



**DEUTSCHE INITIATIVE
FÜR NETZWERKINFORMATION E.V.**

Elektronisches Publizieren an Hochschulen

Inhaltliche Gestaltung der OAI-Schnittstelle

- Empfehlungen -

DINI Arbeitsgruppe Elektronisches Publizieren

DINI-Schriften 2-de

Version 2

September 2005

Inhaltsverzeichnis

1	Das Open Archives Protocol for Metadata Harvesting (OAI-PMH)	4
2	Empfehlungen zur Definition von Sets	6
2.1	Allgemeine Bemerkungen	6
2.3	Gliederung nach dem Publikationstyp	11
2.4	Gliederung nach dem Dokumenttyp	12
2.5	Gliederung nach der Begutachtung	13
3	Datenprovider	14
3.1	Wie werde ich Datenprovider?	14
3.1.1	Wo sollte man sich registrieren?	14
3.2	Dublin Core und andere Metadatenformate	14
4	Serviceprovider	16
4.1	Wie werde ich Serviceprovider?	16
4.1.1	Wo sollte man sich registrieren?	16
4.2	Beispiele für Daten- und Serviceprovider-Dienste	16
4.2.1	Proprint	16
4.2.2	PhysNet	17
4.2.3	Providerdienst des Hochschulbibliotheksentrums NRW (hbz)	17
4.3	Providerdienst des Bibliotheksservicezentrums Baden-Württemberg	18
5	Empfohlene Links und Literatur zu OAI	18
6	Glossar	19

Zusammenfassung

Das *Open Archives Protocol for Metadata Harvesting* (OAI-PMH) ermöglicht es, eigene Metadaten, die der Beschreibung beliebiger Objekte dienen, mit anderen zu teilen oder von verschiedensten Institutionen Daten einzusammeln und selbst weiterführende Dienstleistungen anzubieten.

Dieser Text enthält neben einer kurzen Übersicht über das Protokoll Empfehlungen für die Verwendung von Sets (Mengen) durch deutsche Datenprovider sowie für die Verwendung der Metadatenfelder von *Dublin Core* (DC). Im Mittelpunkt steht dabei das Ziel, einen effizienten Austausch von Metadaten zwischen den Nutzern des OAI-Protokolls zu gewährleisten. Zusätzlich sind einige weiterführende Dienste als Beispiele in dieser Broschüre aufgeführt.

Vorwort

Die *Open Archives Initiative* hat in den letzten Jahren dazu geführt, dass eine große Anzahl von Dokumentenservern entstanden ist, die über das *OAI-Protocol for Metadata Harvesting* Beschreibungen (Metadaten) von Objekten (z.B. wissenschaftliche Artikel) mit anderen Institutionen über das Internet automatisiert austauschen und teilen.

Laut einer Umfrage zu digitalen Repositories, welche im Mai 2005 von der CNI-JISC-SURF Konferenz "Making the strategic case for institutional repositories" (<http://www.surf.nl/en/bijeenkomsten/index2.php?oid=6>) durchgeführt wurde, weist Deutschland (neben den USA) international die höchste Zahl institutioneller Dokumentenserver auf (103). Es wird jedoch eine hohe Dunkelziffer unbekannter Systeme vermutet (in den USA, Kanada, Finnland, aber auch in Deutschland). Während die durchschnittliche Anzahl der Dokumente pro Server in Deutschland, wie auch der internationale Durchschnitt, bei einigen hundert Dokumenten liegt, weisen nur die Niederlande eine höhere Zahl auf (3500). Insgesamt ist der Abdeckungsgrad der publizierten Literatur durch institutionelle Dokumentenserver noch sehr gering. Bezogen auf die Publikationstypen ist der Abdeckungsgrad bei Dissertationen in Deutschland mit am höchsten. Im Bereich Bibliotheks- und Informationswesen liegt er bei 75%, juristische Dissertationen sind jedoch nur zu 2% online verfügbar. In absoluten Zahlen sind die Fächer Biologie, Chemie, Medizin, Physik, Ingenieurwissenschaften und Informatik am stärksten, hier liegen Abdeckungsquoten bei 30-50%. Hier ist die langfristige Wirkung des Projektes [DissOnline](#) festzustellen.

Von den in Deutschland existierenden 103 Dokumenten- und Publikationsservern (<http://www.dini.de/dini/wisspub/dokuserver.php>) besitzen nicht alle eine OAI-Schnittstelle, die es ihnen ermöglicht, ihre Dokumente darüber weltweit anzubieten und über Dienste wie das DINI-Suchportal (http://www.dini.de/oai_suche), zurzeit als (http://edoc.hu-berlin.de/e_suche/oai.php) verfügbar, SCIRUS (<http://www.scirus.com>) oder Proprint (<http://www.proprint-Service.de>) nutzbar machen zu können.

1 Das Open Archives Protocol for Metadata Harvesting (OAI-PMH)

Das *Open Archives Protocol for Metadata Harvesting* (OAI-PMH) ermöglicht einen effizienten Austausch von Metadaten und impliziert eine funktionale Aufteilung in Anbieter von (Dokumenten und) Metadaten, so genannte *Datenprovider*¹, und darauf aufbauende Dienste (*Serviceprovider*²). Das OAI-PMH basiert auf dem grundsätzlichen Prinzip des so genannten *Harvesting*³, bei dem im Gegensatz zum Ansatz des *Cross Searching*⁴ eine asynchrone Suche durchgeführt wird. Das heißt, der Serviceprovider fragt in regelmäßigen Abständen die Metadaten der Datenprovider ab und speichert diese in seiner lokalen Datenbank. Konkrete Suchanfragen (z.B. von Endnutzern) werden im Falle des Harvesting-Ansatzes ausschließlich mithilfe der lokalen Datenbank des Serviceproviders beantwortet.

Das OAI-Protokoll basiert auf weithin bekannten und verbreiteten Standards. Fest definierte Beschreibungssprachen, die ein Objekt beschreiben (Metadatenformate), werden in der *eXtensible Markup Language* (XML) kodiert und dann über das vom World Wide Web bekannten *Hypertext Transfer Protocol* (HTTP) über das Internet transportiert.

Dabei können Metadaten in beliebigen Metadatenformaten ausgetauscht werden. Aus Gründen der Interoperabilität wird jedoch das bekannte und sehr einfache Metadatenformat *Dublin Core* als kleinster gemeinsamer Nenner von allen Daten Providern unterstützt. So sind Kommunikation und der tatsächliche Austausch von Metadaten zwischen beliebigen OAI-kompatiblen Daten- und Service Providern ohne weitere zusätzliche Vereinbarungen sofort möglich.

