2nd Technical Validation Questionnaire
- interim results -

Birgit Matthaei
Humboldt-University, Berlin, Germany
Electronic Publishing Group
Computer- and Mediaservice
birgit.matthaei@cms.hu-berlin.de
Why this technical questionnaire?

1st Technical Validation Questionnaire
- provide an overview on status, experiences and future plans belonging OAI implementations of participants of the 1st OAForum Workshop
- target group: workshop participants

High Interest, Feedback
- to collect experiences of a broader spectrum
- to learn more about starting conditions of planned implementations
  - Is there large common ground?
  - Are requirements so individual that it will be necessary for many isolated solutions to be developed?
  - Should tools and protocols correspond more than now to the needs of different communities?
What are the Goals?

- Extended 2nd Questionnaire
  - extended questions + target audience + duration
  - new subdivision in two questionnaires
    - technical presuppositions of those, which have not yet integrated OAI-PMH
    - experiences of implementers

- to get information about
  - used software
  - implementation costs
  - offered spectrum and interoperability
  - experiences and expectations
    - in different communities
    - in different countries

- to share experiences and information about technical issues related to open archives
Who participated to date?

Countries

- Austria: 1
- Belgium: 1
- Denmark: 2
- Germany: 5
- Italy: 3
- Latvia: 1
- Moldova: 1
- Netherlands: 4
- Norway: 3
- Portugal: 2
- Spain: 1
- Sweden: 1
- Switzerland: 1
- United Kingdom: 6
- Overseas: 4

- Repository that does not have any OAI implementations, but the possibility is being considered (21)
- Already OAI compatible Data Provider and/or Service Provider (test implementations included) (28)
Who participated to date?

- **Data Provider**
  - Active: 17
  - In development: 9
  - Planned: 19

- **Service Provider**
  - Active: 5
  - In development: 10
  - Planned: 12

30% of active DP are also SP
41% of active DP plan or still develop SP implementations
Who participated to date?

- Communities
  - Multiple answers possible

- no specification: 11
- Library: 23
- Archive: 9
- Preprints/Science: 13
- Museum: 4
- Publisher: 2
- Others: 9
Used Software

- Technical infrastructure before OAI-Implementation
  - not many statements to Interface and Collection Systems
  - dominant programming languages:
    ° Perl, XML, also Java, PHP
  - dominant databases:
    ° MySQL, Oracle

- Almost no one changed existing software tools to be OAI compatible
Implementations to be OAI compatible

- about 60% of the used tools were self-developed by both Data- and Service Providers
- most of them make their developments and the source code available for others
- dominant programming languages: Java, Perl, PHP, also XML
- tools like PERL implementations, OAI Cat, EPrints, and OAI Harvester were mentioned 3 or 4 times each

list of OAI-PMH software:
http://www.openarchives.org/tools/
Implementation Costs

Necessary Know How: Data- & Service Provider

focused on various combinations of the following five competence fields:

• system administration (UNIX | Linux)
• web server configuration (Apache)
• knowledge on Databases and SQL (MySQL | Sybase | Oracle)
• programming skills (Perl | Java | PHP | Servlets | CGI | XML)
• experiences with metadata
Implementation Costs

➢ Time and Manpower

- implementations of OAI-specifications:
  ° 75% concluded within a quarter by one programmer
    (span: from 2 to 750 personal days per month)

- reasons for few bigger expenditures:
  ° context of bigger research projects
  ° construction of archives
  ° processing of bigger data amounts

- further maintenance for a stable protocol:
  ° max. 25, mostly 1 personal day per month
Implementation Costs

➢ Expectations of those who haven’t implemented yet

- implementations of OAI-specifications (same):
  ° concluded within a quarter by one programmer

- further maintenance for a stable protocol (higher):
  ° up to 40 personal days per month

- No specific trend recognizable with expectations if
  ° data structures suggested by the OAI-PMH are easy to integrate in existing infrastructure
  ° the adaption of the data to the OAI-PMH will be expensive
  ° the preparation of the data for an internet usage will be expensive
Offers of Data Providers

- **number of documents:**
  - between 5 and several million documents

- **storage space:**
  - between 1 megabytes and 2 Terabyte.

