

## OPEN ARCHIVES FORUM: FINAL RESULTS OF THE TECHNICAL VALIDATION QUESTIONNAIRE

<b>Project Number:</b>	IST-2001-32015
<b>Project Title:</b>	Open Archives Forum

<b>Date of Delivery:</b>	30-09-2003
<b>Title:</b>	Final Results of the Technical Validation Questionnaire
<b>Workpackage contributing to the Document:</b>	WP2
<b>Total Number of Pages:</b>	11
<b>URL:</b>	<a href="http://www.oaforum.org/otherfiles/tv-questionnaire.pdf">http://www.oaforum.org/otherfiles/tv-questionnaire.pdf</a>
<b>Author:</b>	Birgit Matthaei with Susanne Dobratz
<b>Contact Details:</b>	Birgit Matthaei Humboldt-University Berlin, CMS (Computing Centre) Rudower Chaussee 26, 12489 Berlin, Germany

### TABLE OF CONTENTS

1	Technical Validation Questionnaire: .....	2
1.1	1 <sup>st</sup> Questionnaire about Technical Validation .....	2
1.2	2 <sup>nd</sup> Questionnaire about Technical Validation .....	2
1.3	Changes and Developments between the Questionnaires .....	3
1.4	Questions and Results of the Technical Validation Questionnaire: .....	3
1.4.1	Who has participated? .....	4
1.4.2	Software used .....	5
1.4.3	Implementation costs .....	6
1.4.4	Resources offered and issues of interoperability .....	6
1.4.5	Experiences and expectations .....	8
1.4.6	Information sources .....	9
1.4.7	Conclusion .....	9
1.5	Excursion: Comparison with DINI data provider questionnaire .....	10

## 1 TECHNICAL VALIDATION QUESTIONNAIRE:

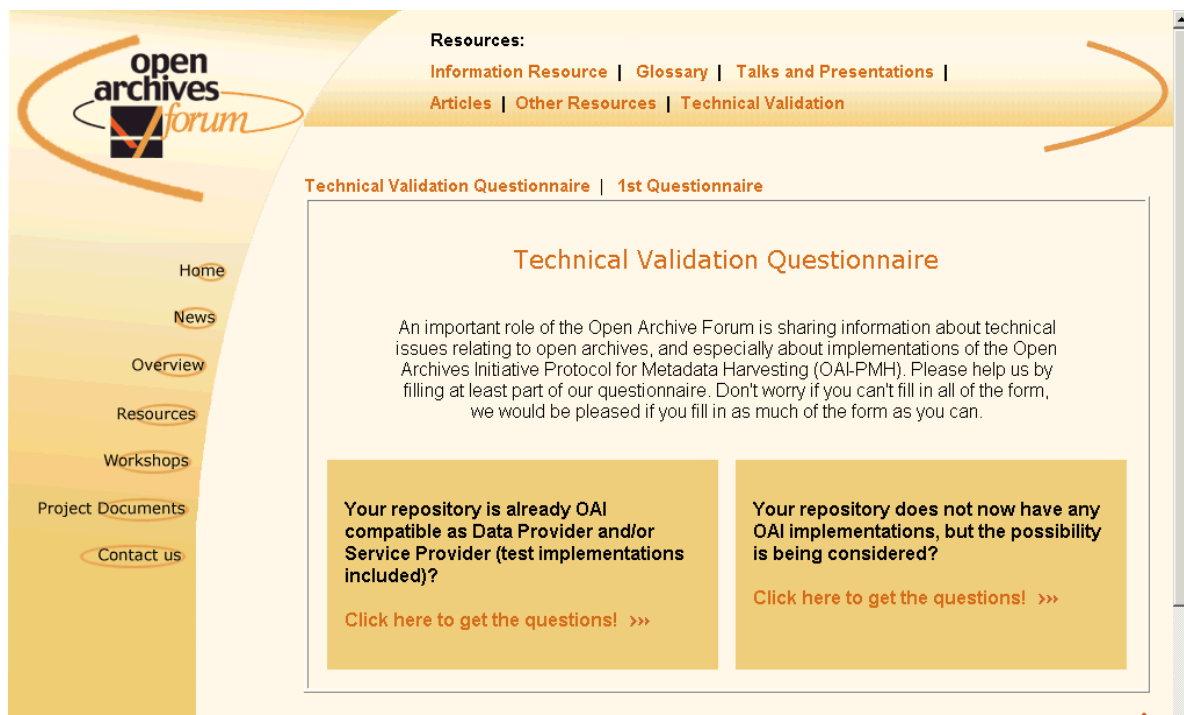
### 1.1 1<sup>st</sup> Questionnaire about Technical Validation

