Migrating eprints.org data to a Fez repository

Report compiled by

Belinda Weaver

on behalf of the

eScholarshipUQ testbed

for the

Australian Partnership for Sustainable Repositories

Contents

1. Scope and Background

This document records the process of migrating eprints.org data to a Fez repository. Fez is a Web-based digital repository and workflow management system based on Fedora (http://www.fedora.info/). At the time of migration, the University of Queensland Library was using EPrints 2.2.1 [pepper] for its ePrintsUQ repository. Once we began to develop Fez, we did not upgrade to later versions of eprints.org software since we knew we would be migrating data from ePrintsUQ to the Fez-based UQ eSpace. Since this document records our experiences of migration from an earlier version of eprints.org, anyone seeking to migrate eprints.org data into a Fez repository might encounter some small differences.

Moving UQ publication data from an eprints.org repository into a Fez repository (hereafter called **UQ eSpace** (http://espace.uq.edu.au/) was part of a plan to integrate metadata (and, in some cases, full texts) about all UQ research outputs, including theses, images, multimedia and datasets, in a single repository. This tied in with the plan to identify and capture the research output of a single institution, the main task of the eScholarshipUQ testbed for the Australian Partnership for Sustainable Repositories project (http://www.apsr.edu.au/). The migration could not occur at UQ until the functionality in Fez was at least equal to that of the existing ePrintsUQ repository. Accordingly, as Fez development occurred throughout 2006, a list of eprints.org functionality not currently supported in Fez was created so that programming of such development could be planned for and implemented.

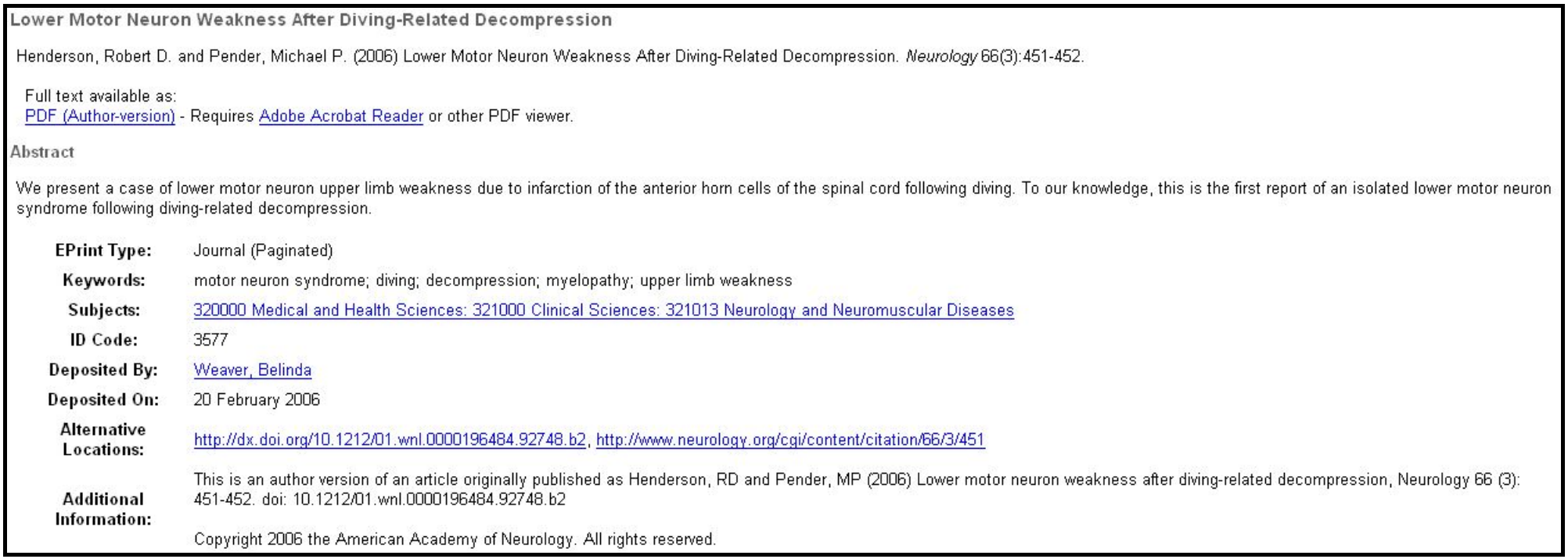
2. Pre-migration tasks

The ‘must-have’ functionality pre-migration checklist included:

- Citation view
- Linking objects and creators
- Download statistics
- ePrintsUQ usernames
- Service providers
- ePrintsUQ copyright information
- ePrintsUQ document types
- Ability to bulk move objects to new collections
- Succession control

2.1. Citation view

In ePrintsUQ, searches retrieve lists of matching results. Clicking on a search result always leads first to an **Abstract View** page, for example:



Lower Motor Neuron Weakness After Diving-Related Decompression

Henderson, Robert D. and Pender, Michael P. (2006) Lower Motor Neuron Weakness After Diving-Related Decompression. *Neurology* 66(3):451-452.

