

OPEN ARCHIVES FORUM: EXPERIENCES – IMPLEMENTATION TESTS: OAI SERVICE PROVIDER

Project Number:	IST-2001-32015
Project Title:	Open Archives Forum

Date of Delivery:	30-09-2003
Title:	Experiences – Implementation Tests: OAI Service Provider
Workpackage contributing to the Document:	WP2
Total Number of Pages:	7
URL:	http://www.oaforum.org/otherfiles/tv-service-impl.pdf
Author:	Uwe Müller, Susanne Dobratz with Matthias Schulz, JingYuan Wang, Birgit Matthaei
Contact Details:	Uwe Müller Humboldt-University Berlin, CMS (Computing Centre) Rudower Chaussee 26, 12489 Berlin, Germany

TABLE OF CONTENTS

1	Own experiences - Implementation tests: OAI service provider	2
1.1	Experiences and further development of the service provider at Humboldt University	2
1.2	ProPrint	3
1.2.1	ProPrint for publishers and libraries	3
1.2.2	ProPrint for printing services providers	4
1.2.3	ProPrint Service	4
1.2.4	ProPrint Technology	6

1 OWN EXPERIENCES - IMPLEMENTATION TESTS: OAI SERVICE PROVIDER

1.1 Experiences and further development of the service provider at Humboldt University

The OAI protocol provides an efficient mechanism to exchange metadata. It implies a functional division into data providers and service providers, the latter of which building up their services on the basis of metadata harvested via the OAI protocol. Data providers offer metadata via OAI interfaces which can be prompted by service providers. Based on the collected metadata service providers may create certain services (e.g. special information service, document delivery service, search machine, subject gateway). The end user only needs to contact the service provider's user interface to find documents on distributed document servers that usually are administered independently. Due to the principal approach of the OAI protocol it is necessary for service providers to have an own database to store the harvested metadata. The queries issued to provide the service itself to the user always relate to this central database administered and maintained by the service provider (see chapter 6). To realise an OAI based service provider the following modules have been developed by the Computer and Media Service of Humboldt University:

- An own database structure to store the metadata,
- an administrative interface to update and store the metadata,
- an automated update mechanism allowing to harvest archives regularly according to a configurable schedule plan, and
- a search machine, with which the user can search the desired metadata of documents on decentralised document servers according to different search criteria.

The service provider uses standard technologies, e.g. PHP4 as programming language and Sybase/MySQL as database management system. The service provider supports the OAI protocol versions 1.0, 1.1, 2.0 and the metadata format Dublin Core as well as the OAI sets (standardised structures for description of archives), which allows a rough selection of metadata and an advanced search.

The service is available at http://edoc.hu-berlin.de/e_suche/oai.php (german language).

Suche

Suchmöglichkeiten

- Metadaten
- Volltext
- RZ-Mitteilungen
- HU Dissertationen
- Dissertationen deutschlandweit
- Dissertationen weltweit
- Die Deutsche Bibliothek
- OAI Suche**

Suche in verteilten OAI Archiven

Suche nach ...

Titel:

Autor:

Schlagwort:

Zusammenfassung:

Suche nur in ...

Sprache:

Datum: .JJJJ

Zeitraum: ab .JJJJ bis .JJJJ

Archiv:

Fachgebiet:

Dokumenttyp:

Sortiert nach Datum

Treffer pro Seite: [Hilfe Statistik](#)

While implementing and using the service provider the following difficulties on the part of the data providers have been identified:

- URLs of some data providers are not available.
- Some data providers do not deliver their data records encoded in UTF-8 standard, thus XML-errors occur during the collection of metadata.
- Some data providers do not support both, the GET and the POST methods.

In order to operate the service provider appropriately the following recommendations are reasonable:

- Data providers should deliver date and language information accordingly to the respective ISO standards in order to provide a more exact filter mechanism of the search results.
- Archives should apply persistent identifiers (e.g. URNs) on the documents, so that a further requirement is ensured for the long-term archives.
- Data providers should agree on consistent and interoperable set definitions for the archives.

