Appendices

EPrints v3 Configuration Notes

Enabling OAI

Enabling OAI in ePrints is very easy to achieve as versions of ePrints since v3.0.0 have included both oai_dc and uketd_dc OAI compliance as part of the standard installation. The only actions required to make an archive OAI-enabled are to configure the archive-specific settings so that OAI calls return the correct information.

These are stored in /opt/eprints3/archives/ARCHIVE_ID/cfg/cfg.d/oai.pl and comments within this file describe the institution-specific data than needs to be added. Most of the data is informational text (such as content- and data-policies) but the crucial, technical items are :

- Archive ID : The name of the archive. This MUST be unique. As this can include full stops, we use the obvious ARCHIVE_NAME.INSTITUTION_DOMAIN (eg. etheses.bham.ac.uk)
- Output Plugins : the metadata formats supported by this archive. These are used by the system to return data when an OAI call is made. See below for more information on this.
- Sets/Filters : Optional configuration so that exports can be made of subsets of the archive.

Once this is configured and the archive reloaded with the command/ opt/eprints3/bin/epadmin reload ARCHIVE_ID

Instructions on testing the compliance of the archive are given below.

Adding/Changing fields.

If any customisation of the ePrints metadata structure is made, new or edited fields may need to be added to the XML returned by an OAI getRecord command.

This XML is generated by the appropriate ePrints Export plugin and are assigned to metadata prefixes in the output_plugins section of
`/opt/eprints3/archives/ARCHIVE_ID/cfg/cfg.d/oai.pl`

by default this contains the lines:

```
# The output plugins must be loaded for the archive and have
# the methods xml_dataobj and properties for xmlns and schemaLocation.
#
# The keys of this hash are the OAI metadataPrefix to use, and the values
# are the ID of the output plugin to use for that prefix.
$oai->{v2}->{output_plugins} = {
        "oai_dc" => "OAI_DC",
       "didl" => "DIDL",
        "uketd_dc" =>"OAI_UKETD_DC",
       "context_object" => "ContextObject",
       "mets" => "METS"
};
```

which show that the eTHOS plugin is named OAI_UKETD_DC and is enabled by default. This plugin can be found at
`/opt/eprints3/perl_lib/EPrints/Plugin/Export/OAI_UKETD_DC.pm`

Although written in perl, little programming experience is necessary for basic administration of this file. Adding and editing simple fields can be done by copying an existing field's code. These are at the end of the file and of the format :

```
   if( $eprint->exists_and_set( "language" )){
      push @etddata, [ "language", $eprint->get_value( "language" ), "dc"];
                }
```

```
if( $eprint->exists_and_set( "sponsors" )){
   push @etddata, [ "sponsor", $eprint->get_value( "sponsors" ), "uketdterms"];
               }
if( $eprint->exists_and_set( "alt_title" )){
   push @etddata, [ "alternative", $eprint->get_value("alt_title" ), "dcterms"];
                    }
```

The last parameter on each line refers to the schema from which the type is derived. The above three fields, for example, will appear in the output XML as

```
<dc:language>English</dc:language>
<uketdterms:sponsors>The Umbrella Corp.</uketdterms:sponsors>
<dcterms:alt_title>Raccoon City Pop. Study</dcterms:alt_title>
```

Testing OAI Compliance.

The basic configuration of an archive (the name and policies) can be tested by pointing a browser to the address

`http://ARCHIVE.WEB.ADDRESS/cgi/oai2?verb=Identify`

e.g.

`http://etheses.bham.ac.uk/cgi/oai2?verb=Identify`

Full OAI compliance validity (the correct record formats, etc.) can be tested by entering the base url of the archive, http://ARCHIVE.WEB.ADDRESS/cgi/oai2, into the OAI-PMH Repository Explorer at http://re.cs.uct.ac.za/

The XML generated by a plugin can be viewed by submitting an OAI call including the Prefix=<prefixname> parameter. For Ethos, this would be :

`http://ARCHIVE.WEB.ADDRESS/cgi/oai2?verb=ListRecords&metadataPrefix=uketd_dc`

This will list all available records in an archive including their uketd_dc stanza. If an archive is large, an individual record's data can be retrieved using the address

`http://ARCHIVE.WEB.ADDRESS/cgi/oai2?`
`verb=GetRecord&metadataPrefix=uketd_dc&identifier=oai:ARCHIVE_ID:ID`

(Where the ID is the ePrint ID)

e.g.