- The first questionnaire can be read at: <http://www.oaforum.org/resources/tecvalquest1.php>
- Results: [http://www.oaforum.org/otherfiles/pisa\\_techvalresult.pdf](http://www.oaforum.org/otherfiles/pisa_techvalresult.pdf)
- Presentation of the results: [http://www.oaforum.org/otherfiles/pisa\\_techvalresult.ppt](http://www.oaforum.org/otherfiles/pisa_techvalresult.ppt)

The first questionnaire was designed to receive information about existing OAI and open archives implementations and usage in Europe. It was primarily aimed towards the participants of the first workshop. 18 people contributed, 6 from Germany, 5 from Italy, 2 from Belgium, 2 from the Netherlands, 1 each from France, Sweden and the UK. The results have been already described in detail in the "Interim Review of Technical Issues" (Deliverable Number D2.2) [http://www.oaforum.org/otherfiles/oaf\\_d22\\_techval1.pdf](http://www.oaforum.org/otherfiles/oaf_d22_techval1.pdf).

### 1.2 2<sup>nd</sup> Questionnaire about Technical Validation

- The second questionnaire can be read at: <http://www.oaforum.org/resources/tecvalq2.php>



The screenshot shows the Open Archives Forum website. The header includes the logo and a navigation menu with links for Resources, Information Resource, Glossary, Talks and Presentations, Articles, Other Resources, and Technical Validation. The main content area is titled "Technical Validation Questionnaire | 1st Questionnaire" and contains the following text:

**Technical Validation Questionnaire**

An important role of the Open Archive Forum is sharing information about technical issues relating to open archives, and especially about implementations of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Please help us by filling at least part of our questionnaire. Don't worry if you can't fill in all of the form, we would be pleased if you fill in as much of the form as you can.

Two yellow boxes are present at the bottom of the main content area:

- Your repository is already OAI compatible as Data Provider and/or Service Provider (test implementations included)?**  
[Click here to get the questions! >>>](#)
- Your repository does not now have any OAI implementations, but the possibility is being considered?**  
[Click here to get the questions! >>>](#)

### 1.3 Changes and Developments between the Questionnaires

The second questionnaire about technical validation is based on the first questionnaire with some changes according to the feedback during the first workshop and to our own experiences with the initial attempt - some questions had been added or changed, the duration extended, the target audience expanded and the form was subdivided to account

- for those projects that have not yet integrated OAI-PMH ([http://www.oaforum.org/registration/lists/tecval2/tvqval.php4?my\\_initglobal=tvq2/tvqinitglobal2.inc](http://www.oaforum.org/registration/lists/tecval2/tvqval.php4?my_initglobal=tvq2/tvqinitglobal2.inc))
- in addition to those who are experienced implementers ([http://www.oaforum.org/registration/lists/tecval2/tvqval.php4?my\\_initglobal=tvq1/tvqinitglobal1.inc](http://www.oaforum.org/registration/lists/tecval2/tvqval.php4?my_initglobal=tvq1/tvqinitglobal1.inc)).

In the sessions at Pisa it was repeatedly brought up for discussion that despite all standardisation and protocols in the long run already available software and systems as well as individual targets and community selective demand determine the implementation of metadata and the pre-harvesting conversion of metadata. - Therefore in the new questionnaire the first emphasis "Technical conversion of the implementation of metadata" had been added by the following subjects: What tools do already exist before implementation? What is the present content? What has to be achieved?

The first questionnaire, designed for a numerable small target group, was based on a simple, easily to install HTML form, the analysis took place manually. For the second, somewhat more extensive questionnaire, we decided to accept a higher expenditure at the beginning – the programming of the binding to a database. Below are summarised the results of the information the participants gave about used software, implementation costs, offered spectrum and interoperability, experiences and expectations in different communities and in different countries.

This second, long-term survey continued through autumn 2003. Workshop participants of Lisbon, Berlin and Bath had been asked to fill out this questionnaire at the end of the online registration process for the workshop. This was the most successful way to encourage people to participate in the survey. Furthermore participants had been recruited by presentations at conferences (e.g. CERN, 2<sup>nd</sup> workshop on the Open Archives Initiative), articles (e.g. D-Lib Magazine, Jan. 2003), flyer at conferences (e.g. ETD 2003, EDCL 2003), info lists, mail invitations and a link on the homepage of the projects website.

### 1.4 Questions and Results of the Technical Validation Questionnaire:

In the appendix you find a complete list of all questions and answers as they were listed in the database.

