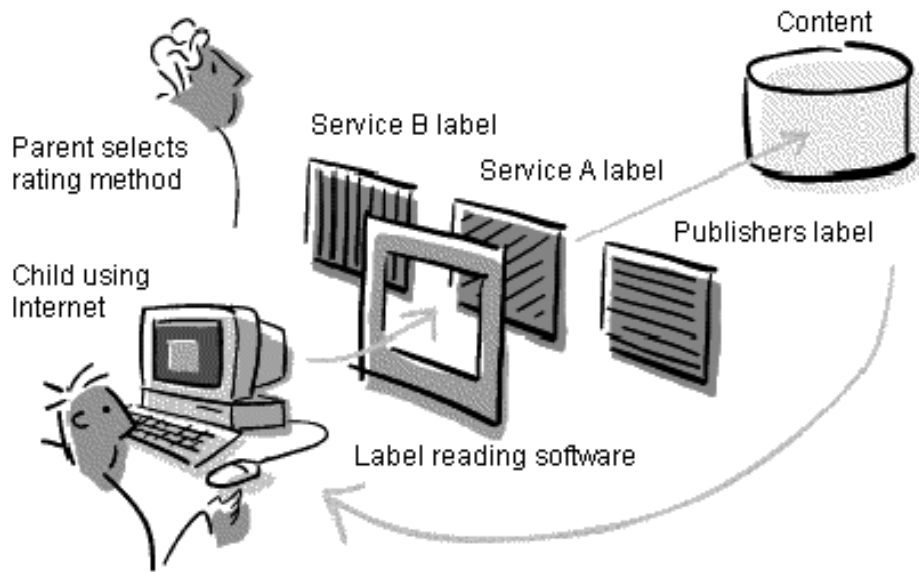


Abbildung 4.1: Anwendung von Blocking Systemen [RM96]

4.1 Blocking Systeme

4.1.1 CYBERSitter von Solid Oak

Solid Oak³ ist eine 1986 in Santa Barbara gegründete Softwarefirma, welche Mitte 1995 den Internetfilter CYBERSitter auf den Markt gebracht hat. Die Software funktionierte ursprünglich nur mit dem eigenen Bewertungsdienst VCR (Voluntary Content Rating). Weil die Kunden es wünschten kann CYBERSitter mittlerweile auch PICS-Einstufungen syntaktisch analysieren. [Dyson97]

Der Filter blockiert viele tausend Sites, die ein Bewertungskomitee bei Solid Oak aufgrund von Benutzermeldungen und systematischen Recherchen ausgewählt hat. Firmengründer Brian Milburn meint 1997, dass nur ungefähr neun Firmen für 50 Prozent der wirklich schmutzigen Inhalte im Internet verantwortlich zeichnen. Diese Firmen halten jedoch Kinder gerne von ihren Sites fern, und stufen sich daher selbst ein. Die Kriterien der aktuellen Version von CYBERSitter sind in Anhang C.1 aufgelistet. [Dyson97]

³Solid Oak Software Inc., PO Box 6826, Santa Barbara, CA 93160, USA. <http://www.solidoak.com>

4.1.2 Cyber Patrol von Microsystems Inc.

Microsystems⁴ tritt als Privatfirma nachdrücklich für die Durchsetzung von PICS ein, begünstigt auch durch die Nähe zu Cambridge, Massachusettes, wo die Entwicklung von PICS stattfand. Im Rahmen dieses Engagements unterstützt die Software auch RSACi (vgl. Kapitel 4.1.3), und sie kann deren Labels ohne Veränderungen oder Anpassungen übernehmen. [Dyson97]

Das Herz von Cyber Patrol bilden die Filtermaschine, die CyberNOTTM und die CyberYESTM Listen. Die Sites auf der CyberNOT Liste werden von einem Team von Internet Spezialisten, Lehrern und Eltern gepflegt. Sie benutzen eine Menge festgelegter Kriterien, welche Internet-Sites und -Ressourcen mit möglicherweise anstößigem Inhalt kategorisieren. Dabei wird berücksichtigt, welchen Effekt die Sites auf ein typisches 12-jähriges Kind haben könnte, welches das Internet ohne elterliche Kontrolle durchstöbert. [CyberPatrol2000]

Die CyberYES Liste dagegen empfiehlt schüler-freundliche, bildende Sites, geeignet für 6- bis 16-jährige. Die Sites werden von einem Team von Eltern und Lehrern ausgewählt. Um die Qualität zu sichern, werden die Ressourcen mindestens alle 60 Tage besucht. Alle Kategorien sind in Kapitel C.2 aufgezählt. [CyberPatrol2000]

4.1.3 RSACi

Der *Recreational Software Advisory Council* (RSAC, Beratungsgremium für Unterhaltungssoftware), eine unabhängige, gemeinnützige Selbstkontrollinstitution für die Bewertung von Computerspielen, schuf 1986 eine neue Abteilung für Bewertungen von Internetinhalten: Der RSACi (i steht für Internet) ist ein objektives, PICS-kompatibles, Inhalte beschreibendes Beratungssystem. Dabei werden die PICS Label direkt auf der bewerteten Website angebracht, da das RSACi kein eigenes Label-Büro unterhält. [Dyson97]

Praktisch funktioniert die Selbsteinstufung so, dass der Inhaber einer Website auf der Homepage von RSACi⁵ einen Katalog mit ja/nein-Fragen beantwortet (z.B.: „Fließt Blut?“), und Einstufungen zwischen 0 und 4 in vier Kategorien (Sex, Nacktaufnahmen, Sprache und Gewalt) vornimmt (vgl. Tabelle 4.1 und Abbildung 4.2). Im Anschluss erhält er die PICS Label und ein Symbol zum Plazieren auf seiner Homepage; zuvor muß der Inhaber der Website die Richtigkeit seiner Angaben bei dem Dienst bestätigen. [Dyson97]

⁴Cyber Patrol Sales, The Learning Company, One Athenaeum Street, Cambridge, MA 02142, USA. <http://www.cyberpatrol.com>

⁵<http://www.rsac.org>

Level	Violence	Language
4	Rape, gratuitous violence	Vulgar, extreme hate speech
3	Aggressive violence	Strong language or hate speech
2	Destruction of realistic objects	Moderate profanity
1	Injury to humans	Moderate expletives or profanity
0	None of the above or sports related	None

Tabelle 4.1: Beispielhafte Auswahl von zwei der vier Kategorien, nach denen Webseiten bewertet werden können. Die beiden anderen Kategorien sind Darstellungen von Sex und Nacktaufnahmen. Man erkennt, dass die Eigenschaften nur einen von fünf Zuständen annehmen können. [RSAC2000]

4.1.4 INCORE

INCORE⁶ steht für *Internet Content Rating for Europe* und ist ein von der Europäischen Union finanziertes Projekt, das ebenfalls Kinder vor Internet-Inhalten schützen soll, die schädlich für sie sein könnten. Beim Aufbau dieses Systems hat man sich dazu entschlossen, sich an Prinzipien anzulehnen, welche von der Internet Content Rating Alliance (ICRA) entwickelt wurden, einer Gruppe Europäischer, Nord-Amerikanischer und Australischer Organisationen. Dazu zählen der Schutz der freien Meinungsäußerung, Objektivität (im Sinne der Unabhängigkeit vom kulturellen Kontext), Benutzer- und Provider-Freundlichkeit, Qualitätskontrollen, ein System von Kategorien und Levels sowie ein gewisses Maß an Anpassungsfähigkeit. Im Zuge der Beratungen, die zur Zeit noch andauern, sollen die „Inhaltsarten ermittelt werden, bei denen die europäischen Verbraucher die meisten Bedenken haben, und die sie gerne herausfiltern möchten.“ Nicht alle Menschen haben bei jeder Kategorie Bedenken, genauso wie jede einzelne Person verschieden hohe Toleranzschwellen pro Kategorien besitzt. [Incore99]

Aus diesem Grund ist es notwendig, das Blockieren von Inhalten den einzelnen Benutzern zu überlassen, und Techniken des Filterns (vgl. Kapitel 2.3.2 'Downstream Filtering') anzuwenden. Dazu benötigt man Vereinbarungen über Metadaten, wie jene des RSACi.

4.2 Zusammenfassung

Das World Wide Web ist der am einfachsten zu navigierende und von Kindern am häufigsten benutzte Teil des Internet. Und Firmen haben zunehmend die Hoffnung, diesen Teil in einen globalen Online-Marktplatz zu verwandeln. Auch wenn die tägliche Erfahrung zeigt, dass das Web nicht von Pornographie überschwemmt

⁶<http://www.incore.org>

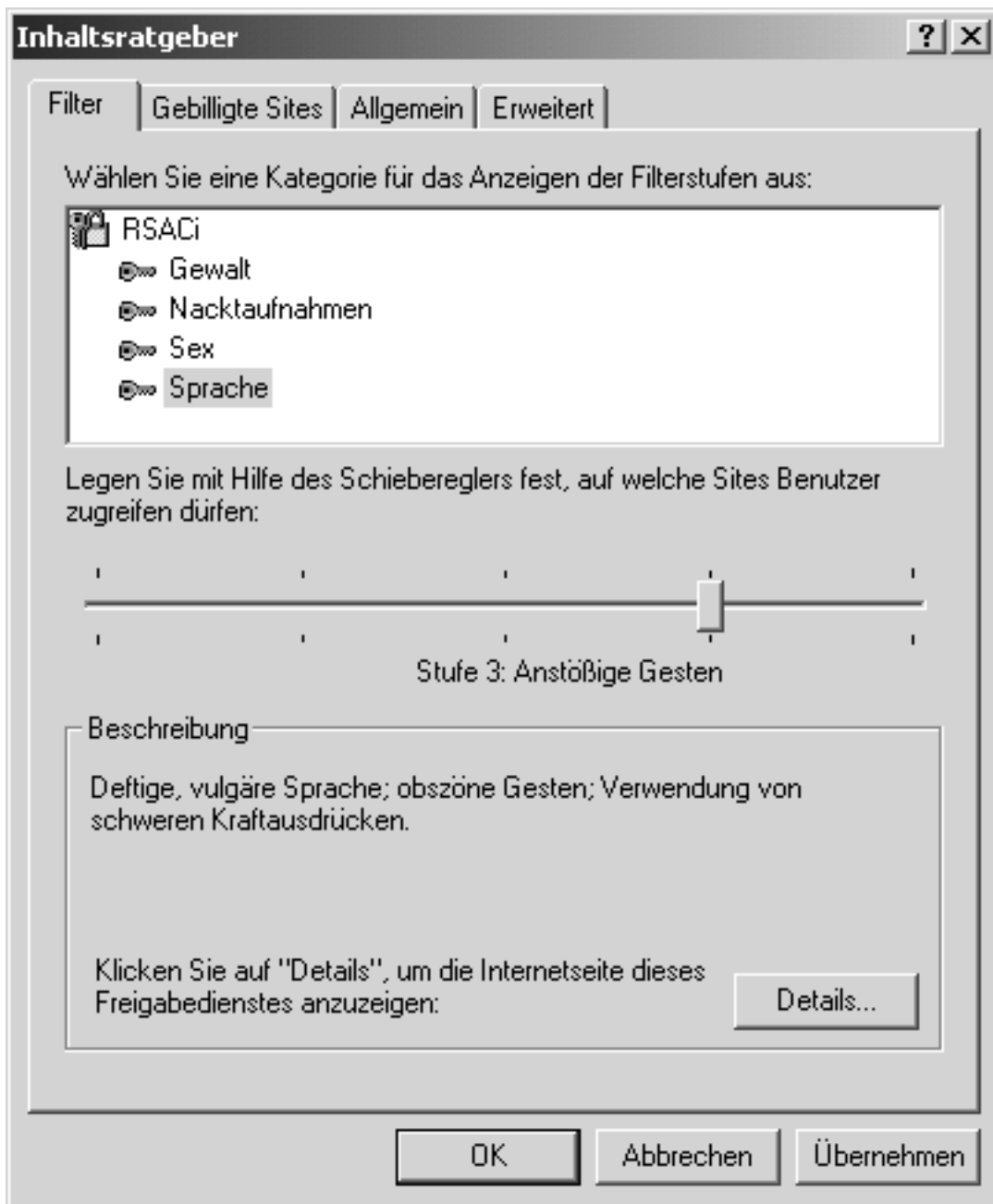


Abbildung 4.2: Im Webbrowser „Microsoft Internet Explorer 5“ lassen sich mit Hilfe des Inhaltsratgebers RSACi Label festlegen. Man kann Einstellungen in den Kategorien Gewalt, Nacktaufnahmen, Sex und Sprache tätigen. Das Beispiel zeigt, dass der Benutzer in Bezug auf die Sprache die Stufe 3 wählt. Man lässt damit eine deftige, vulgäre Sprache, obszöne Gesten, und die Verwendung schwerer Kraftausdrücke zu.

wird, und viele dieser Seiten ohnehin die Angabe einer Kreditkarten-Nummer verlangen, ist ein Schutz der Kinder vor schädlichem Inhalt nötig, ohne dabei Inhalte zensurieren zu müssen. Ein Plan in dieser Richtung kommt vom W3 Consortium, einer Organisation die vom Erfinder des Web, Tim Berners-Lee, geführt wird. Dieser Plan sieht vor, den Protokollen, die dem Web zugrunde liegen, eine neue Schicht hinzuzufügen: Ein Inhalts Bewertungs Feld (*content ratings field*), das es sowohl den Informations-Providern erlaubt, die publizierten Inhalte zu bewerten, als auch den Konsumenten die Möglichkeit gibt, Inhalte mit speziellen Bewertungen auszublenden. [Rosenberg95]

Zur Zeit erlaubt PICS die Bewertung von Ressourcen. Es ist wert-neutral und erlaubt jede Menge von Dimensionen und Kriterien der Bewertung. PICS-kompatible Software kann die Labels von jeder beliebigen Quelle interpretieren, weil jede Quelle eine maschinen-lesbare Beschreibung der Label-Dimensionen mit-schickt. [RM96] Der Internet Explorer von Microsoft hat in der Version 5 standardmäßig das Vokabular des Blockingsystems RSACi integriert, und kann weitere PICS-kompatible Systeme aufnehmen und verarbeiten.

Wie bereits in Abschnitt 3.2.2 erwähnt, sind die Mechanismen um Zutritt zu beschränken und zu erreichen zwei Seiten derselben Münze: Anforderungen, den Zugang zu allen Seiten, welche mit einem speziellen Inhalt zu tun haben, zu unterbinden sind sehr ähnlich mit Anforderungen, Seiten mit demselben Inhalt zu finden [Armstrong97]. In Kapitel 6 wird das Suchen und Auffinden von Inhalten untersucht, die den Konsumenten interessieren. Auch für diesen Vorgang ist der Einsatz von Metadaten, die Ressourcen in einer computer-interpretierbaren Form bewerten, hilfreich. PICS-ähnliche Metadaten würden es einem Anwender erlauben, die für ihn relevanten Maßstäbe und Kriterien an Qualität selbst fest-zulegen.

Das im vorigen Kapitel gezeigte Metadaten Schema Dublin Core (vgl. Abschnitt 3.3.3) ermöglicht die Beschreibung von Dokumenten im Internet. Die hier vorgestellten Systeme erlauben hingegen eine Bewertung von Ressourcen mit Hilfe computer-interpretierbarer Attribute. Wenn es nun gelingt, ein Metadaten Schema zu entwickeln, welches Ressourcen beschreibt und gleichzeitig die inhaltliche Qualität bewertet (mit anderen als den bei Blocking Systemen zum Einsatz kommenden Attributen wie Gewalt, Sex etc.), und dieses Schema in Verbindung mit einem Bewertungssystem bzw. einem Suchdienst zu bringen, so erhält man die Möglichkeit, nach qualitativer Information zu suchen. Möglichst viele der Attribute des Metadaten Schemas dürften dabei nur einen eingeschränkten Wertevorrat besitzen, damit sie computer-interpretierbar sind. Im Gestaltungsbereich der Arbeit wird versucht, ein solches System zu entwickeln.

Kapitel 5

Empfehlungs- und Annotationssysteme

Im vorigen Kapitel wurde gezeigt, wie sich Metadaten zum Blockieren von Inhalten im Internet eignen. In diesem Kapitel wird die Implementierung von Systemen zur Weitergabe von Empfehlungen diskutiert. Auch Bemerkungen und Notizen lassen sich durch sogenannte Annotationssysteme mit Hilfe von Metadaten an andere Benutzer weitergeben.

Wenn wir Entscheidungen ohne ausreichendes Vorwissen oder Erfahrungen zu treffen haben, ist es oft klug, sich so zu entscheiden, wie es andere, ähnlich handelnde Personen zuvor erfolgreich gemacht haben. Man kann die Erfahrungen anderer Leute zum Filtern und Leiten nutzen: Zum Filtern, um möglichen schlechten Entscheidungen auszuweichen, und zum Leiten, um auf eine mögliche gute Wahl hingewiesen zu werden. Im Alltag erfolgt dieser Prozess in Form von Gesprächen, Empfehlungsschreiben, Kino- und Buchrezensionen in Zeitungen oder Restaurantführern. Die gebräulichen Mensch-Computer Schnittstellen¹ ignorieren oft diese Macht von sozialen Strategien. Empfehlungssysteme (vgl. Abschnitt 5.1) sollen diesen natürlichen, sozialen Prozess jedoch unterstützen und verbessern. In einem typischen solchen System stellen Menschen Empfehlungen bereit, die das System dann sammelt und an passende Empfänger weiterleitet. [HSRF95] [RV97]

Wenn man den Aspekt des Filterns weglässt, erhält man ein sogenanntes Annotationssystem. Hierbei ist es möglich, Kommentare und kurze Charakterisierungen zu Texten oder anderen Stellen in elektronischen Ressourcen zu schreiben, die dann andere Mitglieder des Systems lesen können, sobald sie die betreffende Stelle anschauen. Die Systeme werden in Abschnitt 5.2 behandelt.

¹HCI human computer interface

5.1 Empfehlungssysteme

Die Entwickler des ersten Empfehlungssystems, Tapestry, haben den Begriff „collaborative filtering“ geprägt, der von vielen Nachfolgern aufgenommen wurde. In [RV97] wird jedoch der allgemeinere Begriff *recommender systems* (Empfehlungssysteme) favorisiert. Einerseits weil die Empfehlenden nicht immer einer Gemeinschaft angehören, und sich nicht einmal kennen müssen. Andererseits weil Empfehlungssysteme nicht nur interessante Ressourcen vorschlagen können, sondern auch jene anzeigen können, die ausgefiltert werden sollen. [RV97]

In [BS97] werden im Wesentlichen zwei Methoden von Empfehlungssystemen unterschieden: Beim reinen inhalts-basierten Ansatz (Vgl. Abschnitt 5.1.1 *Content-based recommendation*) beruhen Empfehlungen alleine auf einem Benutzerprofil, welches sich aus der Analyse von Texten, die der Benutzer in der Vergangenheit bewertet hat, ergibt. Beim reinen gemeinschafts-basierten Empfehlungssystem (vgl. Abschnitt 5.1.2 *Collaborative recommendation*) werden die Texte überhaupt nicht analysiert. Die Empfehlungen basieren alleine auf der Basis von Ähnlichkeiten zu anderen Benutzern. Neben den beiden reinen Ansätzen, gibt es auch noch Hybrid Systeme.

In Abschnitt 5.1.3 folgt eine Abhandlung darüber, welche sozialen Probleme solche Empfehlungssysteme aufwerfen können, hauptsächlich in Bezug auf den Schutz der Privatsphäre.

5.1.1 Content-based recommendation

Der Inhalts-basierte Ansatz von Empfehlungen (*content-based recommendation*) hat seine Wurzeln im *Information Retrieval (IR)*². Textdokumente werden auf Grund von Vergleichen zwischen deren Inhalten und Benutzerprofilen empfohlen. Dabei werden die Texte oft mit einem Gewichtungsschema analysiert, indem entscheidenden (*discriminating*) Wörtern hohe Gewichtungen gegeben werden. Beispielsweise könnten die drei höchst bewerteten Wörter dieses Kapitels lauten: „recommendation“ (0.33), „basieren“ (0.27) und „Benutzer“ (0.18). Wenn eine Seite für einen Benutzer ausgewählt wurde, wird sie ihm gezeigt, und er sollte sie bewerten. Wenn dem Benutzer der Text gefiel, können die Gewichtungen der Wörter, welche aus dem Text extrahiert wurden, zu den Gewichtungen derselben Wörter im Profil des Benutzers addiert werden. Dieser Prozess nennt sich *Relevance Feedback*. Obwohl viele verschiedene Methoden existieren, die Gewichtungen zu berechnen, oder das Benutzerprofil zu aktualisieren, birgt der inhalts-basierte Ansatz gewisse Nachteile. [BS97]

²In EDV: das Suchen u. Auffinden gespeicherter Daten in einer Datenbank. [Duden]

Ein Nachteil dieses Ansatzes liegt darin, dass die genannten IR-Techniken nur auf gewisse Typen von Ressourcen anwendbar sind. Beispielsweise können bei Webseiten ästhetische Aspekte, die meisten Multimedia Dateien und Netzwerkfaktoren (z.B. die Zeit, welche zum Laden der Webseite benötigt wird) nicht berücksichtigt werden. Ein zweites Problem betrifft die Über-Spezialisierung. Wenn das System nur Ressourcen empfehlen kann, die eine hohe Übereinstimmung mit dem Profil des Benutzers besitzen, bekommt dieser nur Texte vorgeschlagen, die ähnlich zu bereits bewerteten Texten sind. Abhilfe würde hier das Hinzufügen von Zufälligkeit schaffen, beispielsweise durch Mutations-Operationen (genetischer Algorithmus). Drittens sind Empfehlungen bei diesem Ansatz immer sprach-abhängig. Und das vierte Problem liegt darin, dass der Benutzer dazu gebracht werden muss, Bewertungen zu jedem der gelesenen Texte abzugeben, was bald lästig werden kann. Ein Verringern der Bewertungen führt automatisch zu einer verringerten Anzahl guter Empfehlungen. [BS97]

Diese Nachteile entfallen bei gemeinschafts-basierten Systemen, weil dort nicht Ressourcen vorgeschlagen werden, die ähnlich jenen sind, die einem Benutzer in der Vergangenheit gefallen haben. Stattdessen werden Ressourcen empfohlen, die anderen Benutzern mit ähnlichen Interessen gefallen haben.

5.1.2 Collaborative recommendation

Wie bereits im vorigen Abschnitt erwähnt, werden bei gemeinschafts-basierten Empfehlungssystemen (*collaborative recommendation*) Ressourcen empfohlen, die anderen Benutzern mit ähnlichen Interessen und Geschmack gefallen haben. Es wird also nicht die Ähnlichkeit von Ressourcen berechnet, sondern die Ähnlichkeit von Benutzerprofilen. Typischerweise werden für jeden Benutzer eine Menge von „nächsten Nachbarn“ gefunden, mit deren vergangenen Bewertungen die größten Korrelationen bestehen. Die geschätzten Werte für einen ungelesenen Text werden basierend auf den Werten von den nächsten Nachbarn ermittelt. Mit Hilfe dieser Methode werden die Nachteile der inhalts-basierten Systeme überwunden. Weil die Empfehlungen anderer Benutzer verwendet werden, kann jede Art von Ressource behandelt werden, und es können Dinge empfohlen werden, die anders sind als jene, die der Benutzer in der Vergangenheit bereits gesehen hat. Weil die Empfehlungen aller Benutzer verwendet werden, kann die Leistung des Systems aufrecht bleiben, auch wenn die einzelnen Benutzer weniger Bewertungen abgeben. Es entstehen mit der Methode jedoch zwei andere Probleme: Wenn eine neue Ressource in die Datenbank eingefügt wird, kann diese solange nicht empfohlen werden, bis ein Benutzer sie bewertet hat, oder er festlegt, mit welcher anderen Ressource eine Ähnlichkeit besteht. Um ein gut funktionierendes System zu erhalten, darf also die Anzahl der Benutzer relativ zur Anzahl der Ressourcen nicht zu klein sein. Das zweite Problem betrifft Benutzer, deren Geschmack verglichen

mit der restlichen Benutzergruppe ungewöhnlich ist. Diese werden dann nur sehr wenige Empfehlungen erhalten. [BS97] [ME95]

5.1.3 Soziale Folgen

Neben der Frage nach dem technisch besten System ergeben sich prinzipielle soziale Folgerungen und Fragen. Wie aus den vorangegangenen Abschnitten hervorging, hängt die Leistung von Empfehlungssystemen mit der Anzahl geleisteter Bewertungen der Benutzer zusammen. Sobald ein Benutzer ein Profil seiner Interessen eingerichtet hat, könnte er sich zurücklehnen, und nur mehr die Empfehlungen anderer konsumieren, ohne selbst Bewertungen zur Verfügung zu stellen. Auch die Beobachtung von Benutzerverhalten oder die implizite Bewertung von Ressourcen können in so einem Fall ein Absinken der Systemleistung nicht verhindern. Zukünftige Empfehlungssysteme werden wahrscheinlich ein Belohnungssystem einführen müssen. Entweder in Form von Bezahlung für geleistete Bewertungen, oder dadurch, dass man nur im Gegenzug für Bewertungen Empfehlungen bekommt. [RV97]

Wenn jedermann Bewertungen bereitstellen kann, ergibt sich ein weiteres Problem: Autoren könnten haufenweise positive Empfehlungen für ihre eigenen Texte generieren, und die der anderen negativ bewerten. Ein drittes Problem betrifft die mögliche Gefährdung der Privatsphäre. Denn die Menschen möchten üblicherweise nicht, dass ihre Gewohnheiten und Ansichten allgemein bekannt sind. Einige Empfehlungssysteme erlauben die Teilnahme anonym oder unter Verwendung eines Pseudonyms. Doch auch dies ist noch nicht die Lösung des Problems, nachdem es Menschen gibt, die eine Zwischenlösung wünschen. Nämlich eine, die einerseits den Schutz der Privatsphäre und andererseits Anerkennung ihrer Bemühungen bietet. [RV97]

Wie bereits erwähnt, verlassen wir uns im Alltagsleben häufig auf die Empfehlungen anderer. Dabei sind aber die Quellen der Information wesentlich konstanter als im Internet. Aber was sind die Faktoren, die uns auf Empfehlungen vertrauen lassen, und woher bekommen wir sie üblicherweise? Freunde und Bekannte sind häufig Quellen von Empfehlungen. Wenn es um Geld geht, werden viele Menschen den Empfehlungen der Bank oder eines Investment Beraters vertrauen. Für medizinische Ratschläge konsultieren wir einen Arzt oder gehen in ein Krankenhaus. Jeder baut sich ein soziales Netzwerk auf, und bezieht Empfehlungen aus unterschiedlichsten Quellen. [Horwath99]

Oft verlassen wir uns auf eine gut bekannte Quelle der Information, wie z.B. etablierte Zeitungen, Journale oder Publikationsreihen. Wir tendieren dazu, solchen Quellen und deren Weg, Informationen zu filtern, zu vertrauen. Ein wenig rationeller – aber umso menschlicherer Weg – Quellen der Information

auszusuchen, ist es, nach dem Erscheinungsbild zu urteilen. Anzeichen, ob man den Empfehlungen einer Person traut, sind häufig, ob diese einen Anzug, einen weißen Kittel oder kaputte Jeans trägt. [Horwath99]

Diese sozialen Mechanismen fallen bei der Verwendung von Empfehlungssystemen weg. Es ist beispielsweise auch nicht ohne weiteres möglich, den medizinischen Ratschlägen eines „Doktor Sonundso“ zu vertrauen, nur weil er auf dem Foto in seiner Homepage wie ein Arzt aussieht. Denn im Internet kann prinzipiell jeder unter jedem Namen publizieren. Dies ist jedoch eines der grundsätzlichen Probleme des Internet, und wird auch in Kapitel 2.1 behandelt. Abhilfe könnte hier beispielsweise die Verwendung elektronischer Unterschriften schaffen.

5.2 Annotationssysteme

Im Unterschied zu Empfehlungssystemen werden bei reinen Annotationssystemen andere Benutzer nicht explizit auf die Bemerkungen und Charakterisierungen hingewiesen. Man kann sich Annotationen wie „Post-it®“-Notizen auf Dokumenten vorstellen. Jeder Benutzer kann Gelesenes annotieren, und seinen Kommentar oder seine Bewertung hinterlassen. Dies kann öffentlich geschehen, sodass andere Benutzer des Systems – wenn sie das Dokument betrachten – diese Annotationen lesen können, oder privat.

In [RMW95] werden verschiedene Gründe aufgezählt, warum Menschen über Netzwerk Ressourcen miteinander kommunizieren wollen. Das sind unter anderem Kommentare und Annotationen, die eine Arbeitsgruppe über Dokumente von gemeinsamen Interesse teilt. Weiters schaffen Annotationsysteme die Möglichkeit, Newsgruppen-ähnliche Foren über spezielle Artikel im Netz zu betreiben, Rezensionen anzubringen oder Benutzungsrichtlinien zu publizieren. Eine weitere Form von Annotationen, die sogenannten *Seals of Approval* (SOAPs), können nicht von jedermann angebracht werden. Spezielle Organisationen und Serviceeinrichtungen können diese vergeben, um Inhaltsbewertungen zu publizieren. Mit einer entsprechenden Client-Software können sie für ein Blockingsystem (vgl. Kapitel 4) benutzt werden.

Hat man zusätzlich die Möglichkeit, innerhalb von Annotationen Verweise in Form von Links auf andere Ressourcen zu plazieren, dann ergeben sich noch weitere Möglichkeiten: Auf diesem Weg kann man andere Benutzer auf interessante Inhalte, die den Text betreffen, den derjenige gerade liest, hinweisen. Schon 1945 hatte sich Vannevar Bush eines solchen Systems vorgestellt. [RMW95]

Im nächsten Abschnitt wird erklärt wie die Architektur eines Annotationssystems unter Verwendung von Metadaten aussehen könnte. Im darauffolgenden

Abschnitt wird als Beispiel für das Interface eines Annotationssystems jenes von Hyperwave vorgestellt.