Die im OAI-PMH definierte prinzipielle Trennung zwischen Datenprovider und Serviceprovider schließt natürlich die Entwicklung von Diensten nicht aus, die beide Funktionalitäten umfassen. Diese Möglichkeit wird von so genannten *aggregierenden* oder *kumulativen* Daten Providern genutzt. Sie fragen über das OAI-Protokoll die verfügbaren Daten einer bestimmten Menge von Daten Providern ab und halten diese für Anfragen anderer Serviceprovider ebenfalls über eine OAI-Schnittstelle vor.

Für Benutzerinnen und Benutzer von auf dem OAI-PMH basierenden Diensten ist die Technik, auf der die Suchanfragen beruhen, in der Regel transparent. Ihnen steht zum Beispiel eine Web-Schnittstelle zur Verfügung, über die sie mit dem Serviceprovider kommunizieren und dessen Dienste nutzen. Dass sich die letztlich gefundenen digitalen Objekte auf verteilten Servern befinden, wird aus Nutzersicht erst bei deren Abruf bzw. einer Autorisierungsanforderung sichtbar (siehe Abbildung 1).

¹ Als Datenprovider wird eine OAI-kompatible Schnittstelle zu einer Datenbank verstanden, in der sich Metadaten über Dokumente oder andere digitale Objekte befinden und die über eine HTTP-Verbindung erreichbar ist. Sie muss dazu in der Lage sein, OAI-Anfragen entsprechend der Protokolldefinition korrekt zu beantworten.

² Der Serviceprovider bietet mithilfe von Daten, die er unter Nutzung des OAI-PMH gesammelt hat, einen (innerhalb der Protokollspezifikation nicht näher definierten) Dienst an. Der aus Sicht des Protokolls relevante Teil des Serviceproviders, der so genannte *Harvester*, versendet OAI-konforme Anfragen an Datenprovider und wertet die entsprechenden Antworten aus.

³ engl. ernten. Dabei werden regelmäßig alle verfügbaren bzw. relevanten Daten aus den verwendeten Datenquellen abgefragt (geerntet) und in einer zentralen Datenbank gespeichert, innerhalb derer dann die eigentliche Suche erfolgt.

⁴ bekannt auch unter dem Namen *Federated Searching*, bezeichnet die unmittelbare Suche in allen verwendeten Datenquellen, wird beispielsweise häufig von Meta-Suchmaschinen verwendet

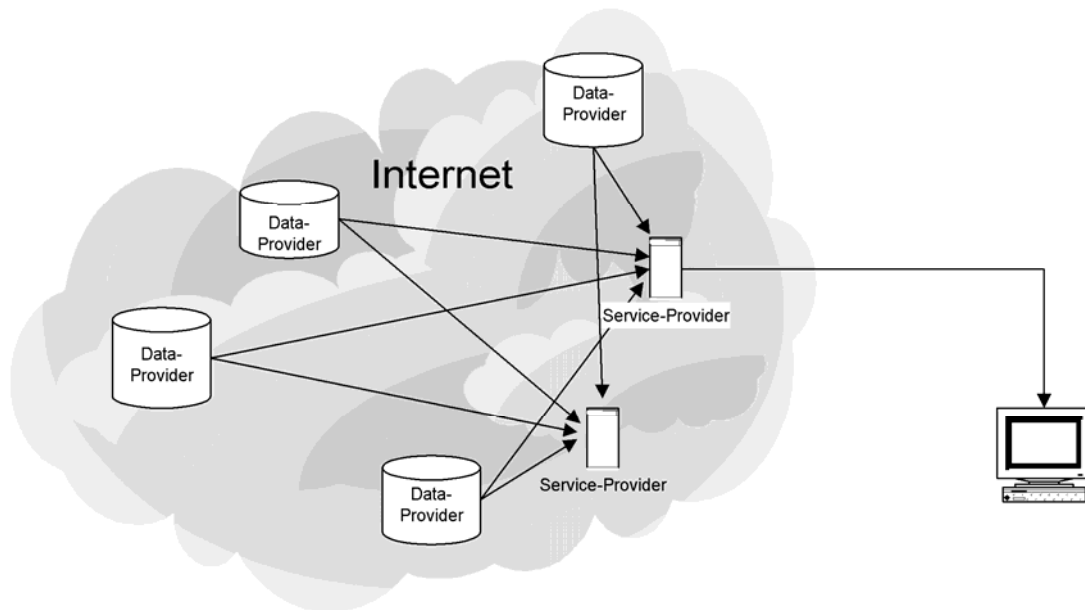


Abbildung 1: Zusammenwirken von Daten- und Service Providern

Das OAI-Protokoll ist kein Suchprotokoll. Qualifizierte Suchanfragen lassen sich über das Protokoll demnach nicht formulieren. Die letztlich dem Endnutzer oder einem Suchprotokoll zur Verfügung gestellte Suche ist Teil des Serviceproviders und bezieht sich immer auf dessen Datenbank. Die Möglichkeiten, Auswahlkriterien bei OAI-Protokollanfragen zu nutzen, beschränken sich auf das letzte Änderungsdatum der Metadaten (**from**- und **until**-Argument) und eine grobe logische Gliederung der Datenbestände in unterschiedliche Mengen (**set**-Argument). Die OAI-PMH Spezifikation beschreibt jedoch nicht die konkrete Aufteilung solcher Mengen, so dass dies den jeweiligen Daten Providern überlassen ist.

Diese Eigenschaft des OAI-PMH ermöglicht es, Metadaten von Daten Providern selektiv abzufragen. Beispielsweise lassen sich auf diese Weise Fachportale effizient realisieren, da der entsprechende Serviceprovider schon auf Protokollebene die Menge der angeforderten Daten grob einschränken kann. Abbildung 2 zeigt schematisch das Zusammenwirken von Daten- und Service Providern auf der Grundlage einheitlicher Definitionen über die logische Struktur, die so genannte Set-Hierarchie. Jeder Serviceprovider fordert nur die für ihn interessanten Daten an. Um ein hohes Maß an Interoperabilität innerhalb der dokumentarischen und bibliothekarischen Anwendungen zu gewährleisten und damit auch den Aufbau von Daten- und Serviceproviderstrukturen zu erleichtern, erscheint es sinnvoll, für die Definition und Verwendung von Set-Hierarchien Empfehlungen und Richtlinien zu entwickeln und für Ihre Anwendung zu werben. Dadurch wird es möglich, dass Serviceprovider gezielt Daten nach bestimmten formalen (z.B. Dissertationen) und inhaltlichen Kriterien (z.B. Physik) sammeln und spezifische Dienste (z.B. ein Nachweissystem für Dissertationen oder eine Suchmaschine für Physik-Dokumente) aufbauen können (siehe Abbildung 2).