- Summary of first results (Lisbon, Dec. 2002): [http://www.oaforum.org/otherfiles/lisb\\_tvq.ppt](http://www.oaforum.org/otherfiles/lisb_tvq.ppt)
- Interim results (Berlin, March 2003): [http://www.oaforum.org/otherfiles/berl\\_tvq.ppt](http://www.oaforum.org/otherfiles/berl_tvq.ppt)
- Technical Validation (Bath, September 2003): [http://www.oaforum.org/otherfiles/bath\\_tvq.ppt](http://www.oaforum.org/otherfiles/bath_tvq.ppt)
- Summarising article: Open Archives Activities and Experiences in Europe. An Overview by the Open Archives Forum • Susanne Dobratz, Birgit Matthaei • D-Lib Magazine, Vol. 9 no 1, January 2003 <http://www.dlib.org/dlib/january03/dobratz/01dobratz.html>
- Final results as listed in the database: <http://www.oaforum.org/otherfiles/tvqdbresults.pdf>
- Final results: <http://www.oaforum.org/otherfiles/tvqfinalresults.pdf>

This second, long-term survey continued during approximately the period of one year – until the end of the project in autumn 2003. We asked for information about software used, implementation costs, coverage of the archive, and interoperability, experiences and expectations in different communities and in different countries. The focus of interest was on fundamental questions such as:

- Is there a common ground and therefore good conditions for cooperating and learning from each other, or are requirements so individual that it will be necessary for many isolated solutions to be developed?
- Do the existing instruments for implementation fulfil all requirements or should tools and protocols be developed to meet the needs of different communities?

### 1.4.1 Who has participated?

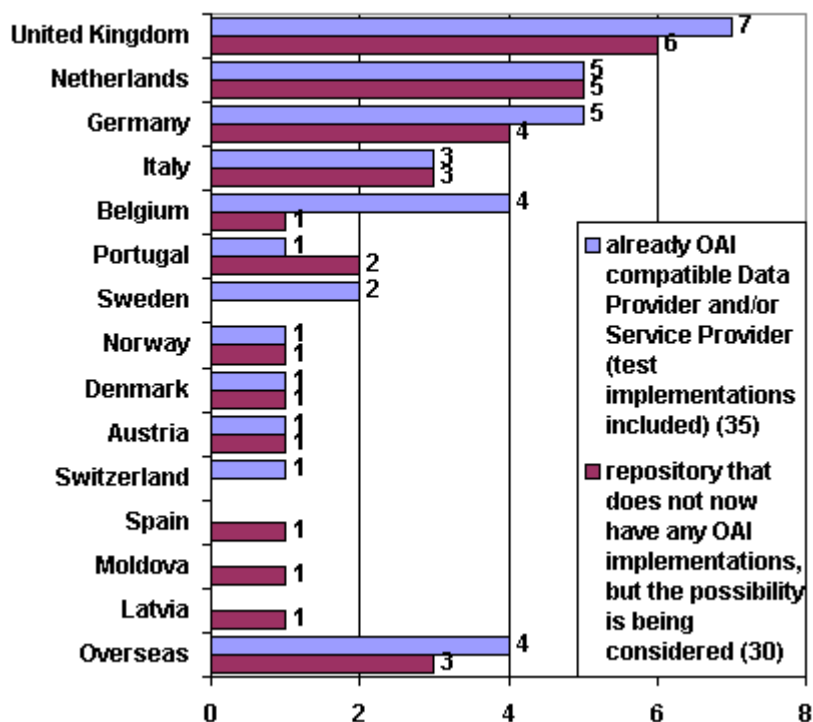
Finally, 65 repositories have participated in the survey. Thirty of the repositories are not yet OAI implementers, but they are considering becoming implementers. The responding repositories are distributed throughout Europe, with also seven participants from overseas. Half of the survey respondents are from UK, the Netherlands or Germany.

The large commitment of the Netherlands is a newer trend, which stands in direct connection with the start of DARE (Digital Academic Repositories) in the beginning of 2003. DARE is a collective initiative by the Dutch universities to make all their research results digitally accessible. This triennial project is funded by the Dutch government.

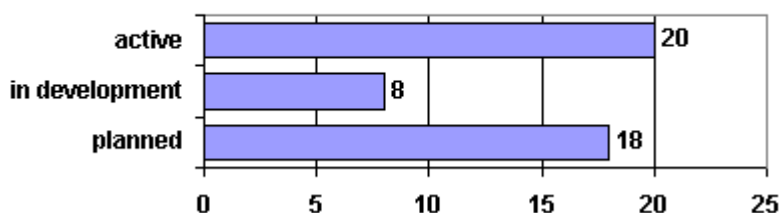
Up to now, clearly more data- than service providers have completed their OAI implementations. If one views the number of implementations under development or being planned, we soon will have many new repositories and services available. Many Data Providers used their implementation experiences to guide them in becoming Service Providers.

- 35 % of active data providers are also service providers.
- 30 % of active data providers plan or are still developing service provider implementations.