`http://etheses.bham.ac.uk/cgi/oai2?`
`verb=GetRecord&metadataPrefix=uketd_dc&identifier=oai:etheses.bham.ac.uk:4`

EThOS persistent id

Field Name	Comments	qDC Element.Qualifier
EThOS persistent id	A persistent ID assigned to each individual element in EThOS. (reverse domain name 'uk.bl.ethos' as its first characters followed by a dot then a running number. e.g. uk.bl.ethos.12345)	identifier

UKETD_DC: The metadata core set recommended by EThOS

Field name	Status	Comments	qDC Element.Qualifier
Title	Mandatory	Full title, including any subtitle	title
Alternative Title	Optional	Other titles for the work, e.g. Translations or abbreviations.	title.alternative (refinement)
Author	Mandatory	The author of the work as on the title page. Separate the surname (or family name) from the forenames, given names or initials with a comma, e.g. Smith, Andrew J.	creator
Supervisor(s)/advisor	Optional	Thesis supervisor, other supervisors, and advisors. Format as for author.	*contributor.advisor*(local refinement)
Subject keywords	Optional	Any keywords that the student or librarian feel belong in the metadata. Populated by student for free text and librarian to verify and add full subject headings based on DDC, LCSH etc. as below	Subject
Abstract	Optional	Include translations in "repeatable" section.	description.abstract (refinement)
DDC	Optional	Dewey Decimal Classification headings as assigned by librarian	subject.DDC (encoding scheme)
LCC	Optional	Library of Congress Classification headings as assigned by librarian	subject.LCC (encoding scheme)
LCSH	Optional	Library of Congress Subject Headings as assigned by librarian	subject.LCSH (encoding scheme)
MESH	Optional	Medical Subject Headings as assigned by librarian	subject.MESH (encoding scheme)
UDC	Optional	Universal Decimal Classification headings as assigned by librarian	subject.UDC (encoding scheme)
Awarding Institution	Mandatory	Name of institution awarding degree (e.g. The Robert Gordon	*publisher.institution* (local refinement)

		University)	
Author Affiliation	Optional	Name of school, department, centre, faculty or college where the author was based. Add the name of the host institution only if this was not the same as the awarding institution. (Free text e.g. School of Computing, Faculty of Design and Technology. Where the author's host institution is NOT the same as the awarding institution enter: School of Computing, Faculty of Design and Technology, The Robert Gordon University)	*publisher.department* (local refinement)
Publisher	Optional	Name of the formal publisher of a thesis	*publisher.commercial* (local refinement)
Sponsors	Optional	Sponsor of student	*contributor.sponsor* (local refinement)
Grant number	Optional	Grant number allocated by funding body	*identifier.grantnumber* (local refinement)
Type	Mandatory	Type = Thesis or dissertation (default on system)	Type
Qualification level	Mandatory	Level = Diploma, Masters, Doctoral, Postdoctoral, etc (locally controlled look up list)	*type.qualificationlevel* (local refinement)
Qualification name	Optional	Name = Specific degree (MPhil ,PhD, DPhil etc). (locally controlled look up list or free text)	*type.qualificationname* (local refinement)
Language	Optional	Primary Language (Controlled look up list) using ISO639-2	language.ISO639-2 (encoding scheme)
Thesis Date	Mandatory	Date appearing on the title page of the thesis in format: YYYY-MM (ISO8601), or YYYY is also acceptable	date.issued (refinement)
Institution item page URL	Optional	The metadata 'jump off' page for the e-thesis at the institutional archive	relation.isReferencedBy (refinement)
Citations	Optional	Citations to previously published sections of this thesis. Applies	relation.hasVersion (refinement) *or, for DSpace (for*

