Digital library projects in the Nordic national libraries

Juha Hakala

Helsinki University Library –

The National Library of Finland

Contents

- Introduction
- Current projects
  - Nordic Web Archive
  - Scandinavian Virtual Union Catalogue
  - Identification of electronic resources
- Some shared challenges
  - Legal deposit
  - Long-time preservation of electronic resources

Introduction

- Nordic national libraries have important roles in their communities
  - Format & cataloguing rules maintenance: all but DK
    - shift towards MARC21
  - National bibliography: all (DK: music)
  - Article index: FI, IS, NO, SE (as part of Libris)
  - Union catalogue host: FI, IS, NO, SE
  - Large-scale digitisation: all (especially NO)
    - project TIDEN

Introduction (2)

- There is a long tradition of co-operation between the Nordic libraries in general
  - Meetings of the national librarians (NORON)
  - Topical meetings since at least 1920's, ranging from library science students to diverse professional communities, including e.g. ILL specialists
  - Joint projects, funded e.g. by Nordinfo & Nordunet2

Nordic Web Archive

- Partners: all Nordic national libraries
- Funding: libraries + Nordunet2
- Aim: archive the freely available Web documents for future generations as a part of each library's legal deposit obligation
- Free access to the index, limited access to the deposited documents

NWA – Background

- Kulturarw3 project, Kungliga biblioteket
  – proved the feasiblity of the Web archiving
  – the Swedish Web space has been archived
    several times with altered Combine harvester
- NEDLIB project, 1997-2000
  – EU-funded initiative, many national libraries
    involved
  – developed the NEDLIB harvester, using the KB
    experiences as the starting point

NWA tools

- Web harvesting and archiving is done by the NEDLIB harvester (except in SE)
  - it is open source, optimised for the Web archiving purposes
    - archiving module, weird scheduling principles, archive metadata (MD5 checksum, time stamp, original URL)
  - Multiple users – better maintenance and development
  - "Combat proven" – strengths and weaknesses are known reasonably well

NWA tools (2)

- Indexing is done by a search engine built by a Norwegian company FAST
  - Can process billions of files
    - www.alltheweb.com
    - Present need: tens of millions files
  - Can handle >200 file formats via conversion to HTML prior to indexing
  - Can recognise large number of languages

NWA tools (3)

- Diverse additional modules are under development in national libraries in order to facilitate access to the archived files

- These tools will most likely be available in the public domain, like the harvester
  - The search engine is the only commercial module in the package

Archiving results: Finland

- Harvesting of *.fi was completed in 3/2002
  - A few weeks of processing with Sun E450
- 29 million URLs, 9.4 million files
  - Same proportion of duplicates as in Iceland
- After compression, the archive is 340 GB
  - Storage on tape robot in CSC (Finnish NCSA)
- Next step: Finnish pages in other domains
  - Co-operation with InfoCenter Finland

Archiving: experiences

- The Internet is a dirty place
  - Quality of data and (some) applications is appalling
  - Any tool dealing with a large number of Web resources must be extremely robust
- Very important to have an encompassing list of start pages

Archiving: problems

- The cost of storing the bits is small, for now
  - Ratio of the size of the Web and storage price in the future?
- Preserving access is easy for HTML, JPEG and GIF (97 % of the archive content)
  - The rest will be a problem in the future
- How to get to the "deep Web"?

Scandinavian Virtual Union Catalogue

- Partners: national libraries, Bibsys (Norway) & Dansk BiblioteksCenter
- Funding: partners + Nordinfo
- Aim: free use of national union catalogues to the consortia maintaining these databases
  - Each partner "pays" with its data for access to all other systems
- 1st contract 1.3.2002-31.12.2003

SVUC – databases

- DK: Danbib

- FI: Linda & Manda

- IS: Gegnir

- NO: Bibsys, Sambok

- SE: Libris

- Approximately15-20 million records

- New databases and partners may be added in the future

SVUC – services

- Searching – directly via Z39.50 connection, possibly also via Web portals

- Copy cataloguing
  - Via Z39.50, using the Bath profile; One-2 profile support also possible

- Future extensions (e.g ILL and document delivery) are likely; no schedule yet

Identification of electronic resources

- Nordic national libraries participate actively in development of e.g. ISBN, ISSN and Uniform Resource Names (URNs)
- Common principles and some shared software development in implementation of URNs based on nat. bibliography numbers
  - Initial development of e.g. the URN generator SW in co-operation with Netlab

Shared challenges: legal deposit

- All Nordic countries have either recently revised their legal deposit acts or are in the midst of the process
  - Lots of sharing of ideas is taking place
- Revision of the Copyright Act (in order to align it with the EU Copyright Directive) is under way
- Legal platform for deposit and preservation of electronic resources will be built; lobbying needed to guarantee its suitability for (national) libraries

Long-time preservation

- NEDLIB provided a good starting point; since then the activities have shifted to domestic level
  - There is a risk of re-inventing the wheel, e.g. in development of preservation metadata
- Need for European / Global co-operation
  - OCLC/RLG Preservation metadata WG
  - development and evaluation of tools

Summary

- Shift from bibliographic data to "full text" is well under way

- This creates legal and technical challenges, which the Nordic national libraries are solving together

- For historical and organisational reasons local priorities differ, but there are a lot of shared activities

Links

- NWA - http://nwa.nb.no/
- NEDLIB - http://www.kb.nl/coop/nedlib/
- NEDLIB harvester - http://www.csc.fi/sovellus/nedlib/
- SVUC - http://www.lib.helsinki.fi/svuc/
- URN - http://www.ietf.org/html.charters/urn-charter.html