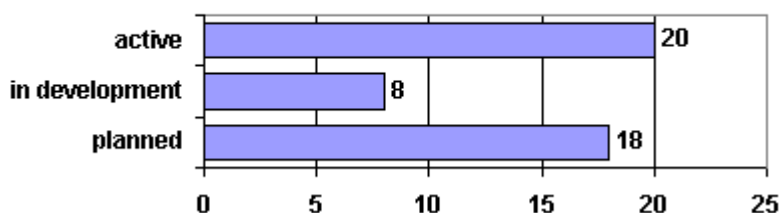
Distribution of OAI repositories by country



Number of responding Data Providers and status of implementation

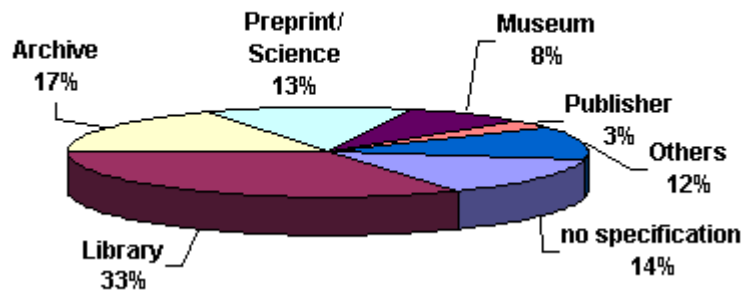


Number of responding Data Providers and status of implementation



If we look at the types of communities represented in the responses, it is remarkable that nearly half of the responders came from libraries or archives. Preprints/Science is also indicated often. These numbers correspond with the high portion of universities as operating organisations.

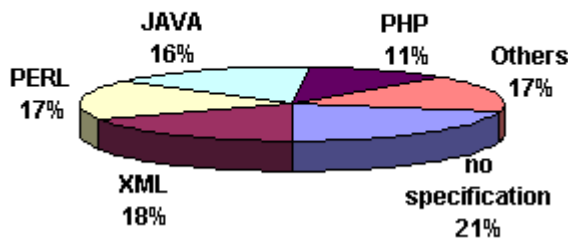
Types of OAI-implementing Communities responding to the survey



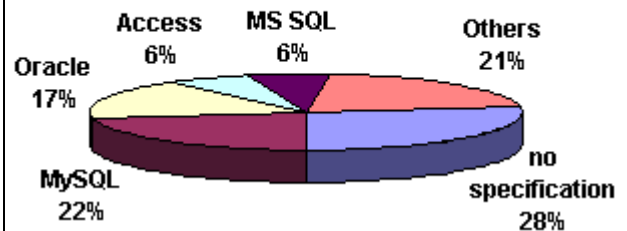
1.4.2 Software used

As mentioned above, the first block of the survey is made up of questions about technical infrastructure and software solutions. Prior to OAI implementation, the dominant programming languages used by responders were XML, PERL, Java and also PHP, and the dominant database was MySQL followed by Oracle. Practically no statements were made to interface and collection systems, so it is not possible to provide relevant information from the survey about those. However, it is significant that almost none of the organisations needed to replace existing software tools in order to become OAI compatible.

Technical infrastructure before OAI-implementation: Programming languages



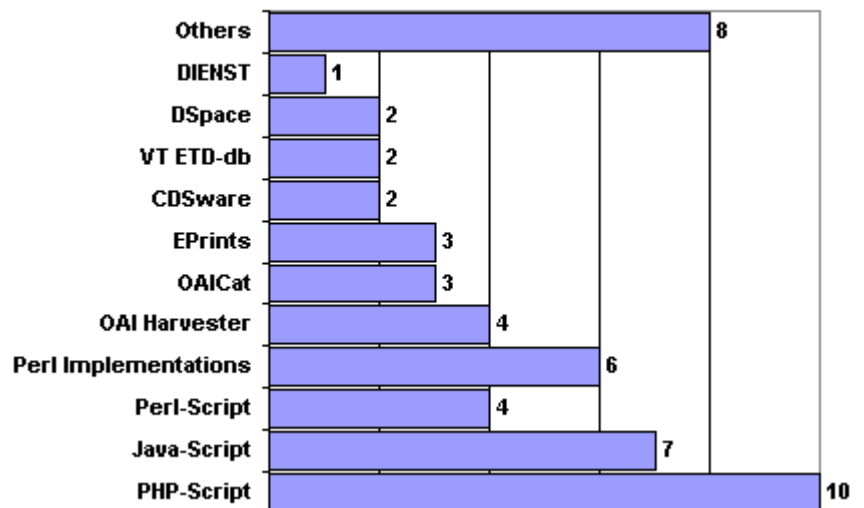
Technical infrastructure before OAI-implementation: Databases



The questionnaire surprisingly shows that about half of the respondents (52%) has implemented their own OAI compliant interfaces instead of using existing tools. Most of the data providers and service providers make their tools and source code available for others to use. The programming languages used to develop these tools are mainly Java and PERL, and also used frequently are PHP and XML. The other half mentioned many different tools, no favourite one is recognisable.