To match the last of these suggestions the DINI (German Initiative for Networked Information) has published recommendations for German OAI data providers defining different hierarchies of sets and their respective semantics. The set definitions describe an allocation of document servers with respect to subject categories, the document type of the digital object and its formal publication type. The service provider developed at Humboldt University integrated these recommendations and for the first time demonstrated the usefulness of the application of widely accepted set recommendations. Among other things subject gateways and browsing mechanisms can be realised very conveniently applying standardised set definitions.

Nevertheless, the service provider has to be enhanced in the near future:

- It should support metadata formats different from Dublin Core.
- The advanced search function should be extended using the relevance ranking.
- The browsing functionality has to be extended.

Within the scope of the considerations for the improvements of the search service the idea was developed to offer search functions in all document archives to the end users without having to store all metadata themselves. This can be realised with the help of a search agent which issues retrieval queries to different search services by an interface, subsequently summarises the answers and finally presents the results to the user. A respective search agent has been developed in the context of a diploma thesis at Humboldt University in 2002/03.

1.2 ProPrint

From 2000 – 2003 the Electronic Publishing Group of Humboldt University and the State and University Library Lower Saxony Göttingen run the project ProPrint (<http://www.ProPrint-service.de>). Within this project a print-on-demand service has been developed. This project was funded by the German Ministry of Education and Research and the Association for the German Research Network (DFN-Verein).

1.2.1 ProPrint for publishers and libraries

ProPrint offers readers, libraries and publishing houses an interconnection of digital archives on a standardised user interface for a print-on-demand service. This offer currently covers more than 2000 monographs, more than 1000 magazines and journals, more than 1000 scientific essays from students and other university members.

ProPrint consist of a LAMP-system (that is: operating system **L**inux, **A**pache-server, **M**ySQL-database, and **P**HP script language), an Allegro database developed for bibliographical purposes, and a commercial pdf-Merge-Program. Based on a given ProPrint Application Profile the ProPrint software collects heterogeneous metadata from different document servers via an OAI-interface. All rights of use are centrally administrated by

the software: It controls the access of anonymous, filed/applied, and registered users, as well as of the operator of the documentation-server, any cooperated printing services, and the system-administrator. The software is built to merge the heterogeneous systems and operates them in a homogenous environment. This allows central management of all account transactions. If a further archive of digital documents is interested, it has the opportunity to participate in ProPrint. For this purpose, we offer the developed interface software as well as the extended OAI-interface.

1.2.2 ProPrint for printing services providers

Institutions outside Germany wishing to implement the ProPrint service should cooperate with a local service provider like a digital printing office or sophisticated copy-shop.

The service provider should feature:

- Internet connection.
- Ability to download, process and issue PDF-files, (at least PDF 1.4).
- Digital printing facilities to reproduce line originals such as text or single tone graphics and "good-enough-quality" for halftone templates, (minimum; gray-scale and color preferably).
- Capacity to handle European paper sizes (DINA4 = 210x297 mm / 8.3x11,7 inch, DINA5 = 148x210mm / 5,8x8,3 inch, and the required sizes for book covers).
- Binding facilities; bindings should be able to hold 280 sheets, 560 pages resp., in a durable fashion (hotmelt glue minimum; glue preferably).

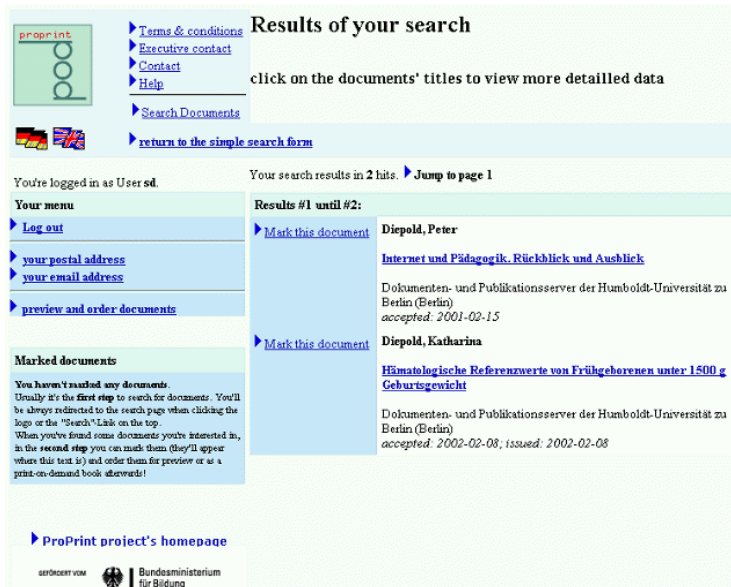
It is highly recommended that the service provider has an internet connection with a bandwidth of at least 500kBit/s as the amount of data per print job might grow up to over hundred megabytes.

