Harvesting the Crop – Voyager and OAI-PMH

IGeLU 2007
Brno
Ere Maijala
The National Library of Finland

Agenda

• Background
• OAI-PMH
• How does the script work
• Features
• Compared to Z39.50 or SRU
• How we use it
• Technical stuff

Background

• Voyager 6 ("repository")
• MetaLib portal with MetaIndex service ("harvester")
• Collection Map project
• Need to highlight collections in library catalogs
• Finding the corresponding records too slow or impossible with search protocols

OAI-PMH

• Open Archives Initiative Protocol for Metadata Harvesting
• A fairly simple protocol for harvesting metadata via HTTP
• Request in URL
• Response in XML
• Supports multiple metadata formats
  – Commonly Dublin Core and MARC-XML

How Does the Script Work

• Alongside WebVoyage on the server
• Connects directly to Oracle
• Queries the database
• Converts results into Dublin Core or MARCXML
• http://server/cgi-bin/oai-pmh.cgi?verb=Identify
• http://server/cgi-bin/oai-pmh.cgi?verb=ListSets
• http://server/cgi-bin/oai-pmh.cgi?verb=ListRecords&set=intermezzo&metadataPrefix=marc21

Features

• Complete support for OAI-PMH 2.0
• Set definitions based on
  – Record format
  – Location
  – Call number
  – Publication place
  – Language
  – Keyword search
  – Custom filter function
• Handling of deletions
• Bibliographic and authority records
• Access Control via IP address list
Technical Stuff

- Written in Perl
- Passes protocol conformance tests
- Can access keyword server directly
- Supports seconds granularity and time zones properly
- Filter function in Perl to create custom rules for filtering records
- No perceivable impact on the server
  - The masses are elsewhere
- Keep-alive for slow fetches
- OAI-PMH:
  - [http://www.openarchives.org/OAI/openarchivesprotocol.html](http://www.openarchives.org/OAI/openarchivesprotocol.html)

Compared to Z39.50 or SRU

- Not a search protocol
- Can be slow
- Results must be processed further in a way or another
- Data can be harvested from multiple sources into one collection

How We Use It

- To harvest collections to our MetaLib Portal (Nelli)
- E-theses (single and combined)
- Music collections (combined from multiple libraries)
- Subject collections
- Collection Map
- Separating (geographically) distinct organizational units in a library catalog
- Could be used to replace periodical exports in some cases

Thank You!

- [http://www.nationallibrary.fi/libraries/linnea/resources.html](http://www.nationallibrary.fi/libraries/linnea/resources.html)