It is also very interesting to observe the changes during this long-term questionnaire: One year ago were even 80% of the participants self-developers. Most likely this is because the first answers were filled in by many early implementers that started work before the release of general tools. Now that the OAI-PMH is well-known, the interest in simple solutions without large development costs is increasing. Contrary to the numbers specified in the 'Overview' (chapter 5.1) however no clear trend to certain tools like Eprints or DSpace is to recognise with the answers on the questionnaire.

Tools used to be OAI compatible (Data- and Service Provider)



### **1.4.3    *Implementation costs***

After the questions regarding the software used, the next questionnaire subject block concerns implementation costs. With regard to the implementation skills needed, data providers as well as service providers focused on various combinations of the following five competencies:

- System administration (esp. UNIX, Linux)
- Web server configuration (esp. Apache)
- Knowledge of databases and SQL (according to the before designated varying databases)
- Programming skill and knowledge (according to the before designated varying programming languages)
- Experiences with metadata

The survey results showed that 76% of the implementations were concluded within one quarter (three months) and most implementations (77%) were managed by only one programmer. The span reaches from 2 to 750 personal days per month. The reason for bigger expenditures by a few of the implementers was not directly connected to the implementation of the OAI-specifications. The higher costs involved larger research projects or were due to construction of archives or the processing of greater amounts of data.

When survey respondents estimated maintenance costs, these were limited to at most 5 person days (one exception numbered the expenditure on 25 person days per month), and most often were estimated to be one person day per month for stable protocol.

These statements of expenditure for the implementation and its preparation were in line with the expectations of those who have not yet become OAI implementers: i.e., implementations concluded within one-quarter year and by one programmer. However, expectations on further maintenance for a stable protocol are higher; they were expected to be up to 40 personal days per month. With the other survey questions, the answers compared to expectations regarding implementation costs differ too much for trends to be recognisable. This includes expectations regarding easy integration of the data structures suggested by the OAI-PMH in existing infrastructure, the costs of adapting data to the OAI-PMH, and expenditure needed for data preparation for internet usage.

70% of the respondents who does not now have any OAI implementations think that the provision of short introductory training courses (like the tutorials at the beginning of the Lisbon and Berlin OA-Forum workshops) would be useful.

### **1.4.4    *Resources offered and issues of interoperability***

The next block on the survey questionnaire regards the range and kind of resources offered by the archive as well as interoperability.

#### **Data Provider**

The range of the number of resources available from data providers includes a wide span of between one test document and seven million documents. The occupied storage space ranged from between one megabyte to two terabytes. Looking at both these ranges, it is important to note that the storage capacity used has less to do with the amount of data than with the type of objects.

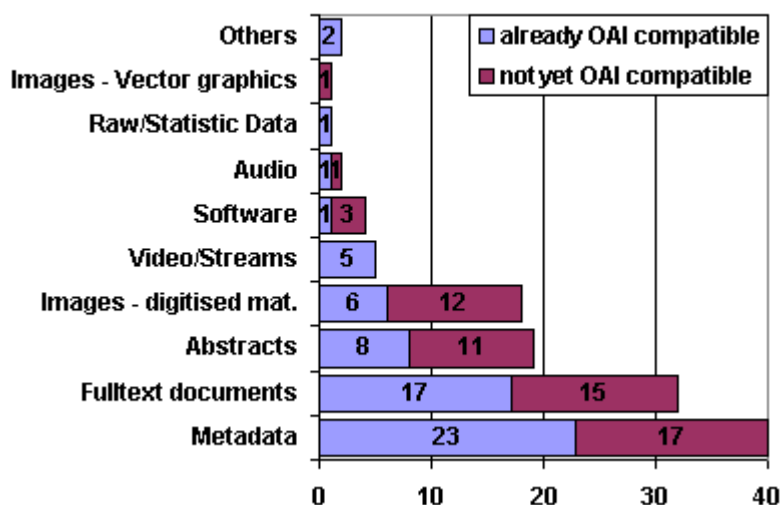
In the list of the object types offered, it strikes one again that metadata and full text documents are what is mainly offered. The reason is due not only to longer experience with storing and evaluating data based on text. Of bigger concern is on the one hand, that if the data is not born digital material there is a high effort to digitise it. And on the other hand the costs of storing pictures and video files, which need stable and efficient databases.

