The Austrian Library Network

*eDOC* - simple open-source software based application for enriching the Aleph Consortium Catalogue with non-bibliographic contents

Victor Babitchev

16th ICAU Meeting

London, 12th - 14th September 2005
The Austrian Library Network

The Austrian Library Network (ALN) - a homogenous* Aleph 500 network that began operation in January 1999.

OBVSG - the Austrian Library Network and Services Ltd. - the central office of ALN (15 staff members, http://www.obvsg.at)

ALN currently includes: 60 institutions using Aleph 500 and Alephino as follows:
- 16 run Aleph on 11 servers outside of OBVSG
- 39 use Aleph 500 operated by OBVSG on its 4 servers
- 5* run ILS Alephino on their local servers

Central cataloguing and data replication
- Each institution does primary cataloguing into the Consortium Database (4.7 m records) supported by the Aleph Cluster module running on the OBVSG Central server
- The bi-directional online data replication transfers data changes between the Central and local Aleph and Alephino systems

The data replication provides a solid support for enriching local ALN catalogs!
eDOC – phases of development - 1

2000 - Phase 1: development of a workflow and of program tools supporting processing of scanned tables of contents, abstracts, reviews as a supplement to Aleph 500 in the ALN consortium

Results: a simple and well accepted workflow; accumulating of scanned documents during the working hours on a library server, followed by automatic data transfer in the evening hours to the central server; update of bibliographic records beginning with the Consortium database and, then, as a result of the data replication, local catalogues having matched titles are updated. Thus, often not only the library “contents provider” but other Consortium libraries get their catalogue records enriched.

The major drawback - no full text searching across scanned contents

2002 - Phase 2: experiments with search engines for indexing of non-bibliographic contents

Swish-e was chosen as an open source search engine that successfully passed our performance tests, including indexing and retrieval of 1 million bib. documents. Development of the first prototype integrating swish-e into the workflow of phase 1 and the reciprocal linking of Aleph OPAC and swish-e display screens (presented at the 14th ICAU Meting). The production implementation was postponed.
eDOC – phases of development - 2

2004 - Phase 3: began in May 2004 due to the lack of functioning alternatives to eDOC for ALN and financial constrains in many ALN libraries. The work on eDOC was resumed

Objectives

- To implement a productive environment based on the results of phase 2 enabling the non-expensive enriching of ALN catalogues for at least the next 2-3 years, supported by modern open source search engine technologies integrated into Aleph 500 OPAC

- To preserve the eDOC simplicity in handling of electronic documents achieved in phase 1

- To position eDOC as a supplementary tool to the Consortium and local libraries’ OPACs for searching in free, mainly non-bibliographic contents

- Implementation of a single search mask for simultaneous searching of Aleph OPACs and eDOC, provision of local library-specific views for integrating eDOC searches into local Aleph OPACs
eDOC – project participants and the contents they provide

**eDOC project participants (Aug. 2005):**

University libraries of:
- Leoben University of Mining and Metallurgy (from April 2005)
- University of Graz (June 2005)
- University of Innsbruck (2000)
- University of Natural Resources and Applied Life Sciences Vienna (2004)
- Vienna Medical University (2004)
- Vienna University of Economics and Business Administration (2001)
- Vienna University of Technology (2003)

Austrian State libraries:
- the Library of Upper Austria (2005)

**What contents do libraries provide to enrich their catalogues?**

- In the selection of materials each library follows its own scanning strategy; e.g. tables of contents are often scanned for conference proceedings, journals and multi-volume works
- Two libraries delivered tables of contents, title pages and indices that they received from publishers
- Some libraries supply such full text documents as: abstracts of theses, tables of contents, full texts of theses etc.