Full text available as:
PDF (Author-version) - Requires Adobe Acrobat Reader or other PDF viewer.

Abstract

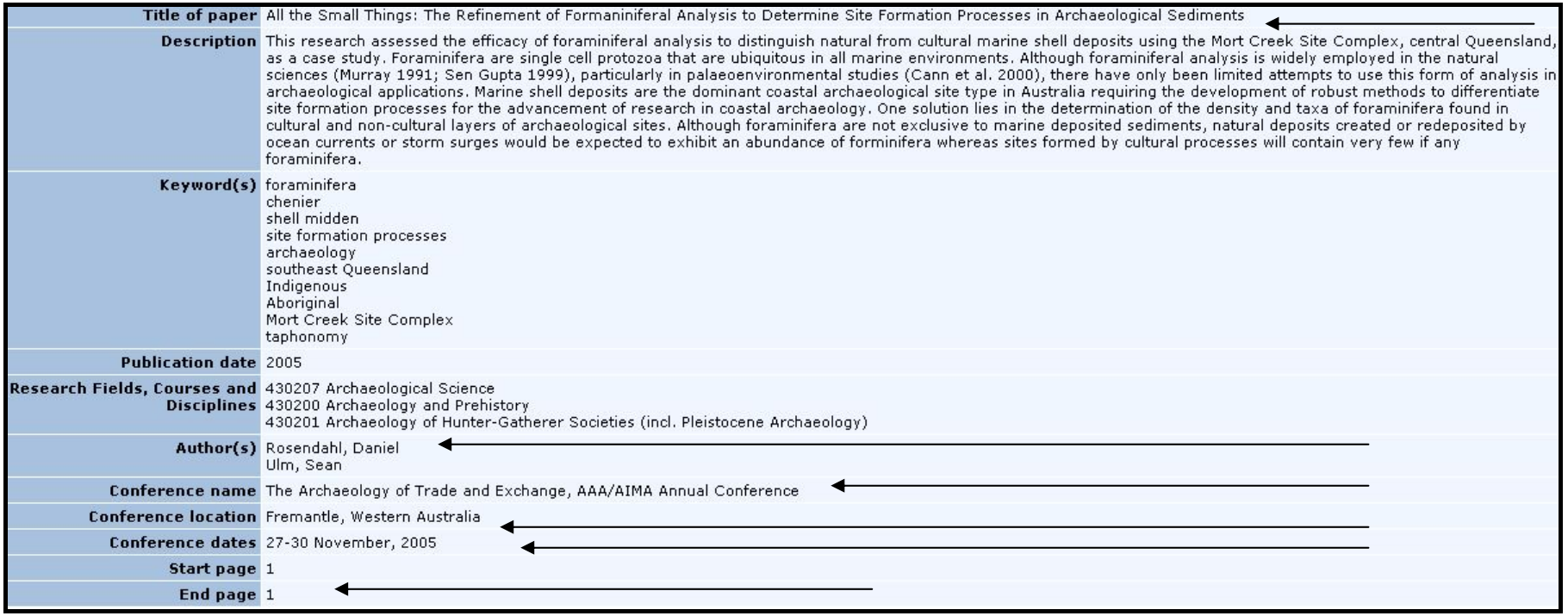
We present a case of lower motor neuron upper limb weakness due to infarction of the anterior horn cells of the spinal cord following diving. To our knowledge, this is the first report of an isolated lower motor neuron syndrome following diving-related decompression.

