

EPrints Architectural Review

Document details

Project: SHERPA DP
Work Package: Kirti Bodhmagé
Version: Version 1.0
Document date: 06/09/2005
Change history:

<i>Date</i>	<i>Version</i>	<i>Author</i>
06/09/2005	First version	Kirti Bodhmagé

Contents

Summary	2
Introduction.....	2
Data Model	2
EPrints and OAI (Open Archives Initiative)	3
Export Interfaces in EPrints:	3
1. export_xml script	3
2. ArchiveOAIConfig.pm module	4
Similar Projects:.....	4
1.WEB Services in JISC CORE Project.....	4
JISC Preserv Project	4
Potential Approaches:.....	5
1. Web Service Interface.....	5
2. Exporting METS by OAI-PMH.....	5
3. Secure Copy Protocol (SCP)	5
4. Download, Upload Servlets	5
Glossary:	6
References:	6

Summary

This document investigates the data export interfaces in the EPrints software. The primary objective is to identify data migration mechanisms that may be suitable for EPrints archive-to-AHDS (Preservation Service) data transfer. The options outlined are assessed according to the functionality that EPrints is likely to offer.

Introduction

The EPrints software enables open access to the research output of scholarly and scientific research institutions. GNU EPrints (<http://www.eprints.org/software/download/>) is open source software developed at Southampton University using PERL as programming language on the UNIX platform. EPrints uses a MySQL database to store the metadata. The actual content files, e-prints are stored in the UNIX file system. The configuration files are a combination of XML and PERL. More than one EPrints archive can be served from a single installation of GNU EPrints.

Data Model

The EPrints software has several data objects within the PERL modules to represent runtime entities such as users, e-prints, sessions, subjects and the metadata fields. The data entities that are related to an e-print are:

- **Archive**
An archive is an institutional repository with its own website configuration and data. A single installation of EPrints software can have many archives running at the same time.
- **Eprint (e-print)**
It is a record in the system, which has one or more documents and some metadata. These can be more than one document for a single record to provide the same information in multiple formats, for example a single document can be available as Postscript format as well as a PDF derivative.
- **Document**
A document is a single format of an e-print like HTML, PDF, RTF or Postscript. A single document can contain more than one file, for example a web page may contain a HTML page and associated images. The files are stored in the file system.
- **Dataset**
A dataset is a collection of the same type that can be searched in an archive. Inbox dataset is the collection of e-prints on which users are still working. Archive dataset is the collection of e-prints that are available to public. Buffer dataset is the collection of e-prints, which are submitted for editorial approval. User dataset is the group of all the users of an archive.