*For text materials PDF is the currently preferred format!*
eDOC – what is it now?

eDOC is:

- an uncomplicated application that from the very beginning was developed as a “ALN consortium-enabled" tool. *Its purpose is to enrich the Union Catalogue with such non-bibliographic contents as tables of contents, reviews, abstracts, images etc.*

- an application that automatically collects data batches from libraries and integrates them into eDOC and Aleph, providing library or data specific Web views to this information

- an in-house developed application based on proven open source tools and INTERNET technologies including the search engine Swish-e that is used both for indexing of data selected from a RDBMS (MySQL) repository and for retrieval

- a tool providing simultaneous searching of the Aleph OPAC and Swish-e indices

- a set of statistical tools for monitoring both non-bibliographic data entry and data use on Consortium or single library levels

- accessed via httpd://media.obvsg.at/suche
eDOC – naming and addressing of digital objects

Naming

Every digital object submitted to eDOC for processing must have a **special name**

Example of a scanned TOC as file in PDF format

```
AC00016679n01in.pdf
```

Bib. record ID - Aleph Tag 001 (unique in ALN), it is also ID of the respective eDOC meta data record

Access

Every object in eDOC is directly addressed by a “stable URL“ of the following structure:

```
<Stable URL>::= <Meta data record ID><-><Contents code><Object sequence>
```

e.g. 
- http://media.obvsg.at/AC00016679-1005  ← access to eDOC object
- http://media.obvsg.at/AC00016679  ← access to eDOC meta record and links to its objects
# List of supported objects

<table>
<thead>
<tr>
<th>Object Type</th>
<th>Suffix</th>
<th>Contents code</th>
<th>Tag in ACC01</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard Objects</strong></td>
<td></td>
<td></td>
<td>655e</td>
<td></td>
</tr>
<tr>
<td>Table of contents</td>
<td>in</td>
<td>10</td>
<td>update *</td>
<td>Preferred format - pdf Option: no 655e-update for TOC objects grouped into the aggregate object (mod. = ‘x’)!</td>
</tr>
<tr>
<td>Abstract</td>
<td>ab</td>
<td>30</td>
<td>update</td>
<td>Expand into bib. display for txt-formatted objects</td>
</tr>
<tr>
<td>Review</td>
<td>rz</td>
<td>33</td>
<td>update</td>
<td></td>
</tr>
<tr>
<td>Title cover</td>
<td>ub</td>
<td>40</td>
<td>update</td>
<td>Expand into bib. display</td>
</tr>
<tr>
<td>Full text</td>
<td>vt</td>
<td>41</td>
<td>update</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Special objects</strong></td>
<td></td>
<td></td>
<td>655e</td>
<td></td>
</tr>
<tr>
<td>Aggregate</td>
<td>is</td>
<td>15</td>
<td>update</td>
<td>html file linking bib. record with multiple objects of the same type – e.g. TOCs of journals</td>
</tr>
<tr>
<td>External object</td>
<td>-</td>
<td>-</td>
<td>read 655e</td>
<td>html file linking external objects</td>
</tr>
</tbody>
</table>

* Update action is set by the ‘modifier’: ‘n’ – update, ‘x’ – no update
eDOC in ALN Consortium – workflow “step by step”

1. Library A
   - Tables of contents, cover pages, reviews...
   - Abstracts
   - Data from publishers
   - eDOC batch client
   - Aleph Local System

2. eDOC Repository
   - e.g. AC00016679n01m.pdf
   - AC00016680n01m.jpg

3. Documents
   - lib.-specific and common Web search masks

4. Swish-e search
   - A
   - Z
   - ALL

5. eDOC Server

6. Central Catalogue
   - ACC01

7. Central OPAC

8. Aleph Consortium Server

9. Central OPAC

10. Data Replication to Aleph and Alephino Local Systems

11. Central OPAC

12. eDOC Record

13. OPAC Library Y

14. OPAC Library Z

15. bib. data update

16. Data Replication from ACC01

17. OPAC of Library A

18. bib. data update
eDOC – some statistics

Production statistics as of 30.08.2005

eDOC repository contents

1. Meta data records: ~ 6550 (html, Dublin Core qualified; xml - possible)

2. Objects

- 4565 – TOCs
- 55 – Aggregates (e.g. TOC-links summaries to journals)
- 1449 – Abstracts
- 450 – Cover pages
- 35 – Title pages
- 650 – Links to external full text documents
- 180 – Reviews

Total ~ 7400 objects (~ 4000 objects in Nov. 2004)

Object data formats

- 6 – application/msword
- 41 – image/gif
- 55 – text/html
- 407 – image/jpeg
- 3139 – application/pdf
- 3746 – text/plain
Swish-e – the eDOC „family“ of Web search screen masks

Search screen masks can be tailored to requirements of libraries or collections.