Die DINI-Arbeitsgruppe Elektronisches Publizieren empfiehlt neben der Benutzung des OAI-Protokolls selbst eine inhaltliche und formale Strukturierung des Archivs, um den Aufbau spezifischer Dienste auf der Basis des OAI-PMH zu erleichtern. Diese Empfehlungen bilden den Schwerpunkt des vorliegenden Papiers und werden im folgenden Kapitel ausführlich dargelegt.

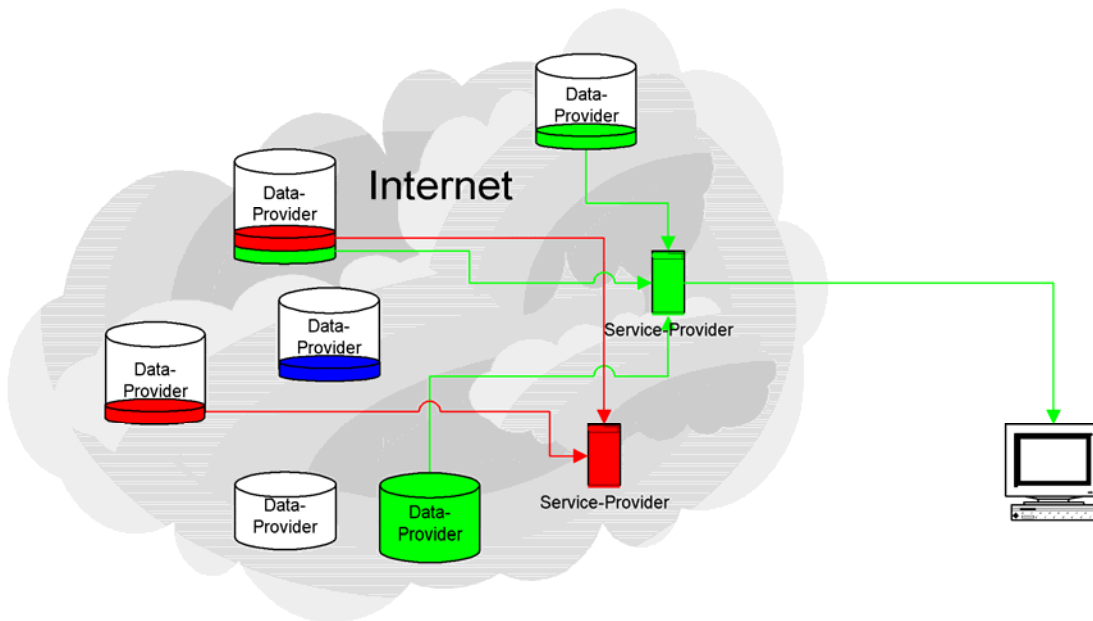


Abbildung 2: Zusammenwirken von inhaltlich oder formal strukturierten Daten- und Service Providern

2 Empfehlungen zur Definition von Sets

2.1 Allgemeine Bemerkungen

Ausgehend von einer Vielzahl möglicher auf dem OAI-Protokoll basierender Dienste ist eine Strukturierung der Archive sowohl nach formalen als auch nach inhaltlichen Kriterien sinnvoll. Die inhaltliche Beschreibung dient einer groben fachlichen Zuordnung der mit Metadaten beschriebenen Dokumente und Objekte und orientiert sich an den Sachgruppen der Deutschen Bibliothek⁵. Die formalen Untergliederungen beziehen sich auf die Publikationsform und den technischen Dokumenttyp des jeweiligen Objekts.

Die durch einen Datenprovider zur Verfügung gestellten Sets sind maschinell abfragbar. Eine entsprechende OAI-Anfrage liefert neben einem eindeutigen Bezeichner, der in dem **setSpec**-Element enthalten ist, eine verbale Beschreibung (**setName**-Element) des jeweiligen Sets. Damit eine Nutzung des Sets unabhängig von der Kenntnis der deutschen Sprache gewährleistet ist, sollte die Beschreibung vornehmlich in englischer Sprache erfolgen. Die hier empfohlenen Sets entsprechen vier unterschiedlichen Gliederungsansätzen:

- einer inhaltlichen Gliederung (ddc),
- einer Gliederung gemäß der Publikationsform (pub-type),
- einer Gliederung nach Dokumenttypen (doc-type) und
- einer Gliederung nach der qualitativen inhaltlichen Begutachtung (status-type).

Sie sind jeweils als zweistufige Hierarchie definiert, wobei die Hierarchieebenen gemäß der Spezifikation des OAI-PMH durch einen Doppelpunkt getrennt werden. Insgesamt werden damit vier Sets der obersten Hierarchieebene empfohlen, und zwar die Sets **ddc**, **pub-type**, **doc-type** und **status-type**. Die Sets der zweiten Hierarchieebene (z.B. **ddc:004**) sind als Teilmengen der jeweiligen Sets der obersten Ebene zu verstehen. Abgesehen von diesem hierarchischen Zusammenhang ist mit der Zugehörigkeit eines

⁵ Die Deutsche Bibliothek hat die Sachgruppen-Gliederung zum Bibliografie-Jahrgang 2004 auf ein auf der Dewey-Dezimalklassifikation (DDC) beruhendes System umgestellt. Die bis dahin geltende Sachgruppen-Gliederung der Deutschen Nationalbibliografie (DNB) wird nicht mehr verwendet.

Objekts zu einem bestimmten Set über die Zugehörigkeit desselben Objekts zu einem anderen Set nichts gesagt. Wünschenswert im Sinne dieser Empfehlungen und logisch durchaus nachvollziehbar wäre es, wenn jeder Metadatensatz in mindestens je einem Set der zweiten Hierarchieebene jedes Gliederungsansatzes vertreten wäre – also beispielsweise eine Dissertation auf dem Fachgebiet der Medizin in den Sets **ddc:610**, **pub-type:dissertation** und **doc-type:text**. Denkbar ist es darüber hinaus auch, dass ein Metadatensatz in mehreren Sets desselben Gliederungsansatzes enthalten ist, also beispielsweise in zwei Sets der inhaltlichen Gliederung (z.B. **ddc:004** und **ddc:610** bei einer Arbeit, die sich mit Informationstechnik in der Medizin beschäftigt). In der Regel wird es darüber hinaus auch Fälle geben, in denen ein Metadatensatz nicht in jeder Kategorie in einem Set vertreten ist. Die Gründe hierfür können in einer unvollständigen Metadatenerfassung oder in der Migration älterer Metadaten liegen.