5.2.1 Architektur eines Annotationssystems unter Verwendung von Metadaten

Die Architektur eines Annotationssystems soll beispielhaft an der „ComMentor“-Plattform³ erläutert werden, welche es Dritten erlaubt, Annotationen und Kommentare zu Ressourcen zu geben.

Der Benutzer interagiert mit einem Browser um sich Dokumente von mehreren Dokumentservern anzusehen. Zusätzlich gibt es Metainformations-Server von denen Annotationen zu den Dokumenten automatisch geladen werden. Die Annotationen sind in „Mengen“ organisiert, zu denen Mitglieder von „Gruppen“ Zutritt haben. Abhängig von der Kontext-Menge des Benutzers kann der Browser entscheiden, von welchem Metainformations-Server Annotationen besorgt werden sollen, welche dann an der richtigen Stelle des originalen Dokuments angezeigt werden (Vgl. Abbildung 5.1). [RMW95]

Die Metainformationen werden einheitlich und maschinen-verstehbar gespeichert. Dazu wurde im Rahmen des Stanford Integrated Digital Library Project⁴ eine eigene Objektbeschreibungs-Sprache entwickelt, PRDM (*Partial Redundant Descriptive Meta-language*). Die Metainformationen werden zum Browser als Teil einer MIME Nachricht des neuen Types „application/x-PRDM“⁵ geschickt. [RMW95]

Im nächsten Abschnitt wird gezeigt, wie das Interface eines Annotationssystems aussehen könnte. Es handelt sich dabei um eine Funktion des Hyperwave Information Server, der es erlaubt, Bemerkungen zu jedem Objekt auf dem Server in mehreren Sprachen zu tätigen.

³ComMentor ist der Name eines Prototypen für ein Annotationssystem im World Wide Web. Technisch gesehen ist es ein NCSA Mosaic Browser und ein NCSA HTTP Server erweitert um ein paar Serverskripts. [RMW95]

⁴Garcia-Molina, Hector; Shoam, Yoav; Winograd, Terry: Stanford Integrated Digital Library Project. Computer Science Department, Stanford University. 1994. <http://www-diglib.stanford.edu/diglib>

⁵Borenstein, Nathaniel; Freed, N.: Multipurpose Internet Mail Extensions. Draft, Internet Engineering Task Force. 1993.

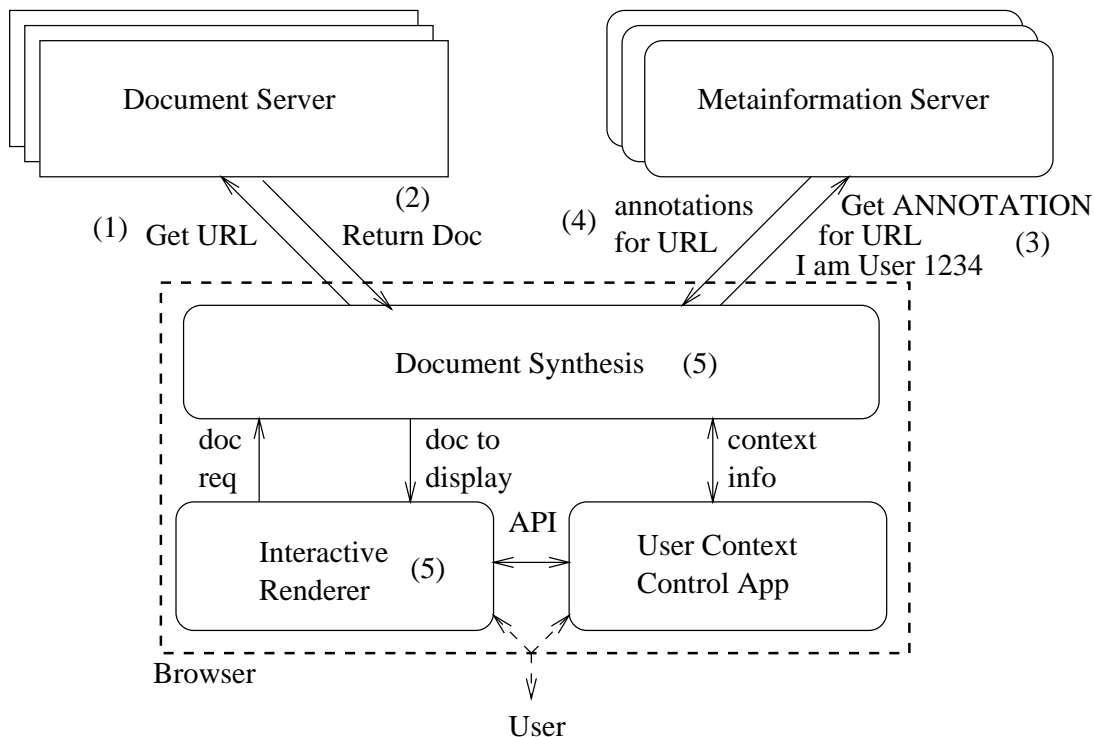


Abbildung 5.1: Das Original Dokument wird über die normale Browser Interaktion geladen (1+2). Daraufhin sendet der Browser eine Anfrage nach Annotationen an einen Metainformations-Server mit der URL und der Benutzerkennung (3). Der Server schickt die Annotationen zurück (4), welche mit dem ursprünglichen Dokument verschmolzen (5) und angezeigt werden. [RMW95]

5.2.2 Das Interface des Annotationssystems von Hyperwave

Als ein Beispiel für ein Annotationssystem sei der Hyperwave⁶ Information Server erwähnt, eine flexible „Entwicklungsplattform, mit der verteilte Organisationen ihre Intranet/Extranet- sowie Wissensmanagement-Lösung realisieren“ können. Die Navigation in den großen Informationspools erfolgt mit Hilfe inhaltsbezogener Hierarchien und der Verarbeitung von Metadaten. Dadurch lassen sich Informationen auf einfache Weise kategorisieren und visualisieren. Mit Hilfe der Volltext- und Index-Suchfunktionen können Recherche-Ergebnisse rasch und effizient erzielt werden [Hyperwave]

Das System erlaubt es unter Anderem, Annotationen an jedes Hyperwave-Objekt zu hängen. Man braucht dazu nur auf das „Annotate“-Icon zu klicken, und kann die Kommentare (auch als HTML Text) einfügen (Abbildung 5.2 zeigt

⁶<http://www.hyperwave.com>

die Annotationsfunktion des Gentle WBT Systems, welches auf Hyperwave basiert). Seit Version 2.5 ist es auch möglich, Textstellen zu annotieren, indem man die entsprechende Passage einfach markiert, und dann auf das „Annotate“-Icon klickt. Über Attribute kann man zusätzlich die Sprache der Bemerkung angeben, einen Titel eingeben und festlegen, ob sie öffentlich oder privat sein soll. [Hyperwave]

Man sieht, dass man mit einfach zu bedienenden und intuitiv verständlichen Interfaces ein relativ komplexes Annotationssystem steuern kann. Der Benutzer kann so mit nur sehr geringem Aufwand anderen Benutzern des Systems seine Meinung über Textstellen mitteilen, und – wenn er will – die Kommentare anderer lesen und wiederum annotieren.

5.3 Zusammenfassung

Im vorigen Kapitel 4 wurde gezeigt, wie Metadaten dazu eingesetzt werden können, Kinder vor dem Zugang zu schädlichen Inhalten zu schützen – ohne dabei die Inhalte zensurieren zu müssen. In diesem Kapitel wurde gezeigt, wie Metadaten dazu eingesetzt werden können, „soziales Verhalten“ nachzubilden: Sie ermöglichen es, Empfehlungen und Ratschläge zu geben und zu bekommen. Sie machen es weiters möglich, jede Stelle in Online-Ressourcen zu annotieren, also zu charakterisieren. Bei dem Projekt „ComMentor“ wurde zu diesem Zweck eine eigene Metadatensprache, PRDM, entwickelt. Sie ähnelt jedoch sehr PICS oder RDF (vgl. Abschnitte 3.2.2 und 3.2.3). Es stellt sich überhaupt heraus, dass die Techniken des Blockierens bzw. Empfehlens/Annotierens gar nicht so unterschiedlich sind.

Auch der Unterschied zwischen *content-based* und *collaborative filtering* wurde in diesem Kapitel erläutert. Der erste Ansatz hat seine Wurzeln im Information Retrieval, und basiert auf dem Vergleich von Texten untereinander. Es wird versucht, anhand von Worthäufigkeiten Ähnlichkeiten festzustellen. Diese Technik wird heute auch von den meisten Suchmaschinen verwendet. Beim collaborative filtering hingegen werden die Ähnlichkeiten von Benutzerprofilen untersucht.

Annotationssysteme können die Qualität von Dokumenten erhöhen, indem von Benutzern des Systems Informationen hinzugefügt werden, und Kritik geäußert wird. Empfehlungssysteme sind ein guter Schritt in Richtung leichteres Auffinden von passenden Informationen, indem man als Benutzer auf Inhalte hingewiesen wird, von denen das System „weiß“, dass sie einen interessieren.

Man kann den Prozess des aktiven Suchens jedoch nicht automatisieren, weil die Informationen über die Qualität des Inhaltes unstrukturiert und nicht computer-interpretierbar gespeichert werden. Dies kann nur gelingen, wenn die

Write a Note - GENTLE-WBT - Microsoft Internet Explorer

Write a Note

Title:
Ich bin anderer Meinung

Content:
Hier kann ich eine Notiz einfügen

Rights :
Public note

Type :
 Remark Question Agree
 Answer Disagree

Attachments:
[Empty box]

V 1.0 Copyright © 1998-2000 by Hyperwave Information Management AG i.G. All rights reserved.

Internet

Abbildung 5.2: Der Annotations Wizard einer WBT-Umgebung (<http://wbt-2.iicm.edu/hts>). Jeder Kommentar erhält einen Titel, der auch erscheint, wenn man mit der Maus auf die annotierte Textstelle zeigt. Weiters kann man Attachments anhängen, und ein Symbol wählen, welches den Kommentar charakterisiert (Zustimmung, Ablehnung, Frage, Ausruf und Bemerkung).

Bewertungen bzw. Beschreibungen anhand vorher festgelegter Kriterien erfolgt. Diese Kriterien müssen z.T. auch noch die Forderung nach einem eingeschränkten Wertevorrat erfüllen, damit sie computer-interpretierbar sind. Im nächsten Kapitel werden Systeme untersucht, welche die gezielte Suche nach gewünschten Inhalten ermöglichen. Durch den Einsatz von Metadaten soll die Suche nach qualitativ hochwertigen Inhalten verbessert werden.

Kapitel 6

Suchen durch Einsatz von Metadaten

In der Einleitung von Kapitel 1 wurde erläutert, wie schnell das Internet wächst, und wie es immer schwieriger wird, passende Inhalte zu finden. Was für den einen jedoch passend erscheint, mag für jemand anderen wenig ausreichend erscheinen. In Kapitel 2 wurde daher der Versuch einer Definition von inhaltlicher Qualität von Online-Ressourcen unternommen, wobei verschiedene Kriterien ausgemacht wurden, die sie beschreiben. Für die Bewertung und Beschreibung von Online-Ressourcen nach diesen Kriterien bietet sich der Einsatz von Metadaten (vgl. Kapitel 3) an, welche sich im Internet beispielsweise durch das Resource Description Format (RDF) beschreiben lassen. Die Kapitel 4 und 5 haben gezeigt, wie man mit Hilfe von Metadaten unpassende Inhalte blockieren und interessante Inhalte empfehlen kann. In diesem Kapitel wird erläutert, welche Möglichkeiten Metadaten beim Einsatz in Suchmaschinen bieten könnten: Ziel soll es sein, jene Dokumente aus dem Internet herauszufiltern, die das gesuchte Thema behandeln und die (Qualitäts-)Erfordernisse der Benutzer erfüllen.

Der Einsatz von Metadaten in Suchmaschinen ist natürlich nur sinnvoll, wenn Such-Roboter diese auch indizieren können. Viele Suchmaschinen haben die Wichtigkeit von Metadaten mittlererweile erkannt, und können zumindest schon einige META-Tags auswerten. Meist sind es die beiden Felder *description* und *keywords*, nicht aber viele der anderen Elemente des Dublin Core (vgl. Abschnitt 3.3.3). [AT99]

Nur wenige stellen den Nutzen von Metadaten bei der Verbesserung von Suchen im Internet an sich in Frage. Einige denken, dass das Indizieren des gesamten Texts (*full text indexing*) informativ genug wäre. Dabei versuchen Suchmaschinen mit Hilfe umfangreicher Algorithmen die relevanten Wörter aus den Dokumenten herauszufinden. Weil diese Arbeit zwar schneller als von Menschen, dafür aber

auch weniger genau ausgeführt wird, kommt es häufig zu einer Flut von Suchergebnissen. Durch den Einsatz von META-Tags könnten die Inhalte in HTML Dokumenten beschrieben werden, und damit die Suche nach spezifischen Feldern ermöglicht werden. Dies erfordert jedoch auch ein komplizierteres Indizieren, und verlangt von den Suchmaschinen größere Ressourcen. [AT99]

Sobald Suchmaschinen Bild-, Ton- oder andere Multimedia Dateien zu indizieren beginnen, wird der Bedarf von Metadaten offensichtlich. Es ist zur Zeit nämlich nur sehr eingeschränkt möglich, Multimedia Files ohne zusätzliche Informationen in den, üblicherweise von Suchmaschinen benutzten, textorientierten Datenbanken zu indizieren. Lycos z.B. versucht Ton- und Bilddateien nur anhand des Dateinamens und des Linktexts, welcher zur Datei führt, zu erkennen. [AT99]

In den folgenden Abschnitten werden drei Prinzipien von Suchdiensten erläutert, nämlich die Index Suchmaschinen, die Metasucher (vgl. Abschnitte 6.1 bis 6.2) und die Katalogdienste (vgl. Abschnitt 6.3). Danach folgt in Abschnitt 6.4 die Beschreibung des xFIND¹ Systems, mit dem es unter anderem möglich sein soll, mit Hilfe von Metadaten Suchergebnisse zu verbessern (vgl. Kapitel 6.5).

6.1 Index Suchmaschinen

Index Suchmaschinen wie Altavista², Alltheweb³, Lycos⁴ oder Infoseek⁵ analysieren Dokumente im Internet automatisch mit Hilfe sogenannter Robots und Gatherer. Typischerweise liest die Maschine zumindest die ersten paar hundert Worte einer Seite, einschließlich des Titels, der Bildunterschriften und anderer Textvorkommen. Dann wird versucht, aufgrund von Worthäufigkeiten oder der Platzierung eines Wortes innerhalb des Texts seine Bedeutung (Relevanz) festzustellen. In einem Index werden anschließend zu jedem Wort jene Internet Adressen gespeichert, in denen das Wort vorkommt, und es wird auch notiert, als wie relevant ein Wort innerhalb der Seite erkannt wurde. In einem Suchergebnis werden dann jene Seiten mit den höchsten Relevanzfaktoren sortiert und aufgelistet. [Legenstein99]

Der Vorteil von Index Suchmaschinen liegt in der großen Menge an indizierten Dokumenten, und darin, dass auch neue Seiten zum Teil relativ rasch erfasst werden. Der große Nachteil liegt in den oft unbefriedigenden – weil oft zu umfangreichen – Suchergebnissen, innerhalb derer man erst wieder das passende Dokument finden muss. [Legenstein99]

¹Extended Framework for Information Discovery, <http://xfind.iicm.edu>

²<http://av.com>

³<http://www.alltheweb.com>

⁴<http://www.de.lycos.de/>

⁵<http://infoseek.go.com/>

6.2 Metasucher

Einzelne Suchmaschinen können jeweils nur einen Bruchteil der Dokumente im Internet abdecken⁶. Es liegt daher nahe, gleichzeitig mehrere Suchdienste in Anspruch zu nehmen. Genau diese Technik verwenden die sog. Metasuchdienste⁷ (andere Bezeichnungen sind Meta-Maschine, MetaCrawler, MultiSearcher oder ParallelSearcher), die mehrere einfache Suchdienste parallel abfragen, und die Ergebnisse sinnvoll aufbereiten. Im Unterschied zu den normalen Suchdiensten, auf denen diese zugreifen, können Metasuchdienste nicht direkt auf die Menge der Webinhalte zugreifen, stattdessen aber auf eine Menge von Suchdiensten. [Sander98]

Metasucher können daher auch nicht die in Dokumenten enthaltenen Metadaten abrufen. Diese stehen nur dann zur Verfügung, wenn die Suchmaschine sie indiziert hat. Nicht zu verwechseln sind Metasucher übrigens mit einfachen All-in-one Formularen. Dahinter verbergen sich einfache Eingabehilfen, die mehrere Suchdienste nacheinander über eine einheitliche Eingabemaske abfragen. Sie können aus einfachen CGI Programmen bestehen. [Sander98]

6.3 Katalogdienste

Maschinen wie Yahoo!⁸ sind die Kataloge des WWW: Sie ordnen die Seiten speziellen Themengebieten zu. Üblicherweise ist dabei menschliches Urteil gefragt. Bei manchen kategorisieren Angestellte; andere erlauben den Seiteninhabern eine Zuordnung ihrer eigenen Seiten; und wiederum andere befragen zufällige Site-Besucher nach ihrem Urteil. [Langa98]

Ein weiterer bekannter Vertreter von Katalogdiensten ist DMOZ⁹. Der Ansatz, nicht das explosive Wachstum des Internet aufhalten zu wollen, sondern im Gegenteil, die damit einhergehende, wachsende Internet-Gemeinde als potentielle Kritiker und Redakteure zu sehen, führt zu einem umfassenden Katalog. Jeder kann auf seinem Spezialgebiet ein freiwilliger Redakteur werden, um als Experte anerkannt zu werden. [DMOZ]

Der Vorteil von Katalogdiensten ist, dass die Seiten gruppiert und daher leichter zu durchsuchen sind als in Keyword-Indizes. Ein menschlich generiertes The-

⁶Zur Zeit werden nur bis zu maximal 20% der vorhandenen Dokumente abgedeckt [BMWVK2000]

⁷Metasucher haben nichts mit Metadaten zu tun. Sie heissen so, weil sie Suchmaschinen über Suchmaschinen sind.

⁸<http://www.yahoo.com>

⁹<http://www.dmoz.org>

menverzeichnis erlaubt weiters feinere und schärfere Abstufungen, sowie die Zuteilung einer Ressource zu mehreren Themengebieten. Sie sollten weiters in der Lage sein, sinnvolle Ratschläge darüber zu geben, wo der Inhalt zu finden ist und wie gut oder schlecht er ist. [Legenstein99]

Jedoch sind Menschen nicht so effizient wie Maschinen, und menschlich generierte Kataloge können niemals so umfangreich und aktuell sein wie maschinell generierte. Eine Analyse von WWW-Adressen in Proxy-Caches hat zutage gefördert, daß bereits nach einem halben Jahr die Hälfte aller Adressen veraltet ist. [Sander98]

Katalogdienste unterscheiden sich durch ihre Organisation, den Inhalt und die angelegten Qualitätsmaßstäbe. Die Grenzen sind jedenfalls verschwommen, und es gibt keine einheitliche Benennung. In [Skov98] werden Katalogdienste wie in den folgenden Abschnitten 6.3.1 bis 6.3.6 eingeteilt.

6.3.1 Subject Catalogues

Die bereits erwähnte Maschine Yahoo! zählt zu den Subject Catalogues (Themenkatalog), und sie ist der wohl am häufigsten eingesetzte Katalogdienst. Die Websites werden hierbei ohne Qualitätskontrolle eingetragen, und die Abdeckung akademischer Ressourcen ist spärlich. Kurzfassungen sind meist nicht vorhanden, wodurch die Suche nach relevanten Dokumenten erschwert wird. [Skov98]

6.3.2 Annotated Directories

Diese Art von Katalogdiensten (kommentierte Verzeichnisse) unterscheiden sich von Themenkatalogen dadurch, dass die Inhalte von Experten auf dem jeweiligen Gebiet gefiltert und kommentiert wurden. Beispiele hierfür sind die Britische Seite BUBL LINK¹⁰ und die Amerikanische Infomine¹¹, die Zugang zu schulischen Themen bieten [Skov98]. Bei BUBL kann man das Thema alphabetisch sortiert suchen, nach Kategorien zusammengefasst oder dem Dewey Themenkatalog folgend. Die einzelnen Sites im Verzeichnis besitzen eine kurze Beschreibung.

¹⁰<http://www.bubl.ac.uk/link/>

¹¹<http://infomine.ucr.edu/>

6.3.3 Annotated Directories with Ratings or Reviews

Kommentierte Verzeichnisse mit Bewertungen oder Rezensionen bieten beispielsweise die Systeme Magellan¹², Excite¹³ und Lycos¹⁴. Bei Lycos werden die Editoren genannt und sollen via E-mail erreichbar sein. Magellan bietet dafür ausführlichere Rezensionen. Ein weiterer Vertreter dieser Gattung ist der Britannica Internet Guide¹⁵, bei dem die Sites eine „Sterne“-Wertung erhalten und kurz beschrieben werden. [Skov98]

6.3.4 Subject Directories with Ratings

Ein Beispiel für ein spezialisiertes Verzeichnis mit umfassenden Rezensionen und Bewertungen ist der Nutrition Navigator¹⁶ der Tufts Universität. Bei dieser Art von Katalogdienst findet man speziell zu einem Thema Ressourcen und Beschreibungen, die die inhaltliche Qualität der Sites in den Vordergrund rücken. [Skov98]

6.3.5 Subject Guides

Subject Guides (Themenführer) werden auch als „Webrings“ bezeichnet, und sind Sammlungen zu Ressourcen verschiedener Themen mit kommentierten Links zu Webseiten, Newsgroups und Journalen. Weil die Autoren üblicherweise Experten auf den jeweiligen Gebieten sind, haben die Ressourcen durchwegs eine hohe Qualität. Vertreter dieser Art sind Argus ClearingHouse¹⁷ und The Mining Corporation¹⁸. Letztere besitzt rund 500 einheitlich gestaltete Führer zu den unterschiedlichsten Themen – vom Schachspielen bis zu Englischer Literatur. [Skov98]

6.3.6 Information Gateways

Eine der wichtigsten Entwicklungen stellen qualitäts-kontrollierte *Information Gateways* („Informations Portale“) dar. Die führenden Initiativen sind das von der EU geförderte DESIRE¹⁹ und das eLib²⁰ Programm. Als Resultat dieser Pro-

¹²<http://magellan.excite.com>

¹³<http://www.excite.com>

¹⁴<http://www.lycos.com>

¹⁵<http://www.britannica.com>

¹⁶<http://navigator.tufts.edu>, Verzeichnis zum Thema Ernährung.

¹⁷<http://www.clearinghouse.net>

¹⁸<http://www.miningco.com>

¹⁹Development of a European Service for Information on Research and Education;
<http://www.desire.org>

²⁰UK Electronic Libraries Programme; <http://ukoln.bath.ac.uk/services/elib>

gramme sind einige qualitativ hochwertige Information Gateways, welche sich auf den Zugang zu Ausbildungsmaterial spezialisieren, entstanden. Dazu zählen unter anderem SOSIG²¹, ADAM²² und OMNI²³. DESIRE verwendet bei der Beschreibung von Ressourcen Standard RDF (vgl. Kapitel 3.2.3), und hat Werkzeuge entwickelt, mit denen man leicht Metadaten erstellen und in andere Formate konvertieren kann. [Skov98]

Alle hier aufgezählten Katalogdienste und die in Abschnitt 6.1 beschriebenen Indexsuchmaschinen haben gemeinsame Nachteile: Die Datenauffindung und Verwaltung erfolgt größtenteils zentral. Dies führt zu einer hohen Netzwerk- und Serverlast. Außerdem kann man mit ihrer Hilfe nicht gleichzeitig nach dem Inhalt und nach Qualitätsmetadaten suchen. Das im nächsten Abschnitt vorgestellte Suchsystem xFIND verteilt diese Aufgaben auf mehrere Server, und besitzt den weiteren Vorteil, dass Daten von externen Systemen – wie z.B. Kommentare oder Bewertungen – in das Suchsystem eingebunden werden können. [Legenstein99]

6.4 xFIND

Eines der Hauptprobleme gegenwärtiger Suchsysteme ist das häufige und unkoordinierte Durchsuchen des Netzes. Dazu kommt, dass sämtliche Rohdaten vom Server geladen, über das Netz gesandt und erst an zentraler Stelle analysiert werden. Um die Informations-Server und das Netz nicht zu überlasten, werden die Updateintervalle auf Kosten der Konsistenz und Aktualität groß gewählt. [BMWVK2000]

Das kooperierende, verteilte Suchsystem xFIND berücksichtigt diese Schwächen. Zum einen kann ein lokales Teilsystem direkt beim Informations-Server Daten auffinden, vorbereiten, komprimieren und bereitstellen. Zum anderen können mehrere Suchdienste von diesen aufbereiteten Daten Gebrauch machen. Der einzelne Informationsserver kann daher wesentlich öfter abgesucht werden, ohne dass es zu Überlastungen kommt. [BMWVK2000]

Das xFIND Suchsystem basiert auf einer verteilten Architektur in Verbindung mit verteilten Suchanfragen. Auf unterster Ebene stehen die lokalen Server, welche die vollständige Erfassung ihrer Daten durch Gatherer sicherstellen müssen. Die Gatherer laufen entweder lokal auf den Servern oder zumindest im Sinne einer guten Netzwerkanbindung „nahe“ dem Server. Darüberliegende Indexer fassen die aufbereiteten Informationen mehrerer Server zusammen. Dies

²¹Social Science Information Gateway; <http://sosig.ac.uk>

²²Art, Design, Architecture and Media Information Gateway; <http://adam.ac.uk>

²³UK's gateway to high quality biomedical Internet resources; <http://omni.ac.uk>

kann unter anderem nach geographischen oder thematischen Gesichtspunkten erfolgen. Die Daten werden indiziert und stehen für Suchanfragen zu Verfügung (vgl. Abbildung 6.1).[BMWVK2000]

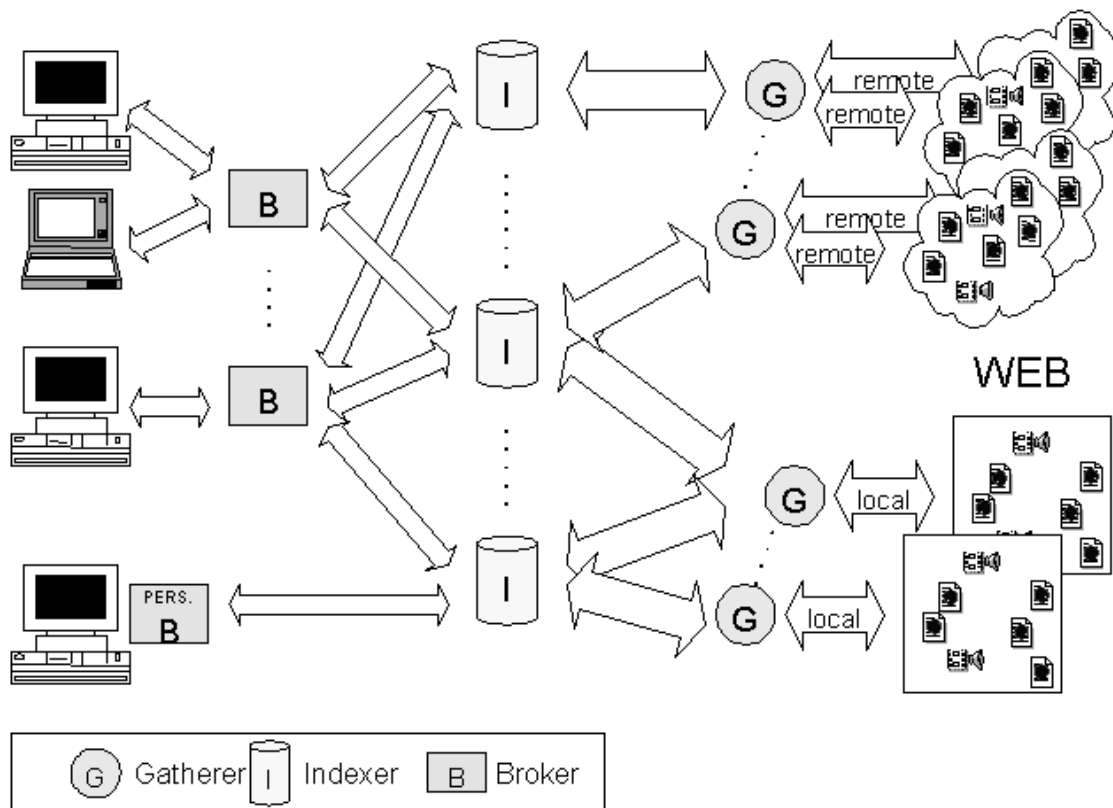


Abbildung 6.1: Verteiltes Konzept des xFIND Suchsystems. [BMWVK2000]

Gatherer sind für die Beschaffung und Aufbereitung der Daten zuständig. So können beispielsweise je nach Konfiguration Titel, Schlüsselwörter, Hyperlinks, eingebettete Multimediaobjekte usw. aus den untersuchten Dokumenten extrahiert und zu Beschreibungsobjekten zusammengefasst werden. Besondere Beachtung finden im Dokument enthaltene Metadaten, die auf unterschiedlichen Metadatenschemata, wie DC, LOM oder xQMS (vgl. Kapitel 3.3.3, 3.3.4 und 7), basieren können. Diese Metadaten werden ebenfalls extrahiert und sollen eine wichtige Rolle bei der Erzielung qualitativ hochwertiger Suchresultate spielen. Ein Gatherer kann entweder direkt am Informationsserver betrieben werden, und auf das lokale Dateisystem zugreifen. In diesem Fall können auch passwort-geschützte Bereiche suchbar gemacht werden, ohne tatsächlich die eigentliche Information zugänglich zu machen. Oder er wird an einem beliebigen Ort gestartet, und fordert die aufzubereitenden Daten über ein Transferprotokoll an. [BMWVK2000]

Der Indexer trägt die von einem oder mehreren Gatherern aufbereiteten und komprimierten Daten in den von ihm verwalteten Index ein. Ein Indexer kann

auf bestimmte Themenbereiche spezialisiert sein, oder auch einer Projektgruppe, Abteilung etc. zugeordnet sein. Hauptaufgabe eines Indexer ist das Verzeichnen von Daten und die Beantwortung von Suchanfragen. [BMWVK2000]

Die Aufgabe des Brokers besteht darin, die an ihn gestellten Suchanfragen – abhängig von dem darin formulierten Problem – auf die passenden Indexer zu verteilen. Wie schon beschrieben, erhält die Indexer die Daten von einem oder mehreren Gatherern, weswegen sie im Normalfall lediglich einen mehr oder minder kleinen Teil des gesamten verfügbaren Wissens verwalten. Jeder Broker kennt eine Reihe von Indexern und weiß auch, welche Wissensbereiche jeder einzelne von ihnen abdeckt. [BMWVK2000]

6.5 Auswahl eines Systems zur verbesserten Wissensauffindung

Das im vorigen Abschnitt vorgestellte Suchsystem xFIND bietet neben der Verteilung von Netzwerk- und Serverlast noch einen entscheidenden weiteren Vorteil: Es erlaubt die Integration von Daten, welche in externen Systemen erzeugt bzw. verwaltet werden, wie beispielsweise Kommentare oder Bewertungen. In diesem Abschnitt soll untersucht werden, welche Daten oder Systeme eine effiziente Suche nach qualitativ hochwertigen Inhalten im Internet erlauben.