1.2.3 ProPrint Service

The screenshot shows the ProPrint search interface. At the top left is the ProPrint logo. To its right are navigation links: Terms & conditions, Executive contact, Contact, Help, Search Documents, and extended search form. The main heading is "Search for Documents" with the subtext "simple form". Below this is a search form with three input fields: "Title", "Author name" (containing "Diebold"), and "Subject". A "Start search" button is located below the Subject field. On the left side, there is a "Your menu" section with links for Log out, your postal address, your email address, and preview and order documents. Below that is a "Marked documents" section with a message: "You haven't marked any documents. Usually it's the first step to search for documents. You'll be always redirected to the search page when clicking the link or the 'Search' link on the top. When you've found some documents you're interested in, in the second step you can mark them (they'll appear where this text is) and order them for preview or as a print-on-demand book afterwards!". At the bottom left, there is a link to the "ProPrint project's homepage" and logos for "eurocert vom" and "Bundesministerium für Bildung".

The ProPrint system regularly collects the (new and changed) metadata from the repositories which hold the documents and reproaches them in its own database. In this way a caching of the archives metadata is realised.

The user can scan these metadata centrally by the ProPrint search interface. If he finds documents of his interest (of which he firstly views the metadata and an abstract contained in it), he can request a preview of the document. By starting this request the ProPrint system begins to examine whether an up to date copy of the document is still present in the cache (e.g. caused by usage of another user). In the negative case the document will be loaded from the offering server.



Results of your search

click on the documents' titles to view more detailed data

Results #1 until #2: [Jump to page 1](#)

Results #1


[Mark this document](#) **Diepold, Peter**
[Internet und Pädagogik. Rückblick und Ausblick](#)
 Dokumenten- und Publikationsserver der Humboldt-Universität zu Berlin (Berlin)
 accepted: 2001-02-15

Results #2

[Mark this document](#) **Diepold, Katharina**
[Hämato-logische Referenzwerte von Frühgebo-renen unter 1500 g Geburtsgewicht](#)
 Dokumenten- und Publikationsserver der Humboldt-Universität zu Berlin (Berlin)
 accepted: 2002-02-08; issued: 2002-02-08

[ProPrint project's homepage](#)

Afterwards the document is examined by the ProPrint system for sufficient technical conditions (embedded writings, colours of the pages, coding) and, if this procedure ends without error messages, shown as a preview to the user. This preview is produced on the fly (at the same time without further preparation).



Place Order

Order book(s) from your document(s)

[back to the list of your available documents](#)

Second step: Managing the distribution of the documents to single books

Here you can manage (if there is more than one document to print) how the documents are combined to books.

Any document ordered by you is listed here. There are **small arrow links** which let you move the documents up and down.

You can **move one document to the far bottom to create a new book**.

There are also switching links with any document that has coloured pages. You can decide this way to print an originally coloured document **in grayscale** to save money for coloured pages.

Single book # 1		
binding cost		3,50 EUR
Part # 1	Diepold, Peter <i>Internet und Pädagogik. Rückblick und Ausblick</i>	
Coloured Pages:	0	EUR
Grayscale pages:	35	1,75 EUR
total amount (incl. VAT)		5,25 EUR

[to next step: Final confirmation of your order and our terms and conditions](#)

print service company: Polyprint, Berlin

On the basis of the provided preview the user can now decide whether he would like to order this document in printed form. In this case, an appropriate cover and a front page for the document are generated and an order will be sent to the printing service according to the choice of the user.

