Persistent identifiers, long-term access and the DiVA preservation strategy

Eva Müller

Electronic Publishing Centre
Uppsala University Library, Sweden
http://publications.uu.se/epcentre/
Outline

• DiVA project and its objectives
• DiVA publishing system
• Persistent identifiers and their roles within the DiVA publishing system
• Conclusions and next steps
DiVA Project

Started 2000 at Uppsala University, Sweden 2004

- ten universities
- three countries
DiVA - Academic Archive Online

(Digitala Vetenskapliga Arkivet)

- Objectives of the DiVA Project:
  - Technical solutions & workflows supporting fulltext publishing, storage and dissemination of university research (theses, dissertations, working and research papers...)
  - Explore ways to ensure future access, use and understanding of digital objects in the archive
DiVA Publishing System
makes it possible to
• reuse and enhance data from source
documents originally created by authors,
both for metadata and a digital master for
electronic & printed versions
• assign a persistent identifier, store &
checksum all files in a local archive
• send a copy to the national library archives
and to other interested parties

Eva Müller, Uppsala University,
Sweden
ERPANET seminarium,
June 2004, Cork
Long term access and the DiVA preservation strategy

• Issues
  – How can we ensure access to documents we produce locally?
  – How can we minimize risks for data loss?
  – What factors increase potential for success?
  – Can these factors be integrated into an automated and low-cost workflow?
How can we ensure access in the future?

– A stable point of reference (persistent identifier)
– Use human-readable, non-proprietary storage format for metadata and if possible even for the content (published documents)
– Storage in several locations
How can we minimize risks for data loss?

- Multiple copies in different locations
- Mechanism to keep track of copies

Can we integrate all these factors into an automated and low-cost workflow?
Long-term access
Stakeholders

– Producers
  • Authors
    – Discovery of their intellectual output
    – Dissemination of their intellectual output
  • University Publishers
    – Increase impact
Long-term access
Stakeholders

- **Consumers**
  - Authors (citation durability)
  - Readers (discovery, bibliography)
  - Universities (track research output)

- **Curators**
  - National Libraries (legal deposit)
  - Archives
  - ? Other parties
Some requirements for PIDs and their resolution

– Easy and reliable maintenance and administration
– Potential to connect a preservation copy to the PIDs (guarantee long-term access)
– Possibility to integrate into automated and low-cost workflows
Which PID and why?

- Cooperation with a trusted, public and non-profit organization
- Management of a resolution service, other metadata services and an archival copy within the same framework
- Possibility to use the same PID for different manifestations of the same content
- Non proprietary solution
Based on that:

- Decision to cooperate with the National Library of Sweden
- Decision to use XML as a primary storage format
- Decision to use URN:NBN as a primary persistent identifier
- Decision to fit all needs into an automated workflow
Assignment of the URN:NBN

The name assigning authority – The Royal Library, the National Library of Sweden – assigns sub domains
Sub domain – manages locally
Structure URN:NBN:se:?:diva

URN:NBN:se:uu:diva+locally managed serial number
URN:NBN is used as identifier for each item – an item is a single publication without consideration of format, where various formats of the item (the identical content) are manifestations
Implementation of URN:NBN Resolution Service

- Version 2.00 released in May
- A new version in cooperation within Nordic countries coming in fall 2004
- Implemented as a java-servlet and contains a harvester which can harvest URN:URL-bindings from many different repositories
Resolution Service Configuration File

URN:NBN Register Format

DiVA

Other

URN:NBN:se to URL mappings

User

URN:NBN:se to URL mappings

response
e.g. http://urn.kb.se/resolve?urn=

user redirected to an URL

Repositories

URN:NBN resolution service

User

Royal Library

URN:NBN Register Format
URN:NBN and its various roles within the DiVA system

- URN:NBN as a unique identifier within the archive
- URN:NBN as a naming convention for files, directories and archival packages
- URN:NBN as a part of disseminated metadata
URN:NBN as a naming convention for files, directories and information packages

Eva Müller, Uppsala University, Sweden

ERPANET seminarium, June 2004, Cork
Information Package

- metadata
- content
- stylesheets
- schemas
- checksums

checksum

name: URN:NBN:se:[specific part]
URN:NBN as a part of disseminated metadata
Central

URN:NBN Resolution Service

Local

Long-term Storage

List of URN:NBN to URL mappings

Repository

Metadata & Content

Long-term storage packages

Library Catalogue

MARC 21

Metadata

XML
Other IDs used within the DiVA

- Within the documents to identify (as pointers to)
  - schemas
  - name authorities
  - authorized names (person name, institutional name), geographical places
  - and other registries and entries in those registries.
- DiVA Document Format supports the concept generically through “Identifier” elements