Bereits in der Einleitung (vgl. Kapitel 1) wurde deutlich gemacht, wie wichtig die Verbesserung von Suchmaschinen ist. Mit dem explosiven Wachstum des Internet geht eine Verminderung der inhaltlichen Qualität einher, weil beinahe jeder im Netz publizieren kann. Suchdienste – welcher Art auch immer – stellen das zentrale Instrument in der Wissensauffindung im Internet dar, und sie müssen den neuen Erfordernissen angepasst werden.

Um qualitativ hochwertige Dokumente erkennen und sie in weiterer Folge von den restlichen Ressourcen trennen zu können, wurden in Abschnitt 2.1 Untersuchungen zum Thema „Qualität“ angestellt. Qualität setzt sich bei Ressourcen im Netz aus der inhaltlichen Datenqualität (DQ), der Kontextuellen-, der Zugangs- sowie der Darstellungs-DQ zusammen. Der Schwerpunkt der vorliegenden Arbeit liegt bei den beiden erst genannten Arten von Qualität, welche bei der Auffindung von interessanten Ressourcen eine wichtige Rolle spielen.

Aus diesem Grund wurde in Abschnitt 2.2 versucht herauszufinden, durch welche Eigenschaften sich im Speziellen die inhaltliche Qualität auszeichnet. Das Ergebnis sind eine Reihe von Attributen, durch die sich Ressourcen bewerten lassen. Diese Eigenschaften lauten u.a. Genauigkeit, Objektivität, Identifikation und Ansehen des Autors, Umfang, Gleichgewicht der Information, Zweck und

Aktualität. Eine mögliche Lösung für das genannte Problem der Suche nach guten Informationen im Internet ist daher die Suche nach Ressourcen mit den genannten Eigenschaften.

In Abschnitt 2.3 wurden aus diesem Grund Bewertungssysteme theoretisch untersucht. Bei der einen Art, dem sog. Upstream-Filtering, bewerten Leute – etwa Angestellte des Bewertungsdienstes, Fachexperten oder Benutzer – Ressourcen. Sie stellen die Informationen dann gesammelt auf einer Website dar, so wie dies bei den bewertenden Katalogdiensten der Fall ist. Benutzer können auf der Website nach qualitativ hochwertigen Dokumenten suchen. Die Nachteile der Methode liegen darin, dass die Anzahl zu bewertender Dokumente zu schnell wächst, um von einigen wenigen Kritikern erfasst werden zu können, und dass man die Qualitätsmaßstäbe nicht selbst bestimmen kann. Der Kontext einer Suche lässt sich nur durch die Verwendung verschiedener Suchdienste bestimmen. Denn nicht immer benötigt man nur die qualitativvollste – und vielleicht teure – Information, oft reicht ein schneller Überblick zu einem Thema. Weiters stellen die Menschen – vom Wissenschaftler bis zum Schulkind – verschiedene Ansprüche an die Qualität von Information.

Zielführender erscheint deshalb das zweite Verfahren von Bewertungssystemen, das Downstream-Filtering (vgl. Abschnitt 2.3.2). Dabei werden die Bewertungen – egal ob von Experten, den Autoren selbst oder Benutzern eines Suchdienstes erstellt – in einer computer-lesbaren Form gespeichert (entweder auf dem Server eines Bewertungsdienstes oder in der Ressource selbst). Die Filterung erfolgt anschließend beim Benutzer, indem er die Qualitätserfordernisse selbst vorgibt. Dies scheint daher die zielführendste Methode in der Verbesserung von Suchmaschinen zu sein: Zuerst sollen Ressourcen mit verschiedenen Attributen beschrieben werden, und nach den genannten Eigenschaften inhaltlicher Qualität bewertet werden. Diese Informationen müssen dann computer-lesbar gespeichert werden, sodass Suchmaschinen in Verbindung mit Volltextsuchen nach für den Anwender relevanten Informationen suchen können.

In Kapitel 3 wurde erläutert, wie Beschreibungen und Bewertungen – also sog. Metadaten – gespeichert werden können (mit Hilfe der Techniken HTML Meta Tag, PICS oder RDF), und welche Vereinbarungen zur Speicherung bereits existieren (MARC, SOIF, Dublin Core und LOM). Natürlich könnte man beliebige Metadaten zu Ressourcen in einer nicht weiter spezifizierten Form speichern. Dadurch ginge aber der Vorteil von Filtersystemen verloren, die die Metadaten nicht nur *lesen* sondern auch *verstehen* können. Dublin Core ist beispielsweise ein Metadaten Schema, bei dem es nur möglich ist, Beschreibungen zu Ressourcen in einer für Menschen verstehbaren Form zu speichern. Hingegen werden Bewertungen in PICS vom Computer interpretiert, sodass er in gewisser Weise darauf reagieren kann (nämlich beispielsweise durch Blockieren einer Webseite). Eine Kombination der beiden Verfahren wäre wünschenswert: Manche Eigenschaften

lassen sich nämlich nicht computer-verstehbar speichern, wie z.B. der Name des Autors oder der Titel einer Ressource. Viele andere jedoch können so spezifiziert werden, dass eine automatische Filterung erfolgen kann.

Ein Attribut kann dann als „computer-verstehbar“ bezeichnet werden, wenn der zulässige Wertebereich eingeschränkt und für jeden Wert eine Bedeutung festgelegt ist. Für manche Eigenschaften existieren bereits Vorschriften und Normen, die man übernehmen kann. Beispielsweise kann sich das Attribut „Sprache“ auf die Konvention RFC1766²⁴ stützen, und die Datumsfelder der ISO-Norm 8601²⁵ folgen. Für einige andere Eigenschaften müssen die zulässigen Wertebereiche mit den Bedeutungen erst bestimmt werden.

Ein weiterer Vorteil von Feldern, deren beschränkter Wertevorrat mit Bedeutungen belegt ist, besteht darin, dass solche Eigenschaften unabhängig von der Landessprache der Benutzer ausgewertet werden können. Für Menschen ändert sich zwar je nach Landessprache die Repräsentation der Bedeutung. Beispielsweise kennzeichnen die verschiedenen Repräsentationen „ausgezeichnet“ und „excellent“ dieselbe Bedeutung. Da festgelegt wurde, dass dieser Bedeutung im Computer beispielsweise die Repräsentation „10“ zukommt, ändert sich für den Computer nichts, egal ob der Bewerter bzw. Anwender Amerikaner oder Österreicher ist.

Ein Nachteil von Attributen mit beschränktem Wertevorrat ist das mögliche Vorkommen von Eigenschaften außerhalb des zulässigen Bereichs. Deshalb muß man für Fälle vorsorgen, in denen die Eigenschaft nicht angegeben oder bekannt ist (dafür könnten die Werte 0 und -1 vorgesehen werden), keiner der erlaubten Werte zutrifft (-2), oder die Eigenschaft für eine Ressource nicht passend erscheint (-3).

Die bisher durchgeführten Überlegungen führen zu einem System, in dem beschreibende und bewertende Metadaten soweit wie möglich computer-verstehbar gespeichert werden. Der Benutzer einer Suchmaschine soll neben Stichwörtern, dem Titel etc. die Eigenschaften einer Ressource eingeben können, wie beispielsweise den Namen des Autors, die intendierte Zielgruppe oder das Thema, und das System filtert die entsprechenden Dokumente aus seiner Datenbank. Die Datenbank wurde zuvor von einem Indexer, der die Metadaten lesen und verstehen kann, erstellt. Auch Blockingsysteme, wie sie in Kapitel 4 beschrieben sind, basieren auf diesen Grundlagen, was den Vorteil hat, dass keine von oben regulierte Zensur eingeführt werden muss, sondern Benutzer festlegen können, was sie lesen und sehen wollen und was nicht. Dazu ist aber noch ein festgelegtes Vokabular notwendig, welches im Falle von RSACi beispielsweise aus den vier Eigenschaften Gewalt, Sex, Nacktaufnahmen und Sprache besteht (vgl. Abschnitt 4.1.3). Denn

²⁴<http://puma.germany.net/internic/rfc/rfc1766.txt>

²⁵<http://www.w3.org/TR/NOTE-datetime>

diese vier Eigenschaften beschreiben im Allgemeinen Dokumente sehr gut, von denen man nicht will, dass sie von Kindern gesehen werden. Im Kontext von Suchmaschinen gelangt man wieder zu den Überlegungen, welche Eigenschaften die inhaltliche Qualität von Ressourcen am besten kennzeichnen, und durch welche Attribute man Dokumente am besten beschreiben kann, sodass die Suche danach erleichtert wird. Es ist auch vorstellbar, dass Elemente der Qualitätsmetadaten von Blockingsoftware verwendet werden könnten.

6.5.1 Auswahl geeigneter Attribute zur Beschreibung und Bewertung von Ressourcen im Netz

Wie beschrieben benötigt man zur Implementierung eines verbesserten Suchsystems unter der Verwendung von Metadaten Vereinbarungen darüber, welche Metadaten gespeichert werden sollen (ein sogenanntes Metadaten Schema). Die im folgenden durchgeführten Überlegungen haben zur Entwicklung des xFIND Quality Metadata Scheme, abgekürzt xQMS, geführt, welches im Gestaltungsbereich der Arbeit in Kapitel 7 näher erläutert wird.

Selbstverständlich wäre es für einen Anwender eines Suchsystems wünschenswert, wenn Ressourcen nach allen nur denkbaren Eigenschaften bewertet und beschrieben sein würden. Doch solche Rezensionen würden den Umfang der eigentlichen Ressource meist übertreffen, und sie wären selbstverständlich unrentabel und teuer. Wenn aber zuwenige Eigenschaften bewertet würden, kann das System keine wesentliche Verbesserung in der Wissensauffindung bringen. Denn schon heute werten viele Suchmaschinen die META Tags „keywords“ und „description“ aus, ohne wesentliche Verbesserung der Leistung. Man muß also einen Kompromiß in der Anzahl der Attribute finden, der sowohl für die Benutzer von der Suchmaschinen als auch für die Bewertenden vertretbar ist.

Wichtige beschreibende Attribute von Ressourcen lassen sich aus dem Buchwesen ableiten. Dazu zählen beispielsweise jene Eigenschaften, die man in einem Literaturverzeichnis über ein Werk festhält: In erster Linie sind dies der Name des Autors und gegebenenfalls der erstellenden Organisation, sowie die Nennung des Titels und des Verlegers. In [Mitrettek97] wird ebenfalls die Quelle – also beispielsweise die Nennung des Autors – als das primäre Kriterium identifiziert, wie sich inhaltliche Qualität kennzeichnen lässt. In Umfragen gaben Hundert Prozent an, dass die Quelle ein essentielles Kriterium in der Beschreibung von Dokumenten darstellt (vgl. Anhang A.1). Weiters werden bei Zitaten das Datum der Erstellung sowie der letzten Auflage und – falls vorhanden – ein eindeutiges Kennzeichen wie beispielsweise eine ISBN Nummer oder URL genannt. Aus dem Bibliothekswesen stammt die Idee, Stichwörter, Schlagwörter und kurze Beschreibungen über Werke zu sammeln. Diese Attribute sind auch in der Elementmenge des Dublin Core

vorhanden. Aus diesen Überlegungen ergeben sich die Attribute Autor, Organisation, Herausgeber, Titel, Datum der Erstellung und letzten Änderung, Identifikation, Schlüsselwörter und Beschreibung. Dem Aufgabengebiet eines Verlegers kommt bei Ressourcen am Netz oft der sogenannte Webadministrator sehr nahe. Die Nennung desselben, beziehungsweise seiner E-Mail- und Kontaktadresse erscheint daher ebenso denkbar. Das Beispiel in der nachfolgenden Abbildung 6.2 soll die Anwendung der bisher ausgemachten Attribute verdeutlichen.

Autor: Weitzer Johann
Organisation: Institut für Informationsverarbeitung
und Computergestützte neue Medien
Herausgeber: TU Graz
Titel: Verwendung von Metadaten zur verbesserten Wissensauffindung
Erstellungsdatum: 1.1.2000
Datum der letzten Änderung: 30.3.2000
Identifikation: <http://www.iicm.edu/DA/Weitzer.html>
Schlüsselwörter: Metadaten, Wissensauffindung, Qualität, Retrieval
Beschreibung: Diplomarbeit zum Thema Wissensauffindung im Internet
Webadmin: webmaster@iicm.edu

Abbildung 6.2: Beispiel für Verwendung der bisher ausgemachten Attributen. Gezeigt wird die Beschreibung der vorliegenden Arbeit.

Der Unterschied zu Printmedien besteht aber darin, dass es leicht möglich ist, hier falsche Angaben zu machen. Deshalb muss der Autor für den Fall, dass die Ressource eine Webseite oder -abschnitt ist, diese elektronisch signieren können (mit Hilfe eines Attributes Signatur). Damit kann man sicherstellen, dass sich ein Dokument nach der Erstellung weder durch den Autor selbst noch durch Fremde (beispielsweise Hacker) verändert hat. [W3C PICS]

Nachdem Informationen im Internet aber oft dynamisch generiert werden, ist es zum Teil schwierig, eine solche Signatur überhaupt durchzuführen. Es wäre denkbar, die Signatur nur über gewisse Teile einer Ressource zu bilden, wie beispielsweise den Titel, die Überschrift, die ersten Absätze und die URL. Oder man geht einen anderen Weg, und stellt mittels einer Signatur nur sicher, dass die Ressource tatsächlich von dem angegebenen Autor stammt. Dann kann man entscheiden, ob die Person vertrauenswürdig erscheint, oder nicht.

In Printmedien sind die zu beschreibenden oder bewertenden Werke entweder Bücher oder Zeitschriften, die aus einer bestimmten Anzahl Seiten bestehen. Bei Referenzen auf eine gewisse Textstelle wird das Werk und gegebenenfalls eine Seitennummer genannt. Bei Ressourcen im Netz sieht diese Einteilung anders aus. Es gibt im World Wide Web Server (z.B. „iicm.edu“), Bereiche innerhalb von Servern („iicm.edu/Lehre“) und Webseiten („iicm.edu/Lehre/xfind/da.html“). Doch es können nicht nur Seiten direkt angesprochen werden, sondern zusätzlich auch

darin enthaltene Abschnitte (Bild, Tabelle oder Absatz innerhalb einer Webseite). Deshalb liegt es nahe, Beschreibungen wahlweise auf eine dieser Größen zu beziehen, was mit einem eigenen Attribut „Art der Ressource“ erfolgen könnte.

Desweiteren ist bei elektronischen Ressourcen der Typ von großem Interesse: Handelt es sich bei einer Ressource um ein Bild, ein Video, einen normalen Text oder um Software? Eine Aufstellung möglicher Typen von Ressourcen findet sich in Anhang B.2.

Bibliotheken bieten nicht nur die Möglichkeit, Bücher über die Stichwörter- oder Schlagwortsuche zu finden. Man kann geeignete Bücher auch über die Themensortierung der Regale finden. Diese Idee führt zu einem Metadaten Attribut, mit dem sich das Thema einer Ressource angeben lässt. Hier wäre es von Nachteil, beliebige Themenangaben zuzulassen, weil dadurch automatische Filterungen unmöglich gemacht würden. Stattdessen kann man standardisierte Themenklassifizierungen einsetzen, die einer definierten Menge verschiedener Themen Werte zuweisen.

Das weitverbreitetste System der Themenklassifikation ist der sogenannte *Dewey Decimal Code (DDC)*²⁶ (Die Klassifizierung der Hauptthemen finden sich im Anhang D.1). Der Code ist ein hierarchisch aufgebautes Verzeichnis, welches jedem Thema eine Dezimalzahl zuweist. Unterthemen werden durch weitere Stellen im Code angegeben. Beispielsweise steht „.500“ für Naturwissenschaft und Mathematik und „.530“ steht für Physik. Der DDC wurde im Jahr 1873 von Melvil Dewey ausgedacht, und wird seither ständig an den Stand des Wissens angepasst. [OCLC].

Weil ein allgemeiner Katalog aber nicht alle Themen bis ins Detail klassifizieren kann, haben sich auf verschiedenen spezielleren Themengebieten eigene Systeme zur Klassifikation herausgebildet, wie z.B. die spezielle Charakterisierung von Computertemen durch ACM²⁷. Und in [ED99] wird beispielsweise ein spezielles Schema zur Themenklassifikation auf dem Gebiet der Medizin eingesetzt. Die *unique concept identifiers (CUIs)* des *Unified Medical Language System (UMLS)* sind numerische Schlüssel mit Werten von 000000 bis 999999, welche alle möglichen medizinischen Fachgebiete kennzeichnen sollen.

In einer Menge von Qualitätsmetadaten scheint daher ein Attribut für die Angabe des weitverbreiteten DDC Themencodes sinnvoll, sowie eventuell einer weiteren standardisierten Themenklassifikation. Dazu benötigt man noch zusätzliche Angaben über das zu verwendende Klassifikationsschema, wie beispielsweise Namen und Internetadresse. Beispielfhaft für die vorliegende Arbeit könnte eine Beschreibung wie in Abbildung 6.3 aussehen.

²⁶<http://www.oclc.org>

²⁷<http://www.acm.org>

DDC: .004 (data processing, computer science),
 .020 (library and information sciences)
 Alternative Themenangabe: H.3.3 (information search and retrieval),
 H.3.4 (systems and software)
 Alternative Themenklassifikation: ACM
 Identifikation der Themenklassifikation: <http://www.acm.org/class/1998>

Abbildung 6.3: Beispiel für Beschreibung des Themas einer Ressource.

Weil Dokumente im Internet von Orten auf der ganzen Welt gelesen werden können, ist es für Leser wichtig zu erfahren, in welcher Sprache sie verfasst sind. Diese Eigenschaft lässt sich, wie bereits erwähnt, mit den Werten in Konvention RFC1766 angeben (Attribut Sprache). Da sich die Sprache aber auch hinsichtlich ihres Schwierigkeitsgrades bzw. Ausdruckes unterscheidet, wäre ein bewertendes Attribut hilfreich, welches die verwendete Sprache diesbezüglich charakterisiert (Attribut Ausdruck). Mit diesem Attribut könnte man festhalten, ob die verwendete Sprache in einer Ressource beispielsweise mühelos lesbar oder anspruchsvoll ist. Die Bewertung der vorliegenden Arbeit könnte hinsichtlich der Sprache wie in Abbildung 6.4 aussehen.

Sprache: Deutsch, neue deutsche Rechtschreibung oder DE [RFC1766]
 Sprachstil: wissenschaftlicher Text
 Informationstiefe: hoch, detailreich
 Informationsbreite: gering, Behandlung eines Spezialthemas
 Alter der Zielgruppe: ab 20 Jahren
 Vorwissen der Zielgruppe: Grundlegende Kenntnisse des Internet und der Wissensauffindung.
 Ansehen: Diplomarbeit

Abbildung 6.4: Beispiel für die Verwendung von Attributen zur Beschreibung der Sprache in der vorliegenden Arbeit.

Bei den Eigenschaften der inhaltlichen Qualität wurden schon die Attribute Genauigkeit und Umfang genannt. Sie lassen sich durch Angaben über die Tiefe und Breite der gebotenen Information klassifizieren. Ein weiteres der in Abschnitt 2.2 genannten Attribute ist der Zweck. Dieser ist immer in Hinblick auf eine Zielgruppe zu beurteilen, weshalb es nötig ist, diese für die Ressource festzuhalten. Das geschieht mit einer Klassifikation hinsichtlich des Alters und bzw. oder des Vorwissens auf dem Gebiet. Vorteil der vier Eigenschaften Informationstiefe und -breite, sowie Vorwissen und bzw. oder Alter ist, dass die Angabe auf vordefinierten Skalen erfolgen kann (beispielsweise von 'detailreich' bis 'überblickhaft' oder von 'Anfängerwissen' bis 'Expertenwissen').

Als weitere Eigenschaft inhaltlicher Qualität wurde in Abschnitt 2.2 'Ansehen/Glaubwürdigkeit' (authority) genannt. Da man aber über die Verlässlichkeit

Zitiervorschrift: Weitzer, Johann: Verwendung von Metadaten zur verbesserten Wissensauffindung, Diplomarbeit an der TU Graz, 2000.
Version: 1.0
aktuellere Version: keine
ältere Version: http://www.iicm.edu/DA/Weitzer_alt.html

Abbildung 6.5: Beispiel für den Einsatz weiterer Attribute zur Beschreibung der vorliegenden Arbeit.

und Qualifikation eines Autors oft keine genauen Aussagen treffen kann, beschränkt man sich bei der Bewertung des Ansehens und der Glaubwürdigkeit auf die Ressource selbst. [SC94] [Legenstein99]

Weil das Ansehen und die Glaubwürdigkeit subjektive Eigenschaften sind, ist es sehr schwierig, eine Skala zu entwerfen, entlang der die Eigenschaften messbar wären. In [Palme98] wird eine Einteilung von Dokumenten aufbauend auf dem System der wissenschaftlichen Bewertung getroffen. Dabei wird beispielsweise zwischen Texten, die für ein wissenschaftliches Journal akzeptiert wurden, oder die populär-wissenschaftlichen Charakter haben unterschieden. In dieses System lassen sich jedoch keine nicht-wissenschaftlichen Dokumente einreihen.

Bei Dokumenten, die in gedruckter Form und online – vielleicht sogar an mehreren unterschiedlichen Orten – verfügbar sind, stellt sich für jemanden, der das Dokument zitieren möchte, die Frage, welches Vorkommen er zitieren soll. Es ist daher hilfreich, wenn der Autor einer Ressource von vornherein angeben kann, wie er sein Werk zitiert haben möchte. Er kann beispielsweise entscheiden, dass sich Referenzen auf sein im Web und in Buchform publiziertes Werk, immer auf das letztere beziehen, weil dieses dauerhafter ist. Eine solche Zitiervorschrift kann auch für Benutzer von Suchmaschinen hilfreich sein kann, wenn sie im Suchergebnis mitangegeben wird. Ein Attribut Zitiervorschrift erscheint daher sinnvoll.

Von Dokumenten im Netz gibt es oft auch mehrere Versionen, von denen meist nur eine dem neuesten Stand entspricht. Der Autor sollte daher für jedes Dokument eine Versionsnummer speichern, und angeben, wann die letzten Änderungen vollzogen wurden. Wünschenswert erscheint es auch, Angaben darüber machen zu können, wo neuere bzw. vorherige Versionen zu finden sind. Solche Angaben könnten mit den Attributen Versionsnummer, Datum der letzten Änderung sowie eindeutige Verweise auf ältere und neuere Versionen gemacht werden. Angaben über die durchschnittliche Änderungsrate von Dokumenten können ebenso interessant sein. Das Beispiel in Abbildung 6.5 soll wieder die Beschreibung der vorliegenden Arbeit demonstrieren.

Bei bewerteten Ressource ist es wichtig, erfahren zu können, wer die Bewertung vorgenommen hat. Eigenschaften dieser Meta-Metadaten sind der Name des Bewerter, die Organisation und der Publizist, weiters die Sprache (für diejenigen

Attribute, die frei sind), das Datum der letzten Bewertung und eine Signatur mit der die Korrektheit der Angaben überprüft werden kann. Weil die Bewertung einer Ressource nicht ewig gelten kann, sollte eine Angabe darüber, wann die Bewertung ihre Gültigkeit verliert, nicht fehlen. Beispielsweise kann man einen Zeitraum ab dem Datum der letzten Änderung der Bewertung angeben, ab welchem dieselbe ungültig wird. [SOIF96] [W3C PICS]

In diesem Abschnitt wurde untersucht, welche Eigenschaften sich bei der Verbesserung von Suchmaschinen dazu eignen, Ressourcen hinsichtlich ihrer inhaltlichen Qualität zu bewerten und zu beschreiben. Es bleibt noch zu klären, wer die Bewertungen durchführen soll.

6.5.2 Wer soll Bewertungen erstellen?

Es wurde bereits in der Einleitung des Kapitels erwähnt, dass die Bewertungen durch Menschen wesentlich genauer sind, als die von Computern automatisch erstellten. Gewisse Eigenschaften lassen sich durch den Computer nur sehr schwierig bis gar nicht feststellen (beispielsweise das Ansehen wissenschaftlicher Dokumente), andere wiederum sehr leicht (wie das Datum der letzten Änderung eines Dokuments). Und das Signatur-Feld lässt sich ausschließlich vom Computer berechnen. Menschliche Bewertung kann durch den Ersteller einer Ressource selbst geschehen, durch einen Mitarbeiter bzw. Experten eines Bewertungsdienstes oder durch die Benutzer des Dienstes (die Leser der Ressource).

Eine automatische Bewertung von Dokumenten hinsichtlich ihrer Qualität wäre sicherlich wünschenswert, weil man so die meisten Dokumente am billigsten bewerten könnte. Manche der bereits genannten Attribute, welche die inhaltliche Qualität von Ressourcen beschreiben, können auch sehr gut durch den Computer automatisch ermittelt werden. Hierzu zählen aus offensichtlichen Gründen das Datum der Erstellung sowie der letzten Änderungen, der Bezeichner (beispielsweise die URL oder der Dateiname des Dokuments) und meist auch der Titel (In HTML und anderen sogenannten *Markup Languages* werden Titel durch Tags gekennzeichnet). Angaben zum Autor, der Organisation dem Verleger und der Sprache lassen sich unter Umständen auch automatisieren, beispielsweise direkt bei der Generierung von Ressourcen durch die Arbeitsumgebung der Autoren. Bereits genannt wurde auch die Signatur, welche sich ausschließlich vom Computer erstellen lässt.

Verschiedene Algorithmen versuchen die relevanten Schlüsselwörter oder das Thema von Texten zu erkennen. Dies geschieht anhand der Position von Wörtern im Text (kommt ein Wort im Titel vor, hat es eine andere Relevanz, als wenn es während eines Absatzes, in einer Bildunterschrift oder einem Fußnotentext vorkommt), sowie deren Häufigkeiten. Allzu häufig vorkommende Wörter haben

geringere Bedeutungen als seltenere Vorkommen. Der Nachteil dieser Algorithmen liegt in der oft unbefriedigenden Genauigkeit (vgl. Abschnitt 6.1). Die Bewertung durch Menschen kann hier weiterhelfen. Zu unterscheiden ist dabei zwischen der Benutzergruppe, den Autoren der Ressourcen und Experten des betroffenen Fachgebietes.

Die Bewertung durch Benutzer hat den großen Nachteil, dass verschiedene Benutzer ein und dieselbe Ressource verschieden bewerten und beschreiben werden. Der Vorteil liegt darin, dass wegen der grossen Anzahl potentieller Benutzer, zumindest theoretisch viele Dokumente bewertet werden könnten. In der Praxis zeigt sich jedoch meist, dass viele Benutzer lieber die Bewertungen anderer konsumieren (in Form eines gut funktionierenden Suchdienstes), als sich selbst die Mühe zu machen, welche zu erstellen (vgl. Abschnitt 5.1.1).

Wenn nur die Autoren von Ressourcen selbst Bewertungen durchführen, sind die Ergebnisse wahrscheinlich häufig zu wenig objektiv. Die Beschreibungen könnten aber größtenteils von den Autoren vorgenommen werden. Dadurch verteilt sich die Arbeit beim Charakterisieren der Ressourcen. Auf jeden Fall ist die Qualität der Beschreibungen besser als jene, die nur vom Computer erzeugt wurden.

Eine Beschreibung und Bewertung alleine durch Experten oder Mitarbeiter von Suchdiensten brächte vielleicht die höchste Qualität, ist aber wegen der enormen Menge von zu bewertenden Dokumenten nicht vorstellbar.