The range of content types includes essentially the entire spectrum of scientific publications. There is a notably high interest in preprints, journal articles and theses. This provides evidence of a big need for a reasonable, fast way to access scientific information beyond conventional scientific publication forms. Other resources offered include library catalogues or video streams of university events.

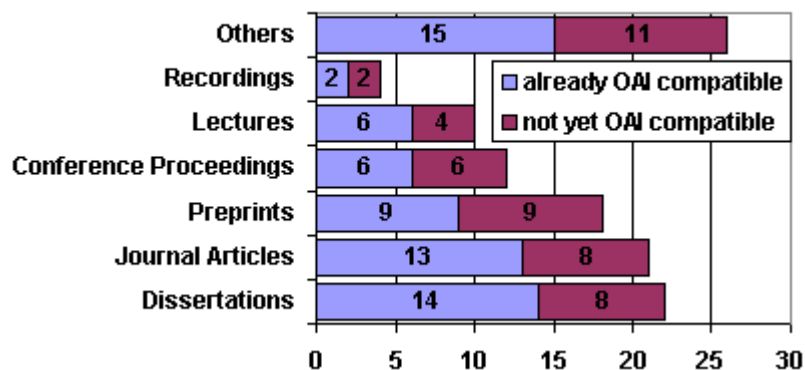
The most-applied metadata format is Dublin Core. In addition, according to the respondents, library-specific formats are used, like MARC 21. However, there are a remarkably high number of formats that are mentioned only once or few times, such as AMF, CEOS CIP, DiTeD, Dublin Core Library Profile, MAB, MIX, MODS, RFC 1807, RIS, SPECTRUM, TEI, and some internal formats. Looking at these numbers it does not wonder that Service Providers - which the questionnaire also asks for their experiences - indicated standardisation as a great problem.

Almost two-thirds of the data providers are offering full text or extracts of documents. If the openness of the interface must be reduced, there are two access-limitation strategies: On the one hand, access control like control of the IP-addresses or licensing can be used, and on the other hand data output is limited.

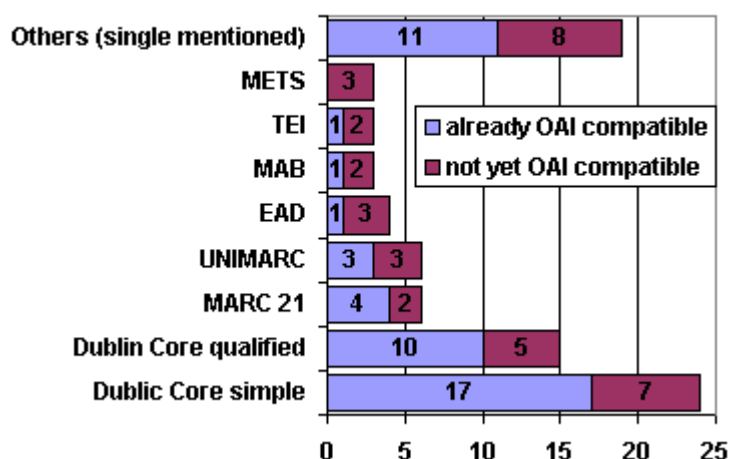
Types of objects offered by Data Providers



Types of content offered by Data Providers



Metadata formats used in OAI implementations



## Service Provider

Most service providers offer local, community or document type specific services to search and browse for information. Some provide portal functions or a workspace for managing documents and metadata, and for collaboration within groups of users. A number of survey answers referred to research projects. For service providers, paging functions are urgent. They search in one or several sources through one search interface. Other services offered comprise cross-linking and annotation.

Strategies to process data harvested from data providers include using no provenance information or filtering harvester output and loading the local database. If service providers comprehend information about data providers in data output, they do so in three ways:

1. When a metadata record is found, the user can also browse information on the archive from which the record came.
2. There is no metadata processing, queries against the portal return data sets as harvested, information about the original data provider is embedded.
3. The metadata is parsed and converted to an intermediate format. The provenance information is encoded in the identifier.

### 1.4.5 Experiences and expectations

#### Data Provider

For data providers, the importance of the OAI technical framework is that it makes it possible to provide additional services to existing services, replace existing services through the OAI interface and offer better retrieval.

OAI is just one way to make data available but the required technical interoperability offers a common ground for service development and gives the chance to establish services that rest upon OAI compliant archives maintained by others. 'Centralised services', 'resource discovery network', 'exchange between research centres / catalogues / institutes', and also 'not to be an isolated silo' are some of the catchwords describing these expectations on OAI compliance. Special impact lies on the cross-search functionalities as base for further offers.