Generell wird empfohlen, auf die **ListSets**-Anfrage, die nach allen verfügbaren Sets eines Datenproviders fragt, nur diejenigen Sets zurückzugeben, in denen auch mindestens ein Metadatensatz enthalten ist. Ein Datenprovider, der die vorliegenden Empfehlungen umsetzt, wird also in der Regel auf eine solche Anfrage nicht eine komplette Liste der in den folgenden Abschnitten aufgeführten Sets zurückgeben, sondern nur die aktuell verwendeten Sets liefern.

Neben den hier empfohlenen Sets ist es jedem Datenprovider überlassen, weitere, für die Einbindung des Servers in Fachgemeinschaften notwendige Sets zu definieren. Die hier empfohlenen Sets stellen eine Normierung dar, die genutzt werden sollte, wenn ohnehin eine inhaltliche, auf dem Publikationstyp basierende oder qualitative Einteilung der auf dem Server bereitgestellten Dokumente erfolgen soll.

2.2 Inhaltliche Gliederung

Als inhaltliche Gliederung OAI-kompatibler Datenprovider werden die von der Deutschen Bibliothek verwendeten Sachgruppen gemäß der Dewey-Dezimalklassifikation (DDC) empfohlen. Diese erlauben eine grobe fachliche Einordnung und bieten eine entsprechende Auswahlmöglichkeit für fachlich ausgerichtete Serviceprovider. Tabelle 1 zeigt die gemäß dieser Gliederung definierten Sets. Die erste Spalte enthält den jeweiligen Bezeichner (**setSpec**). Die zweite Spalte beschreibt den Set, der auf die OAI-Anfrage **ListSets** hin im **setName**-Element enthalten ist. Die letzte Spalte ist hier nur der Klarheit halber eingefügt worden. Ihr Inhalt spielt für die Protokollanfragen und -antworten keine Rolle.

Die erste Zeile von Tabelle 1 enthält das Set der obersten Hierarchieebene. In ihm enthaltene Metadatensätze beschreiben Dokumente, die eine DDC-konforme Klassifikation besitzen. Alle anderen Zeilen enthalten Sets der zweiten Hierarchieebene, die jeweils Teilmengen des Sets **ddc** bilden.

Tabelle 1: Bezeichnung und Beschreibung der Sets gemäß der inhaltlichen Gliederung

setSpec	setName	Deutschsprachige Beschreibung
ddc	DDC classified objects	Gemäß DDC klassifizierte Objekte
ddc:000	Generalities, Science	Allgemeines, Wissenschaft
ddc:004	Data processing Computer science	Informatik
ddc:010	Bibliography	Bibliografien
ddc:020	Library & information sciences	Bibliotheks- und Informationswissenschaft
ddc:030	General encyclopedic works	Enzyklopädien
ddc:050	General serials & their indexes	Zeitschriften, fortlaufende Sammelwerke
ddc:060	General organization & museology	Organisationen, Museumswissenschaft
ddc:070	News media, journalism, publishing	Geografie, Reisen

setSpec	setName	Deutschsprachige Beschreibung
ddc:080	General collections	Allgemeine Sammelwerke
ddc:090	Manuscripts & rare books	Handschriften, seltene Bücher
ddc:100	Philosophy	Philosophie
ddc:130	Paranormal phenomena	Parapsychologie, Okkultismus
ddc:150	Psychology	Psychologie
ddc:200	Religion	Religion, Religionsphilosophie
ddc:220	Bible	Bibel
ddc:230	Christian theology	Theologie, Christentum
ddc:290	Other & comparative religions	Andere Religionen
ddc:300	Social sciences	Sozialwissenschaften, Soziologie
ddc:310	General statistics	Statistik
ddc:320	Political science	Politik
ddc:330	Economics	Wirtschaft
ddc:340	Law	Recht
ddc:350	Public administration	Öffentliche Verwaltung
ddc:355	Military science	Militär
ddc:360	Social services; association	Soziale Probleme, Sozialarbeit
ddc:370	Education	Erziehung, Schul- und Bildungswesen
ddc:380	Commerce, communications, transport	Handel, Kommunikation, Verkehr
ddc:390	Customs, etiquette, folklore	Ethnologie
ddc:400	Language, Linguistics	Sprachwissenschaft, Linguistik
ddc:420	English	Englisch
ddc:430	Germanic	Deutsch
ddc:439	Other Germanic languages	Andere germanische Sprachen
ddc:440	Romance languages French	Französisch, romanische Sprachen allgemein
ddc:450	Italian, Romanian, Rhaeto-Romantic	Italienisch, Rumänisch, Rätoromanisch
ddc:460	Spanish & Portugese languages	Spanisch, Portugiesisch
ddc:470	Italic Latin	Latein
ddc:480	Hellenic languages Classical Greek	Griechisch
ddc:490	Other languages	Andere Sprachen
ddc:500	Natural sciences & mathematics	Naturwissenschaften
ddc:510	Mathematics	Mathematik
ddc:520	Astronomy & allied sciences	Astronomie
ddc:530	Physics	Physik
ddc:540	Chemistry & allied sciences	Chemie
ddc:550	Earth sciences	Geowissenschaften
ddc:560	Paleontology Paleozoology	Paläontologie
ddc:570	Life sciences	Biowissenschaften, Biologie
ddc:580	Botanical sciences	Pflanzen (Botanik)
ddc:590	Zoological sciences	Tiere (Zoologie)
ddc:600	Technology (Applied sciences)	Technik