Aus den genannten Gründen erscheint eine Mischform sinnvoll. Einige Metadaten werden automatisch festgestellt, andere werden vom Benutzer oder dem Autor selbst bestimmt, und wieder andere können von Experten des Gebietes festgelegt werden. Die Qualität der Metadaten nimmt von der automatischen Generierung über die Beschreibung des Autors bis zum Urteil von Experten zu.

Auch bei dem im Gestaltungsbereich in Kapitel 8 gezeigten System erfolgt die Beschreibung und Bewertung auf verschiedenen Ebenen. Wenn ein Autor seine Ressource mit Hilfe von xQMS suchbar machen will, muss er sie erst auf dem Server anmelden. Dazu ist eine erste Angabe der beschreibenden und bewertenden Metadaten nötig. Der Administrator des Suchsystems erhält eine Benachrichtigung über den gewünschten Anmeldevorgang, und kann entscheiden, ob die Ressource im System aufgenommen werden soll oder nicht. Entschliesst er sich zur Aufnahme, wird die Ressource erstmals indiziert. Bei den Metadaten der Ressource wird zu diesem Zeitpunkt vermerkt, dass die Angaben nur vom Autor selbst stammen (Sie besitzen eine niedere Priorität. s. u.). Der Administrator hat aber weiters die Möglichkeit, Experten auf dem jeweiligen Gebiet der Ressource um ihr Urteil zu bitten. Diese können die Angaben zu einer Ressource überprüfen und die Metadaten gegebenenfalls verändern. Sie erreichen dadurch eine höhere Priorität, weil sie nicht mehr nur vom Ersteller selbst stammen.

Die Priorität der Metadaten (Angabe, ob die Metadaten automatisch, vom Ersteller oder einem Experten stammen) hat Auswirkungen auf das Suchergebnis. Wenn bei ähnlichen Dokumenten, die Metadaten über das eine von Experten bestätigt sind, und beim anderen nicht, erhält Ersteres ein besseres Ranking im Suchergebnis.

6.6 Zusammenfassung

Die vorigen Abschnitte behandelten die verbesserte Suche nach qualitativen Inhalten im Netz, weil es durch das enorme Wachstum immer schwieriger und unübersichtlicher wird, gute von schlechter Qualität zu trennen. Die am häufigsten eingesetzten Suchdiensten (Indexsuchmaschinen, Metasucher und Katalogdienste) haben den Nachteil, dass sie Ressourcen kaum nach ihrer Qualität beurteilen, und wenn, dann nur von zentraler Stelle aus. Das Suchsystem xFIND soll in Kombination mit dem xFIND Quality Metadata Scheme, xQMS, an diesem Punkt ansetzen. Es ermöglicht die verknüpfte Suche von Inhalt und Qualitätsmetadaten. Schon bei der Eingabe der Suchkriterien lassen sich auch Qualitätskriterien angeben.

Es wurde untersucht, welche Attribute sich zur Kennzeichnung der Qualität am besten eignen. Dabei wurde zwischen bewertenden und beschreibenden Attributen unterschieden, wie sie in Tabelle 6.1 aufgezählt sind. Ausserdem wurde unterschieden zwischen Attributen, deren Werte man frei wählen kann, und solchen, die einen eingeschränkten Wertevorrat besitzen. Die zweitgenannten Typen von Attributen haben den Vorteil, dass sie leicht vom Computer interpretiert werden können. Anderenfalls wäre es nur schwer möglich, dem Computer mitzuteilen, bei der Suche nach den Begriffen „classification scheme“ beispielsweise nur nach Dokumenten zu suchen, die für *erwachsene Experten* auf dem Gebiet des *Bibliothekswesen* gedacht sind. Mit den Attributen Vorwissen und Alter der Zielgruppe mit den festgelegten Werten, sowie einer Themenklassifikation wie dem Dewey Decimal Code, lassen sich solche Suchanfragen jedoch stellen.

Neben der Auswahl der Attribute wurden auch Überlegungen angestellt, wer die Bewertungen durchführen soll. Dabei kristallisierte sich heraus, dass eine Kombination aus automatisch erstellten Bewertungen sowie Bewertungen durch die Autoren und Fachexperten am sinnvollsten erscheint, weil sich dadurch die Arbeit verteilt, und gleichzeitig die Qualität der Metadaten steigt.

beschreibende Attribute		bewertende Attribute	
Autor		Bewerter	
Organisation		Bewertungsorganisation	
Verleger		Bewertungsverleger	
Titel		Ansehen	*
Zitiervorschrift		Beschreibung	
Identifikation		Sprachstil	*
Signatur		Signatur der Bewertung	
Stichwörter		Gültigkeit der Beschreibung	*
Erstellungsdatum	*	Datum der Bewertung	*
Datum letzter Änderung	*	Gültigkeitszeitraum d. Beschreibung	*
Version	*	Änderungsrate	*
aktuellere Version	*	Alter der Zielgruppe	*
vorige Version	*	Vorwissen der Zielgruppe	*
Sprache	*	Sprache der Bewertung	*
Typ	*		
Thema (DDC)	*		
weiteres Thema	*		
Themenklassifikation			
Themenklassifikation Schema			
Webadministrator			

Tabelle 6.1: Aufzählung von Eigenschaften, die Ressourcen hinsichtlich ihrer inhaltlichen Qualität bewerten und beschreiben. Es ist auch vermerkt, ob sich für das Attribut ein eingeschränkter Wertevorrat bzw. eine Norm zum Ausfüllen festlegen lässt. Dadurch können diese Attribut computer-interpretierbar gemacht werden (mit einem Stern gekennzeichnet).

Diese Überlegungen führen zur Definition des schon erwähnten xFIND Quality Metadata Scheme in der Version 1 in Kapitel 7 des folgenden Gestaltungsbereichs. Für die Vergabe der Metadaten wurde eine Software entwickelt, welche in Kapitel 8 vorgestellt wird. Durch die Kombination mit dem xFIND Suchsystem soll dadurch die verknüpfte Suche nach dem gewünschten Inhalt in entsprechender Qualität möglich werden.

Teil II

Gestaltungsbereich

Kapitel 7

xFIND Quality Metadata Scheme xQMS v1.0

Da durch das rasant wachsende Internet die Suche nach qualitativem Inhalt immer schwieriger wird (vgl. Kapitel 1), wurde im Untersuchungsbereich versucht herauszufinden, mit welchen Mitteln die Suche verbessert werden könnte.

Wie die vorangegangenen Kapitel 2 bis 6 gezeigt haben, erlaubt der Einsatz von Metadaten eine automatische Filterung (vgl. Kapitel 2.3) von Inhalten aus einer Datenbank bzw. dem Internet. Dieser Filterprozess kann den Zweck haben, unerwünschte Inhalte – wie beispielsweise gewaltverherrlichende Webseiten – auszublenden (vgl. Kapitel 4), oder aber erwünschte Inhalte zu finden (vgl. Kapitel 6). Der Benutzer eines solchen Systems soll in der Lage sein, jene Dokumente aus einer Datenbank bzw. dem Internet herauszufiltern, die seinen Kriterien und Erfordernissen entsprechen. Dazu zählen unter anderem gewünschte Qualitätserfordernisse, die Suche nach Stichwörtern, dem Titel, dem Namen des Autors etc. Zu diesem Zweck erscheint die Entwicklung eines Metadaten Schemas sinnvoll, wie es in diesem Kapitel vorgestellt wird.

Bei den im Abschnitt 3.3 vorgestellten Metadaten Schemata, wie z.B. Dublin Core, werden Attribute eingesetzt, die Ressourcen bezüglich ihres Inhaltes beschreiben. Man kann aber auch Bewertungen durchführen, wie es bei den in Abschnitt 4.1 gezeigten Blocking-Systeme geschieht. Bei der Definition des Metadaten Schemas xQMS finden sowohl beschreibende wie auch bewertende Attribute Verwendung. Für Benutzer des Suchsystems soll es in weiterer Folge möglich sein, eine verknüpfte Suche nach gewünschten Inhalten mit den geforderten Qualitätskriterien durchzuführen.

Die Attribute lassen sich weiters bezüglich ihrer Aufgabe klassifizieren. Die bewertenden Attribute bestehen aus der Klasse jener Attribute, welche die Meinung eines Bewerter ausdrücken, und der Gruppe von Attributen, die etwas

über den Bewerter bzw. die Bewertung selbst aussagen (vgl. Abschnitt 6.5.1). Die wichtigsten Gruppen der beschreibenden Attribute beziehen sich auf die Charakterisierung des Inhalts und des Erstellers einer Ressource (vgl. Abschnitt 2.2). Die besonderen Gegebenheiten des Internet machen weiters die Klasse von Attributen, welche die Version beschreiben, und die Klasse, durch deren Attribute technische Gegebenheiten beschrieben werden, nötig (vgl. Abschnitt 6.5.1).

Jede der genannten Kategorien besitzt mehrere Attribute. Dabei ist zwischen solchen Attributen zu unterscheiden, deren Werte frei gewählt werden können, und solchen, die einen eingeschränkten Wertevorrat besitzen, oder einer Norm folgen. Die zweit genannte Gruppe der Attribute lässt sich sehr leicht durch Computer verarbeiten (vgl. Abschnitt 6.5). Für diese Attribute sind aber auch jeweils vier besondere Werte vorgesehen, die es erlauben, ungewöhnliche Fälle zu unterscheiden. Es kann der Fall auftreten, dass gar kein zulässiger Wert auf die Ressource zutrifft, dass die Ressource überhaupt nicht nach diesem Attribut bewertet werden kann oder dass alle Werte gelten. Ein letzter Fall tritt ein, wenn die Ressource nicht bewertet wurde (vgl. Tabelle 7.1).

Bezeichnung Deutsch	Bezeichnung Englisch	Wert
nicht passend	<i>not applicable</i>	-3
unbekannt	<i>unknown</i>	-2
alle	<i>all</i>	-1
nicht angegeben	<i>not specified</i>	0

Tabelle 7.1: Bei jenen Attributen, die einen eingeschränkten Wertevorrat besitzen, ist auf Sonderfälle zu achten.

Aus den Klassen der Eigenschaften, welche Ressourcen beschreiben und bewerten, lassen sich 34 Attribute ableiten. Eine Aufstellung der Attribute befindet sich in Tabelle 7.2, wobei man auch die Trennung zwischen beschreibenden und bewertenden Attributen sowie die Gruppierung nach Thema erkennt. Im folgenden Abschnitt werden die einzelnen Eigenschaften näher erläutert, und am Ende des Kapitels folgen Beispiele, welche die Verwendung der Eigenschaften demonstrieren sollen.

Bezeichnung Deutsch	Bezeichnung Englisch	xQMS
Autor	<i>author</i>	creator.person
Organisation	<i>organistation</i>	creator.organisation
Verleger	<i>publisher</i>	creator.publisher
Bezeichner	<i>identifier</i>	tech.ident
Gültigkeit d. Beschreibung	<i>scope of description</i>	tech.scope (*)
Signatur	<i>signature</i>	tech.signature (*)
Typ	<i>type</i>	content.type (*)
Titel	<i>title</i>	content.title
Thema	<i>topic</i>	content.topic (*)
Alternatives Thema	<i>alternative topic</i>	content.alt_topic
Klassifikationsschema	<i>classification scheme</i>	content.alt_classname
Klassifikationsss. Bezeichner	<i>class. scheme ident</i>	content.alt_classident
Schlüsselwörter	<i>keywords</i>	content.keywords
Beschreibung	<i>description</i>	content.description
Zitiervorschrift	<i>quotation hints</i>	content.quotation
Sprache	<i>language</i>	content.language (*)
Versionsnummer	<i>version number</i>	version.number (*)
Vorversion	<i>previous version</i>	version.prev
Aktuellere Version	<i>next version</i>	version.next
Erstelldatum	<i>creation date</i>	version.create (*)
Datum letzte Änderung	<i>last modification date</i>	version.last_modified (*)
Ausdruck	<i>diction</i>	myop.diction (*)
Alter der Zielgruppe	<i>age of audience</i>	myop.audience.age (*)
Vorwissen der Zielgruppe	<i>knowledge of audience</i>	myop.audience.knowledge(*)
Ansehen	<i>authority</i>	myop.authority (*)
Informationstiefe	<i>depth of information</i>	myop.accuracy.depth (*)
Informationsbreite	<i>width of information</i>	myop.accuracy.width (*)
Bewerter	<i>rater</i>	rating.creator.person
Bewertungs-Agentur	<i>rating service</i>	rating.creator.organisation
Label Büro	<i>label bureau</i>	rating.creator.publisher
Sprache der Bewertung	<i>rating language</i>	rating.content.language (*)
Datum der Bewertung	<i>date of rating</i>	rating.version.last_modified (*)
Lebensdauer d. Bewertung	<i>time to live</i>	rating.version.time_to_live (*)
Signatur	<i>signature</i>	rating.tech.signature (*)

Tabelle 7.2: Die Attribute von xQMS sind in beschreibende und bewertende Attribute eingeteilt (oberer und unterer Bereich der Tabelle). Die mit einem Stern gekennzeichneten Attribute folgen entweder fixen Normen (wie z.B. Datumsangaben), oder haben einen eingeschränkten Wertevorrat.

7.1 Beschreibung der Attribute

7.1.1 Creator

In Abschnitt 6.5.1 wurde gezeigt, dass Angaben zum Ersteller von Ressourcen zu den wichtigsten beschreibenden Attributen zählen. In xQMS besteht die Kategorie *Creator* aus den drei Attributen *Person*, *Organisation* und *Publisher*. *Person* und *Organisation* beschreiben jene Person bzw. Organisation, die primär für die Erstellung des intellektuellen Inhalts der Ressource verantwortlich ist. Dies sind Autoren im Falle eines geschriebenen Texts, und Künstler, Fotografen oder Zeichner bei visuellen Ressourcen. Eine Organisation wäre beispielsweise ein Universitätsinstitut, ein Verein, eine politische Partei oder ein Unternehmen. *Publisher* ist der Name jener Einheit, die für die Veröffentlichung der Ressource verantwortlich ist. [DublinCore99]

Alle drei Eigenschaften sind frei wählbare Textfelder, welche Mehrfachnennungen erlauben. In diesem Fall sind die Angaben durch Semikolons getrennt aufzulisten. Im Personenfeld wird zuerst der Nachname genannt, und dann – durch einen Beistrich getrennt – der oder die Vornamen. Beim Feld zur Angabe der Organisation kann man zuerst die Dachorganisation, und – wieder durch Semikolons getrennt – weitere „Unter“-Organisationseinheiten nennen. Zusätzlich kann auch eine E-Mail Adresse, in spitzen Klammern eingeschlossen, angegeben werden. Das Beispiel in Listing 7.9 soll dies verdeutlichen:

```
xQMS.creator.person="Weitzer, Johann <jweitzer@iicm.edu>"
xQMS.creator.organisation="TU Graz; Institut für
Informationstheorie und Computergestützte Neue Medien"
xQMS.creator.publisher="J.UCS"
```

Listing 7.9: Beispielhafte Verwendung der Creator Eigenschaften.

7.1.2 Identifier

Es liegt auf der Hand, dass die eindeutige Identifikation von Dokumenten wichtig bei der Wiederauffindung ist (vgl. ISBN Buchnummer). Mit der Eigenschaft Identifier werden Angaben über den Ort der Ressource oder eine Methode, welche in einer Ortsangabe resultiert, gemacht (vgl. Abschnitt 6.5.1). Das inkludiert u.a. URLs, eine URN, eine URI oder ein DOI¹ (vgl. Listing 7.10). Die Reihenfolge bei mehreren Ortsangaben (welche durch Semikolons voneinander getrennt sein

¹ *Uniform Resource Locator*, *Uniform Resource Name* und *Uniform Resource Identifier* werden in [W3C Naming] beschrieben. Angaben zum *Digital Object Identifier* finden sich in [DOI] Vgl. Glossar

müssen) kann dazu benutzt werden, den bevorzugten Standort zu kennzeichnen. Das entsprechende Attribut in [LOM98] heißt General-Identifizier.

```
xQMS.tech.ident="http://www.w3.org/Addressing/"
```

Listing 7.10: Beispiel für den Einsatz der Identifizier Eigenschaft.

7.1.3 Scope of description

In Abschnitt 6.5.1 wurde erwähnt, dass es sinnvoll wäre, Beschreibungen nicht nur für Webseiten, sondern auch für Bereiche oder ganze Server tätigen zu können. In xQMS kann man daher den Gültigkeitsbereich der Beschreibung nach Tabelle 7.3 festlegen. So kann man später, wenn ein Dokument keine eigene Bewertung besitzt, die Bewertung einer Ressource heranziehen, in welcher die ursprüngliche als Teilmenge enthalten ist. In der umgekehrten Richtung gilt das selbverständlich nicht.

Interessiert man sich für eine Webpage, findet aber keine Bewertung dieses Ressource Typs, so kann man die Bewertung des Bereichs abfragen bzw. für die Suchverknüpfung heranziehen. Wird man auch dort nicht fündig, sucht man die Bewertung des Servers.

Gültigkeit der Beschreibung	scope of description	Intern
nicht angegeben	not specified	0
Teil einer Seite	page section	1
Webpage	web page	2
Webbereich	web area	3
Server	server	4

Tabelle 7.3: Aufzählung der Gültigkeitsbereiche von Beschreibungen. Es sei darauf hingewiesen, dass hier die Werte 'nicht passend', 'unbekannt' und 'alle' nicht zulässig sind, weil immer einer der Werte zutreffen muss. Bei der Testimplementierung, die in Kapitel 8 beschrieben wird, ist die Auswahl des Teiles einer Seite nicht möglich.

7.1.4 Signatur

Im Abschnitt 6.5.1 wurde auf die Problematik hingewiesen, die entsteht, wenn für ein Dokument eine Beschreibung bzw. Bewertung vorliegt, und das Dokument danach verändert wird. Das kann natürlich völlig rechtlich geschehen, aber

auch vorsätzlich mit einer schlechten Absicht. Solche Veränderungen kann man einerseits am Datum erkennen (vgl. 7.1.16), andererseits dient dazu das Feld Signatur (*tech.signature*), mit dem es weiters möglich ist, den Autor eindeutig zu identifizieren. [W3C PICS]

Um eine Ressource zu signieren, benötigt der Autor ein *Public Key Pair*. Der eine Schlüssel, der *Private Key*, muss vom Autor geheim gehalten werden, während der *Public Key* jedem zugänglich gemacht werden muss, der die Signatur überprüfen will. Nach Fertigstellung der Ressource wird mit einem speziellen Algorithmus (MD5) eine Checksumme berechnet, der sogenannte *Message Digest*, und mit dem *Private Key* verschlüsselt. Dieser verschlüsselte Digest ist die digitale Signatur, welche nach der Konvertierung in eine lesbare Form in einem Metadaten Attribut gespeichert wird. Wenn der Leser einer Ressource die Signatur überprüfen will, wird der Vorgang in umgekehrter Reihenfolge wiederholt. Zuerst wird aus den lesbaren Zeichen die digitale Signatur wiederhergestellt. Dann wird sie mit dem *Public Key*, der dem Empfänger zur Verfügung steht, decodiert. Der resultierende Digest wird schließlich noch mit dem aktuellen Digest verglichen. Stimmen die beiden überein, kann man sich über die Identität und die ursprüngliche und unveränderte Form der Ressource sicher sein. [W3C PICS]

Bei der Erstellung von Signaturen ist zwischen statischen und dynamischen Inhalten zu unterscheiden. Statische Inhalte bleiben über längere Zeit konstant, wohingegen sich der Inhalt dynamischer Ressourcen ständig ändern kann. Über die letztgenannte Art von Ressourcen lassen sich sinnvollerweise keine Signaturen bilden, weil sie kurze Zeit nach Erstellung nicht mehr zum Inhalt passen und als ungültig erachtet würden. Da ganze Server oder Serverbereiche sich üblicherweise ständig verändern, und somit dynamische Ressourcen darstellen, werden sie nicht signiert. Das Attribut ist also nur auf statische Webseiten und -abschnitte anzuwenden. Die Signatur beruht in diesem Fall auf dem HTML Code der zu bewertenden Seite.

Eine Alternative zur Signatur des Inhaltes würde die Authentifizierung bzw. Zertifizierung des Autors darstellen. Benutzer können dann den Autor eindeutig identifizieren, und ein Zertifikat über seine Zuverlässigkeit oder Vertrauenswürdigkeit erhalten. Mangels genügender Verbreitung von Zertifizierungsstellen wird in xQMS zugunsten des beschriebenen Signatur Feldes für statische Ressourcen entschieden.

7.1.5 Type

In Abschnitt 6.5.1 wird erwähnt, dass bei elektronischen Ressourcen der Typ von großem Interesse ist: Handelt es sich bei einer Ressource um ein Bild, ein Video, einen Text oder um Software? Mit Type lassen sich die Art und Gattung des

Inhalts einer Ressource kategorisieren. Dazu wählt man einen Wert aus einem kontrollierten Vokabular (Tabelle 7.4; Mehrfachnennungen sind nicht möglich). Diese Tabelle wurde von Dublin Core übernommen, jedoch ohne die Kategorien 'physical object' und 'interactive resource'. Eine genaue Beschreibung der ursprünglichen Kategorien findet sich in Anhang B.2. Das Beispiel in Listing 7.11 zeigt die Typisierung der Fotografie eines Notenblattes von Mozart.

`xQMS.content.type=12`

Listing 7.11: Auch wenn es sich in dem Beispiel um eine Fotografie handelt, ist der Typ der Ressource dennoch Audio (sound). Denn das Hauptinteresse gilt im Falle Mozarts dem Notenblatt, welches der Musik zugeordnet wird.

Typ	type	Intern
nicht passend	not applicable	-3
unbekannt	unknown	-2
alle	all	-1
nicht angegeben	not specified	0
Sammlung	collection	1
Datenmenge	dataset	2
Ereignis	event	3
Visuell	image	4
Modell	model	6
Person, Organisation, Gruppe	party	7
Ort	place	9
Service	service	10
Software	software	11
Audio	sound	12
Text	text	13

Tabelle 7.4: Typisierung von Ressourcen angelehnt an [DublinCore99]

7.1.6 Title

Neben einer eindeutigen Bezeichnung spielt selbstverständlich der Titel eine wesentliche Rolle in der Beschreibung von Ressourcen (vgl. Abschnitt 6.5.1). Im Kontext eines Abschnitts würde *Title* beispielsweise die Bildunterschrift oder Abschnittsüberschrift sein. Von einer Webpage der Titel (oft die grösste vorkommende Überschrift), bei einem Bereich möglicherweise der Titel einer *default-*bzw. *index-page*. Und im Kontext eines Servers würde dieser Name vom Betreiber vergeben, meist wird er etwas mit dem Domain-Namen zu tun haben. Die

Eigenschaft ist ein Textfeld, welches man frei gestalten kann (vgl. Listing 7.12). Es muß nicht unbedingt mit eventuell schon vorhandenen Metadaten (META Tag oder DC) übereinstimmen, könnte aber auch automatisch von diesen übernommen werden. Solche Daten können aber vom Interface einer Software zum Bewerten herangezogen werden, um dem Benutzer einen Vorschlag zu machen.

```
xQMS.content.title="Österreichisches Wörterbuch" xQMS.tech.scope=2
xQMS.content.title="Technische Universität Graz" xQMS.tech.scope=4
```

Listing 7.12: In der ersten Zeile wird das Attribut Title bezogen auf eine Webseite. Die zweite Zeile zeigt die Verwendung in Bezug auf einen Server.

7.1.7 Version

Wie im letzten Kapitel erwähnt ist es bei Online Ressourcen leicht möglich, dass viele verschiedene Versionen eines Dokuments gleichzeitig zugänglich sind (vgl. Abschnitt 6.5.1). Hier lässt sich eine Angabe über die Version einer Ressource tätigen. Der Vergleich von Datumfeldern ist nicht nur umständlicher, sondern führt mitunter zu Fehlern bei der Feststellung von Versionen. Denn auch ein Objekt mit jüngerem Erstelldatum kann eine veraltete Version des gleichen, aber aktuellen Objekts mit älterem Erstelldatum sein. Die Angabe der Versionsnummer erfolgt als Zahl mit beliebig vielen Kommastellen. Es ist weiters möglich, auf vorherige und folgende Versionen (falls vorhanden) durch Angabe eines Bezeichners, wie z.B. einer URL, zu verweisen. Wobei bei diesen beiden Feldern nur eine Einfachnennung erlaubt ist (vgl. Listing 7.13). So ist es einem Autor möglich, in jedem seiner Dokumente auf die jeweilige neuere Version hinzuweisen. Für den Konsumenten wird dadurch ein Nachvollziehen der Entstehung und Entwicklung einer Ressource möglich.

```
xQMS.version.number="3.141"
xQMS.version.prev="http://www.archiv.com/dok2.html"
xQMS.version.next="http://www.documents.com/dok4.html"
```

Listing 7.13: Demonstration der Verwendung von Versions Attributen. Das der Beschreibung zu Grunde liegende Objekt hat die Versionsnummer 3.141. Mit der Angabe der ersten URL wird auf eine ältere Version verwiesen, mit der Angabe der zweiten URL auf eine neuere Version. Es ist aber nicht gesagt, dass diese neuere Version tatsächlich die Aktuellste ist. Hierzu müsste man den Link verfolgen, und nachsehen, ob in dem Dokument dok4.html eine weitere, neuere Version genannt wird usw.

7.1.8 Subject

Aus den gemachten Untersuchungen der Abschnitte 6.5 bis 6.5.1 geht die Wichtigkeit der Klassifikation des Themas hervor. Das Attribut zur Beschreibung des Themas besteht aus den fünf Feldern *Topic*, *Alternative Topic*, *Classification Name*, *Classification Identifier* und *Keywords* (vgl. Listing 7.14).

Im ersten Feld lassen sich die Themen einer Ressource in einem Textfeld angeben. Da die Auswertung automatisch erfolgen soll, kommt hier eine genormte Themencharakterisierung zum Einsatz, der *Dewey Decimal Code (DDC)*.

Da es aber mitunter vorteilhaft sein kann, auch ein anderes System der Themenklassifikation zu verwenden, wie z.B. die spezielle Charakterisierung von Computerthemen durch ACM², hat man mit den Feldern *Alternative Topic*, *Alternative classification scheme name* und *classification scheme identifier* die Möglichkeit, genau diese zusätzlichen Informationen auszudrücken. Alle drei Felder sind Textfelder, wobei die letzten beiden angeben, wie das Klassifikationsschema heißt und wo eine Referenz desselben zu finden ist (z.B. in Form einer URL).

Das Feld *Keywords* drückt das Thema der Ressource durch frei wählbare Schlüsselwörter aus. Es muß nicht unbedingt mit eventuell schon vorhandenen Metadaten (beispielsweise dem HTML META Tag „keywords“) übereinstimmen, könnte aber auch automatisch von diesen übernommen werden, und dazu dienen, dem Benutzer eines Bewertungssystems einen Vorschlag zu machen.

```
xQMS.content.topic="004, 020"
xQMS.content.alt_topic="H.3.3, H.3.4"
xQMS.content.alt_classname="ACM"
xQMS.content.alt_classident="http://www.acm.org/class/1998/"
xQMS.content.keywords="retrieval, metadata, content, quality,
filtering, scheme"
```

Listing 7.14: Das Beispiel zeigt eine Ressource, welche das Thema *Data Processing* (Dewey Decimal Code 004) und *Library and information sciences* (DDC 020) behandelt. Eine weitere Themenangabe wird über das Klassifikationsschema von ACM gemacht. In diesem Schema bezeichnet H.3.3 *Information Search and Retrieval* und H.3.4 steht für die Thematik *Systems and Software*.

7.1.9 Description

Zur Kategorie der den Inhalt beschreibenden Attribute zählt auch die Beschreibung. Sie soll eine Darstellung des Inhalts der Ressource in einem freien Textfeld

²<http://www.acm.org> [ACM99]

darstellen. Möglichkeiten sind – abhängig vom Gültigkeitsbereich einer Ressource – eine Kurzfassung, ein Inhaltsverzeichnis, Inhaltsbeschreibungen visueller Ressourcen oder eine Darstellung des Inhalts in einer freien Form. [DublinCore99] Ist das Feld im Kontext eines Servers zu setzen, steht hier die Serverbeschreibung. In Listing 7.15 folgt ein Beispiel.

```
xQMS.content.type=4
xQMS.tech.scope=1
xQMS.content.description="Das Bild zeigt die schematische
Darstellung des verteilten Konzepts des xFIND Suchsystems."
```

Listing 7.15: Das Beispiel beschreibt den Inhalt des Teils einer Seite (der Gültigkeitsbereich lautet 'page section (1)'). Dieser Teil ist vom Typ 'image (4)' und die Beschreibung ist im Wesentlichen nichts anderes als die ursprüngliche Bildunterschrift (vgl. Abbildung 6.1).

7.1.10 Date

Wie im üblichen Buchwesen, ist die Angabe eines Datums zur Kennzeichnung von Ressourcen wichtig (vgl. Abschnitt 6.5.1). Dabei ist eine 8 Ziffern Nummer in der Form YYYY-MM-DD wie in ISO 8601³ definiert, zu verwenden. In diesem Schema korrespondiert das Datenelement 1999-08-16 mit dem 16. August 1999. Es sind zwei verschiedene Felder vorgesehen: [DublinCore99]

version.create kennzeichnet das Datum, an welchem eine Ressource ursprünglich erstellt wurde. Im Kontext eines Servers wird es das Datum der Inbetriebnahme sein, ebenso für eine Website und darin enthaltene Dokumente. *version.last_modified* kennzeichnet das Datum der letzten Aktualisierung, und damit gleichzeitig jenes Datum, an welchem die Ressource in der derzeit verfügbaren Form zugänglich gemacht wurde. Daran lässt sich auch ablesen, wie groß die Aktualität einer Ressource ist (vgl. Listing 7.16).

```
xQMS.version.create="1999-09-02"
xQMS.version.last_modified="2000-03-13"
```

Listing 7.16: Die Ressource des Beispiels wurde am 2. November 1999 erstellt, und zuletzt am 13. März 2000 geändert.