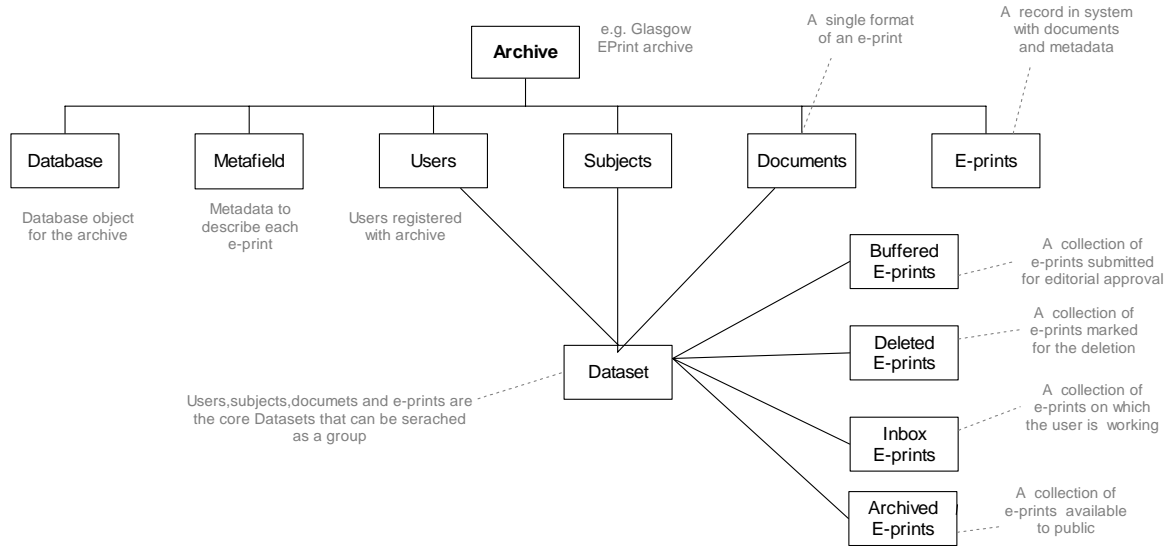


Figure 1: EPrints Data Entities

The EPrints workflow is based upon a web interface that handles the submission process. The Depositor submits an e-print into the user workspace. Papers can be uploaded as a compressed (e.g. zip format) or uncompressed file, which is transferred to a storage buffer within the EPrints software. The submitter can also specify the URL of the existing EPrints site. Once submitted an e-print will be moderated by repository staff, when it may be rejected and returned to the author for amendment or approved. Approved e-prints are moved to a public location where they are accessible by the repository's user community. Subsequent processing may be performed upon the e-print and the accompanying metadata record. This includes migration of the e-print file format, enhancement of the metadata record, or deletion from the public archive

EPrints and OAI (Open Archives Initiative)

OAI-PMH, Open Archive Initiative Protocol for Metadata Harvesting, allows search engines or harvesters to query an EPrints archive and obtain a list of items in the archive. The majority of EPrints archives are OAI-compliant and registered with an OAI search service, such as OAIster. A Dublin Core metadata record about each item is sent to the harvester. Subsequent harvests will only need to query the changes since the last harvest.

Export Interfaces in EPrints:

The EPrints software includes a built-in PERL script to import/export metadata for e-prints stored in the repository. These scripts can be extended to extract content and metadata from the repository.

1. export_xml script

The export_xml script in EPrints enables metadata to be exported for each and every e-print as a single XML document.

The syntax is :

```
export_xml <archive_id> <dataset> >somefile.xml
```

e.g.

```
data/services/eprints2/bin> ./export_xml ahds archive >Test.xml
```

where `archive_id` is the ID configured for the archive and `dataset type` is `archive`. It exports all the metadata related to all the e-prints into a file called `Test.xml`.

This script may be modified to extract metadata information for each individual e-print into a separate record.

2. ArchiveOAIConfig.pm module

This PERL module in the EPrints software configures how the archive exports its data via the OAI protocol. It uses the Dublin Core element set. Few customizations are needed in this file to change the way that an e-print is mapped into Dublin Core.

The EPrints developer has suggested that this module could be customized to retrieve metadata in the METS format. However more investigation is needed to analyze if METS can be exported via OAI-PMH.

Similar Projects:

1. WEB Services in JISC CORE Project

The CORE project aims to develop a collaborative system that will allow medical surgeons to collaborate during multi-centred trials. The majority of work focuses upon the development of a web portal to facilitate communication, through message boards, discussion lists, notification services, etc. It is unlikely that such functionality will have a use in the SHERPA DP project. However, the services they will be implementing to enable disparate communication may be repurposed. The project is developing web services to perform the following tasks:

- Input services to create and update metadata and content files like multimedia files
- Input services to get or retrieve and remove an e-print's link.
- Input service for searching the keywords associated with metadata
- Output services for a SOAP message that indicates the successful completion of a process, and metadata/multimedia attachment for retrieval.
- Output service for search

The web services implemented by this project for search and retrieval of the documents in the EPrints can be customized to fetch an e-print record from the repository. Initial contact with project staff suggests that the source code for web services will be made available in September 2005.

