



Enhanced E-theses Project

Deliverable 9.1

Requirements for Management & Storage to support complex objects & ORE in DSpace

Document Description

Project	
Title:	Enhanced E-theses
Start date:	1 st December 2007
Funding Agency:	Knowledge Exchange
Document	
Deliverable number:	D9.1
Deliverable title:	Requirements for Management & Storage to support complex objects & ORE in DSpace
Actual Date of Delivery:	January 2009
Author:	Peter Ruijgrok (Utrecht University Library) Martin Slabbertje (Utrecht University Library)
Editor(s):	Ene Rammer Nielsen (Roskilde University Library)
Reviewer(s):	Martin van Luijt (Utrecht University Library)
Workpackage:	WP9
Workpackage title:	Management and preservation of complex digital objects within Institutional Repositories
Workpackage leader:	University of Utrecht
Version/Revision:	1.1
Draft/Final:	Final
Total number of pages: (including cover)	5
File name:	Enhanced Etheses project_deliverable 9.1_v1.1.doc

This work is made available under a Creative Commons attribution 3.0 licence. For details please see <http://creativecommons.org/licenses/by/3.0/>



Introduction

Nowadays, most of the institutional repositories make use of open source repository software like DSpace, Eprints or Fedora. If libraries want to successfully use OAI-ORE to create world-wide networks of aggregations, the first requirement is that the architecture of the repository software is compliant with OAI-ORE. Utrecht University Library makes use of DSpace. To investigate what should be done to make OAI-ORE support possible, we carried out a research on a small scale. As a reference point for currently available functionality DSpace 1.4.x / 1.5 is used.

Submission (& Workflow)

In addition to the currently available minimum features in repository software, the submission process, in order to support complex objects and ORE, should at least support the following additional features:

- Upload (a batch of) files or URLs structured in a logical (i.e. hierarchical) way, comparable with for instance the Windows file structure;
- During the submission process it must be possible to define and change the sequence and relationship between files.
- Submission of metadata-only records;
- Validate submitted files against the recommended file formats of the repository;
- Use a specific metadata-scheme for specific types of resources (datasets, software, multimedia etc.) at Bitstream level;
- Describe and verify (!!) semantic relations of the resource with other resources at Item level.

The majority of the features mentioned above is not supported in the current DSpace versions.

Storage

The storage of resources must support:

- Store & maintain Aggregations (in Items content) in a structured (tree-based) way (Community, Collection or Subcollection mechanism is insufficient !!);
- Store & maintain Aggregations (in the Bitstreams content) in a hierarchical (tree-based) way (other than the current 'all on one level' method);
- Store & maintain additional metadata about the (Aggregated) Resource (in the Bitstream content);
- Support the concept of manifestations, where an aggregation for a group of manifestations will have a Persistent Identifier. Resolving the Persistent Identifier should return a list of manifestations of the affected resource. Examples of manifestations are

- The original Word document and the generated PDF document
- The original TIFF bitmap and a generated JPEG and thumbnail bitmap
- The original WordPerfect 4.1 document and the transformed Word2007 document. (digital preservation aspects!).
- Maintain a persistent identifier for each (!!) aggregation. Retrieving Aggregation information is crucial using ORE and therefore aggregations must be identifiable and retrievable using persistent identifiers and resolvers.
- Create a persistent identifier for each Aggregated Resource (=Bitstreams) within an Item;
- Support for a variety of access rules (Open Access, Embargoes, limited access to a specific audience, etc.) to ensure that all types of data can be described and stored while respecting individual rights and to address/avoid legal issues. Access rules should be able to vary from file to file.

The majority of the features mentioned above is not supported in the current DSpace versions.

Relations

Relationships between aggregations and between bitstreams must be supported and relations must periodically be checked on validity. (Does the link work and does it return the expected result?)

It must be possible to maintain links for:

- Links to resources stored in the same repository, like
 - Previous / Next versions,
 - Appendixes stored as separate items in the repository,
 - Chapters of a publication,
 - Datasets (textual or multimedia materials).
- Links to resources stored in external repositories, like
 - datasets stored in another datastore like Archiv-x, DVN from MIT, DANS from the Royal Dutch Academy in the Netherlands
 - links to “Jump Off Pages”,
 - author-pages using the Digital Author Identification (DAI),
 - etc.
- Reference-lists making it easy to present the referred resource “with a mouse click”.

The majority of the features mentioned is not supported in the current DSpace versions.

Persistent Identifiers

Crucial in the ORE concept is a thorough support for persistent identifiers at each level

- Maintain a persistent identifier for each aggregation. Aggregation information is crucial for using ORE and therefore aggregations must be identifiable and retrievable using persistent identifiers and resolvers.
- Create a worldwide resolver infrastructure which resolves various types of identifiers (handle, doi, urn-based, etc) to accommodate that discovery software can throw an identifier to a central resolver which knows how to transform an identifier of whatever type into the actual URL which will return the result.

The majority of the features mentioned above is not supported in the current DSpace versions.