

Questions about the technical validation for the 1st Open Archives Forum Workshop:

Participants: 21

Country: 6 Germany • 6 Italy • 2 Belgium • 2 Netherlands • 1 France • 1 Sweden • 1UK • 1 Portugal • 1 Norway • 1 Schweiz

1) **Since when has your archive been OAI compatible ... (month/year eg. 04/02)**

a) **as Data Provider?**

02/01 • 03/01 • 06/01 • 08/01 • 09/01 • 01/02

In planning -, development -, test phase: 7

b) **as Service Provider?**

03/01 • 09/01 • 10/01 • 02/02

In planning -, development -, test phase: 8

2) **Questions about the software:**

Data Providers:

a) **Which software tools does your Archive use to be OAI compatible?**

- ETD-db software with OAI extension
- Eprint software (2)
- PHP4 Script
- Java Servlet (2)
- Java Applikation
- Perl Data Provider Template (Hussein Suleman - Virginia Tech); accompanied by local Perl developments
- Elektra
- php-oai-dp, make use of mysql and php
- DIENST (package from Cornell)
- OAICat (from OCLC)
- CDSware

b) **Are these tools developed by your organisation?**

Yes: 9 • No: 5

c) **Are these tools also available for other organisations?**

Yes: 12 • No: 2

If yes, where?

- <http://scholar.lib.vt.edu/ETD-db/>
- <http://www.eprints.org>
- <http://edoc.hu-berlin.de/oai2.0/oai-huberlin-2.0.tar>
- <http://www.dlib.vt.edu/projects/OAI/software/altperl/altperl.html>
- Sisis GmbH, Grünwalderweg 28b, D-82041 Oberhaching
- On request, complete rewrite will be available soon
- for an Oracle database
- TBA
- From Cornell and OCLC
- <http://cdsware.cern.ch>

d) **Is the sourcecode open source?**

Yes: 10 • No: 3

e) **If the software is not self developed: Who is the vendor?**

- Sisis GmbH, Grünwalderweg 28b, D-82041 Oberhaching
- Cornell University and OCLC

f) **Is this vendor situated in Europe?**

Yes: 2 • No: 2

g) **Which programming language(s) were used to develop the tools to make it compatible?**

Perl: 5 • Php: 2 • Java: 6 • C: 1 • Python: 1

Service Providers:

a) **Which software tools does your Archive use to be OAI compatible?**

- ODL – DBUnion
- CYCLADES Access Service
- Java Servlet, Java ServerPages, XML, XSLT
- OMNIS/II
- php-oai-sp, make use of mysql, expat and php
- Java Servlet
- Cheshire IR software + java application
- CDSware

b) **Are these tools developed by your organisation?**

Yes: 9 • No: 2

c) **Are these tools also available for other organisations?**

Yes: 5 • No: 6

If yes, where?

- <http://oai.dlib.vt.edu/odl/>
- <http://www.dini.de/oaisuche/>
- On request, since not generally installable and not usable by others. Will be completely rewritten.
- Cheshire from <http://cheshire.lib.berkeley.edu/> --- Java = TBA
- <http://cdsware.cern.ch>

d) **Is the sourcecode open source?**

Yes: 4 • No: 6

e) **If the software is not self developed: Who is the vendor?**

f) **Is this vendor situated in Europe?**

Yes: 2 • No: 0

g) **Which programming language(s) were used to develop the tools to make it compatible?**

Perl: 3 • Php: 2 • Java: 5 • C: 1 • others: Visual Basic: 1 • Tcl: 1 • Python: 1

3) **Questions concerning the implementation costs:**

Data Providers:

a) **Which know-how must the involved persons have (eg. programming languages experiences)?**

- Perl understanding / Linux system administration / Apache configuration
- Experiences with HTML, Perl, Linux, database MySQL and web server Apache
- - basic knowledge on databases and SQL - basic skills in programming (PHP4)
- Java, ServletContainer, DB
- In our case : Perl
- JAVA, C, Databases, CGI
- some Perl experience; some Oracle experience
- Knowledge of php,sql and xml required
- experiences with metadata, java programming
- Undergraduate student
- - Software engineer skilled in databases, programming (Perl, Java, etc.) and