Currently no broadly agreed upon recommendations in the many fields
DiVA Document Format

Identifier component

identifier agnostic

The identifier name is specified in a property element. Currently valid identifiers are internal, isbn, issn, local, uri, iso639-1, iso3166-1
Comprehensive identifiers for the document. Identifiers specified here belong to all manifestations. The property internal is used to link this document to other external descriptions. The value with the property uri contains for example the URN:NBN identifier of the document.
Identifiers for the serial publication. The property issn is used for the ISSN identifier. The property internal is used to link this serial publication to a more detailed external description.

... implemented
Container element for organisation identifiers. The property internal is used to link the name of organisation to a more detailed external description. Identifiers can for example link the organisation to an authority data register (identifier name not implemented yet).

... partly implemented
Container element for person identifiers. Identifiers can be used to link the person to an authority data register (identifier name not implemented).
Archiving workflow to the National Library

Infrastructure

– Local producer
– Central archive
– Solutions and methods for addressing and identifying the resources
– Methods for transmission of data (information packages)
– (Temporary) File format registry
Infrastructure

Consumers
metadata – local services, Union Catalogue, OAI-based services ….

Local archive
(university, other)

Producers

Resolution Service

Available at local a.?

Y

N

Archive
(documents and metadata)

Format registry

Metadata & PI

URN:NBN

Information Packages

metadata

Eva Müller, Uppsala University, Sweden

ERPANET seminarium, June 2004, Cork
Infrastructure/producer

• Local producer
  – Follows recommendations on:
    • Metadata
    • Storage formats
    • Persistent Identifiers
    • Organization of the local archive
  – Implements solutions and routines for storage of the data and transmission of the data to the central archive
Infrastructure/archive

• Central Archive
  – sets up requirements for the producer regarding quality of the data delivered to the archive
  – provides quality control of the delivered package at ingest event
Infrastructure

- Methods for addressing and identifying resources
  - provides conditions for long-term access
    Primary
    - URN:NBN
    - URN:NBN resolution service
  Secondary identifiers (e.g., Handle, DOI, ARK)
Infrastructure

- Transmission of data (information packages)
  - Provides guarantees for access in the long term
    - Verifiable agreement
    - Quality control on both the producer side and on the central archive side
Infrastructure

• (Temporary) File format registry
  – Provides additional information about formats submitted to the archive
  Methods
    • **Persistent identifiers** for format information
    • Populate format metadata on ingest

Using format registry information increases probability of longevity of the archived documents by providing more technical metadata in uniform form

• Relation to other format registry projects
Identifiers for the manifestation. Here can identifiers pointing to a file format register/dictionary can be specified (not yet implemented).

...not yet implemented

Pointer to format registry/format dictionary

Eva Müller, Uppsala University, Sweden
DiVA project experience

Conclusions:

• Low-cost system that supports an semi automated workflow from the point of submission works well
  – Automated creation of metadata
  – Workflow to the National Library Archive

• Using harvesting model for updates to the mapping registry makes the management of URN:NBN simple, reliable and economic

• Long-term access to institutional research can be assured with cooperation from national libraries
Next steps

On the national (Swedish) level:
2003-2005 project “Coordination of electronic academic publishing at Swedish Universities”. Subproject “Long-term access and preservation” with goal to develop and implement an generalized archiving workflow between a local repository and a national archive focusing on the variety of publishing platforms and systems.

On the Nordic level:
Additional development of the resolution service is being undertaken as a cooperative effort amongst the Nordic countries within a by NORDINFO granted project “Access to documents now and in the future”.

Further development of the URN:NBN resolution service as international cooperative effort.
… but is the international cooperation within URN:NBN community enough?

No!

There is a need for a global resolution mechanism which can accommodate different types of identifiers!
More information

- Electronic Publishing Centre, Uppsala University
  [http://publications.uu.se/epcentre/](http://publications.uu.se/epcentre/)
- DiVA – Academic Archive Online
  [http://www.diva-portal.org/about.xsql](http://www.diva-portal.org/about.xsql)
- SVEP (Coordination of electronic publishing at Swedish universities)
- NORDINFO granted project “Access to documents now and in the future”
  [http://epc.ub.uu.se/niwiki/pmwiki.php/Main/HomePage](http://epc.ub.uu.se/niwiki/pmwiki.php/Main/HomePage)