Project URL: <http://www.core.ecs.soton.ac.uk/>

Project Contact: Gari Wills

JISC Preserv Project

The Preserv project is currently investigating the OAI-based preservation services that may be implemented at the British Library and the Southampton University archives using EPrints software. The project staffs are currently examining additional metadata elements that may be added to the base schema (e.g. they request the depositor to rate the suitability of the paper for preservation). They also intend to link EPrints software to the National Archives' PRONOM software, in order to identify and verify file formats. There is a preliminary plan to package metadata in the METS format and export through the OAI-PMH. The actual content files will be extracted using the file URL embedded in the METS.

Project URL: <http://preserv.eprints.org>

Contact: Tim Brody

Potential Approaches:

The following approaches may be considered to extract and transfer relevant metadata and contents to the AHDS preservation system.

1. Web Service Interface

The `export_xml` utility discussed above may be modified to extract metadata for each e-print in the repository. Additional functionality is required to transform these metadata records into METS format, either at the institutional repository or the preservation server.

METS has a file-location parameter, which is a web URL to the content files in the EPrints repository. The preservation server can use this parameter to fetch the content files.

Alternatively metadata or METS and the content files can be package as a Zip file and sent to the preservation server using SOAP/WSDL as a SOAP attachment.

A web interface using SOAP/WSDL can be written for the institutional repository to facilitate the data transfer. A similar web interface has been implemented by the JISC CORE project.

2. Exporting METS by OAI-PMH

The Preserv project is planning to export the metadata in METS format over OAI-PMH. The details of the implementation are not fixed.

As each metadata record has a file location in the form of URL, these URLs can be used to fetch the actual content files.

3. Secure Copy Protocol (SCP)

SCP is a secure equivalent of FTP. The `rsync` or `SCP` utility can be used to synchronise the data between an EPrints repository and the preservation system. A crontab for every night can be setup on institutional repositories, which will initiate the data transfer automatically using SCP to the preservation system. The `export_xml` Interface can be used to extract the metadata data. The content extraction tool, which will group all the documents of an e-print together, has to be implemented separately.

The AHDS is using SCP as one of the methods to transfer the collection into the repository. The technology is simple and has proven record for the data transfer. Network traffic is key to the reliability and performance of the data transfer.

4. Download, Upload Servlets

The preservation system implements a Data Upload facility that allows an institutional repository administrator to upload data to a specific directory on the preservation server.

A utility can be developed for EPrints software, which will extract the metadata and content files for an e-print and group the contents and metadata together. An institutional repository administrator runs the extraction utility, logs onto the preservation server and requests the file upload to the preservation server.

Glossary:

Archive

A term for an EPrints based repository, which stores the research papers and metadata related to it.

E-print

A record in the EPrints software for the document submitted by an author. It may consist of several computer files.

Harvest

It is a method to collect the metadata from the repositories by issuing OAI-PMH request. Harvester is a client application that issues the request to server repositories.

METS

METS is a metadata framework, which provides an XML document format for encoding metadata necessary for both management of digital library objects within a repository and exchange of such objects between repositories.

SOAP/WSDL

SOAP, Simple Object Access Protocol is a lightweight framework for exchanging XML-based information in a decentralized, distributed environment.

WSDL (Web Service Description Language) is an XML-based language for describing network services, which communicates using SOAP.

Servlet

An interactive web interface implemented using JAVA technology that receives requests and generates a response based on the request.

FTP/SCP

File Transfer Protocol and Secured Copy Protocol are the utilities that enable the file transfer over the net.

Rsync

The rsync remote-update protocol allows transferring just the differences between two sets of files across the network connection, using an efficient checksum-search.

References:

- Preserv Project: <http://preserv.eprints.org>
- EPrints software: <http://www.eprints.org/>
- JISC Core Project: <http://www.core.ecs.soton.ac.uk/>
- OAI-PMH: <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- METS: <http://www.loc.gov/standards/mets/METSOverview.v2.html>
- OAIster: <http://oaister.umdl.umich.edu/o/oaister/>