- Internet/web technology in general. - Librarian, skilled in bibliographic metadata
- C, Java
 - web applications, knowledge about used database
- b) **How long did the actual implementation take?**
< 1 week (3) • 2 weeks (2) • 3 weeks (2) • 2 month (1) • 3 month (1) • 3 years • several years (1)
- ETD-db and Eprints are still test installation. They took respectively one and two weeks. / The next big work will be to customize these softwares for our use.
 - To develop the OAi interface itself took 2 days. The interface is a module in larger project.
 - ca 80 hours (2 weeks)
 - OAI est. 5 person days (incl. opt.features)
- c) **How many programmers were involved?**
1 Programmer: 10 • 2 Programmer: 2 • 20 Programmer: 1
- d) **How much has to be done to keep the OAI implementation running (person days per month)?**
- We don't have enough experience to answer this question
 - for stable protocol: < 1person day
 - a strict minimum; let's say 1 person day per month
 - we are not operational yet. I guess 1 day per month
 - No time - everything runs without any problems. But we will need to work with OAI 2.00 compliance and we will also adapt a ListIdentifiers and a possibility for different Metadataformat. We didn't do it yet because the lack of time.
 - Unknown
 - one person.day/month
 - five person.day/month
 - no OAI run-time regular maintenance required

Service Providers:

- a) **Which know-how must the involved persons have (eg. programming languages experiences)?**
- Perl understanding / Linux system administration / Apache configuration
 - Apache Web Server, PHP, Sybase
 - general computer system administration for installing the database, web server, Cyclades software, and configuring
 - Java, ServletContainer, DB (.NET/C# in the near future)
 - JAVA, Databases, Servlets
 - Knowledge of php, sql and xml required
 - experiences with metadata, java programming
 - Undergraduate student
 - depends on the service provided
- b) **How long did the actual implementation take?**
2 month (1) • 6 months (1)
- DBUnion is still a test installation. It took 2 or 3 days.
 - it is still in progress
 - 3 months for integrating the OAI harvester of OCLC into OMNIS II
 - approx 2 weeks, but no full implementation
 - ca 80 hours (2 weeks)
 - unknown, being developed
 - depends on the service provided
- c) **How many programmers were involved?**
1 Programmer: 5 • 2 Programmer: 2
- in Dortmund, 3, but we have external project partners who implement other parts of the system

- 1 programmer for 3 years
 - depends on the service provided
- d) **How much has to be done to keep the OAI implementation running (person days per month)?**
- We don't have enough experience to answer this question
 - We cannot estimate that yet
 - We use OAI (similar) harvesting protocol for local services. And we need to configure the source data (as dataProvider) for each new local data resource...
 - unknown
 - no OAI run-time regular maintenance required

4) Questions regarding content type, structure and integration of archive/ of service:

Data Providers:

- a) **How many documents/ metadata sets does your archive contain?**
 1 • 490 • ca. 800 • 1600 • 12.000 • 400000 • 7.000.000 • several million
- As it's still in test mode, not more than 10 documents per software
 - Our archive is not yet live
 - +/-600.000 metadata records; at this moment only 1 set defined
 - The archive will start with 3000 documents
 - external approx. 1000, internal data-provider 140.000
 - 550000 (about 15% harvestable only)
- b) **How much disc space covers your archive?**
 15 MB • 150 MB • 1.4 GB • +/- 2.5 GB • about 10 GB (2) • 300 GB • 2 TB
- 500 MB by fulltext archive, 1 GB by an Oracle database
 - a very small amount
 - 280MB compressed data
- c) **Which type of objects does you archive contain?**
 fulltext documents: 9 • just metadata: 10 • image files: 5 • video files/ streams: 4
- audio files: 0 • others: 0
 - mixuter of fulltext and metadata sets with abstracts (about 360 fulltext) and 1600 metadatasets with abstracts
 - will be extended to full documents and sound
- d) **Which content type does you archive contain?**
 preprints: 6 • journal articles: 7 • dissertations: 7 • lectures: 4
- Conference Proceedings (1)
 - Short Articles (1)
 - Video streams of University Events (1)
 - Library Catalogue (metadata records for books, periodicals, video,...) (1)
 - All publications produced by the academic staff of the university (1)
 - Recordings (1)
 - Earth Observation Satellite Images (1)
- e) **Which metadata formats are associated with them?**
- Dublin Core (6)
 - qualified DC, stripped down to oai_dc
 - Dublin Core Library Profile
 - MARC21 (3)
 - German MAB and Dublin Core, with XML exchange
 - An internal format is used that can be converted to any standard metadata format
 - RIS
 - UNIMARC, DiTeD (internal format for thesis and dissertations), Dublin Core
 - CEOS CIP