- **object types:**

<table>
<thead>
<tr>
<th>Object Type</th>
<th>0</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td>19</td>
<td>15</td>
<td>13</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fulltext documents</td>
<td>15</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images - digitised mat.</td>
<td>6</td>
<td>11</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images - Vector</td>
<td>1</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abstracts</td>
<td>7</td>
<td>8</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video/Streams</td>
<td>5</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audio</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Raw/Statistic Data</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **already OAI compatible**
- **not yet OAI compatible**
Offered spectrum - DP

- Content types

- Dissertations
  - already OAI compatible: 12
  - not yet OAI compatible: 8

- Journal Articles
  - already OAI compatible: 11
  - not yet OAI compatible: 7

- Preprints
  - already OAI compatible: 9
  - not yet OAI compatible: 8

- Lectures
  - already OAI compatible: 5
  - not yet OAI compatible: 4

- Conference Proceedings
  - already OAI compatible: 4
  - not yet OAI compatible: 6

- Recordings
  - already OAI compatible: 2
  - not yet OAI compatible: 11

- Others
  - already OAI compatible: 13
  - not yet OAI compatible: 9

Legend:
- Green: already OAI compatible
- Blue: not yet OAI compatible
- Metadata formats

Dublin Core simple: 14 already OAI compatible, 5 not yet OAI compatible
Dublin Core qualified: 7 already OAI compatible, 4 not yet OAI compatible
MARC 21: 4 already OAI compatible, 2 not yet OAI compatible
UNIMARC: 3 already OAI compatible, 2 not yet OAI compatible
MAB: 1 already OAI compatible, 2 not yet OAI compatible
EAD: 1 already OAI compatible, 2 not yet OAI compatible

Others (single mentioned): 11 already OAI compatible, 2 not yet OAI compatible

Single mentioned formats:
- Dublin Core Library Profile, DiTeD, CEOS CIP, AMF, RIS, MODS, METS, SPECTRUM, TEI, internal format, self developed
Offered spectrum - DP

- Dissemination

  - more than half of the Data Providers are offering all parts or rather extracts of the documents
  - if the openness of the OAI interface is reduced due to several reasons, people use two limitation strategies:
    - access control
      (control of the IP-addresses, licensing, agreements)
    - limitation of the data output
Offered spectrum - SP

- **Kind of Services**
  - OAI-Service / Portal
  - local or community specific services
  - searching and browsing for information
  - search in different sources through one search interface
  - cross-linking, annotations, harvesting
  - workspace for managing documents and metadata, collaboration within groups of users
  - document management
Offered spectrum - SP

➢ Strategies to process with harvested data from DP

- use no provenance information
- filter harvester output and load local database
- strategies to include information about DP in data output:
  ° when a metadata record is found, the user can also browse information on the archive the record came from
  ° queries against the portal return data sets as harvested, including information about the original data provider
  ° provenance information is encoded in the identifier
Experiences - DP

➢ Importance / Advantages of OAI

- provide additional services to existing services
- replace existing services through OAI interface
- better retrieval, make Metadata exchange available
- share scientific knowledge, harvest other knowledge databases, cross-search in institutional assets
- major dissemination of researchers' results
- simple and cheap in implementation
- easy adaption for project internal usage
- simple to implement facility of exchanging metadata in comparison to more complex protocols

→ „provide access to all of human knowledge“
→ „nothing other than political expediency“
Problem: Standardisation

- heterogeneity of the content of the metadata records requires the service provider to expend a lot of effort in normalizing the data in order to make it more comparable and usable
  - could be done at lesser cost by the individual data provider
  - development of middleware tools that service providers could use for data normalization
Experiences - SP

Future Plannings

- extend search & browse functions
- export in other formats such as XML
- document delivery services, print on demand
- collaboration environment for users and groups of users, discussion forums, annotations, awareness
- extend existing services, building distributed services
- establish an exchange of different library catalogues and the integration into a virtual union catalogue for the whole country
- create a single catalogue of all library's catalogues: library opac, archives database, image database, Internet gateways
Problems to find useful informations?
- Many of those who haven't implemented yet made the experience that it is laborious to find good informations about metadata and especially technical support
- Some asked for a gentle introduction to the protocol ("too jargonish")

Recommendations of the questionnaire participants
- Online journals eg. Ariadne, D-Lib Magazine • www.ariadne.ac.uk • www.dlib.org
- Conferences and workshops
- Informal discussions with other gateway managers
- Test programs eg. http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai
Thank You!

➢ Please contribute!
- Information about your projects
- Your implementation and usage experience

Technical Validation Questionnaire
http://www.oaforum.org/resources/tecvalq2.php

Information Resource Database
http://www.oaforum.org/oaf_db/

Humboldt-University, Berlin, Germany - Electronic Publishing Group - CMS / University Library