setSpec	setName	Deutschsprachige Beschreibung
ddc:610	Medical sciences Medicine	Medizin
ddc:620	Engineering & allied operations	Ingenieurwissenschaften
ddc:630	Agriculture	Landwirtschaft, Veterinärmedizin
ddc:640	Home economics & family living	Hauswirtschaft
ddc:650	Management & auxiliary services	Management
ddc:660	Chemical engineering	Technische Chemie
ddc:670	Manufacturing	Industrielle Fertigung
ddc:690	Buildings	Hausbau, Bauhandwerk
ddc:700	The arts	Künste, Bildende Kunst allgemein
ddc:710	Civic & landscape art	Landschaftsgestaltung, Raumplanung
ddc:720	Architecture	Architektur
ddc:730	Plastic arts Sculpture	Plastik, Numismatik, Keramik, Metallkunst
ddc:740	Drawing & decorative arts	Zeichnung, Kunsthandwerk
ddc:741.5	Comics, Cartoons	Comics, Cartoons, Karikaturen
ddc:750	Painting & paintings	Malerei
ddc:760	Graphic arts Printmaking & prints	Grafische Verfahren, Drucke
ddc:770	Photography & photographs	Fotografie, Computerkunst
ddc:780	Music	Musik
ddc:790	Recreational & performing arts	Freizeitgestaltung, Darstellende Kunst
ddc:791	Public performances	Öffentliche Darbietungen, Film, Rundfunk
ddc:792	Stage presentations	Theater, Tanz
ddc:793	Indoor games & amusements	Spiel
ddc:796	Athletic & outdoor sports & games	Sport
ddc:800	Literature & rhetoric	Literatur, Rhetorik, Literaturwissenschaft
ddc:810	American literature in English	Englische Literatur Amerikas
ddc:820	English & Old English literatures	Englische Literatur
ddc:830	Literatures of Germanic languages	Deutsche Literatur
ddc:839	Other Germanic literatures	Literatur in anderen germanischen Sprachen
ddc:840	Literatures of Romance languages	Französische Literatur
ddc:850	Italian, Romanian, Rhaeto-Romanic literatures	Italienische, rumänische, rätoromanische Literatur
ddc:860	Spanish & Portuguese literatures	Spanische und portugiesische Literatur
ddc:870	Italic literatures Latin	Lateinische Literatur
ddc:880	Hellenic literatures Classical Greek	Griechische Literatur
ddc:890	Literatures of other languages	Literatur in anderen Sprachen
ddc:900	Geography & history	Geschichte
ddc:910	Geography & travel	Geografie, Reisen
ddc:914.3	Geography & travel Germany	Landeskunde Deutschlands
ddc:920	Biography, genealogy, insignia	Biografie, Genealogie, Heraldik
ddc:930	History of the ancient world	Alte Geschichte, Archäologie
ddc:940	General history of Europe	Geschichte Europas
ddc:943	General history of Europe Central Europe Germany	Geschichte Deutschlands

setSpec	setName	Deutschsprachige Beschreibung
ddc:950	General history of Asia Far East	Geschichte Asiens
ddc:960	General history of Africa	Geschichte Afrikas
ddc:970	General history of North America	Geschichte Nordamerikas
ddc:980	General history of South America	Geschichte Südamerikas
ddc:990	General history of other areas	Geschichte der übrigen Welt

Das folgende Beispiel zeigt einen Ausschnitt aus einer möglichen Antwort eines Datenproviders auf eine **ListSets**-Anfrage, der den Empfehlungen über die inhaltliche Gliederung folgt.

```

<set>
  <setSpec>ddc</setSpec>
  <setName>DDC classified objects</setName>
</set>
<set>
  <setSpec>ddc:004</setSpec>
  <setName>Data processing Computer science</setName>
</set>
<set>
  <setSpec>ddc:610</setSpec>
  <setName>Medical sciences Medicine</setName>
</set>

```

2.3 Gliederung nach dem Publikationstyp

Als zweite Möglichkeit der Gliederung eines OAI-kompatiblen Archivs mittels Sets wird eine Einteilung der Objekte nach deren formalem Publikationstyp empfohlen.

Tabelle 2 zeigt die verwendeten Publikationstypen (Spalte 3) mit den entsprechenden Bezeichnungen (Spalte 1) und Beschreibungen (Spalte 2) der dazugehörigen Sets.

Tabelle 2: Bezeichnung und Beschreibung der Sets nach dem Publikationstyp

SetSpec	SetName	Deutschsprachige Beschreibung
pub-type	Objects having a formal publication type	Objekte mit einem formalen Publikationstyp
pub-type:monograph	Books, Monographs	Bücher, Monographien
pub-type:article	Journal Articles	Zeitschriftenartikel
pub-type:dissertation	Dissertations and Professional Dissertations	Dissertationen und Habilitationen
pub-type:masterthesis	Diploma Theses	Diplomarbeiten
pub-type:report	Reports	Berichte
pub-type:paper	Papers	Papers
pub-type:conf-proceeding	Conference Proceedings	Tagungs- und Konferenzbeiträge
pub-type:lecture	Lectures	Vorlesungen
pub-type:music	Music	Musik
pub-type:program	Programs / Software	Programme / Software
Pub-type:play	Plays	Schauspiele / Theaterstücke
Pub-type:news	News	Nachrichten
Pub-type:standards	Standards	Standards

Die nachfolgende XML-Sequenz ist ein Ausschnitt aus einer möglichen OAI-Antwort auf die ListSets-Anfrage an einen Datenprovider.

```
<set>
  <setSpec>pub-type</setSpec>
  <setName>Documents having a formal publication type</setName>
</set>
<set>
  <setSpec>pub-type:monograph</setSpec>
  <setName>Books, Monographs</setName>
</set>
<set>
  <setSpec>pub-type:dissertation</setSpec>
  <setName>Dissertations and Professional Dissertations</setName>
</set>
```

2.4 Gliederung nach dem Dokumenttyp

Als dritte Kategorie für die Strukturierung der Daten eines Datenproviders wird der Dokumenttyp der digitalen Objekte zur Unterscheidung verwendet. In Tabelle 3 sind die unterstützten Dokumenttypen (Spalte 3) mit den jeweiligen Bezeichnern (Spalte 1) und Beschreibungen (Spalte 2) der dazugehörigen Sets dargestellt. Sie basieren im Wesentlichen auf den MIME Media Types (<http://www.iana.org/assignments/media-types/>).

Tabelle 3: Bezeichnung und Beschreibung der Sets nach dem Dokumenttyp

SetSpec	SetName	Deutschsprachige Beschreibung
doc-type	Objects having a formal document type	Objekte mit einem formalen Dokumenttyp
doc-type:text	Text	Text
doc-type:notes	Notes	Noten
doc-type:image	Images	Bilder
doc-type:audio	Audio files	Audiodateien
doc-type:video	Video files	Videodateien
doc-type:multimedia	Multimedia files	Multimediateien
doc-type:data	Data	Daten
doc-type-binary	Binary data, (executable) programs	Binärdaten, (ausführbare) Programme

Die nachfolgende XML-Sequenz ist ein Ausschnitt aus einer möglichen OAI-Antwort auf die ListSets-Anfrage an einen Datenprovider, der Metadaten über Videodateien besitzt.

```
<set>
  <setSpec>doc-type</setSpec>
  <setName>formal document-type</setName>
</set>
<set>
  <setSpec>doc-type:video</setSpec>
  <setName>Video files</setName>
</set>
```

2.5 Gliederung nach der Begutachtung

Peer Review bezeichnet allgemein die Bewertung oder Begutachtung eines Objekts oder Prozesses durch unabhängige Gutachter, die sog. „Peers“ (engl. für „Ebenbürtige“). Da in unterschiedlichen Wissenschaftsdisziplinen auch unterschiedliche Verfahren der inhaltlichen Bewertung und Qualitätssicherung von Publikationen angewandt werden, wurde für die vorliegenden Empfehlungen nur eine allgemeine Unterscheidung verwendet, die sowohl Peer Review als auch andere Begutachtungsverfahren (z.B. in den Geisteswissenschaften durch den oder die Herausgeber einer wissenschaftlichen Zeitschrift) mit einbezieht. Auch Dissertationen gelten in diesem Sinne durch das Prüfungsverfahren als begutachtet.