f) **Do you disseminate all parts of the archive (metadata)?**

Yes: 7 • No: 3

- In different formats. F.E. we produce a Marc record which is directly included in our Local/national library system
- Only to the TEL project at present
- ca 15%

g) **Is the OAI interface open for all service providers?**

Yes: 8 • No: 5

h) **If no, which restrictions exist?**

- It's OK for ETD-db software. / The OAI layer of Eprints software does not handle 'from' and 'until' parameters with a 'ListRecord' verb.
- IP address controlled
- The service is not yet fully implemented
- Swets licenses required
- Only part of the data is accessible through the data-provider. Most of the data is only accessible through a search form
- Awaits funding and management approval for wider access when the service is developed

Service Providers:

a) **Which kind of services did you develop?**

- OAI service
- OAI portal
- a searching and browsing facility for information retrieval, other project partners add a workspace for managing documents and metadata, organizing the data, collaboration within groups of users etc.
- cross linking, annotations
- Search all different kinds of journals and other publications through one form. Very different sources.
- Local service for search of the local information in structured form
- search for documents in particle physics and related areas

b) **Is the harvesting level of interoperability sufficient for your purposes or do you would need more? What?**

Yes: 2

- Better metadata
- we would like to see more standardized content in the DC metadata, for example, standard ways to specify names, dates, languages, preferably more structure than unqualified DC
- OK. We did some adaption of the protocol for local service. F. E. we use different unique keys etc.
- The technical interoperability of harvesting is best, but at cost of the higher vel interoperability. - sets do not use uniform classification - 'dublincore' information loss within the transfer - how to deal with oai identifiers in different topologies

c) **How do you process with harvested data from other data providers? Do you use any provenance information?**

- I have used no provenance information.
- we index the data and make it searchable, when a metadata record is found and displayed to the user, the user can also browse information on the archive the record came from
- no metadata processing; queries against the portal return data sets as harvested, including information about the original data provider
- filter harvester output, load local database
- The metadata is parsed and converted to an intermediate (close to DC) format. The provenance information is somewhat encoded in the identifier.
- added-value activities, enriching metadata

d) **In your experience is there any weak point in the OAI approach to interoperability? If yes, explain?**

No: 1

- no experience.
- The protocol is easy to implement for data providers, but the heterogeneity of the content of the metadata records requires the service provider to invest a lot of effort in normalizing the data in order to make it more comparable and usable, thus ultimately again writing wrappers. Much of this standardization could be done at lesser cost by the individual data providers, as the data within an archive usually shows less heterogeneity than between archives. A possible solution might be the development of middleware tools that service providers could use for data normalization instead of each service provider inventing the wheel again. Another suggestion is to define additional metadata formats (or use existing ones) and convincing data providers to export them, too.
- Not a problem of OAI itself, but the quality of metadata is just a big mess.
- no weak point in technical interoperability (see point b)

5) Questions about experiences and future planning

Data Providers:

a) **How important is your OAI compatible data provider for your institution and your service? Were existing services replaced or completed?**

- We intend to build a catalog of theses, dissertations, scientific papers in order to help students in their work and to share the knowledge. Only a paper catalog exists now.
- The experience with our oai compatible archive will help us answer to this question. Actually, we don't know if this service is able to respond to the real needs of our institution. However, the archive is intended to replace the single, non-searchable lists of eprints available on the websites of our departments and research centres with a centralized service and to develop the practice of alternative means of dissemination of the scholarly communication.
- - harvesting of our own archive via OAI (replacement of search engine) - interface for metadata exchange within several projects (new service)
- - Exchange of our library catalog with other university in Brussels - Integration into virtual union catalog of Belgium - Delivery of electronic holdings info to an OpenUrl resolver system
- OAI is one way of making the data available. The main reason for being OAI compliant is that we want to be able to develop services based on OAI compliant archives maintained by others.
- Research project. We used to run a search-engine which tried to combine different HTML-outputs from different sources. This has been replaced by an compatible OAI search-engine
- It's very important because the possibilities of dissemination of information about scientific results of our researcher
- Not known at this stage. No services have been replaced
- The implementation stills in the evaluation phase. No definitive service is being supported or announced based on it
- oai-compatibility provides required technical interoperability

b) **What advantage is there for you in participating in the OAI? Which advantages offers the OAI interoperability framework in contrary to other interfaces and logs (eg. Z3950)?**

- OAI gives the chance to share scientific knowledge and to harvest other knowledge databases. It will also give the opportunity to import metadata from theses,.. in our Libraries software (Virtua).