³<http://www.w3.org/TR/NOTE-datetime>

7.1.11 Quotation hints

In Abschnitt 6.5.1 wurde erwähnt, dass es für Autoren hilfreich und sinnvoll sein kann festzuhalten, wie sein Werk zitiert werden soll (vgl. Listing 7.17). So kann ein Autor beispielsweise auch angeben, dass eventuelle Zitate sich statt auf den Text im Internet auf die gedruckte Version beziehen sollen. Andere Metadaten Schemata (vgl. [DublinCore99] und [LOM98]) besitzen stattdessen ein Attribut *Rights*, das Informationen über die Rechte am intellektuellen Inhalt speichert. Weil damit aber kein Mechanismus einhergeht, der Missbrauch vorbeugt, wird in xQMS auf dieses Attribut verzichtet. Das Zitate-Feld ist ein freies Textfeld, und eignet sich auch hervorragend dazu, im Suchergebnis angezeigt zu werden.

```
xQMS.content.quotation="Resnick, Paul, Miller, James,
PICS: Internet Access Controls Without Censorship,
Communications of the ACM, Vol. 40, No. 3, March 1997, p. 56-58,
http://www.w3.org/PICS/iacwcv2.htm"
```

Listing 7.17: Beispielhafte Verwendung der Zitiervorschrift

7.1.12 Language

Weil Dokumente im Internet – wie in Abschnitt 6.5.1 erwähnt – von Orten auf der ganzen Welt gelesen werden können, muss der Leser erfahren, in welcher Sprache sie verfasst sind, bzw. er muss bereits bei der Suchanfrage spezifizieren können, in welcher Sprache die gesuchten Inhalte vorliegen sollen. Die Sprache wird in zwei Dimensionen angegeben: Einerseits gibt der Autor hier an, in welcher Landessprache (*content.language*) der intellektuelle Inhalt der Resource verfasst ist. Der Inhalt des Feldes soll der Konvention RFC1766⁴ folgen [DublinCore99].

Andererseits können sowohl der Autor als auch der Kritiker zusätzliche Angaben über den Stil (*myop.diction*) der Sprache machen: Dazu werden Zahlen von 1 bis 5, Tabelle 7.5 folgend, verwendet (vgl. Listing 7.18). So lässt sich jeder Text von mühelos lesbar bis anspruchsvoll charakterisieren. Auf diese Weise erhält der Leser einen Hinweis darüber, für welchen Leserkreis die Ressource bestimmt ist (vgl. Attribut Audience in Abschnitt 7.1.13).

```
xQMS.content.language="en-uk"
xQMS.myop.diction="4,5"
```

Listing 7.18: Der dem Beispiel zu Grunde liegende Text ist in Britischem Englisch geschrieben, und der Ausdruck wird als 'schwierig (4)' bis 'anspruchsvoll (5)' charakterisiert.

⁴<http://puma.germany.net/internic/rfc/rfc1766.txt>

Sprachstil	diction	Intern
nicht passend	not applicable	-3
unbekannt	unknown	-2
alle	all	-1
nicht angegeben	not specified	0
müheless	effortless	1
einfach	easy	2
normal	normal	3
schwierig	difficult	4
anspruchsvoll	exacting	5

Tabelle 7.5: Typisierung des sprachlichen Ausdrucks

7.1.13 Audience

Die Charakterisierung des Leserkreises wurde in Abschnitt 2.2 als weiteres Merkmal zur Bewertung von Ressourcen ausgemacht. Die Klassifikation erfolgt in den zwei Dimensionen Alter (*myop.audience.age*) und bzw. oder Vorwissen (*myop.audience.knowledge*). Es soll hier zum Ausdruck kommen, für welchen Empfängerkreis die Ressource gedacht ist. Die Tabellen 7.6 und 7.7 lassen eine Einteilung vom Kind bis zum Senioren und vom Anfänger bis zum Experten zu (vgl. Listing 7.19).

Dabei ist zu beachten, dass diese Klassifikation im Zusammenhang mit dem *Hauptthema* der Ressource zu sehen ist, welches den Feldern *content.topic* und *content.keywords* sowie *content.description* zu entnehmen ist. Weiters ist zu beachten, dass die Felder nicht voneinander abhängen. Meist ist es sogar sinnvoll, nur eines der beiden Attribute anzugeben. Der Kritiker bringt hier zum Ausdruck, für welchen Leserkreis nach seiner Meinung die Ressource geeignet ist (appropriateness). [Rettig96]

```
xQMS.myop.audience.age="-3"
xQMS.myop.audience.knowledge="3"
```

Listing 7.19: Das Beispiel zeigt, dass die betreffende Ressource primär für Experten auf dem betrachteten Gebiet gedacht ist. Das Alter spielt keine Rolle.

7.1.14 Authority

Informationen, welche über das Internet bereitgestellt werden, stammen oft von einer fragwürdigen Herkunft. Wie bereits in Abschnitt 2.2 ausgeführt, durchlaufen sie nicht die gleichen rigorosen Rezensions-prozeduren wie dies bei anderen

Alter	age	Intern
nicht passend	not applicable	-3
unbekannt	unknown	-2
alle	all	-1
nicht angegeben	not specified	0
Vorschulalter	preschool age	1
Kinder	children	2
Jugendliche	youth	3
Junge Erwachsene	twens	4
Erwachsene	adult	5
Senioren	senior citizens	6

Tabelle 7.6: Alter der Zielgruppe

Vorwissen	knowledge	Intern
nicht passend	not applicable	-3
unbekannt	unknown	-2
alle	all	-1
nicht angegeben	not specified	0
Anfänger	beginner	1
Fortgeschrittene	advanced	2
Experte, Spezialist	expert	3

Tabelle 7.7: Vorwissen der Zielgruppe

Publikationskanälen üblich ist. Aber da man über die Verlässlichkeit und Qualifikation eines Autors oft keine genauen Aussagen treffen kann, beschränkt man sich bei der Bewertung der Glaubwürdigkeit auf die Ressource selbst. [SC94] [Legenstein99] Es ist schwierig bis unmöglich, für jede Art von Ressource eine geeignete Klassifikation des Ansehens zu finden. In xQMS werden nur Dokumente mit wissenschaftlichem Charakter hinsichtlich ihres Ansehens bewertet. Die Einteilung erfolgt nach Tabelle 7.8

7.1.15 Accuracy

In Abschnitt 2.2 wird erläutert, wie die Genauigkeit einer Ressource durch die Breite (*width*) und Tiefe (*depth*) der Information charakterisiert werden kann. Die Breite von Information ist ein Maß für die Anzahl verschiedener Themen innerhalb eines Themengebietes. Die Tiefe spezifiziert gleichzeitig den Grad an Details innerhalb eines Themas (vgl. Abbildung 7.1). [Legenstein99] [Smith97]

Texttyp / type of text	Intern
nicht passend	-3
unbekannt	-2
alle	-1
nicht angegeben	0
inhaltlos	1
Diskussionspunkt in einer Newsgroup oder Mailinglist	2
Allgemeine Zeitungs- oder Journal Artikel	3
Populär-Wissenschaft von Wissenschaftler verfasst	4
Andere schulische, wissenschaftliche oder techn. Texte	5
Wissenschaftlicher Forschungsbericht	6
Akzeptiert für Präsentation auf wissenschaftl. oder techn. Konferenz	7
Akzeptiert für Publikation in wissenschaftl. oder techn. Journal	8
Doktorarbeit oder Vergleichbares	9
<i>not applicable</i>	-3
<i>unknown</i>	-2
<i>all</i>	-1
<i>not specified</i>	0
<i>junk</i>	1
<i>common newspaper or journal article</i>	2
<i>discussion item in newsgroup, mailing list or other on-line forum</i>	3
<i>popular science written by a scientist</i>	4
<i>other scholarly scientific or technical text</i>	5
<i>scientific research report</i>	6
<i>accepted for presentation at scientific or technical conference</i>	7
<i>accepted for publication in peer-reviewed scientific or technical journal</i>	8
<i>ph.d. thesis or equivalent</i>	9

Tabelle 7.8: Charakterisierung der Glaubhaftigkeit von wissenschaftlichen Dokumenten nach [Palme98] in Deutsch und Englisch.

In xQMS beziehen sich die Angaben von Informationstiefe und -breite auf das Thema der Ressource, welches den Feldern *content.topic*, *content.keywords* und *content.description* zu entnehmen ist.

Angenommen die Themenangabe einer Ressource lautet „Naturwissenschaft - Physik - Magnetismus“, was dem Dewey Decimal Code 538 entspricht. In diesem Fall würde eine hohe Tiefe bedeuten, dass das Thema Magnetismus sehr detailliert und genau behandelt wird. Umgekehrt zeigt eine geringe Tiefe an, dass die Thematik oberflächlich behandelt wird. Eine hohe Informationsbreite bedeutet auf das Beispiel bezogen, dass auch Themen rund um das Thema Magnetismus behandelt werden, wohingegen eine sehr geringe Breite aussagt, dass vielleicht

nur ein Bereich der Thematik behandelt wird (vgl. Listing 7.20). Man beachte, dass große Breite nicht Detailreichtum ausschließt, oder geringe Breite gleichzeitig ein hohes Maß an Tiefe bedeutet! Die Angabe der Genauigkeit erfolgt anhand der Tabellen 7.9 und 7.10.

```
xQMS.myop.accuracy.width="1, 2"
xQMS.myop.accuracy.depth="3, 4"
```

Listing 7.20: In dem Beispiel wird eine Ressource beschrieben, welche sich mit einem Detail oder speziellen Thema befasst, und dieses (sehr) ausführlich behandelt. Bezogen auf die Themenangabe durch den Dewey Decimal Code 538 (Physik) würde es beispielsweise bedeuten, dass die Ressource den Spin der Elektronen (Detail) mit vielen Formeln und Berechnungen (sehr ausführlich) behandelt.

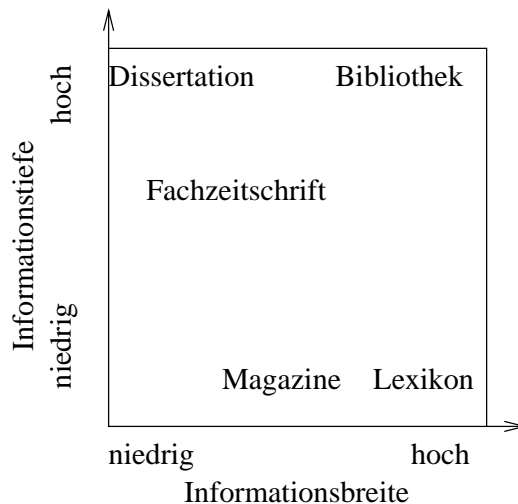


Abbildung 7.1: Einteilung verschiedener Ressourcen in eine Matrix von Informationstiefe und -breite.

7.1.16 Rating

In Abschnitt 6.5.1 wurde festgehalten, dass man bei Bewertungen auch wissen sollte, von wem sie stammen. Auch in xQMS muß der Bewerter Angaben über die Bewertung selbst vornehmen, also Metadaten über Metadaten (Meta-Metadaten) eingeben. Der Bewerter kann beim Ausfüllen der Felder sehr leicht von der Benutzeroberfläche unterstützt werden, weil sich der Inhalt der Felder kaum ändert, und vom System automatisch feststellbar ist. Die Attribute über das Rating folgen denselben Regeln wie in den oben aufgeführten Feldern.

Breite	width	Intern
nicht passend	not applicable	-3
unbekannt	unknown	-2
alle	all	-1
nicht angegeben	not specified	0
geringste Breite, Detail	lowest width, detail	1
geringe Breite, spezielles Thema	low width, special topic	2
mittlere Breite, Themenauswahl	average width, selection of topics	3
große Breite, sehr viele Themen	high width, many topics	4

Tabelle 7.9: Typisierung der Informationsbreite

Tiefe	depth	Intern
nicht passend	not applicable	-3
unbekannt	unknown	-2
alle	all	-1
nicht angegeben	not specified	0
oberflächlich	superficial	1
guter Überblick	good overview	2
ausführlich	detailed	3
sehr ausführlich	very detailed	4

Tabelle 7.10: Typisierung der Informationstiefe

Das Attribut *rating.creator.person* bezeichnet die Person, welche die Bewertung durchführt, *rating.creator.organisation* gibt die Bewertungs-Agentur an, und *rating.creator.publisher* jene Organisation, welche die Bewertung zugänglich macht, also ein Label-Büro. Oft werden die beiden aber ident sein.

Die Eigenschaft *rating.version.last_modified* kennzeichnet das Datum, an welchem die Bewertung zugänglich gemacht wird. Liegt das Datum vor jenem in *version.last_modified* (das Datum der letzten Änderung der zu bewertenden Resource), kann davon ausgegangen werden, dass die Bewertung veraltet ist, und möglicherweise nicht mehr zutrifft. Dasselbe ist der Fall, wenn die Bewertung eine gewisse Zeit alt ist (Nämlich *rating.version.time_to_live* Tage nach der letzten Änderung der Bewertung). *rating.content.language* schließlich stellt fest, in welcher Sprache die Bewertung erfolgt.

Eine Schwierigkeit bei Bewertungen ist, dass sich der Endnutzer sicher sein muss, von wem die Bewertung durchgeführt wurde, und dass die Bewertung seit ihrer Erstellung nicht manipuliert wurde. Dies kann wieder mit einer digitalen Signatur (*rating.tech.signature*) geschehen. [W3C PICS]

7.2 Verwendung der xQMS Attribute

Die Verwendung der xQMS Attribute (ausgenommen der Signaturen) wird anhand von drei Beispielen demonstriert. Beim ersten Beispiel erstreckt sich die Bewertung über eine Webseite (s. Tabelle 7.11), beim zweiten über einen Bereich (s. Tabelle 7.12), und beim dritten Beispiel über einen Server (s. Tabelle 7.13).

7.3 Zusammenfassung

In Abschnitt 2.3.2 und in Kapitel 6 wurde erläutert, wie man mit Hilfe von Metadaten den Prozess des Filterns von relevanten Informationen aus einer großen Datenmenge bewerkstelligen kann. Ebenso wurde in den Kapiteln 1 und 2 dargestellt, wie wichtig die Suche nach qualitativen Inhalten wird.

Aus diesem Grund war es das Ziel dieses Kapitels, ein Qualitäts-Metadatenchema aufzustellen, mit dessen Hilfe die verknüpfte Suche nach den gewünschten Inhalten in entsprechender Qualität möglich wird. Auf Grund der in den vorangegangenen Kapiteln angestellten Untersuchungen, wurde ein System von 34 Attributen aufgestellt. Diese lassen sich grob in Eigenschaften unterteilen, die Ressourcen beschreiben, und solche, die sie bewerten.

Bei der Auswahl der Attribute war es wichtig, möglichst viele Eigenschaften davon computer-interpretierbar zu gestalten. Das bedeutet, dass die Werte dieser Eigenschaften speziellen Normen folgen müssen, oder einen vorgegebenen, eingeschränkten Wertevorrat besitzen. Dadurch wird eine automatische Verarbeitung der Metadaten möglich.

Die Definition der Attribute erfolgte hierarchisch, indem jedes Attribut einer Kategorie angehört. Diese Kategorien charakterisieren den Ersteller, technische Einzelheiten, den Inhalt und die Version. Sie bewerten weiters die Sprache, den Leserkreis, die Genauigkeit und das Ansehen. Zuletzt ist die Kategorie der Attribute zu erwähnen, die die Bewertung an sich beschreiben.

Im nächsten Kapitel der Arbeit wird die Entwicklung einer Testimplementierung beschrieben, welche die Anmeldung und Bewertung von Ressourcen erlaubt, und diese Qualitätsmetadaten über eine Schnittstelle dem xFIND System zur Verfügung stellt.

Attribut	Wert
creator.person	Rettig, James <jrettig@mail.swem.wm.edu>
creator.organisation	College of William and Harry
creator.publisher	Online Inc. <webmaster@onlineinc.com>
tech.ident	www.onlineinc.com/onlinemag/SeptOL/rettig9.html
tech.scope	2 [Webseite]
content.type	13 [Text]
content.title	Beyond „Cool“
content.topic	020 [library and information sciences]
content.alt_topic	H.3.7 [digital libraries]
content.alt_classname	ACM
content.alt_classident	http://www.acm.org/class/1998/
content.keywords	Bibliotheken, Digitale Ressourcen, Metadaten
content.description	Der Autor ist der Ansicht, dass...
content.quotation	Rettig, J.:Beyond Cool. Analog Models for...
content.language	en-us
version.prev	www.onlineinc.com/onlinemag/SeptOL/rettig8.html
version.last_modified	1996
myop.diction	3 [normal]
myop.audience.age	-3 [nicht passend]
myop.audience.knowledge	2, 3 [Fortgeschritten, Experte]
myop.authority	5 [Wissenschaftl. Text]
myop.accuracy.depth	2 [guter Überblick]
myop.accuracy.width	2 [geringe Breite, spezielles Thema]
rating.creator.person	Weitzer, Johann <jweitzer@iicm.edu>
rating.creator.organisation	TU Graz
rating.creator.publisher	TU Graz, Bibliothek
rating.content.language	de [Deutsch]
rating.version.last_modified	2000-03-13
rating.version.time_to_live	365

Tabelle 7.11: Beispiel zur Verwendung der xQMS Attribute für Webseiten. Bei der Ressource handelt es sich um die Arbeit [Rettig96]. Nach Ansicht des Bewerters ist diese Ressource für Menschen ab ca. 20 Jahren geeignet, die über ein fortgeschrittenes oder Experten-Wissen auf dem Gebiet des Bibliothekswesens verfügen.

Attribut	Wert
creator.person creator.organisation creator.publisher	Giordano, Gillian; Weinberg Tracy Eastchester Middle School EMS
tech.ident tech.scope	http://www.westnet.com/rickd/AIDS/ 3 [Webbereich]
content.type content.title content.topic content.alt_topic content.alt_classname content.alt_classident content.keywords content.description content.quotation content.language	13 [Text] The AIDS Handbook 614 [Incidence and prevention of disease] AIDS, kids, prevent, HIV Written By Middle School Kids For Middle School Kids http://www.westnet.com/rickd/AIDS/ en
version.number version.prev version.next version.create version.last_modified	1997
myop.diction myop.audience.age myop.audience.knowledge myop.authority myop.accuracy.depth myop.accuracy.width	1 [müheles] 2, 3 [Kinder und Jugendliche] 1 [Anfänger] -3 [nicht passend] 2 [guter Überblick] 3 [mittlere Breite, Themenauswahl]
rating.creator.person rating.creator.organisation rating.creator.publisher rating.content.language rating.version.last_modified rating.version.time_to_live	Weitzer, Johann <jweitzer@iicm.edu> TU Graz TU Graz, Bibliothek de [Deutsch] 2000-03-30 200

Tabelle 7.12: In dem Beispiel handelt es sich um einen Webbereich, der über das Thema AIDS aufklärt. Die enthaltenen Informationen sind für Kinder und Jugendliche gedacht, welche noch wenig über das Thema wissen. Der Text ist in Englisch geschrieben, und müheles lesbar.

Attribut	Wert
creator.person	Gütl, Christian <cguetl@iicm.edu>
creator.organisation	Graz University of Technology; IICM
creator.publisher	Graz University of Technology; IICM
tech.ident	http://xfind.iicm.edu
tech.scope	4 [Webserver]
content.type	11 [Software]
content.title	Extended Framework for Information Discovery
content.topic	004, 020 [data proc., library/information sciences]
content.alt_topic	H.3.3, H.3.4 [information search/retrieval]
content.alt_classname	ACM
content.alt_classident	http://www.acm.org/class/1998/
content.keywords	xFIND, distributed search, metadata
content.description	A search engine is developed that. . .
content.quotation	http://xfind.iicm.edu
content.language	en-us
version.number	1
version.last_modified	1999-09-20
myop.diction	3, 4 [normal, schwierig]
myop.audience.age	-3 [nicht passend]
myop.audience.knowledge	2, 3 [Fortgeschritten, Experte]
myop.authority	6 [Wissenschaftl. Forschungsbericht]
myop.accuracy.depth	2, 3 [guter Überblick, ausführlich]
myop.accuracy.width	2 [geringe Breite, spezielles Thema]
rating.creator.person	Weitzer, Johann <jweitzer@iicm.edu>
rating.creator.organisation	TU Graz
rating.creator.publisher	TU Graz, Bibliothek
rating.content.language	en [Englisch]
rating.version.last_modified	2000-03-30
rating.version.time_to_live	100

Tabelle 7.13: Beispiel zur Verwendung der xQMS Attribute für Server. Die Sprache der Bewertung ist Englisch, und sie wurde am 30. März 2000 erstellt. Sie soll weiters 100 Tage gültig sein.

Kapitel 8

Testimplementierung QMRatingSystem

Im Gestaltungsbereich der Arbeit wurde in Kapitel 7 das Qualitäts-Metadaten-Schema xQMS entworfen, mit dessen Hilfe man Ressourcen hauptsächlich hinsichtlich ihrer inhaltlichen Qualität bewerten kann. Es wurden dabei 34 Attribute definiert, die Ressourcen beschreiben und bewerten können sollen. Zwischen einem tatsächlich durchgeführten Suchvorgang mit Hilfe der Qualitätsmetadaten durch Benutzer eines Suchsystems und dem Aufstellen eines Bewertungsschemas liegt aber noch der eigentliche Prozess des Bewertens. Im Abschnitt 6.5.2 wurde bereits erläutert, wie man die Arbeit der Bewertung sinnvollerweise auf die Autoren, Administratoren von Suchsystemen und Fachexperten verteilen kann. Einige Attribute sollten auch automatisiert vergeben werden können.

In diesem Kapitel wird eine Testimplementierung eines Bewertungssystems vorgestellt, das auf den Attributen von xQMS basiert, und in das xFIND Suchsystem (vgl. Kapitel 6.4) integriert werden kann. Mit Hilfe dieses Systems soll es Autoren von Webseiten oder Besitzern von Servern möglich sein, ihre Ressourcen am Suchsystem anzumelden, und gleichzeitig Qualitätsmetadaten zu vergeben. Die Administratoren des Systems sollen eventuelle Änderungen vornehmen, und die Ressourcen von dem Suchdienst indizieren lassen können. In einem letzten Schritt können auch noch Experten die Bewertungen überprüfen und überarbeiten. Auf diese Weise soll die Qualität der Metadaten schrittweise verbessert werden. Die Bewertungen werden über eine Schnittstelle dem xFIND Suchsystem zur Verfügung gestellt, wo sie verarbeitet werden, und letztendlich eine verknüpfte Suche nach Inhalt und Qualitätsmetadaten möglich machen.

Die Testimplementierung – QMRatingSystem genannt – ist, wie das xFIND Suchsystem selbst, in Java¹ programmiert. Diese Programmiersprache erlaubt ein einfaches Übertragen der Anwendung auf die verschiedensten Plattformen. Das Programm wird über einen Internet Browser bedient, und kann in deutscher oder englischer Sprache betrieben werden.

8.1 Das QMRatingSystem aus der Sicht der unterschiedlichen Benutzergruppen

In Kapitel 6.5.2 wurde erläutert, dass die Bewertung von Ressourcen aus Gründen der Effizienz und Qualität auf verschiedenen Ebenen bzw. von unterschiedlichen Benutzergruppen erfolgen sollte. Das QMRatingSystem erlaubt deshalb eine dreistufige Bewertung: Autoren bzw. Webserverbetreiber können Ressourcen anmelden und vorbereiten, wonach Administratoren die Ressourcen in das System aufnehmen und die Bewertungen überarbeiten können. In der dritten Stufe bewerten Fachexperten die Ressource, und erhöhen damit die Qualität der Metadaten ein weiteres mal.

8.1.1 Anmeldung

Jeder Benutzer des Systems muss sich pro Sitzung einmal anmelden (s. Abbildung 8.1). Hierzu genügt für bereits registrierte Benutzer die Angabe der Email Adresse und des Passwortes. Benutzer, die noch nicht registriert sind, wählen den Menüpunkt 'Ich bin neu'. Man gelangt dann zu einer Maske, bei der man neben seiner Email Adresse ein paar persönliche Daten, wie Vor- und Nachname, Postleitzahl, Ort und Land eingeben muss (vgl. Abbildung 8.2). Wenn man die Daten vollständig eingegeben hat, erhält man eine Email, in der einem ein zufällig generiertes Passwort zugewiesen wird.

Nach erfolgter Anmeldung gelangt der Benutzer zum Hauptmenü. Dort kann man zwischen der Änderung seiner persönlichen Daten und den verschiedenen zulässigen Modi für den jeweiligen Benutzer wählen. Bei der Änderung der persönlichen Daten kann auch das Passwort geändert werden. Man muss hierzu nur zweimal dasselbe Passwort eingeben. Anderenfalls wird man erneut zur Eingabe aufgefordert.

¹Java ist ein eingetragenes Warenzeichen von Sun Microsystems, Inc. <http://java.sun.com>



Willkommen bei der Ressourcen Anmeldung

Um sich am System anzumelden, geben Sie bitte Ihre Email Adresse und das Passwort, welches Sie via Email erhalten haben, ein. Klicken Sie anschliessend auf 'Anmelden'. Wenn Sie neu sind, klicken Sie bitte nur auf 'Ich bin neu'.

Login:

Passwort:

Abbildung 8.1: Die Anmeldung am QMRatingSystem.

8.1.2 Der Benutzer-Modus

Im Benutzermodus kann man zwischen einer Neuanmeldung von Ressourcen und der Bearbeitung seiner bereits getätigten Anmeldungen wählen. Man kann seine Anmeldungen auch löschen. Der Administrator des Systems sieht die Ressource dann als zum Löschen markiert, und kann sie endgültig aus dem System entfernen.

Führt man eine neue Anmeldung aus, oder bearbeitet eine bereits vorhandene, gelangt man in ein Menü zur Auswahl der Ressource (s. Abbildung 8.3). Dort muss man in erster Linie den Namen des Servers angeben. Die Nummer des Ports ist mit 80 voreingestellt, lässt sich aber auch ändern. Weiters muß man wählen, ob man einen Server, einen Bereich oder eine Webseite bewerten möchte. Wenn man einen Server oder einen Webbereich anmelden will, kann man im Feld Homepage zusätzlich eine Einstiegsseite spezifizieren. Mit dem Eingabefeld Verzeichnis kann man einen Bereich wählen.

Bei diesen Feldern ist zu beachten, dass aus Sicherheitsgründen ein “..“ nicht eingegeben werden darf. Ein Benutzer könnte damit auf Verzeichnisse am Server zugreifen, die nicht zum Schreiben von Dateien vorgesehen sind. Es ist nicht möglich, Ressourcen anzumelden, die bereits von jemand anderem angemeldet wurden. Das heisst, es gibt für jede Ressource nur eine Anmeldung und Bewertung. Weiters kann man als Besitzer von Ressourcen auswählen, wie oft sie durch den Suchdienst abgesucht werden soll, und aus wievielen Seiten sie ungefähr besteht.



The image shows a web registration form with the following elements:



- Language selection: UK flag, DE flag, 'Deutsch' dropdown, and 'OK' button.
- Title: **Willkommen bei der Ressourcen Anmeldung**
- Text: 'Bitte nennen Sie uns die folgenden persönlichen Daten. Sie erhalten umgehend eine Email an die angegebene Email Adresse mit ihrem Passwort. Mit diesem Passwort können Sie sich am System anmelden.'
- Fields: Email (jweitzer@iicm.edu), Nachname (Weitzer), Vorname (Johann), Postleitzahl (8000), Ort (Graz), Staat (Österreich dropdown).
- Buttons: 'Anmelden' and 'Zurücksetzen'.

Abbildung 8.2: Neue Benutzer müssen ein paar persönlich Daten eingeben, und erhalten im Anschluß eine Email mit ihrem Passwort.

Das nächste Formular fordert zur Eingabe der xQMS Qualitätsmetadaten auf (vgl. Kapitel 7 und Abbildung 8.4). Die Anmeldung wird im Anschluss im Verzeichnis für Voranmeldungen gespeichert, wobei noch zwei weitere Kennzeichen gespeichert werden. Das eine gibt Auskunft über den Status der Anmeldung (neu, gelöscht oder akzeptiert), das andere über den Status der Bewertung (Benutzer, Administrator oder Experte).

8.1.3 Der Administrator-Modus

Der Administrator hat die umfangreichsten Rechte im System. Er erhält in seinem Menü eine Liste aller Ressourcen im Verzeichnis für Anmeldungen mit einem Symbol, das etwas über den Status der Anmeldung aussagt. Bewertungen können den Status einer Neuanmeldung haben, oder sie sind zum Löschen markiert oder akzeptiert (s. Abbildung 8.5). Er kann nun eine Ressource auswählen, und diese löschen, bearbeiten oder akzeptieren.



 Deutsch ▼

Anmeldung einer Webressource

Bitte füllen Sie das folgende Formular aus. Sie gelangen dann zur Bewertung Ihrer Ressource.

Servername: Bsp.: www.tu-graz.ac.at

Port: wichtig wenn ungleich 80!

Gültigkeit der Beschreibung:

Verzeichnis(*): z.B. /my-path/

Homepage(*): z.B. index.html

*) nur bei Webbereich

Änderungsrate:

täglich

wöchentlich

monatlich

Anzahl der Seiten:

1 bis 100

100 bis 500

500 bis 1000



1000 bis 5000

mehr als 5000

Abbildung 8.3: Die Anmeldung von Ressourcen am QMRatingSystem. Das Beispiel zeigt, wie ein Server angemeldet wird. Das Feld für den Bereich bleibt daher frei. Nur im Feld Homepage könnte man – wenn erforderlich – eine Einstiegsseite nennen.