Tabelle 4: Bezeichnung und Beschreibung der Sets nach der Begutachtung

SetSpec	SetName	Deutschsprachige Beschreibung
status-type	Objects having a formal state	Objekte mit einem formalen Status
status-type:draft	Draft	Entwurf
status-type:not-reviewed	Not reviewed	Nicht begutachtet
status-type:reviewed	Reviewed	Begutachtet

Die nachfolgende XML-Sequenz ist ein Ausschnitt aus einer möglichen OAI-Antwort auf die ListSets-Anfrage an einen Datenprovider, der Metadaten über begutachtete Dokumente besitzt.

```
<set>
  <setSpec>status-type</setSpec>
  <setName>Objects having a formal state</setName>
</set>
<set>
  <setSpec>status-type:reviewed</setSpec>
  <setName>Reviewed</setName>
</set>
```

3 Datenprovider

3.1 Wie werde ich Datenprovider?

Obwohl die Daten gemäß dem OAI-PMH in XML übertragen werden, ist es nicht notwendig, die eigenen Daten auch in XML-Form zu speichern. Wichtig ist jedoch, dass die eigenen Metadaten in strukturierter Form (z.B. innerhalb einer SQL-Datenbank) vorliegen.

Verschiedene Datenprovider, die unter <http://www.openarchives.org/tools/tools.html> zu finden sind, unterstützen eine on-the-fly (also im Moment der Anfrage/Antwort) Übersetzung der aus einer Datenbank extrahierten Daten nach XML. Diese Softwarewerkzeuge setzen also in der Regel nur einen Webserver und eine Datenbank voraus. Dank der großen Auswahl an Tools werden verschiedenste Programmiersprachen unterstützt.

3.1.1 Wo sollte man sich registrieren?

Falls man einen eigenen Datenprovider aufgesetzt hat, empfiehlt es sich, diesen unter <http://www.openarchives.org/community/index.html> zu registrieren. Dabei wird dann auch geprüft, ob der eigene Datenprovider sich konform zu dem OAI-Protokoll verhält.

3.2 Dublin Core und andere Metadatenformate

Jeder Datenprovider soll entsprechend dem OAI-PMH seine Metadaten zumindest nach dem unqualifizierten Dublin-Core-Standard anbieten. Dies stellt keinerlei Beschränkung bezüglich der Verfügbarmachung beliebiger anderer Metadatenformate dar. Um einen qualitativ hochwertigen Metadaten austausch zwischen Daten- und Service Providern zu erreichen, ist es sinnvoll, sich – zumindest innerhalb von Communities – auf die Verwendung spezialisierter Metadatenformate zu einigen. Ähnlich wie bei der Festlegung von Sets, wie sie in diesem Papier empfohlen werden, ist es auch auf diesem Gebiet notwendig, entsprechende Einigungen zu erreichen. Diese haben allerdings einen weniger übergreifenden Charakter als die vorliegenden Empfehlungen zur Set-Hierarchie. Stattdessen müssen Empfehlungen zu Metadaten Sätzen und spezialisierten Set-Definitionen auf der Ebene von (fachlichen) Communities erarbeitet und verbreitet werden.

Das Metadatenformat Dublin Core definiert 15 Elemente, die im Einzelnen aufgeführt werden. Diese Felder können praktisch mit beliebigem Inhalt gefüllt werden. Um eine bessere Interoperabilität zu gewährleisten, sollte man sich jedoch an den Empfehlungen der Dublin-Core-Arbeitsgruppe orientieren.

Diese Empfehlungen werden in Tabelle 5 in verkürzter Form wiedergegeben, die ausführlichen Beschreibungen finden sich in den Empfehlungen der Arbeitsgruppe⁶.

⁶ <http://www.ietf.org/rfc/rfc2413.txt>

Tabelle 5: Empfehlungen für den Inhalt der DC-Elemente

Dublin-Core-Element	Empfehlungen
Title	beliebiger Text
Creator	beliebiger Text
Subject	beliebiger Text
Description	beliebiger Text
Publisher	beliebiger Text
Contributor	beliebiger Text
Date	Empfohlene Praxis ist ein Unterformat von ISO 8601 [W3CDTF] ⁷ und schließt Datumsangaben der Form YYYY-MM-DD ein.
Type	kontrolliertes Vokabular (z.B. das <i>DCMI Type Vocabulary</i> [DCT1] ⁸)
Format	kontrolliertes Vokabular (z.B. die Liste der <i>Internet Media Types</i> [MIME] ⁹ , in der digitale Medienformate definiert werden)
Identifizier	Zu den häufig verwendeten formalen Identifikationssystemen zählen der <i>Uniform Resource Identifier</i> (URI), der <i>Digital Object Identifier</i> (DOI) und die <i>International Standard Book Number</i> (ISBN). Der Inhalt dieses Feldes ist jedoch nicht auf diese Systeme beschränkt.
Source	beliebiger Text
Language	RFC3066 ¹⁰ in Verbindung mit ISO639 ¹¹ , welches zwei oder drei Buchstaben für die Sprachen benutzt. Beispiele sind "en" oder "eng" für englisch.
Relation	beliebiger Text
Coverage	beliebiger Text
Rights	beliebiger Text. Empfehlenswert ist jedoch die <i>Creative Commons License</i> ¹² , die automatisch ausgewertet werden kann.

Mittels des *DC Checker*¹³ lässt sich überprüfen, ob die Metadaten, die man über seinen eigenen Datenprovider anbietet, diesen Empfehlungen entsprechen.

⁷ <http://www.w3.org/TR/NOTE-datetime>

⁸ <http://dublincore.org/documents/dcmi-type-vocabulary/>

⁹ <http://www.isi.edu/in-notes/iana/assignments/media-types/media-types>

¹⁰ <http://www.ietf.org/rfc/rfc3066.txt>

¹¹ <http://www.loc.gov/standards/iso639-2/langhome.html>

¹² <http://creativecommons.org/worldwide/de/>

¹³ <http://harvest.physik.uni-oldenburg.de/dc/dcchecker.php>

4 Serviceprovider

4.1 Wie werde ich Serviceprovider?

Es gibt bereits viele (freie) Softwarepakete, mit dessen Hilfe man einen Serviceprovider aufsetzen kann. Unter <http://www.openarchives.org/tools/tools.html> lässt sich eine aktuelle Übersicht finden.

