Deutsche Initiative für Netzwerkinformation (DINI)

Electronic Publishing in Higher Education
How to design OAI interfaces

- Recommendations -

Working Group Open Archives Initiative in Germany

October 2003

Table of Contents

Abstract

The *Open Archives Protocol for Metadata Harvesting* (OAI-PMH) allows sharing metadata serving to describe arbitrary objects with others. In addition to a short overview of the protocol the paper on hand contains recommendations for the application of *Sets* by German data providers and for the proper usage of the metadata elements of *Dublin Core* (DC). Thereby the target is pursued to ensure an efficient metadata exchange between the different users of the OAI protocol.

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

The *Open Archives Protocol for Metadata Harvesting* (OAI-PMH) allows for an efficient metadata exchange. It implies a functional split-up between providers of (documents and) metadata, so-called *data providers*[1], and services based thereon (service providers[2]). The OAI-PMH is based upon the fundamental principle of the so-called *harvesting*[3], which in contrast to the *cross searching*[4] approach uses an asynchronous search model. That means that the service provider queries the metadata of the data providers at regular intervals und stores them in its local database. When using the harvesting approach concrete (user) search enquiries are answered with the aid of this database exclusively.

The OAI protocol is based on widely spread and accepted standards. It rests upon the *Hypertext Transfer Protocol* (HTTP) and uses the *eXtensible Markup Language* (XML) to encode the metadata and other information implied within the responses. While the OAI protocol is suited for transmission of metadata in arbitrary formats (defined by an XML schema) *Dublin Core* has been included in the protocol specification as least common denominator for reasons of interoperability. OAI compliant data providers have to be able to deliver at least Dublin Core for their metadata. Thus, communication and an effective exchange of metadata between arbitrary OAI compliant data and service providers are possible at once.

Naturally, the split-up between data providers and service providers defined in the OAI-PMH specification does not exclude the development of services which contain both functionalities. This possibility is exploited by so-called *aggregating* data providers. On the one side they use the OAI protocol to harvest the available data of a certain set of data providers. Following, they hold these data ready via an OAI interface for queries of other service providers.

For users of services based on the OAI-PMH the underlying technology is normally invisible. E.g., they use a web interface with which the can communicate with the service provider and use its services. The fact that the found digital objects are located on distributed services becomes visible for the users only when retrieving them or when being prompted for authentication (see Figure 1).

---

[1] The term data provider describes an OAI compliant interface to a database containing metadata about documents or other digital objects. The interface is accessible via an HTTP connection. It must be able to correctly answer OAI requests according to the protocol specification.

[2] With the aid of data collected using the OAI-PMH a service provider offers services not elaborated on in the protocol specification. The harvester which from the protocol's point of view is the relevant part of a service provider posts OAI compliant requests to data providers and accordingly analyzes the received answers.

[3] Thereby all available / relevant data are inquired (harvested) regularly und stored in a central database which the actual user driven search is based upon.

[4] Also known as *Federated Searching* describes the immediate search in all utilized databases, e.g. often used within meta search engines.
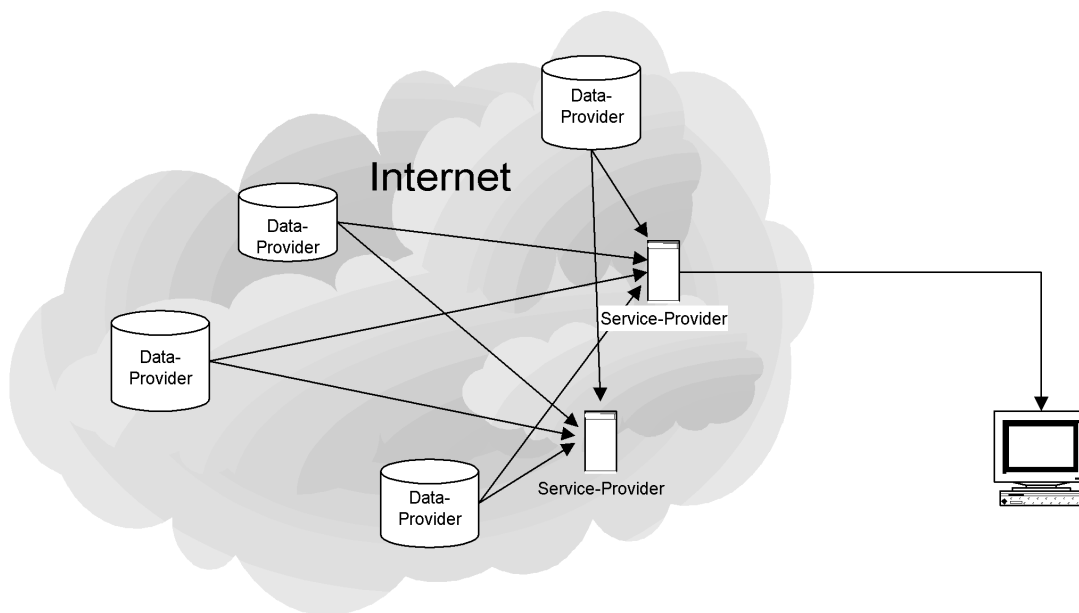
Figure 1: Cooperation of data providers and service providers