Löscht ein Administrator eine Anmeldung, die bereits zum Löschen markiert ist, werden die Dateien endgültig entfernt, und zwar aus den Verzeichnissen für Anmeldungen und gegebenenfalls auch aus dem Verzeichnis für akzeptierte Sites. War die Datei noch nicht zum Löschen markiert, wird sie nicht sofort entfernt, sondern nur zum Löschen markiert.

Die Möglichkeit der Bearbeitung dient hauptsächlich dazu, dass der Administrator die Bewertung einer Site überprüfen kann, bevor er sie akzeptiert. Nimmt er Änderungen vor, wird auch vermerkt, dass diese durch einen Administrator erfolgt sind. Der Status der Anmeldung bleibt unverändert (entweder 'accept', 'del' oder 'new').



Deutsch

Bewertung einer Webressource

Für die optimale Performance der Suchmaschine geben Sie bitte so viele Daten und so genau wie möglich ein. Möchten Sie einen Unterbereich anmelden, so geben Sie bitte beim Bezeichner einen weiteren Pfadnamen ein.

Bezeichner	<input type="text" value="http://xfind.icm.edu/"/>		
<hr/>			
Autor	<input type="text" value="Gütl, Christian <cguetl@ic"/>		
Organisation	<input type="text" value="TU Graz; IICM"/>		
Verleger	<input type="text" value="IICM"/>		?
<hr/>			
Typ	<input type="text" value="Software"/>		?
Titel	<input type="text" value="xFIND Extended Framew"/>		?
Thema	<input type="text" value="004, 020"/>		?
Alternatives Thema	<input type="text" value="H.3.3, H.3.4"/>		
Klassifikationsschema	<input type="text" value="ACM"/>		
Klass.schema Bezeichner	<input type="text" value="http://www.acm.org.class"/>		?
Schlüsselwörter	<input type="text" value="search, knowledge, inform"/>		?
Beschreibung	<input type="text" value="Entwicklung eines Suchsy"/>		?
Sprache	<input type="text" value="Englisch"/>		
Zitiervorschrift	<input type="text" value="xfind.icm.edu"/>		?
<hr/>			
Versionsnummer	<input type="text" value="1"/>		
Vorversion	<input type="text"/>		
Aktuellere Version	<input type="text"/>		
Erstelldatum	<input type="text"/>		
Datum der letzten Änderung	<input type="text" value="1999-09-20"/>		?
<hr/>			
Sprachlicher Ausdruck	<input type="text" value="normal"/>		
Alter der Zielgruppe	<input type="text" value="Erwachsene"/>		
Vorwissen der Zielgruppe	<input type="text" value="Fortgeschrittene"/>		?
Ansehen	<input type="text" value="Wissen. Forschungsbericht"/>		?
Informationstiefe	<input type="text" value="geringe Breite, spezielles Thema"/>		
Informationsbreite	<input type="text" value="sehr ausführlich"/>		?
<hr/>			

Vielen Dank für Ihre Bewertung!

Abbildung 8.4: Zur Bewertung von Ressourcen gibt der Benutzer die xQMS Qualitätsmetadaten ein.

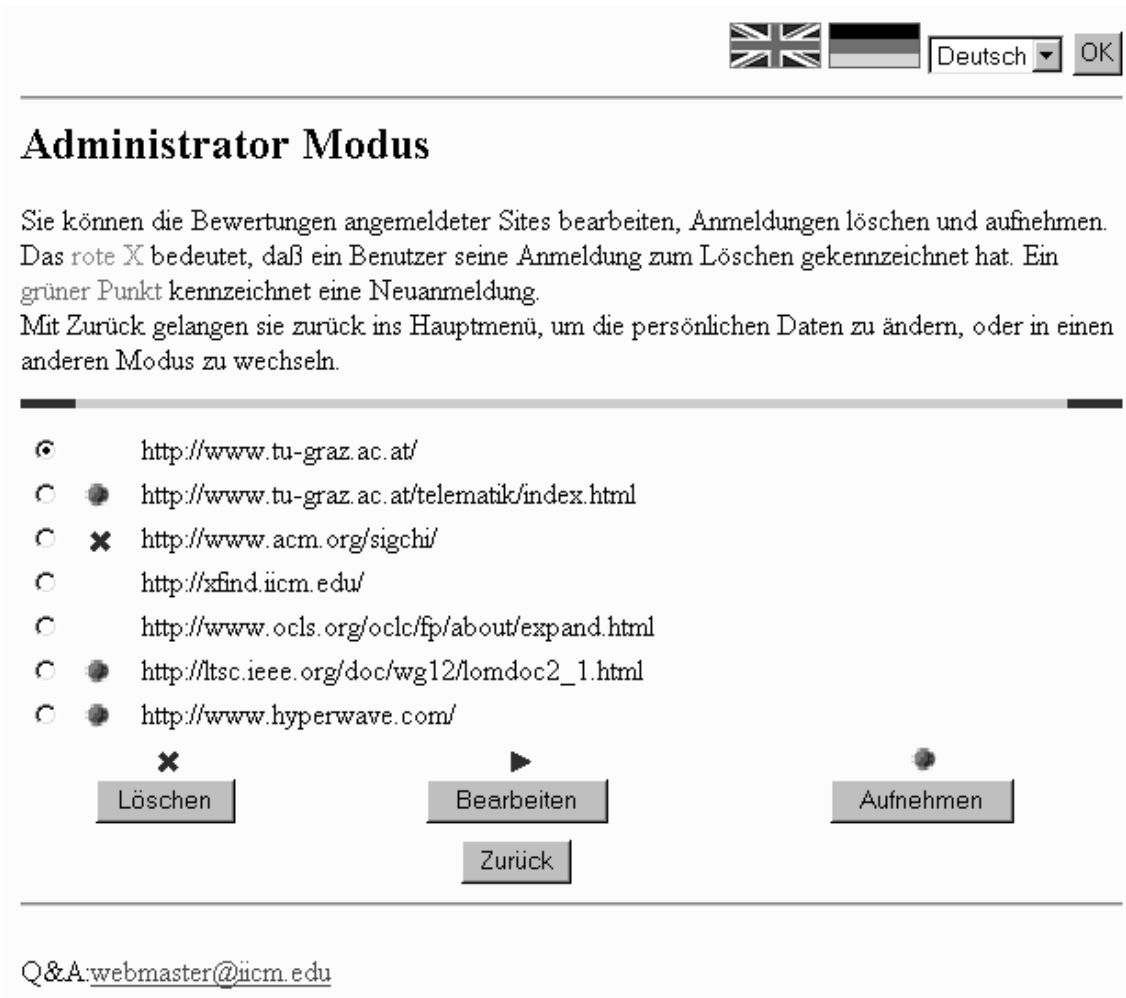


Abbildung 8.5: Im Administrator-Modus kann man Ressourcen löschen, ins System aufnehmen oder die Bewertungen ändern. Ressourcen, die mit einem Punkt gekennzeichnet sind, sind neu; die zum Löschen markierten Sites sind durch ein kleines Kreuz gekennzeichnet.

Erst durch den Knopf 'Akzeptieren' nimmt der Administrator Ressourcen ins System auf. Der Status wird als 'accept' gespeichert, und eine Kopie der Anmeldung wird im Verzeichnis für angemeldete Sites gespeichert. Dort kann sie dem Suchsystem zur Verfügung gestellt werden.

Wenn ein Benutzer seine Anmeldung im nachhinein ändert, bleibt seine bisherige Anmeldung solange aufrecht, bis auch die Änderung akzeptiert wurde. Sonst könnte ein Benutzer seine Bewertungen im nachhinein fälschlicherweise zu seinen Gunsten „verbessern“. Ändert jedoch ein Experte etwas an der Bewertung, wird dies sofort akzeptiert und gespeichert, weil man vom Experten annimmt, dass er nur objektive Bewertungen bereitstellt.

8.1.4 Der Experten-Modus

Die Dialoge im Experten-Modus sind z.T. analog zu jenen im Benutzer-Modus. Der Experte erhält in seinem Menü eine Liste jener Anmeldungen, die von einem Administrator akzeptiert wurden. Er sieht also keine zum Löschen markierten oder neuen Anmeldungen. In der Liste wird auch angezeigt, ob die Ressource bereits durch einen Experten bewertet wurde, und um welches Thema es sich handelt (vgl. Abbildung 8.6).



Abbildung 8.6: Beim Auswählen der Ressourcen erkennt der Experte, ob ein Dokument schon von einem Experten bewertet wurde, und um welches Thema es sich handelt: Der Dewey Decimal Code ist zu jeder Ressource in Klammern angeführt.

Im Experten-Modus ist es nicht möglich, Anmeldungen zu löschen oder neue hinzuzufügen. Er kann stattdessen Bewertung von Ressourcen analog zum Benutzermodus bearbeiten, und gegebenenfalls Unterbereiche anmelden und bewerten. Auf dem Bewertungsformular des Experten befinden sich zusätzliche Felder, mit denen er die Bewertung selbst charakterisieren kann (s. Abbildung 8.7). Beim Abschicken der Bewertung wird gespeichert, dass sie durch einen Experten erfolgte. Die Dateien werden sowohl im Verzeichnis für Neuameldungen, wie im Verzeichnis akzeptierter Sites aktualisiert. Das bedeutet, dass Bewertungen von Experten nicht erst von einem Administrator bestätigt werden müssen, wie dies im Benutzermodus nötig ist.

Bewerter	<input type="text" value="Weitzer, Johann"/>	
Bewertungs-Agentur	<input type="text"/>	
Label Büro	<input type="text"/>	?
Sprache der Bewertung	<input type="text" value="Deutsch"/>	
Datum der Bewertung	<input type="text" value="2000-03-31"/>	
Lebensdauer der Bewertung	<input type="text" value="50"/>	?

Vielen Dank für Ihre Bewertung!

Abbildung 8.7: Ein Experte muß beim Bewerten zusätzlich Angaben über die Bewertung selbst vornehmen. Das Bild zeigt diesen Ausschnitt aus dem Bewertungsformular für Experten, welches ansonsten analog zum Formular für Benutzerbewertungen ist (vgl. Abbildung 8.4)

8.2 Aufbau des Programms

Die Applikation besteht prinzipiell aus zwei Komponenten: Einem Perl-CGI Skript, welches über einen Webserver angesprochen wird, und dem Java Programm, das auf einem Port „lauscht“ und die Verbindung zum Skript aufnimmt. Das Skript übernimmt nur die Aufgabe, Anforderungen, die von einem Browser an den Webserver geschickt werden, an die Applikation weiterzuleiten und umgekehrt (s. Abbildung 8.8).

Da die Bedienung des Programms über Webbrowser erfolgen soll, müssen Formulare ausgewertet werden. Formulare, die der Benutzer in seinem Browser ausfüllt, und anschließend über die POST-Methode an den Webserver schickt, werden über die Standardeingabe von dort an die CGI Anwendung weitergeleitet. Diese öffnet ein Socket auf einem bestimmten Port (voreingestellt ist der Port 1234), und stellt so die Verbindung mit der Java Applikation her. Erst dort wird die HTTP-Anforderung ausgewertet. Die Daten liegen zu diesem Zeitpunkt in einem einzigen String von Name-Wert Paaren vor, wobei die einzelnen Paare durch Ampersands voneinander getrennt werden. Außerdem sind die Daten *URL-codiert*. Hat man alle Attribute mit ihren Werten aus der Anforderung herausgefiltert, können diese ausgewertet werden.

Weil die Applikation nicht linear abläuft, sondern nach jeder Benutzereingabe ein neuer Thread erzeugt wird, muß die Anwendung Informationen darüber erhalten, welches Formular gerade ausgewertet und bearbeitet werden soll. Zu diesem Zweck wird auf den HTML Seiten ein sogenanntes hidden-field eingefügt.

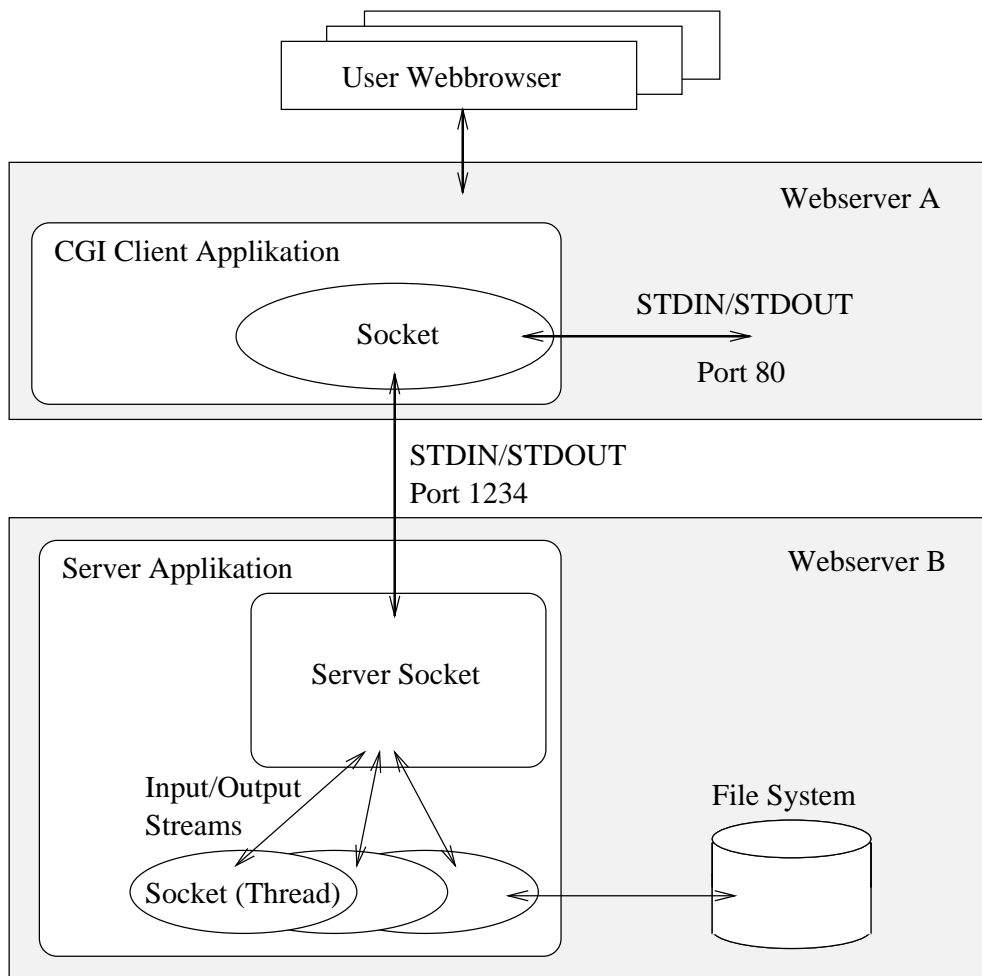


Abbildung 8.8: Aufbau der Client-Server Anwendung. Man erkennt die zwei Komponenten CGI-Skript und Java Applikation, die über ein Socket miteinander kommunizieren. Das CGI Skript nimmt die Verbindung zum Webserver A, und damit indirekt zum Browser des Benutzers auf. Die Server Applikation behandelt Benutzeranfragen in sog. Threads, und greift auf das Filesystem des Server B zu.

Das sind Felder, die der Benutzer nicht sieht, die aber beim Senden von Formulardaten mitgesendet werden. In diesen Feldern wird der Name des Formulars, die Email Adresse des Benutzers und der Modus gespeichert.

Von der Applikation wird also erkannt, um welches Formular es sich handelt, und die Daten können ausgewertet und gespeichert werden. Im Anschluß wird wieder ein Formular oder eine andere HTML Seite erstellt, und über den umgekehrten Weg (Socket - CGI - Webserver - Browser) zum Anwender geschickt (s. Listing 8.21).

Daten, welche die Java Applikation speichern muss, werden in Dateien im sogenannten SOIF-Format gespeichert. Darin werden pro Zeile jeweils der Name des Attributes, die Länge des Wertes in geschwungenen Klammern, und – getrennt durch ein Tabulatorzeichen – der Wert des Attributes, gespeichert (vgl. Listing 8.22). Die Daten werden in den Verzeichnissen für Benutzerdaten, Voranmeldungen und akzeptierte Sites gespeichert, die im folgenden beschrieben werden. In weiteren Entwicklungen können die Verzeichnisse mit den Dateien durch Datenbanken ausgetauscht werden.

```
Content-Type: text/html\n\n<HTML>\n  <HEAD>\n    <BASE href= ...>\n    <TITLE>QMRating System</TITLE>\n  </HEAD>\n  <BODY>\n    <FORM method="POST" action="http://...">\n      <INPUT type="text" name="server">\n      ... \n      <INPUT type="hidden" name="form_name" value="login">\n    </FORM>\n  </BODY>\n</HTML>
```

Listing 8.21: Die Anwendung schickt HTML Formulare an den Browser des Anwenders. In die Seite werden hidden-fields eingebaut, die der Benutzer nicht sieht, die aber von der Anwendung ausgewertet werden.

8.2.1 Verzeichnis der Benutzerdaten

Das Verzeichnis für Benutzerdaten trägt den voreingestellten Namen 'users'. Darin werden für jeden Benutzer die persönlichen Daten gespeichert, gemeinsam mit dem Passwort und dem Benutzerstatus, der 'User', 'Expert' oder 'Admin' lauten kann. Weiters wird festgehalten, für welches Thema der Benutzer gegebenenfalls Experte ist. Für jeden Anwender wird ein Datei im SOIF-Format erstellt, deren Name die Email Adressen des Anwenders ist. In Tabelle 8.1 werden alle Felder aufgelistet, die in der Datei gespeichert werden (vgl. Listing 8.22).

```

lastname{7} Weitzer
firstname{6} Johann
zip{4} 8020
place{4} Graz
country{10} Österreich
password{6} DEX-355
domain{3} 020
status{11} User, Admin

```

Listing 8.22: Beispiel eines SOIF Fileformat.

Feldname	Beschreibung
lastname	Nachname
firstname	Vorname
zip	Postleitzahl
place	Ortsangabe
country	Land
password	Passwort
domain	Fachbereiche
status	Benutzerstatus

Tabelle 8.1: Aufbau der Dateien im Benutzer-Verzeichnis. Der Dateiname entspricht der Email Adresse des Benutzers. Mehrfachnennungen bei Vornamen und Status sind möglich. Die Form der Mehrfachnennung ist beim status-Feld nicht von Bedeutung (Trennung durch Beistrich etc.), wohl aber die exakte Schreibweise von 'User', 'Expert' oder 'Admin'.

8.2.2 Verzeichnis für Voranmeldungen

In diesem Verzeichnis, welches den voreingestellten Namen 'booked' trägt, werden alle Voranmeldungen gespeichert. Die Namen der Dateien setzen sich aus der Email Adresse des anmeldenden Benutzers und der URL der Site zusammen. Zuerst kommt die Email Adresse, dann eine öffnende runde Klammer, gefolgt von der URL, in der alle Doppelpunkte durch öffnende eckige Klammern, und alle Schrägstriche durch @-Symbole ersetzt wurden (vgl. Listing 8.25).

Die SOIF-Dateien enthalten die Attribute von xQMS, sowie die in Tabelle 8.2 aufgezählten Felder (vgl. Listing 8.23). Das Attribut `rating_status` kennzeichnet, ob die Bewertung zuletzt von einem Benutzer (dem Autor oder Webserverbetreiber), Administrator oder Experten vorgenommen oder überprüft wurde. Das Flag `record_status` kann ebenso drei verschiedene Werte annehmen: 'del' markiert eine

Anmeldung zum Löschen, 'new' kennzeichnet neue Anmeldungen, welche sich nur im Verzeichnis für Voranmeldungen befinden können, und 'accept'. Diesen Status erhalten Anmeldungen, wenn sie von Administratoren akzeptiert werden. Sie werden damit gleichzeitig in das Verzeichnis der Bewertungen kopiert, und der Suchmaschine über eine Schnittstelle zugänglich gemacht. Außerdem ist es Experten möglich, die Bewertungen akzeptierter Anmeldungen zu überprüfen und zu korrigieren.

```
server{18} www.tu-graz.ac.at
port{2} 80
dir{0}
homepage{0}
frequ{1} w
number{1} c
record_status{3} new
rating_status{4} User
creator.person{0}
creator.organisation{7} TU Graz
...
myop.accuracy.width{3} 3
```

Listing 8.23: Beispiel einer SOIF-Datei im Verzeichnis für Voranmeldungen. Der Server soll wöchentlich abgesucht werden, und er besitzt laut Bewertung rund 500-1000 Seiten. Die Bewertung wurde durch einen Benutzer erstellt.

Feldname	Beschreibung
server	Der Name des Servers
port	Die Portadresse (voreingestellt 80)
dir	Angabe eines Verzeichnisses (nur bei Bereichen)
homepage	Angabe nur wenn es sich um Bereich handelt
frequ	Wie oft Server abgesucht werden soll (day, week, month)
number	Ungefähre Anzahl Seiten (a, b, c, d, e)
record_status	new, del oder registered
rating_status	User, Admin oder Expert

Tabelle 8.2: Aufbau der Dateien im Verzeichnis für Voranmeldungen. Der Dateiname entspricht der Email Adresse des anmeldenden Benutzers und der URL.

8.2.3 Verzeichnis der Bewertungen

In diesem Verzeichnis mit dem voreingestellten Namen 'rated' werden alle vom Administrator akzeptierten Anmeldungen mit den Vorbewertungen der Benutzer gespeichert. Ein Experte kann nun die Vorbewertungen der Sites überprüfen und ggf. ändern und ergänzen, womit sich auch das Feld `rating_status` von 'User' oder 'Admin' zu 'Expert' ändert. Die Qualität der Metadaten wird damit höher eingestuft, als wenn die Bewertung nur vom Besitzer einer Ressource stammen würde. Diese können die Daten im Verzeichnis für Bewertungen nicht ändern, sondern müssen Änderungen vom Administrator bestätigen lassen. Wie im folgenden Abschnitt beschrieben, werden diese Daten über eine Schnittstelle dem xFIND Suchsystem zur Verfügung gestellt, und werden auf diese Weise suchbar gemacht.

8.2.4 Schnittstelle zum xFIND Suchsystem

Wie am Beginn des Kapitels bereits geschildert, soll das Bewertungssystem QM-RatingSystem die Bewertungen dem Suchdienst xFIND zur Verfügung stellen können. Schon beim Design von xFIND wurde die leichte Anbindung an externe Systeme bedacht. So lassen sich externe Suchdienste und Qualitätsbewertungen integrieren (vgl. Kapitel 6.4 und Abbildung 8.9).

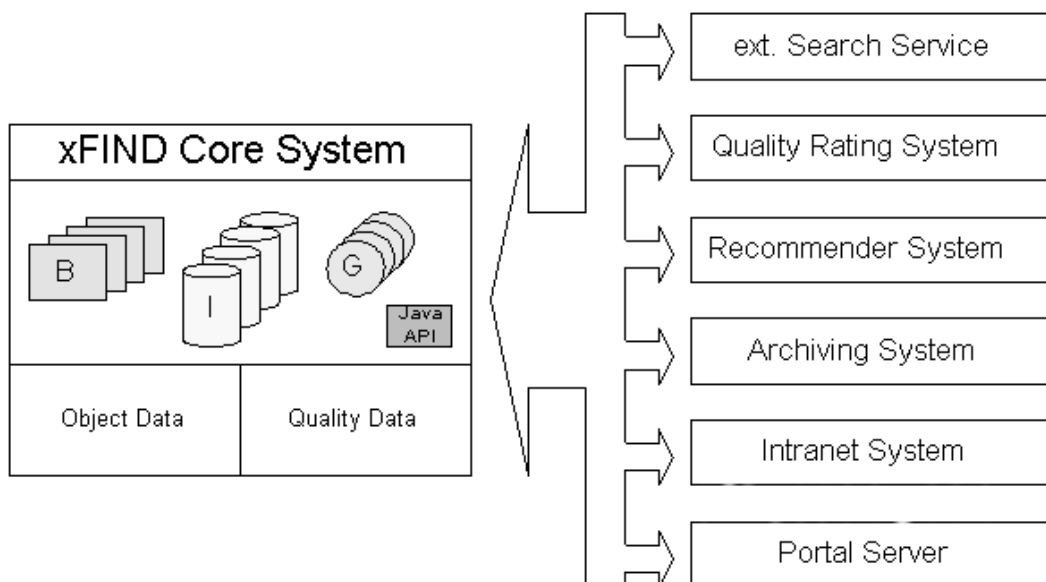


Abbildung 8.9: Der Kern des xFIND Suchsystems mit der Anbindung an externe Systeme [BMWVK2000].

Die Qualitätsmetadaten werden vom Indexer eines xFIND Suchsystems übernommen. Die Aufgabe des Indexers ist das Beschaffen der Dokumente von ein oder mehreren Gatherern, das Indizieren der Texte sowie die Speicherung der extrahierten Informationen in einem leicht suchbaren Format in einem Dokumenten Cache. Umgekehrt können für Suchanfragen Daten von Dokumenten aus dem Cache abgerufen werden. Der Indexer ist auch in der Lage, Dokument-, Bereichs- und Serverbeschreibungen zu behandeln. Er besteht aus vier Komponenten: Die Komponenten der Dokumenten Beschaffung und Speicherung, sowie jenen Teilen, welche Metadaten beschaffen und speichern. [BMWVK2000]

Solche Metadaten stellt das QMRatingSystem dem xFIND Suchsystem zur Verfügung. Die Schnittstelle zwischen den Systemen bilden die SOIF-Dateien einerseits, und der SOIF Parser andererseits (vgl. Abbildung 8.10). Wie in Abschnitt 8.2.3 erläutert, werden die Bewertungen in einem Verzeichnis abgelegt. Für jede Ressource wird eine eigene SOIF-Datei angelegt, in der nicht nur die Attribute des Qualitäts-Metadatenschemas xQMS, sondern auch Zusatzinformationen gespeichert sind. Diese Felder sind in Tabelle 8.2 aufgezählt. Dazu zählen u.a. Angaben über die Einstiegsseite, wenn es sich bei der Ressource um einen Server oder Bereich handelt, Informationen über die Häufigkeit des gewünschten Absuchens, die Anzahl der Seiten in der Ressource, sowie Angaben zum Bewertungsstatus. Der Bewertungsstatus sagt aus, ob die Ressource vom Autor, einem Administrator oder Experten gemacht wurde.

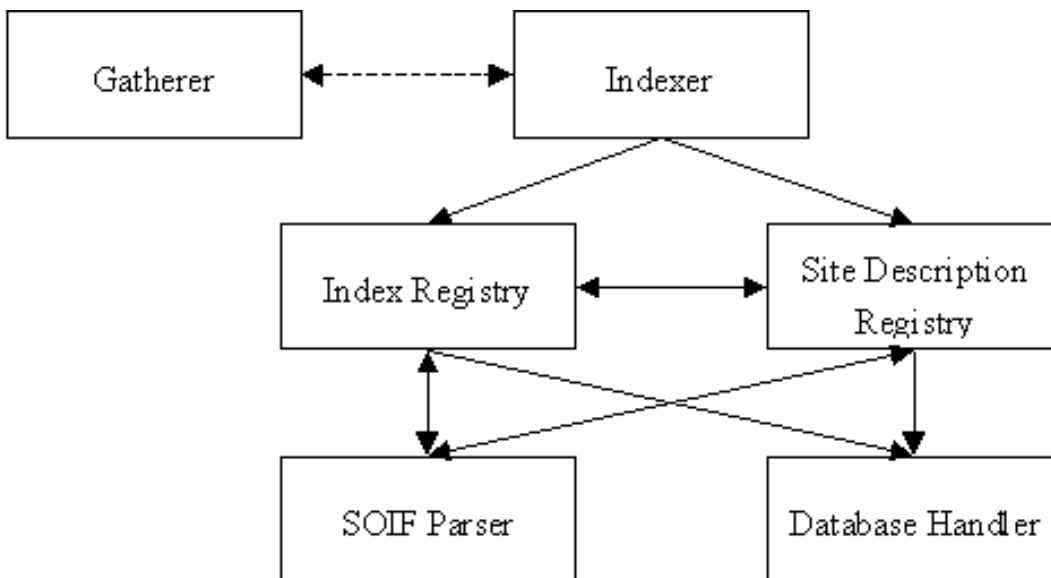


Abbildung 8.10: Die SOIF Dateien des QMRatingSystem werden vom SOIF-Parser eingelesen und dem Indexer zur Verfügung gestellt [BMWVK2000].

Der SOIF Parser liest die Informationen aus den Dateien aus, und speichert sie über den Database Handler in einem suchbaren Format ab. Nachdem die Daten

indiziert wurden, überprüft die Index Registry, ob Bewertungen veraltet sind, und löscht sie gegebenenfalls aus dem Index und der Datei. Im Anschluß stellt die Site Description Registry die Verbindung zwischen den Ressourcen Beschreibungen und den registrierten Sites her. [BMWVK2000]

Es kann nun eine verknüpfte Suche nach dem gewünschten Inhalt in der geforderten Qualität erfolgen. Die Abbildung 8.11 zeigt einen Ausschnitt aus dem Suchformular einer xFIND Implementierung im Gentle WBT System². Man kann eine einfache Stichwortsuche durchführen, eine erweiterte Suche oder die Profisuche mit der Möglichkeit, Qualitätsmetadaten anzugeben. Im Suchergebnis können zu jeder Ressource die Bewertungen aberufen werden, wie in Abbildung 8.12 dargestellt.

8.2.5 Struktur der Java Applikation

Die Klassen-Struktur der Java-Applikation (Zusammengefasst im package QM-RatingSystem) folgt im Wesentlichen dem Fluss der Bedienung des Programms, ergänzt durch Klassen, welche Hilfsfunktionen erfüllen, und einem Teil, der die Verbindung zum CGI-Skript hält. In Abbildung 8.13 sind nur die wichtigsten Komponenten schemenhaft dargestellt, die im Anschluß erläutert werden.