Unter <http://www.openarchives.org/documents/> findet man weitere interessante Texte wie Tutorials und FAQs.

4.1.1 Wo sollte man sich registrieren?

Falls man einen eigenen Serviceprovider aufgesetzt hat, empfiehlt es sich, diesen unter <http://www.openarchives.org/community/index.html> zu registrieren. Dabei wird dann auch geprüft, ob der eigene Serviceprovider sich konform zu dem OAI-Protokoll verhält.

4.2 Beispiele für Daten- und Serviceprovider-Dienste

4.2.1 Proprint

Der Print-on-Demand-Service ProPrint¹⁴ ist ein Gemeinschaftsprojekt der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) und des Computer- und Medienservices der Humboldt-Universität zu Berlin. Der Service bietet die Möglichkeit, wissenschaftliche Aufsätze, elektronische Dissertationen oder Beiträge aus digitalisierten Büchern online zu recherchieren, individuell zusammenzustellen und als Paperback oder Skript von Druckdienstleistern produzieren zu lassen. ProPrint bietet bereits einen Zugriff auf über viertausend Monographien und Hochschulschriften an. Dabei wird das Netzwerk von Dokumentenservern und die Einbindung lokaler Druckunternehmen mit Print-on-Demand-Ausstattung kontinuierlich erweitert. Ziel ist es, die individuelle Publikation mit dem Pro-Print-Service¹⁵ in jeder wichtigen deutschen Universitätsstadt verfügbar zu machen.

Die zentrale ProPrint-Suchmaschine ist im Sinne von OAI der Serviceprovider, und die einzelnen Dokumentenserver sind OAI-Datenprovider.

Voraussetzung für die OAI-Kompatibilität eines Dokumentenservers ist dessen Fähigkeit, Dublin Core als Metadatenformat auszuliefern. Für die Beschreibung weiterer Metadaten, die für den ProPrint-Dienst erforderlich sind, wurden innerhalb dieses Metadatenformates weitere Elemente mit einem gesonderten Namensraum definiert¹⁶. Dieser Namensraum enthält auch Elemente des DIEPER-Metadatensatzes.

Die ProPrint-Erweiterung des Metadatenformates, der ProPrint-Metadatensatz, umfasst Informationen für besondere Seitenformate, Vertriebsinformationen und Gliederungen eines Dokumentes (Kapitel, Unterkapitel).

Über die zentrale ProPrint-Suchmaschine haben die Wissenschaftler und Studierenden so einen unbegrenzten Zugang zu elektronischen Dokumenten. Der ProPrint-Dienst kann aber nicht nur über den zentralen Zugang genutzt werden. Jeder angeschlossene Dokumentenserver kann einen so genannten Pro-Print-Button auf seinen Internetseiten

¹⁴ <http://rdd.sub.uni-goettingen.de/wiki/ProPrint/HomePage>

¹⁵ <http://www.proprint-service.de/>

¹⁶ http://edoc.hu-berlin.de/epubDocs/edocDocs/material/pdf/PP_Application_Profile_20021015.pdf

anbringen¹⁷. Damit erweitert sich das Dienstleistungsangebot der Bibliothek um einen Print-on-Demand-Dienst.

4.2.2 PhysNet

Physiker publizieren relevantes Wissen je nach Forschungsschwerpunkt über sehr verschiedene Publikationskanäle. Auch ist in der Physik sehr verschiedenes Material wissenschaftlich relevant; neben Buch- und Zeitschriftenpublikationen sind dies Tagungsbeiträge, Instituts- und Forschungsberichte, Messdaten, Auswertungsalgorithmen, Dissertationen, und insbesondere in der Theoretischen Physik Preprint-Publikationen wie z.B. in ArXiv¹⁸. Dies spiegelt den Bedarf der Physiker nach einer schnellen und möglichst instantanen Kommunikation wider.

Das PhysNet-Angebot PhysDoc bietet deshalb einen Serviceprovider, der einen gemeinsamen Zugang zu vielen Physik-relevanten Datenquellen ermöglicht, die einen OAI-Datenprovider anbieten. Zusätzlich bedient sich dieser Serviceprovider des Index-Levels von Suchmaschinen um auch jene Publikationen zu erschließen, die nicht über OAI-Datenprovider publiziert werden, sondern auf den Webservern von Physik-relevanten Institutionen von Wissenschaftlern im Rahmen des „Self Archiving“ publiziert und dort mit Dublin Core Metadaten beschrieben wurden. Die Qualitätskontrolle erfolgt dabei durch die Auswahl der Webserver und die Anforderung einer qualifizierten Metadatenerschließung.

Dieses Beispiel eines fachspezifischen Serviceproviders zeigt, dass eine nachvollziehbare und dokumentierte Beschreibung der Objekte mittels „sets“ im OAI-PMH notwendig ist, so dass der Serviceprovider auf einfache Weise lediglich die Physik-relevanten Sets importieren und anbieten kann.

Auf die Suche kann über <http://www.physnet.de/PhysNet/physdoc.htm> zugegriffen werden.

4.2.3 Providerdienst des Hochschulbibliothekszentrums NRW (hbz)

MeIND (Metadata on Internet Documents)¹⁹ ist der vom hbz angebotene aggregierende OAI-Serviceprovider. Dem Nutzer werden zur Recherche elektronische Hochschulschriften und weitere hochschulrelevante Open-Access-Inhalte im Einzugsgebiet des hbz zugänglich gemacht. Des Weiteren werden alle an der Deutschen Bibliothek (DDB) gemeldeten, außerhalb des Einzugsgebiets des hbz veröffentlichten Online-Dissertationen über die Suchoberfläche angeboten. Hierzu wird der Deposit Server der DDB täglich abgefragt. Die dritte Säule des Metadatenbestands des Serviceproviders MeIND bilden ausgesuchte, weltweite Datenprovider, die über eine hohe Informationsqualität für die deutsche Universitäts-/Forschungsgemeinschaft verfügen.