1.2.4 ProPrint Technology

In order to integrate another document server into the ProPrint system, the local server administrator has to fulfil the following requirements:

1. Provide an OAI-Interface for harvesting the metadata.
2. Provide an OAI+-Interface for realising the download of documents.
3. Be able to react to error messages, coming from the ProPrint system.

The OAI-Interface was extended in two ways: (this is the OAI+-Interface)

1. A ProPrint metadata set.
2. One additional verb "disseminate" for the protocol, which was taken from the DIENST protocol, where OAI-PMH was derived from.

ProPrint Metadata Set

All metadata according to the ProPrint-Schema are valid XML documents. The XML-Schema can be found at: <http://www.ProPrint-service.de/xml/>.

The following information are enclosed within the ProPrint metadata set:

- Identifier (<DCTERMS.Identifier>);
- Title of the document (<DC.Title language="...">), different entries for different languages possible;
- Subtitle (<DCTERMS.Title.Alternative language="...">), for different languages possible;
- Author information, in 3 possible ways:
 - As unspecified name
(<DC.Creator>Harry Smith</DC.Creator>);
 - As logically ordered name
(<DC.Creator PPQ.Creator.Affiliation="Institution">
 <PPQ.Creator.FirstName> Harry</PPQ.Creator.FirstName>
 <PPQ.Creator.LastName> Smith</PPQ.Creator.LastName>
 </DC.Creator>);
 - Name of an associated institution
(<PPQ.Creator.CorporateName>Institution </PPQ.Creator.CorporateName>);
- Keywords can belong to different schemas (Attribute „scheme“ becomes a valid value¹) and can be used for different languages
(<DC.Subject scheme="..." language="...">Stichwort</DC.Subject>);
- Table of Contents (non standardised form)
(<DCTERMS.Description.TableOfContents>);
- Abstract as free text, several abstracts for different languages possible
(<DCTERMS.Description.Abstract language="...">);
- Publishing institution, with location
(<DC.Publisher DIEPER.PlaceOfPublication="...">);
- In addition to the author, persons related to the document.
(<DC.Contributor>...</DC.Contributor>);
- Dates:
 - Date of creation (<DCTERMS.Date.Created>);
 - Date of Acceptance (for scientific works like dissertations) (<METADISS.Date.Accepted>);
 - Date of Publication (<DCTERMS.Date.Issued>);

¹ Valid values for scheme see <http://www.proprint-service.de/xml/schemes/v1/TypeDCTERMSSubjectScheme.xsd>

- Type of Resource – for ProPrint-System use „Text“ (<DCTERMS.Type>);¹
- Content type (<PPQ.Type>);²
- Static link to the document. (<PP.Origin DCTERMS.Format.Medium=“...”>);³
- Language of the resource (<DCTERMS.Language>);⁴
- Relation to other documents on a higher hierarchy level (<DCTERMS.Relation.IsPartOf id=“...” sort=“...”>);⁵
- Information about geographical maps (<DC.Coverage>, <DCTERMS.Coverage.Temporal>, <DCTERMS.Coverage.Spatial>, <PP.MapProjection>, <PP.MapScale>);
- Rights information for the document (<DC.Rights>);
- Number of pages (at present not in use, is being extracted from the binary document) (<PPQ.Format.NumberOfPages>).

¹ Valid values for DCTERMS.Type see: <http://www.proprint-service.de/xml/schemes/v1/TypeDCTERMSType.xsd>

² If the tag name starts with „PP“ instead of „PPQ“ this is an orthography error, which will be improved within the final version of the ProPrint metadata schema. Valid values for PPQ.Type see <http://www.proprint-service.de/xml/schemes/v1/TypePPQType.xsd>

³ Valid values for DCTERMS.Format.Medium see: <http://www.proprint-service.de/xml/schemes/v1/TypeDCTERMSFormatMedium.xsd>

⁴ Valid values for DCTERMS.Language see: <http://www.proprint-service.de/xml/schemes/v1/TypeDCTERMSLanguage.xsd>

⁵ The id parameter must include the identifier of one document on a higher hierarchy level of this document. The parameter sort can contain a numeric chapter or another sequence number.