8.2.6 QMStart

Die einzige Aufgabe von QMStart ist das Starten des Java Servers (s. Listing 8.24). Als Parameter lässt sich beim Start die Nummer des Ports angeben, auf dem der Server „lauscht“.

```
Kommandozeile> java QMStart 2000
Stop server by pressing Ctrl-C
```

Listing 8.24: Start des Java Server auf Port 2000. Ohne Angabe des Parameters ist der Port 1234 eingestellt.

8.2.7 QMServer

Innerhalb der Klasse QMServer wird ein ServerSocket Dämon gestartet, der auf einem Port auf Client-Verbindungen wartet (lauscht). Sobald eine Verbindung

²http://wbt-1.iicm.edu/wbt/v1/core/app/hwt/mod/ce2;↔↔course=Wissens&oid=0x811bc838_0x00125e15

WBT Suche

Profi Suche

Geben Sie bitte den gewünschten Suchausdruck ein:

Geben Sie hier Qualitätsattribute ein:

Geben Sie hier die gewünschten Topics ein:

Geben Sie hier **A**uthor, **P**ublisher und **O**rganisation ein:

Zusätzliche Kriterien (alle mit AND verbunden):

Welche Document Sprache

Welches Kriterium	VON (beginnend bei)	BIS (bis einschließlich)
Zeitraum	<input type="text" value="[YYYY/MM/DD]"/>	<input type="text" value="[YYYY/MM/DD]"/>
Sprachanforderung	<input type="text" value="normal"/>	<input type="text" value="anspruchsvoll"/>
Altersgruppe	<input type="text" value="Jugendliche"/>	<input type="text" value="Senioren"/>
Vorwissen	<input type="text" value="Fortgeschrittener"/>	<input type="text" value="Experte, Spezialist"/>
Informationstiefe	<input type="text" value="ausführlich"/>	<input type="text" value="detailliert"/>
Informationsbreite	<input type="text" value="mittlere Breite, Themenauswahl"/>	<input type="text" value="geringe Breite, spezielles Thema"/>
Glaubhaftigkeit von wissenschaftl. Dokumenten	<input type="text" value="Wissenschaftlicher Forschungsbericht"/>	
Type	<input type="text" value="Doktorarbeit oder Vergleichbares"/>	

Abbildung 8.11: In das Suchformular der Profisuche des Gentle WBT Systems kann man Inhaltskriterien verknüpft mit Qualitätskriterien eingeben.

zustandekommt, wird ein Thread von QMApplication gestartet, und der Server kann wieder auf neue Anforderungen warten.

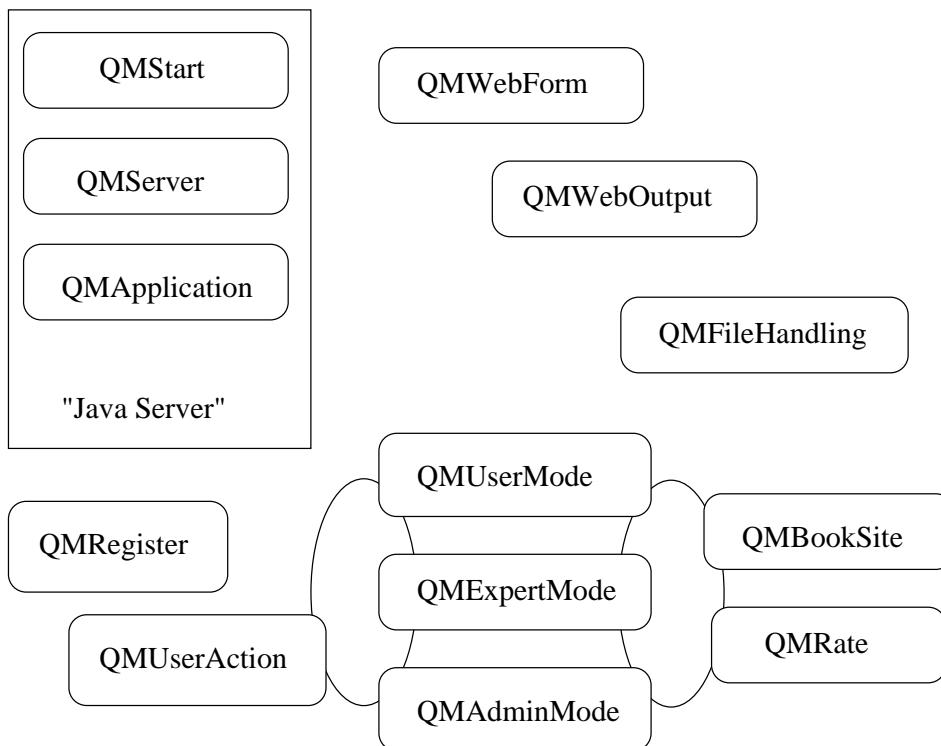


Abbildung 8.13: Schema der zentralen Komponenten. Die genaue Beschreibung aller Klassen befindet sich in der technischen Dokumentation auf der beigelegten CD-ROM.

8.2.9 QMWebForm

Diese Klasse übernimmt die Dekodierung der Formulardaten aus dem Byte-Strom (String), der der Anwendung über das CGI Skript geliefert wird. Dabei wird der String zuerst in Name-Wert Paare aufgespalten, indem er bei jedem Ampersand (&) Symbol aufgetrennt wird. Rekursiv werden anschliessend die Namen von den Werten getrennt, indem eine Spaltung beim Gleichheits (=) Symbol erfolgt. Und zuletzt müssen die Werte URL-decodiert werden. Die so erhaltenen Daten werden in eine Hashtable eingetragen, und der Anwendung zur Verfügung gestellt.

8.2.10 QMWebOutput

In dieser Klasse befinden sich Methoden zum Schreiben der HTML Formulare und Dateien auf einen Ausgabe Strom. Es können ganze HTML Seiten, die in Dateien gespeichert sind, übertragen werden, oder einzelne Zeilen von HTML Code.

8.2.11 QMFileHandling

Alle Aktivitäten, die mit dem Schreiben und Lesen von Dateien zu tun haben, werden über die Klasse QMFileHandling erledigt. Dazu zählen unter anderem das Löschen von Dateien, das Auslesen von Verzeichnissen und das Schreiben und Einlesen von SOIF-Dateien.

Beim Einlesen von SOIF-Dateien werden die Attribut-Wert Paare in eine Hashtable geschrieben. Genauso kann man der Methode einfach eine Hashtable zur Speicherung übergeben. Es gibt auch Methoden, die einzelne Feldeinträge aus SOIF-Dateien auslesen und verändern können. Weiters wird die Bildung der Dateinamen von dieser Klasse übernommen. Dateinamen setzen sich, wie in Abschnitt 8.2.2 schon erwähnt, aus einer Email und der codierten URL zusammen. Doppelpunkte in der URL werden durch öffnende spitze Klammern ersetzt, und Schrägstriche durch @-Symbole (s. Listing 8.25).

```
Email: j0nny@iicm.edu
URL: http://www.test.com/home1.html
Dateiname: j0nny@iicm.edu(http[@www.test.com@home1.html
```

Listing 8.25: Die Bildung von Dateinamen aus Email Adresse und URL.

8.2.12 Userinteraktion

Die Interaktion mit dem Benutzer erfolgt über verschiedene, gleichartig aufgebaute Klassen. Diese Klassen produzieren ein HTML Formular, welches sie über die Klasse QMWebOutput zum Browser schicken können. Die Formulare werden dynamisch aufgebaut, abhängig von den Daten, die sie via hidden-fields bekommen. Jede dieser Klassen besitzt auch die Methode, mit der die Antwort der User auf die Formulare verarbeiten wird.

8.3 Zusammenfassung

Zur Beschreibung und Bewertung von Ressourcen ist nicht nur eine Vereinbarung über die Eigenschaften nötig, die man bewertet (vgl. Kapitel 7). Es muß auch ein System zur Verfügung gestellt werden, mit dem Bewertungen tatsächlich vorgenommen werden, und an einen Suchdienst weitergeleitet werden können.

Die Testimplementierung QMRatingSystem erfüllt diese Aufgaben: Betreiber von Webservern oder Autoren von Webseiten können ihre Ressourcen an diesem System voranmelden, und mit den vorgestellten Qualitätsmetadaten beschreiben und bewerten. Die Administratoren des Systems können diese Bewertungen

im Anschluß bearbeiten, und über eine Schnittstelle dem xFIND Suchsystem zur Verfügung stellen, wo die Anwender schließlich eine verknüpfte Suche nach Inhalt und Qualitätsmetadaten durchführen können. Der Austausch der Daten erfolgt in Form von SOIF-Dateien. Eine dritte Gruppe von Benutzern sind die Fachexperten, welche ebenfalls die Bewertungen überprüfen und bearbeiten können, um die Qualität der Metadaten zu erhöhen.

Damit das System auf möglichst vielen Plattformen und Konfigurationen lauffähig ist, wurde das Programm in Java entwickelt, und die Kommunikation mit dem Webserver erfolgt über ein CGI Skript. Die Interaktion mit den Anwendern erfolgt über einen Webbrowser in deutscher oder englischer Sprache. Da es sich beim QMRatingSystem nur um einen Prototypen handelt, bleiben einige Wünsche offen, die im nächsten Kapitel behandelt werden.

Kapitel 9

Ausblick

In vorigen Kapitel wurde die Implementierung eines Prototypen vorgestellt. Die Aufgabe des Programmes ist es, die Verwendung des in Kapitel 7 vorgestellten Metadatenschemas xQMS in der Praxis zu testen. Es ist mit dem Programm möglich, Ressourcen mit den vorgestellten Attributen zu bewerten. Diese Bewertungen können im Anschluss durch das xFIND Suchsystem verarbeitet und suchbar gemacht werden. Die Vergabe der Metadaten erfolgt in drei Stufen: Zuerst meldet der Besitzer seine Ressource mit Angabe von Qualitätsmetadaten am System an. Diese werden dann von einem Systemadministrator überprüft, und in das System aufgenommen. In einem letzten Schritt können auch noch Fachexperten Änderungen vornehmen, und Metadaten ergänzen.

Da es sich bei der Implementierung lediglich um einen Prototypen – und kein kommerziell vertreibbares Produkt – handeln kann, mussten einige Einschränkungen schon im Design berücksichtigt werden. Damit das Programm auf jedem Webserver lauffähig ist, wurde die Kombination aus CGI-Skript und Java Applikation gewählt. In einer zukünftigen Version ist jedoch sicherlich über die Verwendung server-seitiger Java-Servlets anstatt des Skripts nachzudenken. Aber da die eigentliche Applikation von der Schnittstelle zum Webserver weitestgehend getrennt ist, ist eine derartige Adaptierung leicht möglich. Weitere Verbesserungsvorschläge werden punktuell aufgezählt:

- Mehrfachbewertungen sind nicht möglich. Jede Ressource kann nur einmal am System angemeldet werden, und nur vom Besitzer, den Administratoren oder Experten bewertet werden. In einer weiteren Entwicklung sollte es auch anderen Benutzern möglich sein, Bewertungen zu ergänzen.
- In der vorliegenden Version wird zwar bei Bewertungen vermerkt, ob diese von Autoren, Administratoren oder Experten erfolgte. Es wird jedoch

bei der Suche nicht weiter berücksichtigt. Weil die Metadaten von Dokumenten, die von Experten bewertet wurden, eine höhere Qualität haben als andere, könnte diesen Bewertungen und Dokumenten eine höhere Priorität zukommen.

- Die Speicherung der Benutzer- und Anmeldungs- bzw. Bewertungsdaten erfolgt in Dateien direkt am Filesystem des Servers. Den nächsten Schritt würde hierbei die Anbindung an eine leistungsfähige Datenbank darstellen.
- Experten können noch unabhängig von ihrem Fachbereich alle angemeldeten Sites bewerten. In Zukunft soll es möglich sein, die Ressourcen je nach Thema bestimmten Fachexperten zur Bewertung zu übermitteln. Bereits im bestehenden System werden für jeden Benutzer Fachgebiete gespeichert, wodurch eine künftige Erweiterung leicht möglich ist.
- In der vorliegenden Version werden nur wenige Attribute automatisch festgelegt. Wünschenswert wäre beispielsweise eine automatische Themenklassifikation.
- Es ist weiters eine Verknüpfung zu einer Autoren-Datenbank vorstellbar. Damit wäre es unter anderem möglich, die Eigenschaft „Herkunft, Glaubwürdigkeit und Ansehen“ exakter zu bestimmen. Das Signatur-Feld, welches zur Zeit nur auf statische Ressourcen anwendbar ist, könnte in diesem Fall zur Authentifizierung und Zertifizierung der Autoren dienen, wodurch auch das Vertrauen in dynamische Ressourcen objektivierbar wäre.

Kapitel 10

Zusammenfassung

Diese Arbeit gliedert sich in die zwei Teile Untersuchungs- und Gestaltungsbereich. Die Einleitung des ersten Teils schildert das explosive Wachstum des Internet. Damit erhöht sich zwar die weltweit abrufbare Informationsmenge, aber gleichzeitig wird wegen der nicht vorhandenen Kontrolle im Internet die Suche nach qualitativen, relevanten Information immer schwieriger. Diese wird ausschließlich durch entsprechend leistungsfähige Suchdienste erfolgen können, welche auch Qualitätsaspekte bei der Suche berücksichtigen.

In Kapitel 2 wird der Versuch unternommen, den Begriff „Qualität“ in Zusammenhang mit dem Inhalt von Ressourcen im Netz zu bringen. Zu diesem Zweck werden Eigenschaften ausgeforscht, die gute inhaltliche Qualität von Online Dokumenten auszeichnen. Weiters wird diskutiert, wie man relevante Informationen aus dem Netz „heraus-filtern“ kann. Bei der Methode des Downstream-Filterings werden hierzu beschreibende Metadaten verwendet.

Aus diesem Grund wird in Kapitel 3 die Verwendung von Metadaten dargestellt. Nach einer ersten Definition wird untersucht, wie man Metadaten speichern kann. Zu erwähnen sind hier hauptsächlich die sogenannten Frameworks RDF und PICS. Desweiteren werden Metadaten Schemata vorgestellt. Diese sind eine Sammlung von Attributen, welche bei der Beschreibung von Ressourcen helfen, und ein Austauschen der Information mit anderen möglich machen.

Die Kapitel 4, 5 und 6 erläuterten den Einsatz von Metadaten in Bereichen des Blockierens (also das Abhalten von Kindern vor schädlichem Inhalt), des Empfehlens (andere auf gute oder schlechte Inhalte aufmerksam machen) und der Suchdienste. In diesem Kapitel werden die üblichen Suchdienste vorgestellt, und modernere Ansätze zur Verbesserung von Suchergebnissen gezeigt. Dazu zählt unter anderem das verteilte Suchsystem xFIND.

Die durchgeführten Untersuchungen führen zu der Aufstellung eines Metadaten Schemas, welches sowohl beschreibende wie bewertende Attribute besitzt.

Möglichst viele dieser Attribute sollen computer-interpretierbar gestaltet sein, was bedeutet, dass man deren erlaubten Wertebereich einschränkt und definiert. Die Kombination von Suchkriterien für Inhalt und Qualität in Verbindung mit einer Suchmaschine verspricht eine Verbesserung der Wissensauffindung in Hinblick auf die inhaltliche Qualität.

Im Gestaltungsbereich der Arbeit wird daher in Kapitel 7 ein solches Metadaten Schema, mit dem die inhaltliche Qualität von Ressourcen im Netz sowohl beschrieben wie auch bewertet werden soll, definiert. Dieses Schema xQMS, bestehend aus sechs Klassen von Eigenschaften und der Beschreibung von 34 Attributen, findet dann in Kapitel 8 bei der Implementierung eines Testsystems zur Bewertung von Ressourcen seinen ersten Einsatz. In diesem System wird zusätzlich unterschieden, wer Bewertungen durchführt: Dies können die Autoren selbst machen, die Administratoren der Suchsysteme und schließlich auch noch Fachexperten. Mit dieser Dreiteilung erhöht sich die Qualität der Metadaten. Da das Bewertungsprogramm in die Suchmaschine von xFIND integriert wird, ist auf ein System zu hoffen, mit dem es für Benutzer rasch möglich sein soll, gesuchte Informationen in einer gewünschten Qualität zu finden.

Teil III
Anhang

Anhang A

Qualitätskriterien

A.1 Kriterienkatalog nach [Mitretek97]

Ziel des Projekts von Mitretek Systems ist eine Verbesserung der Qualität von im Internet angebotenen medizinischen Informationen. Basierend auf Umfragen und Übereinkommen auf Konferenzen, schlägt die Organisation die in Tabelle A.1 aufgezählten Kriterien vor. In der letzten Spalte der Tabelle ist verzeichnet, welcher Prozentsatz der Befragten die entsprechenden Kriterien als essentiell bezeichnen würden.

Nummer			Kriterium	Priorität
C1	C1.1		Credibility	100%
			Source	
		C1.1a	Source	
		C1.1b	Credentials	
		C1.1c	Conflict of Interest	
		C1.1d	Bias	
	C1.2		Context	
	C1.3		Currency	93%
	C1.4		Relevance/Utility	82%
	C1.5		Editorial Review Process	68%
C2			Content	
	C2.1		Accuracy	97%
	C2.2		Hierarchy of Evidence	86%
	C2.3		Original Source Stated	93%
	C2.4		Disclaimer	75%
	C2.5		Omissions Noted	
C3			Disclosure	97%
	C3.1		Purpose of the Site	
	C3.2		Profiling	
C4			Links	
	C4.1		Selection	
	C4.2		Architecture	
	C4.3		Content	75%
	C4.4		Back Linkages and Descriptions	
C5			Design	
	C5.1		Accessibility	
	C5.2		Logical Organization (Navigability)	
	C5.3		Internal Search Engine	
C6			Interactivity	
	C6.1		Mechanism for Feedback	
	C6.2		Chat Room and Bulletin Board	
	C6.3		Tailoring	
C7			Caveats	
	C7.1		Alerts	

Tabelle A.1: Qualitätskriterien für medizinischen Inhalt [Mitretek97]

Anhang B

Metadaten

B.1 PICS Optionen

Die Label Optionen lassen sich in drei Gruppen unterteilen: Optionen der ersten Gruppe stellen Informationen über das Dokument bereit, jene der zweiten Gruppe über das Label selbst, und die Optionen der letzten Gruppe informieren über Verschiedenes.

- Informationen über das Dokument, für welches das Label gilt.
 - at *quoted-ISO-date*
Das Datum der letzten Modifikation des Dokuments, zum Zeitpunkt als das Label vergeben wurde. Diese Option kann als günstigere, dafür aber weniger zuverlässige, Alternative zum *message integrity check* angegeben werden.
 - md5 *Base64-string*
MIC-md5 *Base64-string*
Ein *message integrity check*, kurz MIC, der Resource. Zur Berechnung des MIC wird der MD5 Algorithmus verwendet.
- Informationen über das Label selbst.
 - by *quoted-name*
Der Name der Person innerhalb der Bewertungs-Agentur, der für die Erstellung des Labels verantwortlich ist.
 - for *quotedURL*
Die URL (oder der Prefix String einer URL) des Artikels, zu welchem die Bewertung gehört.

- generic *boolean*
gen *boolean*
Wenn diese Option auf true gesetzt ist, kann das Label auf alle URLs bezogen werden, welche mit dem Prefix beginnen, das in der `for` Option enthalten ist. Dies wird benutzt um Bewertungen für ganze Sites oder Untermengen einer Site zu machen.
 - on *quoted-ISO-date*
Das Datum, an welchem das Label erzeugt wurde.
 - signature-RSA-MD5 *Base64-String*
Eine RSA digitale Signatur, die das Label umfasst. Die Signatur wird durch den MD5 Algorithmus durch das Bewertungs Service berechnet.
 - until *quoted-ISO-date*
exp *quoted-ISO-date*
Das Datum, an welchem das Label seine Gültigkeit verliert.
- Andere Informationen
 - comment *quotedname*
Informationen für Menschen, die das Label sehen.
 - complete-label *quotedURL*
full *quotedURL*
Die hier angegebene URL führt zu einem vollständigen Label, welches anstelle des aktuellen Label verwendet werden kann. Das vollständige Label besitzt Werte für so viele Attribute wie möglich. Es kann benutzt werden, wenn aus Performance Gründen nur ein kurzes Label übertragen wird, aber weitergehende Informationen verfügbar sind.
 - extension
Für zukünftige Erweiterungen kann hier eine URL angegeben werden.

[W3C PICS]

B.2 Liste von Typen in Dublin Core

Die Liste der Ressource Typen stammt von [DublinCore99], und ist dort ausführlich beschrieben. Die Gruppe wird an der Erforschung von Untertypen arbeiten.

collection Eine Ansammlung von Dingen. Der Ausdruck *collection* bezeichnet, dass die Ressource eine Gruppe darstellt; die Teile der Gruppe können unabhängig beschrieben und angesteuert werden.

- dataset** Strukturierte Informationen kodiert in Tabellen, Datenbanken etc., und meist direkt maschinen-lesbar. Unstrukturierte Zahlen und Worte fallen in die Kategorie text.
- event** Ein nicht-anhaltendes, zeit-abhängiges Ereignis. Metadaten für ein event stellen beschreibende Informationen für den Zweck, die Dauer, den Ort etc. des Ereignisses bereit.
- image** Der Inhalt ist primär eine symbolische, visuelle Repräsentation. Beispielsweise Bilder und Fotografien von physischen Objekten, Zeichnungen, Drucke, Grafiken, Animationen, Film etc.
- interactive resource** Eine Ressource, die Interaktion mit dem Benutzer sucht. Z.B. Webpages, Applets, Virtual Reality etc.
- model** Die Abstraktion einer realen Sache, z.B. durch Generalisierung und Interpretation. Beispiele sind Kostenmodelle, Simulationen usw.
- party** Eine Person, Organisation, kulturelle Gruppe oder Institution.
- physical object** Ein nicht-menschliches Objekt oder Substanz, wie Computer, eine Skulptur oder ein Gebäude.
- place** Eine geografische Region.
- service** Ein System das dem Benutzer ein oder mehr wertvolle Funktionen bereitstellt.
- software** Ein Computerprogramm im Sourcecode oder einer ausführbaren Form.
- sound** Eine Ressource, die primär zum Hören gedacht ist - Musik, Sprache und Geräusche. Inkludiert sind aber auch Notenblätter u.Ä.
- text** Eine Ressource, deren Inhalt hauptsächlich Wörter zum Lesen sind. Es ist zu beachten, dass auch Kopien oder Bilder von Text in diese Kategorie fallen.

B.3 Kompatibilität von Dublin Core zu LOM

Dublin Core	LOM
DC.Title	General.Title
DC.Creator	LifeCycle.Create.Contribute
DC.Subject	Disciplin.Keywords
DC.Description	Characteristic.Description
DC.Publisher	LifeCycle.Publish.Organization
DC.Contributor	LifeCycle.Create.Contribute
DC.Date	LifeCycle.Create.Date
DC.Type	Characteristic.Type
DC.Format	Technical.Format
DC.Identifier	General.Identifier
DC.Source	Relation
DC.Language	Characteristic.Language
DC.Relation	Relation
DC.Coverage	Characteristic.Coverage
DC.Rights	RightsManagement

Tabelle B.1: Die Abbildung von Dublin Core auf LOM

Anhang C

Blockingsoftware

C.1 CYBERSitter

Beim Blocking System von Solid Oak (vgl. Kapitel 4.1) kann der Benutzer folgende Gruppen von Sites ausblenden. Für die nahe Zukunft sind auch noch die beiden Kriterien „*Firearms and Weapons*“ sowie „*Violent Games*“ eingeplant. [CYBERSitter99]

Adult/sexually oriented Blockt Websites, die für Erwachsene gedacht sind.

PICS Ratings adult topics Verbirgt alle Themen, die für Kinder unter 13 Jahren nicht geeignet sind.

Sites advocating illegal/radical activities Verhindert das Aufrufen von Themen wie Bombenbasteln, Waffen, Drogen etc. Im Grunde alles, das für einen unter 13-jährigen als verboten angesehen wird.

Sites promoting gay and lesbian activities

Sites advocating hate and/or intolerance

Site promoting cults and/or occult activities

Chatrooms, -sites and -servers

Popup Windows

On-line Games and Game Sites

Gambling

Online Auctions

WWW Advertising

Sports and Leisure Activities Verbirgt Themen wie Sport Nachrichten, MTV-ähnliche Sites, Spiele etc. Gedacht für den Gebrauch in Unternehmen oder wenn Kinder ihre Hausaufgaben machen sollten.

MS Macro Files Verhindert Macro Viren Dateien, welche via e-mail empfangen werden. Sie können zwar empfangen, aber nicht ausgeführt werden.

C.2 Die Listen von CyberPatrol

Folgende Arten von Sites lassen sich bei Cyber Patrol ausblenden (CyberNO) bzw. explizit suchen (CyberYES). Für die jeweiligen Kategorien existieren Listen mit URLs auf entsprechende Seiten. [CyberPatrol2000]

CyberYESTM-Liste Diese Liste enthält, nach den folgenden Kategorien sortiert, Internet Ressourcen, die besonders schüler-freundlich und bildend sind, geeignet für 6- bis 16-jährige (vgl. Kapitel 4.1).

Art, Books & Music Diese Kategorie enthält Informationen über Kunst, Bücher und Musik (inklusive Künstler-, Autoren- und Musiker-Biografien). Texte, welche obszöne Sprache beinhalten, werden ausgeschlossen.

Games & Toys Hierunter fallen Bilder, Texte und Sounds, welche Informationen darüber beinhalten, wie man Spiele spielt oder herstellt. Weiters sind Download-Bereiche für online oder offline Spiele enthalten.

Reference Materials Jede Art von Referenz Material wie z.B. Wörterbücher, Thesauri, Atlanten und Enzyklopädien werden in dieser Kategorie gesammelt.

Movies & TV

Outdoors & Sports

Pets, Animals & Dinosaurs

Vacations & Travel

Puzzels & Hobbies

School Work

Volunteer & Help

Schools on the Net

Parents & Teachers

CyberNOTM In dieser Liste finden sich Internet-Sites und -Ressourcen mit möglicherweise anstößigem Inhalt. Dabei wird berücksichtigt, welchen Effekt die Sites auf ein typisches 12-jähriges Kind haben könnte, welches das Internet ohne elterliche Kontrolle durchstöbert (vgl. Kapitel 4.1).

Violence/Profanity

Partial Nudity

Full Nudity

Sexual Acts

Gross Depictions

Intolerance

Satanic/Cult

Drugs/Drug Culture

Militant/Extremist

Sex Education

Questionable/Illegal & Gambling

Alcohol & Tobacco

Anhang D

Verbesserte Wissensauffindung im Internet

D.1 Themenklassifikation des Dewey Decimal Code

Dezimal Code	Themen
000	Generalities
100	Philosophy & psychology
200	Religion
300	Social sciences
400	Language
500	Natural sciences & mathematics
600	Technology (Applied sciences)
700	The arts
800	Literature & rhetoric
900	Geography & history

Tabelle D.1: Die Hauptthemeneinteilung des Dewey Decimal Code (DDC) [OCLC]

Glossar

Attribut engl. attribute. Eine Eigenschaft einer →Ressource. Sie wird dieser durch eine →Bewertung vom Besitzer der Resource oder einem →Bewertungs-Service vergeben.

Bewertung engl. rating. Dies ist eine Methode, wie Benutzer des Internet die Urteile über die Qualität von Nachrichten, Artikeln oder Webseiten eingeben. Die Bewertungen werden gespeichert und dazu benutzt, andere Leser anzuleiten, was sie lesen sollen. [Palme98]

Bewertungs-Service/Agentur engl. rating service. Ein Individuum oder eine Organisation, welche nach einem speziellen System →Bewertungen durchführen und verteilen. Oft sind die Agenturen zugleich →Label Büros. [Palme97]

Blocking dt. sperren, abblocken. Eine Form des →Filterns, bei der der Benutzer bzw. Leser vor unerwünschtem Inhalt geschützt wird. Dokumente mit →Bewertungen, die die Maßstäbe des Konsumenten über- bzw. unterschreiten, werden nicht angezeigt.

Collaborative Filtering Ein System das es Gleichgesinnten erlaubt, sich gegenseitig zu helfen, interessante Dokumente zu finden.

Downstream Filtering So wie das →Upstream Filtering eine Möglichkeit des →Filtern. Vgl. Kapitel 2.3.2

Filtern Das sind Methoden und Tools, die automatisch Dokumente und →Ressourcen des Internet untersuchen, bevor sie an den Leser geleitet werden. Das Resultat kann eine Sortierung der Dokumente sein, sodass der Benutzer entscheiden kann, was er lesen möchte. Dieses Sortieren kann auf →Bewertungen aufbauen. [Palme98]

Information Retrieval dt. das Suchen und Auffinden gespeicherter Daten in einer Datenbank [Duden].

Subject Guides dt. Katalogdienste. Das sind Dienste, welche die Suche nach Internet Ressourcen in Katalogen (meist themenorientiert) ermöglichen. [Skov98] unterscheidet Subject Catalogues, Annotated Directories, Annotated Directories with Ratings or Reviews, Subject Directories with Ratings, Subject Guides und Information Gateways (Vgl. Kapitel 6.3).

Label dt. Etikett. Mit Labels lassen sich Eigenschaften von \rightarrow Ressourcen beschreiben, wobei diese Charakterisierung nicht beliebig erfolgen kann. Vielmehr wählt man einen Pegel (\rightarrow Level) auf einer zuvor ausgemachten Skala aus. Ein Label verbindet Text Phrasen oder Icons mit Werten einer Zahlenskala. Spezielle Form eines Attributes.