The OAI protocol is not a search protocol. Thus, qualified search queries cannot be formulated within the protocol. The actual search service provided for end users or for a search protocol is part of the service provider and always refers to its database. The possibilities to use selection criteria within the OAI protocol requests are limited to the date of the recent metadata change (**from** and **until** argument) and a rough logical classification of the available data into different collections and sets (**set** argument). The set definitions are not included in the OAI-PMH specification – they are left to the individual data providers.

This property of the OAI-PMH allows for a selective harvesting of the data provider's metadata. For example, by this means it is possible to realize subject gateways very efficiently because the service provider can roughly confine the data to be harvested already on the protocol level. Figure 2 schematically shows the coaction of data and service providers on the basis of a uniform definition of the logical structure – the so-called set hierarchy. Each service provider only asks for the data which are relevant for its service.

In order to ensure a high degree of interoperability within documentary and library applications and to facilitate the constitution of structures of data and service providers it appears to be reasonable to develop recommendations and guidelines for the definition of set hierarchies and its application. Thus, it becomes possible for service providers to selectively collect data according to certain formal (e.g. dissertation) and subject (e.g. physics) criteria and to build up specific services (e.g. a search engine for documents on physics) (see Figure 2).

Besides the usage of the OAI protocol itself the DINI working group *Electronic Publishing in Higher Education* recommends a subject and a formal structuring of the repositories in order to facilitate the development of specific services based on the OAI-PMH. These recommendations form the main focus of the paper on hand. They are presented extensively in the following chapter.
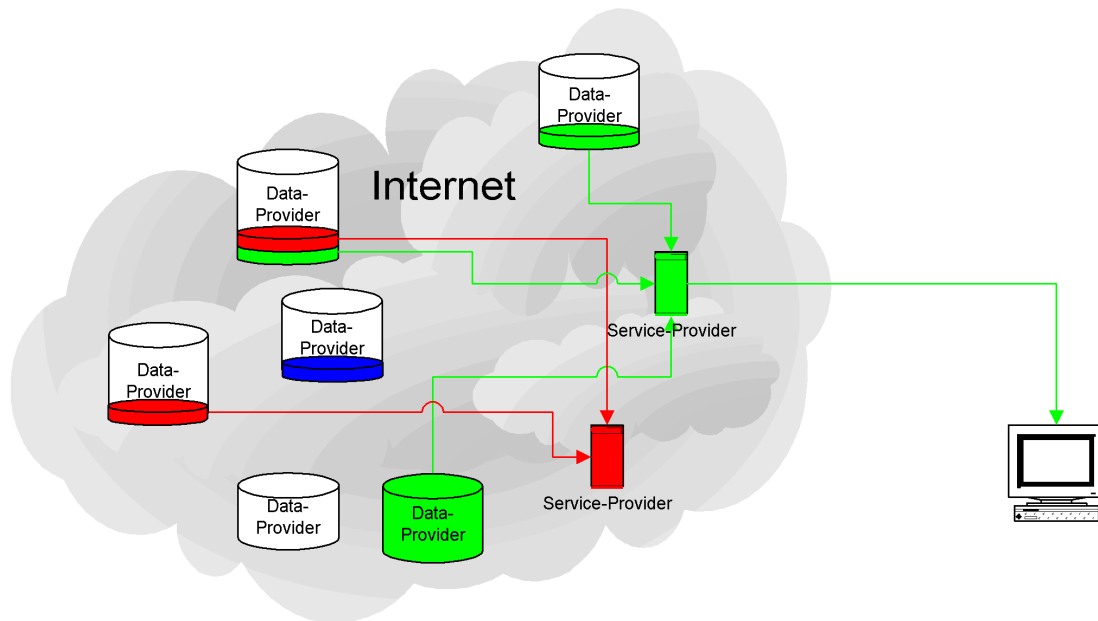
Figure 2: Coaction of logically structured data providers and service providers

