The PURE Institutional Repository: Ingestion, Storage, Preservation, Exhibition and Reporting

Bo Alroe

Atira A/S and 10 Danish Universities, Niels Jernes Vej 10, NOVI, DK-9220 Aalborg OE, Denmark
e-mail: ba@atira.dk

Abstract

Jointly developed over 5 years by Atira A/S and a number of university libraries, the commercial repository system PURE is a tool in the research administration and dissemination effort at 11 Danish and Swedish universities. Also the hospital sector, the pharmaceutical industry and other research institutions use PURE.

Keywords: meta-data models; long-term preservation; OAI-PMH; research portal; bibliometric analysis

1 Introduction

PURE is a commercial modular standard software platform for building institutional repositories at universities and other research institutes. The term CRIS [1] also applies. It supports the publication registration process in full - from ingestion and storage to publishing and reporting. PURE is a J2EE application running on most server operation systems and SQL environments and in connection with DSpace and FEDORA. This document offers a complete overview of PURE by addressing four basic repository issues: Data modeling, data ingestion, data storage and data exhibition.

2 Application Features and Architecture

2.1 Data Modeling

Apart from publications, the following content types can also be handled in PURE: Person, Organization, Project, Student Thesis, Activity, News Clip [2] and Clinical Trial. To some extent, these types are related to separate modules of the application, which are available as integral extensions to the basic module. The separate modules are: Reports, External Publications, Student Thesis, Bibliometry, News and Clinical Trials.

The terminology above is from the PURE meta-data model, which can be delivered with PURE - either as is or adapted to the individual research institution. But any meta-data model can be implemented in PURE; part of the application architecture is development frameworks that facilitate such implementations. During the last 5 years, 4 universities have specified their own meta-data model and have had it implemented.

Currently, the PURE meta-data model is the only default meta-data model available. However, a default implementation of the CERIF2006 [3] meta-data model is currently being planned. It is expected to be completed by the end of 2007. A key specification with both meta-data models mentioned is the use of relations objects and many-to-many relations between all primary content types.

2.2 Data Ingestion

PURE's user interface is browser-based. Both Firefox and Internet Explorer are supported on Windows and Mac, though only in relatively recent versions. Users are assigned roles, which define two things: First, what functionality a user have access to. The individual users interface is adapted accordingly. Next, what the user's rights are. About 10 pre-defined roles comes with PURE and custom roles can be defined and added.

Workflows in PURE allow several users to participate in different parts of the same work process - for example the process of registering a publication. Workflows allow different users to take part in a registration process by modifying, enriching, validating manually entered data.

Data from existing systems can be used in PURE by means of one or more dynamic integrations. Usually, data for Person-objects - e.g. a person's name, title, room number, employment number, direct telephone number, e-mail address, etc. - already exists in one or more systems, and using such systems as a data source for PURE will be desirable in many situations. Organizational data would be a typical example, too, originating from systems such as LDAPs or Active Directories. Integration between PURE and any number of local sources is possible. Also user authentication and Single Sign-On integration is possible.

Imports are different from dynamic integration in that they deal with historic data and are carried out only once per data set. Publication registrations from an old repository would be a good example. To facilitate import of such data, the PURE XML Archive format (PXA) was specified. PXA files are created outside of PURE and imported via an XML I/O unit. Relations between a) publications to be imported and b) organizations and persons already in PURE is possible, depending on available data.

2.3 Data Storage

Data access is encapsulated using an object/relational persistence and query service. This allows the use of most SQL-database environments with PURE. Microsoft SQL-Server, Oracle and PostgreSQL are the most used SQL-environments among current PURE-users. PURE also maintains an index to support searches.

A so-called file connector interfaces PURE with the OS file system. The connector stores full-text files to the OS file-system while still retaining relations to the relevant meta-data objects. Further, two additional connectors are available for storing to a DSpace environment and for storing to a FEDORA environment.

2.4 Data Exhibition

PURE's two web services are a Document/Literal web service, which will make XML easily available for re-styling, and an RPC/Encoded web service, which allows requisition of data from PURE as whole objects. In both cases rich libraries of methods are available. Parameters such as date-range or organization can be added to each methods. Further, PURE has a portal-framework called PUREportal, which is an internal framework for building customized websites for exhibiting data from PURE. The framework itself comes with each PURE license. Together, the two web services and the PUREportal framework are how PURE exhibits data to websites.

Two more services help exhibit data from PURE. One is an OAI-PMH data providing services, the other is a Z39.50 based service. Different formats ca be defined under OAI-PMH; Dublin Core and DDF-MXD is supplied by default. Z39.50 was added to PURE because many library systems interfaces nicely with it. SRU/SRW will be implemented as demand rises. In addition, Reference Manager exporting capabilities allows export of any data set from PURE in native RefMan format, saving double work in some cases.

Finally, the report generator in PURE is a reporting and statistics tool, that will respond to all data in the entire PURE repository. A number of standard reports are supplied and available in three categories: Lists, Analyses and Bibliometrics. To run these standard reports, only a time interval and an organization must be chosen. To that, each standard report can be customized. Finally, custom reports can be build from scratch.

Notes and References

[1] See http://www.eurocris.org

[2] Clips from media where researchers are mentioned. Such clips are usually supplied as an XML feed from a 3rd party supplier. Upon import to PURE, clips can be related to the appropriate researchers.

[3] See http://www.eurocris.org:8080/lenya/euroCRIS/live/