Label Büro engl. label bureau. Ein Label Büro ist ein HTTP Server, der einer speziellen Suchsyntax folgt. Es stellt Label für Dokumente bereit, die auf einem anderen Server liegen. Es ist nicht identisch mit einem \rightarrow Bewertungs-Service, aber diese können ihre eigenen Labels auch selbst hosten, und sind somit auch ein Label Büro. [W3C PICS]

Level dt. Pegel, Stufe, Niveau, Grad. Vgl. Label.

Metadaten engl. metadata. Zusammengefasst sind es „Daten über Daten“. Beispielsweise ist die Information „Der Autor des Dokuments X heisst Y“ ein Metadatum zum Dokument X.

Profil engl. profile. \rightarrow Schema.

Ressource engl. resource. Objekt oder Dokument im Netz, das bewertet werden kann, wie z.B. eine Web Seite, ein Newsgroup Artikel oder herunterladbare Software. [Palme97]

Retrieval dt. das Suchen und Auffinden gespeicherter Daten in einer Datenbank.

Schema engl. scheme. Die Definition einer Menge von Eigenschaften und deren zulässiger Werte, zur Definition von \rightarrow Metadaten (\rightarrow Wortschatz). Das bekannteste Schema ist Dublin Core, welches Metadaten zur Beschreibung von Büchern definiert. Die Bezeichnung Schema wird in Zusammenhang mit RDF und PICS verwendet, und ist gleichbedeutend mit der Bezeichnung Profil in HTML.

Upstream Filtering So wie das \rightarrow Downstream Filtering eine Möglichkeit des \rightarrow Filtern. Vgl. Kapitel 2.3.1

URI Uniform Resource Identifier. Generische Menge aller Namen und Adressen, die kurze Strings sind, welche auf Ressourcen verweisen. [W3C Naming]

URL Uniform Resource Locator. Eine URL beschreibt den Ort und die Art des Aufrufes eines einzelnen Dokuments. Es besteht aus den drei Komponenten „scheme“ (Protokoll zum Aufruf wie z.B. ftp oder http), „host name“ und einem hierarchischen Dokumentnamen innerhalb des Hosts. Beschrieben sind URLs in RFC-1738¹. [W3C PICS]

URN Uniform Resource Name. 1. Eine URI, mit zusätzlicher Festlegung der Verfügbarkeit etc. 2. Ein spezielles Namensschema, welches von der IETF zur Zeit entwickelt wird. [W3C Naming]

Wortschatz engl. vocabulary. Eine in einem Schema definierte Menge von Wörtern. Diese Wörter sind die Namen von Eigenschaften (→Metadaten).

¹<http://puma.germany.net/internic/rfc/rfc1738.txt>

Abbildungsverzeichnis

1.1	Die Anzahl der Hosts im Internet von 1991 bis 2000. Bis Jänner 1998 hat man diese Anzahl mit einer anderen Methode berechnet (unterer Zweig der Kurve), weshalb man bis zu diesem Datum interpolieren musste (oberer Zweig). [ISC2000]	2
2.1	In [WS96] werden vier Arten der Datenqualität unterschieden. Jede dieser Qualitäten wird durch eine Reihe von Eigenschaften repräsentiert.	8
2.2	Normalerweise finden Benutzer Websites mit herkömmlichen Suchmaschinen (1,2,3). Alternativ könnten sie den Katalog einer Bewertungs-Agentur A verwenden (4). Bewertungs-Agenturen treffen aufgrund festgelegter Qualitätskriterien eine Vorauswahl für ihre Benutzer. Der Nachteil dabei ist, dass bereits die Wahl einer Bewertungs-Agentur für das Suchergebnis ausschlaggebend ist. Möglicherweise kennt der Benutzer die Agentur B nicht, obwohl diese für ihn passendere Ergebnisse liefern würde.[ED99] . . .	14
2.3	Autoren können zusätzlich zum Inhalt ihrer Website standardisierte, beschreibende Labels ihrer Sites bereitstellen, die von einer Suchmaschine indiziert werden können (1,3). Zusätzlich beschreiben und/oder bewerten <i>Rating Services</i> die Site, und speichern ihre Labels auf separaten Servern, „Label Büros“ genannt (2). Der Benutzer abonniert einen oder mehrere Bewertungs-Dienste, denen er vertraut, und bekommt automatisch deren Labels zu einer Site, die er abrufen (4). [ED99]	15
3.1	Darstellung des Beispiel 3.8 als Graph [Lassila98]	25
4.1	Anwendung von Blocking Systemen [RM96]	34

4.2	Im Webbrowser „Microsoft Internet Explorer 5“ lassen sich mit Hilfe des Inhaltsratgebers RSACi Label festlegen. Man kann Einstellungen in den Kategorien Gewalt, Nacktaufnahmen, Sex und Sprache tätigen. Das Beispiel zeigt, dass der Benutzer in Bezug auf die Sprache die Stufe 3 wählt. Man lässt damit eine deftige, vulgäre Sprache, obszöne Gesten, und die Verwendung schwerer Kraftausdrücke zu.	37
5.1	Das Original Dokument wird über die normale Browser Interaktion geladen (1+2). Daraufhin sendet der Browser eine Anfrage nach Annotationen an einen Metainformations-Server mit der URL und der Benutzerkennung (3). Der Server schickt die Annotationen zurück (4), welche mit dem ursprünglichen Dokument verschmolzen (5) und angezeigt werden. [RMW95]	45
5.2	Der Annotations Wizard einer WBT-Umgebung (http://wbt-2.iicm.edu/hts). Jeder Kommentar erhält einen Titel, der auch erscheint, wenn man mit der Maus auf die annotierte Textstelle zeigt. Weiters kann man Attachments anhängen, und ein Symbol wählen, welches den Kommentar charakterisiert (Zustimmung, Ablehnung, Frage, Ausruf und Bemerkung).	47
6.1	Verteiltes Konzept des xFIND Suchsystems. [BMWVK2000] . . .	55
6.2	Beispiel für Verwendung der bisher ausgemachten Attributen. Gezeigt wird die Beschreibung der vorliegenden Arbeit.	60
6.3	Beispiel für Beschreibung des Themas einer Ressource.	62
6.4	Beispiel für die Verwendung von Attributen zur Beschreibung der Sprache in der vorliegenden Arbeit.	62
6.5	Beispiel für den Einsatz weiterer Attribute zur Beschreibung der vorliegenden Arbeit.	63
7.1	Einteilung verschiedener Ressourcen in eine Matrix von Informationstiefe und -breite.	83
8.1	Die Anmeldung am QMRatingSystem.	91
8.2	Neue Benutzer müssen ein paar persönlich Daten eingeben, und erhalten im Anschluß eine Email mit ihrem Passwort.	92

8.3	Die Anmeldung von Ressourcen am QMRatingSystem. Das Beispiel zeigt, wie ein Server angemeldet wird. Das Feld für den Bereich bleibt daher frei. Nur im Feld Homepage könnte man – wenn erforderlich – eine Einstiegsseite nennen.	93
8.4	Zur Bewertung von Ressourcen gibt der Benutzer die xQMS Qualitätsmetadaten ein.	94
8.5	Im Administrator-Modus kann man Ressourcen löschen, ins System aufnehmen oder die Bewertungen ändern. Ressourcen, die mit einem Punkt gekennzeichnet sind, sind neu; die zum Löschen markierten Sites sind durch ein kleines Kreuz gekennzeichnet. . .	95
8.6	Beim Auswählen der Ressourcen erkennt der Experte, ob ein Dokument schon von einem Experten bewertet wurde, und um welches Thema es sich handelt: Der Dewey Decimal Code ist zu jeder Ressource in Klammern angeführt.	96
8.7	Ein Experte muß beim Bewerten zusätzlich Angaben über die Bewertung selbst vornehmen. Das Bild zeigt diesen Ausschnitt aus dem Bewertungsformular für Experten, welches ansonsten analog zum Formular für Benutzerbewertungen ist (vgl. Abbildung 8.4) .	97
8.8	Aufbau der Client-Server Anwendung. Man erkennt die zwei Komponenten CGI-Skript und Java Applikation, die über ein Socket miteinander kommunizieren. Das CGI Skript nimmt die Verbindung zum Webserver A, und damit indirekt zum Browser des Benutzers auf. Die Server Applikation behandelt Benutzeranfragen in sog. Threads, und greift auf das Filesystem des Server B zu. . .	98
8.9	Der Kern des xFIND Suchsystems mit der Anbindung an externe Systeme [BMWVK2000].	102
8.10	Die SOIF Dateien des QMRatingSystem werden vom SOIF-Parser eingelesen und dem Indexer zur Verfügung gestellt [BMWVK2000].	103
8.11	In das Suchformular der Profisuche des Gentle WBT Systems kann man Inhaltskriterien verknüpft mit Qualitätskriterien eingeben. .	105
8.12	Im Ergebnis einer Suche innerhalb des Gentle WBT System können zu jeder Ressource die Qualitätsmetadaten angezeigt werden. Das Bild zeigt einen Ausschnitt aus dem Suchergebnis. . . .	106
8.13	Schema der zentralen Komponenten. Die genaue Beschreibung aller Klassen befindet sich in der technischen Dokumentation auf der beigelegten CD-ROM.	107

Tabellenverzeichnis

2.1	Qualitätskriterien aus der Studie von [OWB97]	9
2.2	Kriterien zur Beschreibung elektronischer Ressourcen, aufgestellt durch das OMNI Konsortium [SC94]	11
2.3	In der Arbeit von [Mitretek97] werden Dokumente nach ihrem „Ansehen“ bewertet. Es wird beispielsweise darauf geachtet, ob im Dokument persönliche Meinungen ausgedrückt werden, oder Aussagen mit Beweisen unterlegt werden (beispielsweise durch Stichproben).	11
3.1	Standardattribute bei der Verwendung von SOIF [SOIF96]	28
3.2	Dublin Core Metadata Element Set, Version 1.1 [DublinCore99]	29
4.1	Beispielhafte Auswahl von zwei der vier Kategorien, nach denen Webseiten bewertet werden können. Die beiden anderen Kategorien sind Darstellungen von Sex und Nacktaufnahmen. Man erkennt, dass die Eigenschaften nur einen von fünf Zuständen annehmen können. [RSAC2000]	36
6.1	Aufzählung von Eigenschaften, die Ressourcen hinsichtlich ihrer inhaltlichen Qualität bewerten und beschreiben. Es ist auch vermerkt, ob sich für das Attribut ein eingeschränkter Wertevorrat bzw. eine Norm zum Ausfüllen festlegen lässt. Dadurch können diese Attribut computer-interpretierbar gemacht werden (mit einem Stern gekennzeichnet).	67
7.1	Bei jenen Attributen, die einen eingeschränkten Wertevorrat besitzen, ist auf Sonderfälle zu achten.	70

7.2	Die Attribute von xQMS sind in beschreibende und bewertende Attribute eingeteilt (oberer und unterer Bereich der Tabelle). Die mit einem Stern gekennzeichneten Attribute folgen entweder fixen Normen (wie z.B. Datumsangaben), oder haben einen eingeschränkten Wertevorrat.	71
7.3	Aufzählung der Gültigkeitsbereiche von Beschreibungen. Es sei darauf hingewiesen, dass hier die Werte 'nicht passend', 'unbekannt' und 'alle' nicht zulässig sind, weil immer einer der Werte zutreffen muss. Bei der Testimplementierung, die in Kapitel 8 beschrieben wird, ist die Auswahl des Teiles einer Seite nicht möglich.	73
7.4	Typisierung von Ressourcen angelehnt an [DublinCore99]	75
7.5	Typisierung des sprachlichen Ausdrucks	80
7.6	Alter der Zielgruppe	81
7.7	Vorwissen der Zielgruppe	81
7.8	Charakterisierung der Glaubhaftigkeit von wissenschaftlichen Dokumenten nach [Palme98] in Deutsch und Englisch.	82
7.9	Typisierung der Informationsbreite	84
7.10	Typisierung der Informationstiefe	84
7.11	Beispiel zur Verwendung der xQMS Attribute für Webseiten. Bei der Ressource handelt es sich um die Arbeit [Rettig96]. Nach Ansicht des Bewerter ist diese Ressource für Menschen ab ca. 20 Jahren geeignet, die über ein fortgeschrittenes oder Experten-Wissen auf dem Gebiet des Bibliothekswesens verfügen.	86
7.12	In dem Beispiel handelt es sich um einen Webbereich, der über das Thema AIDS aufklärt. Die enthaltenen Informationen sind für Kinder und Jugendliche gedacht, welche noch wenig über das Thema wissen. Der Text ist in Englisch geschrieben, und mühelos lesbar.	87
7.13	Beispiel zur Verwendung der xQMS Attribute für Server. Die Sprache der Bewertung ist Englisch, und sie wurde am 30. März 2000 erstellt. Sie soll weiters 100 Tage gültig sein.	88
8.1	Aufbau der Dateien im Benutzer-Verzeichnis. Der Dateiname entspricht der Email Adresse des Benutzers. Mehrfachnennungen bei Vornamen und Status sind möglich. Die Form der Mehrfachnennung ist beim status-Feld nicht von Bedeutung (Trennung durch Beistrich etc.), wohl aber die exakte Schreibweise von 'User', 'Expert' oder 'Admin'.	100

8.2	Aufbau der Dateien im Verzeichnis für Voranmeldungen. Der Dateiname entspricht der Email Adresse des anmeldenden Benutzers und der URL.	101
A.1	Qualitätskriterien für medizinischen Inhalt [Mitretek97]	116
B.1	Die Abbildung von Dublin Core auf LOM	120
D.1	Die Hauptthemeneinteilung des Dewey Decimal Code (DDC) [OCLC]124	

Listingsverzeichnis

3.1	Einfache HTML Anweisung	20
3.2	Demonstration des Parameter lang	20
3.3	Beispiel für http-equiv	20
3.4	Allgemeine Form von PICS Labels	22
3.5	Beispiel für ein spezielles Label	22
3.6	Einbettung von Labels in HTML	22
3.7	Neuer Header PICS-Label	23
3.8	Beispiel für RDF Metadaten [Lassila98]	25
7.9	Beispielhafte Verwendung der Creator Eigenschaften.	72
7.10	Beispiel für den Einsatz der Identifier Eigenschaft.	73
7.11	Auch wenn es sich in dem Beispiel um eine Fotografie handelt, ist der Typ der Ressource dennoch Audio (sound). Denn das Hauptinteresse gilt im Falle Mozarts dem Notenblatt, welches der Musik zugeordnet wird.	75
7.12	In der ersten Zeile wird das Attribut Title bezogen auf eine Webseite. Die zweite Zeile zeigt die Verwendung in Bezug auf einen Server.	76
7.13	Demonstration der Verwendung von Versions Attributen. Das der Beschreibung zu Grunde liegende Objekt hat die Versionsnummer 3.141. Mit der Angabe der ersten URL wird auf eine ältere Version verwiesen, mit der Angabe der zweiten URL auf eine neuere Version. Es ist aber nicht gesagt, dass diese neuere Version tatsächlich die Aktuellste ist. Hierzu müsste man den Link verfolgen, und nachsehen, ob in dem Dokument dok4.html eine weitere, neuere Version genannt wird usw.	76

- 7.14 Das Beispiel zeigt eine Ressource, welche das Thema *Data Processing* (Dewey Decimal Code 004) und *Library and information sciences* (DDC 020) behandelt. Eine weitere Themenangabe wird über das Klassifikationsschema von ACM gemacht. In diesem Schema bezeichnet H.3.3 *Information Search and Retrieval* und H.3.4 steht für die Thematik *Systems and Software*. 77
- 7.15 Das Beispiel beschreibt den Inhalt des Teils einer Seite (der Gültigkeitsbereich lautet 'page section (1)'). Dieser Teil ist vom Typ 'image (4)' und die Beschreibung ist im Wesentlichen nichts anderes als die ursprüngliche Bildunterschrift (vgl. Abbildung 6.1). . . 78
- 7.16 Die Ressource des Beispiels wurde am 2. November 1999 erstellt, und zuletzt am 13. März 2000 geändert. 78
- 7.17 Beispielhafte Verwendung der Zitiervorschrift 79
- 7.18 Der dem Beispiel zu Grunde liegende Text ist in Britischem Englisch geschrieben, und der Ausdruck wird als 'schwierig (4)' bis 'anspruchsvoll (5)' charakterisiert. 79
- 7.19 Das Beispiel zeigt, dass die betreffende Ressource primär für Experten auf dem betrachteten Gebiet gedacht ist. Das Alter spielt keine Rolle. 80
- 7.20 In dem Beispiel wird eine Ressource beschrieben, welche sich mit einem Detail oder speziellen Thema befasst, und dieses (sehr) ausführlich behandelt. Bezogen auf die Themenangabe durch den Dewey Decimal Code 538 (Physik) würde es beispielsweise bedeuten, dass die Ressource den Spin der Elektronen (Detail) mit vielen Formeln und Berechnungen (sehr ausführlich) behandelt. 83
- 8.21 Die Anwendung schickt HTML Formulare an den Browser des Anwenders. In die Seite werden hidden-fields eingebaut, die der Benutzer nicht sieht, die aber von der Anwendung ausgewertet werden. 99
- 8.22 Beispiel eines SOIF Fileformat. 100
- 8.23 Beispiel einer SOIF-Datei im Verzeichnis für Voranmeldungen. Der Server soll wöchentlich abgesehen werden, und er besitzt laut Bewertung rund 500-1000 Seiten. Die Bewertung wurde durch einen Benutzer erstellt. 101
- 8.24 Start des Java Server auf Port 2000. Ohne Angabe des Parameters ist der Port 1234 eingestellt. 104
- 8.25 Die Bildung von Dateinamen aus Email Adresse und URL. 108

Literaturverzeichnis

- [AC96] Anagnostelis, Betsy; Cox, John: Data on the Internet: Evaluating the Quality or „Less is More“. In: C. J. Armstrong and R. J. Hartley (eds.), UKOLUG State-of-the-Art Conference, Warwick 17-19 Juli 1996, London: UKOLUG, S.59-69, 1996. <http://omni.ac.uk/agec/ukolug.html>
- [ACM97] Anagnostelis, B.; Cooke, A.; McNab, A.: Thinking critically about information on the Web. In: Vine 104, S.21-28, 1997.
- [ACM99] The ACM Computing Classification System [1998 Version] Valid in 1999. 1999. <http://www.acm.org/class/1998/>
- [Armstrong97] Armstrong, Chris: Metadata, PICS and Quality. In: Ariadne, Nr. 9, Mai 1997. <http://www.ariadne.ac.uk/issue9/pics/>
- [AT99] Altman, Lee; Tuomela, Sanna: WWW Metadata & Collaboration. 1999. <http://ils.unc.edu/altml/collaboration/>
- [BMWVK2000] Bundesministerium für Wissenschaft und Verkehr: Intelligente Wissenserfassung und Wiederauffindung in künftigen WWW-Systemen. Forschungsendbericht GZ.61.090/2-V/B/9/98, Graz, Februar 2000.
- [BS97] Balabanović, Marko; Shoham, Yoav: Fab: Content-Based, Collaborative Recommendation. In: Communications of the ACM, Bd. 40, Nr. 3, S.66-72, 1997.
- [Brockhaus56] Der Grosse Brockhaus, Bd. 9, Verlag F. A. Brockhaus, Wiesbaden 1956.
- [Ciolek96] Ciolek, Matthew T.: The Six Quests for The Electronic Grail: Current Approaches to Information Quality in WWW Resources. In: Review Informatique et Statistique dans les Sciences humaines (RISSH), Centre Informatique de Philosophie et Lettres, Universite de Liege, Belgium, S. 45-71, Nr. 1-4, 1996. <http://www.ciolek.com/PAPERS/QUEST/QuestMain.html>

- [CyberPatrol2000] Cyber Patrol, The Learning Company, Framingham MA 01702 USA. 2000. <http://www.cyberpatrol.com>
- [CYBERSitter99] Solid Oak Software Inc., St. Barbara CA 93160 USA. 1999. <http://www.solidoak.com>
- [Daniel97] Daniel, Ron: Extending the Warwick Framework: From Metadata Containers to Active Digital Objects. In: D-Lib Magazine, November 1997. <http://www.dlib.org/dlib/november97/daniel/11daniel.html>
- [DMOZ] Open Directory Project. 2000. <http://www.dmoz.org>
- [DOI] International DOI Foundation: The Digital Object Identifier System. 2000. <http://www.doi.org/index.html>
- [DublinCore99] Dublin Core Metadata Initiative, Dublin Core Metadata Element Set, Version 1.1: Reference Description. 1999. <http://purl.org/dc>
- [Duden] DUDEN - Die deutsche Rechtschreibung © Bibliographisches Institut & F.A. Brockhaus AG, Mannheim 1996.
- [Dyson97] Dyson, Esther: Release 2.0, Die Internet-Gesellschaft, Spielregeln für unsere digitale Zukunft, aus dem Amerikanischen von Hennig Thies, Droemer Knauer, München 1997.
- [ED98] Eysenbach Gunther, Diepgen Thomas L.: Towards quality management of medical information on the internet: evaluation, labelling, and filtering of information. In: British Medical Journal, Bd. 317, Nr. 7171, S. 1496-1502, 28 November 1998. <http://www.bmj.com/cgi/content/full/317/7171/1496>
- [ED99] Eysenbach Gunther, Diepgen Thomas L.: Labeling and filtering of medical information on the Internet. Methods of Information in Medicine Bd. 38 S. 80-88, 1999.
- [Fielding94] Fielding, Roy T.: Maintaining Distributed Hypertext Infostructures: Welcome to MOMspider's Web. Chapter 6, The Need for Visible Metainformation. 1994. <http://www.ics.uci.edu/↔↔pub/websoft/MOMspider/WWW94/meta.html>
- [Heery96] Heery, Rachel: Review of Metadata Formats. In: Program, Bd. 30, Nr. 4, S. 345-373, October 1996. <http://www.ukoln.ac.uk/metadata/review.html>

- [Heise99] heise online news: Philosoph warnt vor Wissensverlust durch das Internet. Verlag Heinz Heise 1999. <http://www.heise.de/newsticker/data/cp-28.12.99-000/>
- [HSRF95] Hill, Will; Stead, Larry; Rosenstein, Mark; Furnas, George: Recommending and Evaluating Choices in a Virtual Community of Use. In: Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, Bd. 1, S. 194-201, 1995. http://www.acm.org/sigs/sigchi/chi95/Electronic/documnts/↔papers/wch_bdy.htm
- [Horwath99] Horwath, Jürgen: Personalised Recommender System. Diplomarbeit an der TU Graz, 1999.
- [Hyperwave] Hyperwave AG: Hyperwave User's Handbook. 1999. <http://www.hyperwave.com>
- [Incore99] Internet Content Rating For Europe (INCORE). 1999. <http://www.incore.org>
- [ISC2000] Internet Software Consortium: Internet Domain Service. 2000. <http://www.isc.org/ds/>
- [IMS99] IMS Project: IMS Meta-Data Specification Draft. 1999. <http://www.imsproject.org/metadata.html>
- [KW95] Kjartansdóttir, Ásgerdur; Widenius, Marja: The Quality of Business Information on the Internet: Evaluation Criteria Applicable to Internet Resources. In: Swedish Library Research, Nr. 3/4, S. 43-50, 1995.
- [Langa98] Langa, Fred: Track it down. In: Windows Magazine, Bd. 9, Nr. 7, S. 158ff, July 1998.
- [Lassila98] Lassila, Ora: Web Metadata: A Matter of Semantics, IEEE Internet Computing, S.30-37, July/August 1998.
- [Legenstein99] Legenstein, Herbert: Qualitätsaspekte zur Wissensauffindung und Testimplementation für xFIND. Diplomarbeit an der TU Graz, 1999.
- [LOM98] IEEE Learning Technology Standards Committee (LTSC), Learning Object Metadata (LOM) Draft Document v2.1, 1998. http://ltsc.ieee.org/doc/wg12/LOMdoc2_1.html

- [ME95] Maltz, David; Ehrlich, Kate: Pointing the Way: Active Collaborative Filtering. In: Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems, Bd. 1, S. 202-209, 1995. http://www.acm.org/sigs/sigchi/chi95/Electronic/documnts/↔papers/ke_bdy.htm
- [Mitretek97] Mitretek Systems; HITI; AHCPR: Criteria for Assessing the Quality of Health Information on the Internet. Working draft white paper, Oktober 1997. <http://hitiweb.mitretek.org/docs/criteria.html>
- [McMurdo98] McMurdo, George: Evaluating Web information and design. In: Journal of Information Science, Bd. 24, Nr. 3, S.192-204, 1998.
- [OCLC] OCLC: Expanded Introduction to the Dewey Decimal Classification. 1999. <http://www.oclc.org/oclc/fp/about/expand.htm>
- [OWB97] Oliver, Kevin M.; Wilkinson, Gene L.; Bennett, Lisa T.: Evaluating the Quality of Internet Information Sources. Updated version of paper presented at EDMEDIA/ED-TELECOM 97, international conference of the Association for the Advancement of Computing in Education, Calgary, Canada, June 1997. <http://itech1.coe.uga.edu/Faculty/gwilkinson/AACE97.html>
- [Palme97] Palme, Jacob: Choices in the Implementation of Rating, revised version July 1997. 1997. <http://www.dsv.su.se/~jpalme/select/rating-choices.html>
- [Palme98] Palme, Jacob: Select Project Overview: Rating and filtering of scientific, technical and other network documents, 1998. <http://www.dsv.su.se/~jpalme/select/select.html>
- [Resnick97] Resnick, Paul: Filtering Information on the Internet. Scientific American, März 1997. <http://www.sciam.com/0397issue/0397resnick.html>
- [RM96] Resnick, Paul; Miller, James: PICS: Internet Access Controls Without Censorship. In: Communications of the ACM, Bd. 39, Nr. 10, S. 87-93, Oktober 1996. <http://www.w3.org/PICS/iacwcv2.htm>
- [RV97] Resnick, Paul; Varian, Hal R.: Recommender Systems. In: Communications of the ACM, Bd. 40, Nr. 3, S.56-58, März 1997. <http://www.acm.org/cacm/MAR97/resnick.html>

- [Rettig96] Rettig, James R.: Beyond 'Cool': Analog Models for Reviewing Digital Resources. September 1996. <http://www.onlineinc.com/onlinemag/SeptOL/rettig9.html>
- [RMW95] Röscheisen, Martin; Mogensen, Christian; Winograd, Terry: Shared Web Annotations as a Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples. Technical Report STAN-CS-TR-97-1582, Stanford Integrated Digital Library Project, Computer Science Dept., Stanford University, April 1995. <http://www.diglib.stanford.edu/diglib/pub/reports/commentor.html>
- [Rosenberg95] Rosenberg, Scott: The Web: Heading for Content Ratings? San Francisco Examiner, 12. Juli, 1995. <http://www.sfgate.com/net/rosenberg/0712.html>
- [RSAC2000] Internet Content Rating Association. 2000. <http://www.icra.org/>
- [Sander98] Sander-Beuermann, Wolfgang: Schatzsucher. Die Internet-Suchmaschinen der Zukunft. In: c't Heft 13, S 178-184, 1998.
- [SC94] Stoker, David; Cooke, Alison: Evaluation of Networked Information Sources. In: Ahmed H. Helal & Joachim W. Weiss (eds.), Information Superhighway: the Role of Librarians, Information Scientists and Intermediaries: Proceedings of the 17th International Essen Symposium, 24.-27. Oktober 1994, Essen: Universitätsbibliothek Essen, S. 287-312, 1995. <http://omni.ac.uk/agec/essen.html>
- [SJ95] Shelley, E.P.; Johnson, B.D.: Metadata: Concepts and models. In: Proceedings of the Third National Conference on the Management of Geoscience Information and Data, organised by the Australian Mineral Foundation, Adelaide, Australia, pp 4.1-5, 18.-20. Juli 1995.
- [Skov98] Skov, Annette: Separating the Wheat from the Chaff, Internet Quality. In: Database, Band 21, Heft 4, S. 38-40, August/September 1998.
- [Smith97] Smith, Alastair G.: Testing the Surf: Criteria for Evaluating Internet Information Resources. In: The Public-Access Computer System Review, Bd. 8, Nr. 3, 1997. <http://info.lib.uh.edu/↔pr/v8/n3/smit8n3.html>
- [SOIF96] Wessels, Duane: The Summary Object Interchange Format (SOIF). 1996. <http://www.tardis.ed.ac.uk/harvest/docs/↔old-manual/node151.html>

- [W3C Activity] The World Wide Web Consortium: Metadata Activity Statement. 1999. <http://www.w3.org/Metadata/Activity.html>
- [W3C Meta] The World Wide Web Consortium: The global structure of an HTML document. 1999. <http://www.w3.org/TR/REC-html40/struct/global.html>
- [W3C Naming] The World Wide Web Consortium: Naming and Addressing: URIs, URLs,... 1999. <http://www.w3.org/Addressing/>
- [W3C PICS] The World Wide Web Consortium: PICS Label Distribution, Label Syntax and Communication Protocols. Oktober 1996. <http://www.w3.org/TR/REC-PICS-labels-961031>
- [Weinberg97] Weinberg, Jonathan: Rating the Net. In: Hastings Communications and Entertainment Law Journal, Bd. 19, März 1997. <http://www.law.wayne.edu/weinberg/rating.htm>
- [Wood97] Wood Andrew: Metadata - The Ghosts of Data Past, Present and Future. 1997. <http://archive.dstc.edu.au/RDU/reports/↔↔Sympos97/metafuture.html>
- [WS96] Wang, Richard Y.; Strong, Diane M.: Beyond Accuracy: What Data Quality Means to Data Consumers. In: Journal of Management Information Systems, Bd. 12, Nr. 4, S.5-33, Frühling 1996.