2 Recommendations for the Definition of Sets

2.1 General Remarks

Starting from a variety of possible services based on the OAI protocol a structuring of the repositories according to subject as well as to formal criteria appears to be sensible. The subject based classification serves as a rough textual subsumption of the documents and objects described by the metadata. This classification is oriented on the functional groups used by *Die Deutsche Bibliothek*[5] (DDB, The German National Library). The formal classification refers to the publication types and the technical document types of the described objects.

The sets made available by a data provider can be inquired automatically. The answer to an adequate OAI request includes a unique identifier implied in the **setSpec** element and a verbal description (**setName** element) of the respective sets. To allow for the usage of sets independent of the knowledge of the German language the description should be primarily in English.

The sets recommended here accord to three different classification approaches – a subject classification, a classification according to the publication type and a classification corresponding to document types. They are defined as a two-stage hierarchy in each case, whereas the hierarchy levels are separated by a colon according to the protocol specification. Altogether, there were defined three top level sets, i.e. the sets **ddc**, **pub-type** und **doc-type**. The sets of the second hierarchy level (e.g. **ddc:004**) are defined as subsets of the respective top level set. Except from this hierarchical correlation there cannot be assumed anything about the membership of an item to a certain set given the membership of this item to another set. Certainly, it would be desirable and absolutely comprehensible from a logical point of view if each item would be member in at least one set of the second hierarchy level of each classification approach. For example, a dissertation in medical sciences should appear in the sets **ddc:610**, **pub-type:dissertation** and **doc-type:text**. Furthermore, it is also supposable that an item is member of more than one set of one classification approach, for example in two sets according to the subject classification (e.g. **ddc:004** and **ddc:610** for a document dealing with IT matters in medical sciences). Normally, there will also be cases where an item is not classified according to each classification approach. This may result from

---

[5] Die Deutsche Bibliothek changes its classification starting with the bibliography volume 2004 to a system which is based on the Dewey Decimal Classification (DDC). The Deutsche Nationalbibliografie (DNB) applied before will not be used anymore.

incomplete metadata compilation or from migrating older metadata.

When answering a **ListSets** request, it is generally recommend to solely delivering those sets which at least have one member item. Thus, normally a data provider who has implemented the available recommendations would not return a complete list of the sets described in the following sections when answering a **ListSets** request but only a list of the sets actually used.

2.2 Subject Classification

As a rough subject classification for OAI compliant data providers the classification used by Die Deutsche Bibliothek which accords to the Dewey Decimal Classification (DDC) is recommended. This allows for the application of a subject oriented selection criterion. Table 1 shows the sets defined according to this classification. The first column contains the respective identifier (**setSpec**), in the second column the English descriptions of the sets are implied, which are contained in the **setName** element when answering the OAI request **ListSets**. The last column implies the German descriptions which have no impact on protocol requests and resposnes.

The first row of Table 1 contains the top level set. Items which are member to this set describe documents having a DDC conformable classification. All the other rows contain sets of the second hierarchy level which represent subsets of the set **ddc**.