The advantages of OAI are to share scientific knowledge and to harvest other knowledge databases regardless the platform. OAI also enables the import of metadata into library software and major dissemination of the results of research. The OAI implementation is simple (for many even simpler than Z39.50), cheap, easy to adapt for internal project usage, and simplifies extension. In comparison to more complex protocols, it is an easy-to-implement facility for exchanging metadata with the advantage of aggregation and re-presentation of data by using a fairly simple protocol.

While for the most OAI offers the vision to 'provide access to all of human knowledge', there are in addition also critical voices which state that OAI is 'nothing other than political expediency'.

#### Service Provider

Concerning the experiences of service providers, some survey respondents indicated that standardisation presented a problem: 'The heterogeneity of metadata record content requires the service provider to expend a lot of effort to normalise the data to make it usable.' This is only one quotation on behalf for some more answers belonging this topic. The participant quoted before mentioned that solving these difficulties to establish joint services based on open archives could be done at lesser cost by the individual data provider. Another idea for a possible solution was the development of middleware tools that service providers could use for data normalisation.

Some more examples of statements concerning the lacking interoperability: 'quality of metadata is just a big mess', also 'the semantic heterogeneity of OAI sets'. Different terminologies ('more standardised content in the DC metadata, for example standard ways to specify names, dates, languages, preferably more structure than unqualified DC') conducted respondents to the conclusion that 'some best practices about sets and vocabularies would be useful'.

In addition to listing those problems, the service providers who replied to the survey stated that they have future plans to do the following:

- Extend search and browse functions,
- Export data in other formats such as XML,
- Build document delivery services like print on demand,
- Establish collaborative environments for users and groups of users such as discussion forums, annotations, awareness,
- Extend existing services and build distributed services, and
- Establish an exchange of different library catalogues in order to integrate the information into a virtual union catalogue for the whole country,
- Full text indexing of documents

One library is creating a single catalogue of all its library catalogues: library OPAC, archives database, image database and Internet gateways.

#### **1.4.6 Information sources**

Another of the survey questions focused on the quality of information sources. Many of the respondents who are not yet OAI compliant say it takes too much effort to find good information about OAI implementation, and especially difficult to find technical support. Some asked for an easy introduction to OAI-PMH.

Other participants recommended the following ways to find good information:

- Search Websites<sup>1</sup> like that of OAI or, for the museums community, CIMI.
- Read online journals like *Ariadne* and *D-Lib Magazine*<sup>2</sup>.
- Participate at conferences and workshops.
- Initiate informal discussions with other gateway managers.
- Take part on various mailing lists ( e.g. oai-implementors list).
- Experiment with test programs<sup>3</sup>

#### **1.4.7 Conclusion**

Overall it is remarkable that the highest engagement in OAI compliance proceed in countries where national initiatives encourage exchange between the institutions and agreements on organisational and technical solutions (UK: JISC-Joint Information Systems Committee/ RDN-Resource Discovery Network • Netherlands: DARE-Digital Academic Repositories • Germany: DINI-German Initiative for Networked Information).

In addition, also beyond these countries it is to be expected in the near future a strongly growing number of data providers and services which will particularly be offered by libraries and archives from universities and research institutes. Like already the at present most frequent kind of the object and content types shows there exists especially for these organisations a big need for a reasonable fast way to access scientific information beyond conventional scientific publication forms.

It is consensus that OAI-PMH is easy and cheap to implement. The emerging costs are not directly connected to the implementation of OAI specifications. They rather due to the construction of archives, processing of greater amounts of data, digitising data, etc.. In this sense also the use of software tools might have to be evaluated.

---

<sup>1</sup> Web information sources: [www.openarchives.org](http://www.openarchives.org) • [www.ndltd.org](http://www.ndltd.org) • [www.cimi.org](http://www.cimi.org) • [www.eprints.org](http://www.eprints.org) • [www.rlg.org](http://www.rlg.org) • [www.oaforum.org](http://www.oaforum.org) • [www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai](http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai) • <http://library.cern.ch/heplw/4/papers/3/> • <http://dublincore.org/>

<sup>2</sup> Online journals: <http://www.ariadne.ac.uk/> • <http://www.dlib.org>

<sup>3</sup> Test programme: <http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai>

Obviously in the beginning the point was just to extend existing databases by OAI functionalities, which was often realised by the addition of a self-programmed script. Now with the growing number of digitisation activities and born digital material the need of tools which allow to build archives (with OAI compliance as only one of many functionalities) with less development costs becomes obvious.

In the community Eprints and DSpace obtain high attention, nevertheless within the questionnaire participants only a trend for the application of already developed tools is to be recognised, not however special favourites. One possible conclusion from that fact could be that so far the existing instruments for implementation do not fulfil the range of different requirements.