		particularly to "thesis by publication". Where possible, citation information entered should conform to a recognised citation standard	*historical reasons)* identifier.citation
Included/Quoted work	Optional	References to other works	relation.references (refinement)
Rights	Optional	e.g. Copyright/IPR statement regarding rights management, or URI of Creative Commons license. Possibility of a change in rights agreement after a specified time.	Rights
Embargo type	Optional	Valid values are: • None • Partial • Complete • Cannot supply	*Rights.embargotype* (local refinement)
Embargo date	Optional	Date before which a thesis may not be released	*Rights.embargodate* (local refinement)
Embargo reasons	Optional	Reasons that a thesis is embargoed	*Rights.embargoreason* (local refinement)
Identifier	Optional	ID for electronic object(s)	identifier.URI (encoding scheme)
Identifier	Optional	ID for physical object(s)	identifier

The following fields were included in earlier versions of the core metadata set but their use is now deprecated:

Field Name	Comments	qDC Element.Qualifier
File Format	File type for preservation information (For each of these there may be multiple files)	format.IMT (encoding scheme)
File Size	Size of file for preservation information and integrity checking (For each of these there may be multiple files)	format.extent (refinement)
Checksum	Checksum of file for preservation information and integrity checking (For each of these there may be multiple files). MD5 and SHA1 encoding schemes supported.	*format.checksum* (local refinement)
File Version	Version of file format for preservation (For each of these there may be multiple files). Possible problem determining file version, or with consistent expression.	relation.requires (refinement)

UKETD_DC Application Profile

The uketd_dc application profile consists of:

- standard DC elements and recommended qualifiers
- uketd_dc 'namespace' (domain specific) extensions – 'uketdterms'

EThOS OAI uketd_dc XML schema

In order to use uketd_dc for OAI harvesting – and other methods of metadata transfer – it is necessary to implement it using an XML schema that defines the uketd_dc records format. The schema we have developed conforms to the Guidelines for implementing Dublin Core in XML. It provides a qualified DC application, supporting all DCMI terms and the following EThOS additions:

- Sets the container element to 'uketd_dc:uketddc'
- Adds the following elements:
  - uketdterms:advisor
  - uketdterms:sponsor
  - uketdterms:grantnumber
  - uketdterms:institution
  - uketdterms:department
  - uketdterms:commercial
  - uketdterms:embargotype
  - uketdterms:embargodate
  - uketdterms:embargoreason
  - uketdterms:qualificationname
  - uketdterms:qualificationlevel

EThOS XML schemas

The EThOS XMl schema definitions are as follows:

- uketd_dc.xsd
- uketddc.xsd
- uketdterms.xsd

Local copies of DCMI XML schemas

For reasons of convenience and performance, we have used local copies of the DCMI XML schemas:

- dc.xsd
- dcmitype.xsd
- dcterms.xsd

Test instance metadata

- uketd.xml - validate
- oai-uketd.xml - validate

Digital Preservation Statement for UK Theses Digitisation Project / EThOS

Introduction

This document provides a description of the digital preservation provision at the British Library for digitised and born-digital theses. It comprises an introduction to the activities, focus and capabilities of the British Library Digital Preservation Team, an overview of the preservation issues to be considered for e-theses content and a summary of the preservation strategy and techniques which will be applied to ensure their long-term preservation.

Background

British Library Digital Preservation Team

In 2005 the British Library established a dedicated Digital Preservation Team (DPT) who were tasked with ensuring the longevity of the BL's digital collections. The team has grown rapidly and now includes 14 staff (10 full time) working on digital preservation projects and internal activities. The team includes experts in digital file formats, preservation lifecycles, content management and preservation planning. The team has established strategies and procedures to preserve the BL's digital collections, and is working to embed digital preservation skills across the organisation. The British Library Digital Preservation Strategy outlines the broad approach the BL is taking.

Development and Collaboration

Digital Preservation is a cutting edge field that has little or no established best practice, and is still very much the subject of research and development. It is widely accepted that preservation processes will evolve over time, as new tools and techniques become available. The British Library is actively involved in pushing forward the boundaries of understanding and capability in the digital preservation field.