- A major dissemination of our researchers' results.
 - - simple implementation - easy adaptation for project internal usage
 - Attending the OAI Seminar for the purpose of giving recommendations in the Feasibility Study for Italian Digital Library Project: I am in charge of the updating of the Section of the Study on METADATA ISSUES and INTEROPERABILITY
 - OAI provides a simple to implement facility of exchanging metadata. / Z39.50 is probably too complicated to implement. In most cases you need to pay a commercial party to do it for you. This is not the case with OAI.
 - research
 - We are heavily Z39.50 oriented. In many cases the archives are too small for a Z39.50 service. A Z39.50 database must have a reasonable size. By harvesting relatively small archives it is possible to maintain a Z39.50 service. It is a question of scale.
 - Easy (and quick) implementation, minimal maintenance
 - Provides a standard approach for metadata harvesting which will simplify extension of a pilot system. Software for prototype freely available.
 - OAI is easy to implement...
 - oai-compatibility provides required technical interoperability; OAI-PMH simpler than Z39.50
- c) **What are your experiences with being an OAI compatible data provider?**
- We are still testing the products so we did not yet register in the union catalogs
 - No experiences
 - Multiple services have become available, with only ONE SIMPLE implementation on our catalog. / We have been implementing 'OAI extensions', to be used in applications such as access control to our library buildings.
 - No experience. I have experience in being a RePEc compatible data provider. Updating the RePEc archive is done automatically.
 - It just runs.
 - Very good impact on usage of information in our archive
 - Too little experience to comment
 - Good!
 - in test phase

Service Providers:

- a) **Which services do you support with the OAI functionality?**
- get metadata records, search engine
 - harvesting, annotation, cross linking
 - Search-engine
 - We use the information for an metadata catalogue (<http://publications.uu.se/metadata>). Currently there are metadata from 5 local databases - about 5 000 records.
 - None (so far)
 - search for documents in particle physics and related areas
- b) **Which kind of problems do you have concerning the technical implementation and use of the protocol?**
- We have the problem about the different semantic by defining set
 - none (2)
 - As a harvester of RePEc archives for building a Z39.50 service, our main problem is the quality of metadata.
 - Mostly xml related problems, and different formats of metadata...
- c) **What do you plan in the future? (e.g. search engine with compatible OAI integration, printing on demand, document delivery services, alerting services)**
- We plan to integrate the search engine with compatible new OAI protocol in der

future.

- search & browse, collaboration environment for users and groups of users, discussion forums, annotations, awareness
 - We are planning on implementing a virtual union catalog for Belgium using OAI (several millions of metadata records). We will be trying out software like Open Digital Libraries (<http://oai.dlib.vt.edu/odl/>) and ALCME (<http://alcme.oclc.org/index.html>) in the near future.
 - proof of concept, successful
 - Z39.50 services integrated in an information portal (iPort of OCLC|Pica). This will allow searching, browsing, document delivery services, current awareness etc (no printing on demand - printing is for the user).
 - The search engine is already running, but more sources will be included. Implementation of version 2.0.
 - Search engine with compatible OAI integration and document delivery
- We intend to develop a resource discovery service for contents related with our mission (the Portuguese science and technology, culture, history and society in general). An alerting service, coordinated with the national union catalogue, is also under consideration.
- extend existing services based on the user feedback, explore potential for building distributed services

Please contribute!

- Information about your projects
- Your implementation and usage experience

Questionnaire:

<http://www.oaforum.org/resources/tecvalquest1.php>

Database:

http://www.oaforum.org/oaf_db/

Humboldt-University, Berlin, Germany
Electronic Publishing Group
Project Open Archives Forum

Susanne Dobratz – susanne.dobratz@rz.hu-berlin.de
Birgit Matthaei – birgit.matthaei@rz.hu-berlin.de
Jing Yuan Wang – jingyuan.wang@rz.hu-berlin.de