Table 1: Identification and description of the sets according to the subject classification

setSpec	setName	German Description
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
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

setSpec	setName	German Description
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
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

setSpec	setName	German Description
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
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

The following example shows a section of a possible data provider's answer to a **ListSets** request.

```
<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   Classification According to the Publication Type

The second possibility to classify an OAI compliant repository is based on the differentiation according to the formal publication type.

Table 2 shows the different publication types with the respective identifiers (column 1) and descriptions (column 2) of the sets.

Table 2: Identification and description of the sets according to publication types

SetSpec	SetName	German Description
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

The following part of an XML file represents an example of a part of an OAI answer to the a **ListSets** request.

```
<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 Classification According to the Document Type

As the third approach to structure a data repository a classification according to the document type is used. In Table 3 the supported document types are shown with the identifiers (column 1) and descriptions (column 2) of the respective sets.

Table 3: Identification and description of the sets according to the document type

SetSpec	SetName	German Description
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	Multimediadateien
doc-type:data	Data	Daten
doc-type-binary	Binary data, (executable) programs	Binärdaten, (ausführbare) Programme

The following XML sequence is a part of possible OAI response answering a **ListSets** request to a data provider offering metadata of video files.

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

3 Further Recommendations

3.1 Dublin Core und other Metadata Formats

According to the OAI protocol specification each data provider should be able to deliver its metadata at least according to the Dublin Core standard. This requirement does not result in any restrictions concerning other metadata formats and its transmission via the OAI-PMH. In order to achieve a highly valuable metadata exchange between data providers and service providers it is sensible to come to agreements on specialized metadata formats – at least within in communities. Unlike the recommendations for sets as proposed in the previous chapter these agreements on metadata formats have a less general scope. Recommendations for metadata formats and more specialized set definitions have to be developed and distributed by (subject) communities.

The metadata format Dublin Core defines 15 elements. These fields can virtually be filled with arbitrary content. To ensure a better interoperability, however, it is suggested to orient on the recommendations of the Dublin Core working group.

These recommendations are shown in an abbreviated form in Table 4. The detailed descriptions are contained in the recommendations of the Dublin Core working group[6].

---

[6] http://www.ietf.org/rfc/rfc2413.txt

Table 4: Recommendations for the content of the Dublin Core elements

Dublin Core Element	Recommendations
Title	arbitrary text
Author	arbitrary text
Subject	arbitrary text
Description	arbitrary text
Publisher	arbitrary text
Contributor	arbitrary text
Date	Recommended is a subformat of ISO 8601 [W3CDTF][7]. This includes date statements in the format YYYY-MM-DD.
Type	controlled vocabulary (e.g. the *DCMI Type Vocabulary* [DCT1][8])
Format	controlled vocabulary (e.g. the list of *Internet Media Types* [MIME][9], in which digital media formats are defined)
Identifier	Formal identification systems, e.g. *Uniform Resource Identifier* (URI), *Digital Object Identifier* (DOI) and the *International Standard Book Number* (ISBN). This content of this field is not restricted to theses systems.
Source	arbitrary text
Language	RFC3066[10] in conjunction with ISO639[11], using two or three letters for the languages (e.g. "en" or "eng" for English).
Relation	arbitrary text
Coverage	arbitrary text
Rights	arbitrary text. It is recommended to use the *Creative Commons License*, which can be analyzed automatically.

Using the *DC Checker*[12] it can be checked whether metadata exposed by a data provider accord to these recommendations.

---

[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://harvest.physik.uni-oldenburg.de/dc/dcchecker.php

Imprint

These recommendations were developed by the DINI working group *Open Archives Initiative in Germany*. They are available at the DINI web server at http://www.dini.de/. Hints, corrections and other proposals are highly appreciated by the authors. In order to allow a better coordination of possible discussions on the content please use the email address of the DINI office (*gs@dini.de*) to send us your annotations.

Name	Forename	Institution	Email
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
Scholze	Frank	Universität Stuttgart, Universitätsbibliothek	scholze@ub.uni-stuttgart.de
Dr. Stamerjohanns	Heinrich	Institute for Science Networking, Carl-von-Ossietzky-Universität Oldenburg	stamer@uni-oldenburg.de