Mostly difficulties with the OAI implementation are not seen as technical problem, but concern rather to the organisation of interoperability: metadata formats, access strategies, ... ('technically positive, still challenges in data-service provider relationships regarding metadata, terms of use, images, etc.'). To realise cross-search, the mostly mentioned value of OAI, it is necessary to simplify complex data by standardisation and normalisation. Differentiated data samples and special needs in different communities and different countries oppose this base to share scientific knowledge. Now, since the technical definition of the OAI-PMH achieved for the present its final state with version 2.0, discussions approximately around organisational issues of interoperability seem to be in the centre of attention, in order to get some best practices or agreements about sets and vocabularies to accomplish a successful networking.

### **1.5 Excursion: Comparison with DINI data provider questionnaire**

Here we would like to refer to another survey, initiated by the DINI working group (German Initiative for Networked Information) and realised by substitution of the OA-Forum questionnaire technology, which shows still more comprehensively (even apart from the technical and metadata problems), which topics must be considered and pursued, when digital documents are offered. The goal of this questionnaire was, to get an overview on technology, systematics and general structure used by German data providers and to receive information about the general structure of document servers at current time.

The development of the particularly university document servers still stay in an early stage. With this questionnaire the status quo should be recorded: How is the situation? Is it ensured that the electronic documents can be read on different systems? Do standardised communication channels exist, which are able to transport metadata and binary files? Are there content, technical and formal controls to ensure scientific relevance of the documents and its classification into the research context?

The questionnaire was developed in connection with the project ProPrint (more chapter 7.2) and is to detect in particular, which co-operation possibilities exist regarding the implementation of a print-on-demand service as well as which interfaces and mechanisms are available or planned for this purpose. With 47 participants, consisting of document servers which are distributed on the entire federal territory, a relatively informative tendency can be determined - also for each further additional service, which would like to process digital documents. The topics of interest for ProPrint, are to a large extent also of immanent importance for standardisation measures in the range of digital documents.

Elements in document server systems, which apply to pro print as a requirement:

- Offered documents are available in pdf-format (exchange format for the supply of the local print service).
- Exchange of the metadata by an OAI-interface.
- Allocation of persistent identifiers to the documents.
- Contractual regulation of copyrights as well as appropriate entries in the metadata.
- Existing metadata in a format of international standard, ideally Dublin Core.
- Content, technical and formal control of the documents.

Only seven of the 47 received answers could be constituted as potential partners, which can be merged into the web service without larger expenditure. On the other hand 26 would require substantial efforts to be integrated. This result shows the pressing need of substantial discussions, which are still necessary at present, in order to obtain common standards, which enable an easier developing of services built on distributed archives .

Here some excerpts from the general evaluation of the questionnaire:

**Document format:** The format of the documents is indicated nearly in principle as pdf. Here it appears that the Adobe format evolved to a quasi-standard. The presentation of contents by XML and XSL, which is open for new developments, takes place only in four cases.

**Server technology:** Concerning server technology the LAMP systems (Linux operating system, Apache web server, MySQL database and PHP and/or Perl as script language) enjoy of large spreading.

A huge heterogeneity of the document server structure becomes visible: OPUS is used frequently, it exists however also numerous local self-developments and commercial products.

Concerning the exchange of data, the clear winner is OAI. Whether version 1.1 or 2.0, whether already implemented or planned, it is thought of OAI, when an interface is desired for communication with other providers. A third of the participants have and wish no interface. If OAI is present, in all rule it is aimed at no further interface.

The application of persistent identifiers has a high value with the operators of document servers. But answers regarding policies about a clear identity of the servers and the access to the servers are showing that still much development is necessary, in order to ensure the quoting ability of documents.

**Service:** With exception of one commercial service all basic service achievements are free and usable without registration. Only for additional services fees are raised. Basic services are predominantly searching and browsing in existing documents as well as regarding and download of the documents. Further services are offered rarely. For investigation with half of the answers searching in the metadata and in the full texts of the documents is possible.

**Rights:** The multiplicity of contradictory statements in the answers about copyrights shows a general uncertainty to this topic. With considerations for the development of new services frequently the necessity contractual bases to secure copyright issues is ignored.

**Metadata:** With the metadata most dispose of bibliographic and administrative ones, rarely however of technical, structure, archiving or right metadata. With the question about international standards Dublin Core is strongly represented, but it is not in such a way dominant that one could mention it as generally accepted minimum standard.

**Content:** With content control partially the decision whether scientific requirements are fulfilled is committed to the authors themselves. Technical and formal control takes place to certain extent, but is not marked as checked, so that a later filtering by quality criteria is not possible.

**Statistics:** About half of the participants raise the hits to their document servers statistically. None of the document server uses the access statistics for an appraisal of the documents.