- the combination of free text and fielded searches is also supported (i.e. “CCL”-like)

- one common index with filters, or separate indices can support such searches in Swish-e

Subject headings TOC Year
Simultaneous searching of Aleph and eDOC

1. The user begins searching in the Aleph OPAC with both checkboxes (OPAC and eDOC) selected, enters a search string as usual and then presses the START button…

2. The search results appears in two separate windows, for both Aleph (2) and Swish-e (3) in brief formats.

Note that the first title in Swish-e is not in the Aleph list, because the search words came from the TOC attached to this title (a sample how the catalogue enrichment may work).
1. A monograph title with attached title page and TOC. The title page is expanded into the bibliographic display.

2. A sample of a journal title with 70 TOCs attached

Documents containing text information are full-text searchable in Swish-e!
Access to a special collection of historical journals of the Austrian National Library. The journal articles with links to objects ‘harvested’ and indexed by eDOC. Access to external objects is provided by eDOC, objects control is done by an external application (pilot project).
eDOC – Web access statistics examples

Access to objects of all eDOC participants for March - August 2005, (from http://media.obvsg.at/edocstat/all/)

Statistics of access to eDOC objects of the Vienna University of Economics and Business Administration Library for March-August 2005.

*Individual Web statistics pages are available for each eDOC participant*
**eDOC – processing overview**

1. **Batch client.** Processes scanned objects prepared by the library and sends them to OBVSG.

2. **eDOC Server.** Gets uploaded objects, builds from each object a bibliographic record, DC and technical meta data and loads them into MySQL database (eDOC repository). Prepares Aleph 500 “655e update file”.

3. **eDOC Server.** Data extraction from the repository and indexing through the Swish-e search engine.

4. **Aleph Consortium server.** Gets Aleph sequential “655e update file” and updates the Central catalogue. Aleph Cluster replicates changes to other Aleph and Alephino systems.
eDOC indexing and retrieval – library and collection specific search masks

Data selection and indexing

- filter / select
- Swish-e indexer

Swish-e indices

- Common Consortium Index
- Diss.DB
- ANNO-A ext. collect.
- Univ. A
- Univ. Z

Swish-e search masks for libraries and collections

- eDOC-Consortium
- ANNO-A
- Univ. A
- Univ. Z

External data source

Conf.files

ANNO-A ext. objects
Summary – 1

The current phase of eDOC development has been accomplished.

eDOC runs stable, fully automatic, and is highly parameterized. Thus, the implementation of a new library to eDOC takes usually less than 3 hours. Because of its simplicity, the training hours are also minimal. However, our Web implementation of Swish-e is not that fancy and is relatively simple (lack of our resources 😞)

eDOC is not a panacea for the management of all types of digital documents, it is a practical tool developed by OBVSG providing a way for the ALN libraries to begin with the enrichment of their catalogues. As a result of this, our users have a chance to access and search data that were otherwise not available in the catalogue-card-like records of our OPACs

Enriching of catalogues requires resources and therefore, it is imperative to look for cooperation of exchanging data with other libraries and consortium!
And to be frankly, under its “hood”, eDOC is not that simple as it looks from outside, because it cannot be so simple to integrate open source tools and technologies into one application, and then to tune it, especially, when Aleph 500 is the main player in this mosaic. Excellent librarians and the IT expertise of our colleagues have made eDOC possible, and we are proud of our “creature” 😊

The time will come when better, maybe commercial products will replace our eDOC. When will this happen? Maybe in a couple of years, or even sooner (Digitool, ADAM, FAST, X, Y, Z)!