Als aggregierender Datenprovider stellt MeIND seinen Datenbestand über die OAI-Schnittstelle zur Verfügung. Zusätzlich zum unqualifizierten Dublin-Core-Standard werden die Metadaten als MARCXML über die OAI-Schnittstelle präsentiert. MeIND ist daher gleichzeitig Daten- und Serviceprovider und bietet neben dem Zugang auf die Metadaten über eine einfache und erweiterte Suchoberfläche die Suchmöglichkeit über einen sammlungs-basierten Zugang sowie über die DigiBib²⁰. Weiterhin stellt MeIND ein browsen nach Dokumentenarten und ein browsen nach DDC-Kategorien zu Verfügung. Hierbei sind die Metadatensätze über 100 Sammlungen zugeordnet. MeIND bietet den Service eines "Warenkorbsystems" an. Für registrierte Benutzer werden Suchergebnisse

¹⁷ Die Dokumentenserversoftware OPUS wird in der Version 3.0 diese Funktion und ein entsprechendes Lizenzmodul z.B. integriert haben.

¹⁸ www.arxiv.org

¹⁹ <http://www.meind.de>

²⁰ <http://eris.hbz-nrw.de/>

archiviert, selektierte Daten können selbstdefinierten Körben zugeordnet werden. Es gibt einen Benachrichtigungsservice, bei dem der Benutzer per Mail über neu eingestellte, ihn betreffende Metadaten informiert werden kann. Einem Gastbenutzer stehen diese Funktionalitäten nur eingeschränkt zur Verfügung. Als Zusatzinformationen zu den Suchergebnissen kann sich der Benutzer eine Rankingliste ähnlicher Datensätze anschauen und eine Liste von Datensätzen betrachten, die sich andere Benutzer angeschaut haben, die den gesuchten Datensatz aufgerufen haben (Recommender System).

4.3 Providerdienst des Bibliotheksservicezentrums Baden-Württemberg

Das Bibliotheksservicezentrum Baden-Württemberg speichert in seinem virtuellen Medienserver unter anderem die Metadaten elektronischer Ressourcen aus den angeschlossenen Hochschulen²¹. Sofern es sich um elektronische Hochschulschriften handelt, werden die Daten über eine OAI-Schnittstelle mit einem auf Dublin Core Simple basierenden, jedoch erweiterten Datenformat aus den lokalen Dokumentspeichern der Hochschulen (meist OPUS-Systeme) abgerufen. Das BSZ stellt nun die Metadaten im Rahmen des virtuellen Medienservers wiederum über eine OAI-Schnittstelle für Serviceproviderdienste zur Verfügung. Es handelt sich dabei wie im Falle des hbz um einen aggregierenden oder kumulativen Datenproviderdienst, bei dem eine wesentlich größere Menge von Daten angeboten werden kann (Stand September 2005: ca. 30.000 Datensätze). Für Serviceprovider hat dies wiederum den Vorteil, dass die Zahl der einzubindenden Quellen und der damit verbundene Pflegeaufwand überschaubarer wird. Aufgrund der Heterogenität der Quellen auch im Einzugsbereich des BSZ werden die vorliegenden Empfehlungen bezüglich der inhaltlichen Gliederung gemäß DDC durch den Medienserver des BSZ derzeit noch nicht unterstützt. Diese Erweiterung der OAI-Schnittstelle des Medienservers wird jedoch 2006 für die Hochschulschriften realisiert.

5 Empfohlene Links und Literatur zu OAI

Auf dem Webserver der Open Archives Initiative (<http://www.openarchives.org>) finden Sie Informationen zu dem Protokoll selbst sowie eine Übersicht über Software, die Sie in die Lage versetzt, selbst Daten- oder Serviceproviderdienste anzubieten.

²¹ <http://www.bsz-bw.de/diglib/objekte.html>

6 Glossar

Datenprovider

Als Datenprovider wird eine OAI-kompatible Schnittstelle zu einer Datenbank verstanden, in der sich Metadaten über Dokumente oder andere digitale Objekte befinden und die über eine HTTP-Verbindung erreichbar ist. Sie muss dazu in der Lage sein, OAI-Anfragen entsprechend der Protokolldefinition korrekt zu beantworten.

Serviceprovider

Der Serviceprovider bietet mithilfe von Daten, die er unter Nutzung des OAI-PMH gesammelt hat, einen (innerhalb der Protokollspezifikation nicht näher definierten) Dienst an. Der aus Sicht des Protokolls relevante Teil des Serviceproviders, der so genannte Harvester, versendet OAI-konforme Anfragen an Datenprovider und wertet die entsprechenden Antworten aus.

Harvest

Harvest, engl. ernten. Dabei werden regelmäßig alle verfügbaren bzw. relevanten Daten aus den verwendeten Datenquellen abgefragt (geerntet) und in einer zentralen Datenbank gespeichert, innerhalb derer dann die eigentliche Suche erfolgt.

Cross Search

bekannt auch unter dem Namen Metasuche oder Federated Searching, bezeichnet die parallele oder sequentielle Suche in verschiedenen Datenquellen, wird beispielsweise häufig von Meta-Suchmaschinen verwendet

Set

set, engl. = Menge

Impressum

Diese Empfehlungen wurden von Mitgliedern der DINI-Arbeitsgruppe "Elektronisches Publizieren" erarbeitet. Sie sind auf dem DINI Server unter <http://www.dini.de/> veröffentlicht und stehen im Print-On-Demand-Verfahren zur Verfügung²².

Für kritische Hinweise, Korrekturvorschläge und ergänzende Bemerkungen sind wir dankbar. Zur besseren Koordination einer möglichen Diskussion bitten wir Sie, Ihre Anmerkungen per E-Mail an die DINI-Geschäftsstelle (gs@dini.de) zu senden.

<i>Name</i>	<i>Vorname</i>	<i>Institution</i>	<i>E-Mail</i>
Diekmann	Bernd	Carl-von-Ossietzky-Universität Oldenburg, Bibliotheks- und Informationssystem	diekmann@bis.uni-oldenburg.de
Dobratz	Susanne	Humboldt-Universität zu Berlin, Universitätsbibliothek	dobratz@cms.hu-berlin.de
Dr. Klotz-Berendes	Bruno	Hochschulbibliothek Münster	klotz-berendes@fh-muenster.de
Müller	Uwe	Humboldt-Universität zu Berlin, Computer- und Medienservice	u.mueller@cms.hu-berlin.de
Schulz	Matthias	Humboldt-Universität zu Berlin, Computer- und Medienservice	matthias.schulz.1@cms.hu-berlin.de
Scholze	Frank	Universität Stuttgart, Universitätsbibliothek	scholze@ub.uni-stuttgart.de
Dr. Stamerjohanns	Heinrich	International University Bremen, Electrical Engineering & Computer Science	h.stamerjohanns@iu-bremen.de

²² <http://edoc.hu-berlin.de/series/dini/>