DPT is working collaboratively with other organisations on a number of digital preservation projects. Planets , Preservation and Long-term Access through Networked Services, is a four-year project co-funded by the European Union under the Sixth Framework Programme to address core digital preservation challenges. The primary goal for Planets is to build practical services and tools to help ensure long-term access to our digital cultural and scientific assets. The LIFE2 Project is a collaboration with UCL and is examining a lifecycle approach to digital preservation costing. LIFE has developed a lifecycle model and associated methodology which enhances an organisation's ability to plan, cost and evaluate the digital preservation lifecycle. The BL is also involved in collaborative projects with a number of other HE institutions, including: PRESERV2, developing preservation services, INSPECT developing techniques for assessing the significant properties of digital objects and Digital Lives which is examining the preservation of personal digital collections.

Continual enhancement of digital preservation provision

Developments within the digital preservation community will be assessed and where appropriate fed into the operational support for digital preservation applied to the digital collections held within the BL. To facilitate this, the BL Digital Preservation Team operates a Technology Watch activity which assesses new standards, techniques and tools and provides recommendations for their uptake.

Preservation Approach

Lifecycle Management

The BL follows a lifecycle approach to the management and preservation of the digital content it holds. Each content stream is managed, tracked and documented through each stage of the digital preservation lifecycle. This approach is supported by developments from the LIFE and LIFE2 Projects, described briefly above. Notes on key aspects of the lifecycle are provided below.

Creation / Acquisition

Digitised theses comprise uniform content which will present a relatively straight forward preservation challenge. With the creation of the content falling within the remit of the BL, it is possible to ensure the content is constructed and delivered for preservation using formats, standards and techniques that will facilitate subsequent preservation activities.
The BL Digitisation Approvals Process involves assessment of new digitisation activities to ensure execution and delivery to an appropriate standard. As part of this process the DPT characterises and assesses digital production content. This provides an opportunity to identify any issues with the construction of the digital objects that may be problematic in their subsequent preservation. Feedback to the production process will ensure that problems are quickly rectified and if necessary, production can be re-started.
Born-digital theses are created outside the control of the BL. Creation will be performed by a multitude of different authors who will not usually consider future preservation requirements when choosing technical standards such as file formats. This content is therefore likely to present a greater preservation challenge. If problems or issues are encountered with particular types of content it may be possible to liaise with the HE institutions where the authors are based, and provide guidance and advice on these issues. For the moment however, it will be assumed that consideration for preservation requirements will not have been made when the born-digital theses were created.

Bit-stream Preservation

Bit stream preservation for the digital content held by the BL is provided by the BL Digital Library Programme which is developing a large scale, long term digital repository. Digital content will initially be stored and backed up within the EThOS system, which is based on a development of the ePrints repository software. Ingest to the repository provided by the BL Digital Library Programme will later provide the basis for the bit-stream preservation of this content. This activity will ensure that the underlying bitstream of the digital objects is preserved.

Content Preservation

The DPT is responsible for content preservation activities and will ensure effective preservation planning for the content preservation of e-theses. This activity will ensure that the intellectual content of the preserved digital objects remains accessible and understandable over time.

Expected content profile

Digitised theses will comprise digital TIFF masters, in addition to PDF format access copies. These will be accompanied by basic metadata, as outlined in the UK Digitised Theses Proposal document.
Born-digital theses are likely to comprise a large number of relatively straight forward document formats and a smaller number (expected to grow over time) of more complex formats, possibly including differing media types (video, sound) and interactivity. Document formats may include embedded components.

Preservation Planning

Current technology support for automated preservation processes is limited but is expected to improve rapidly over time. In particular, technology outputs from the Planets Project will greatly improve the ability to characterise and assess a greater variety of file formats, perform preservation planning based on those assessments, and then to execute preservation actions based on those planning activities.
In the meantime DPT performs characterisation and planning activities on content samples in order to successfully identify and remedy preservation issues.
A range of digital preservation strategies are currently being pursued by the team, including migration and emulation. It is expected that the non-uniform and in some cases more diverse formats that will be encountered as part of the born digital theses will require a range of different preservation strategies to ensure effective preservation. In some cases it may be necessary to give individual attention to particularly problematic content.

BL Digital Preservation Team
April 2008