EPrint Type:	Journal (Paginated)
Keywords:	motor neuron syndrome; diving; decompression; myelopathy; upper limb weakness
Subjects:	320000 Medical and Health Sciences: 321000 Clinical Sciences: 321013 Neurology and Neuromuscular Diseases
ID Code:	3577
Deposited By:	Weaver, Belinda
Deposited On:	20 February 2006
Alternative Locations:	http://dx.doi.org/10.1212/01.wnl.0000196484.92748.b2, http://www.neurology.org/cgi/content/citation/66/3/451
Additional Information:	This is an author version of an article originally published as Henderson, RD and Pender, MP (2006) Lower motor neuron weakness after diving-related decompression, Neurology 66 (3): 451-452. doi: 10.1212/01.wnl.0000196484.92748.b2
	Copyright 2006 the American Academy of Neurology. All rights reserved.

The full citation is listed along the top of the **Abstract View** screen, with other selected fields appearing below. In addition to the full citation, ePrintsUQ generally displayed

- Abstract
- ePrint Type, e.g. journal article, conference paper
- Keywords
- Subjects (ASRC/RFCD codes)
- ID Code (unique object identifier within the repository)
- Deposited By (personal name)
- Deposited On (date – which became the record creation date)
- Alternative locations, e.g. DOIs, personal web sites
- Additional information – we used this field within ePrintsUQ to record publisher-required copyright statements and other rights related data.

To facilitate proper citation and acknowledgment of the work, we believed that displaying the full citation was an important task for all bibliographic records. Pre-migration, UQ eSpace had a database view of all the fields in the record (even including empty fields), but that view did not include a full citation. People would have had to construct the citation themselves from the displayed fields (see arrows).



Title of paper	All the Small Things: The Refinement of Formaniniferal Analysis to Determine Site Formation Processes in Archaeological Sediments
Description	This research assessed the efficacy of foraminiferal analysis to distinguish natural from cultural marine shell deposits using the Mort Creek Site Complex, central Queensland, as a case study. Foraminifera are single cell protozoa that are ubiquitous in all marine environments. Although foraminiferal analysis is widely employed in the natural sciences (Murray 1991; Sen Gupta 1999), particularly in palaeoenvironmental studies (Cann et al. 2000), there have only been limited attempts to use this form of analysis in archaeological applications. Marine shell deposits are the dominant coastal archaeological site type in Australia requiring the development of robust methods to differentiate site formation processes for the advancement of research in coastal archaeology. One solution lies in the determination of the density and taxa of foraminifera found in cultural and non-cultural layers of archaeological sites. Although foraminifera are not exclusive to marine deposited sediments, natural deposits created or redeposited by ocean currents or storm surges would be expected to exhibit an abundance of forminifera whereas sites formed by cultural processes will contain very few if any foraminifera.
Keyword(s)	foraminifera chenier shell midden site formation processes archaeology southeast Queensland Indigenous Aboriginal Mort Creek Site Complex taphonomy
Publication date	2005
Research Fields, Courses and Disciplines	430207 Archaeological Science 430200 Archaeology and Prehistory 430201 Archaeology of Hunter-Gatherer Societies (incl. Pleistocene Archaeology)
Author(s)	Rosendahl, Daniel Ulm, Sean
Conference name	The Archaeology of Trade and Exchange, AAA/AIMA Annual Conference
Conference location	Fremantle, Western Australia
Conference dates	27-30 November, 2005
Start page	1
End page	1

Fez accordingly had to be adapted to generate this citation view before migration could occur. UQ eSpace records now have headers which give the full citation. The citation view can be 'tweaked' to suit different citation style requirements in different collections.

Citation: Rosendahl, Daniel and Ulm, Sean (2005) All the Small Things: The Refinement of Formaniniferal Analysis to Determine Site Formation Processes in Archaeological Sediments. In *The Archaeology of Trade and Exchange, AAA/AIMA Annual Conference*, 27-30 November, 2005, pages 1 - 1, Fremantle, Western Australia.

2.2. Linking objects and creators

Items created in ePrintsUQ 'belonged' to their creator, even if the creator was not the submitting author. Accordingly, a creator's name would be associated with a set of records. Pre-migration, Fez had no such facility. Now creator is listed, along with the date of creation (actually date of

deposit/creation of record), at the base of each record. (The date of *publication* is part of the bibliographic data.) For records migrated into UQ eSpace from ePrintsUQ, the creator's name was also brought across and displays in exactly the same way as those items directly created first in UQ eSpace. The record creation date –an important feature for ePrintsUQ because of its use in establishing priority and protecting intellectual property – was also transferred.

Access Statistics:	8 Abstract Views, 0 File Downloads Detailed Statistics
Created:	Wed, 07 Jul 2004, 00:00:00 EST by Belinda Weaver Detailed History

2.3. Download statistics

One of the most popular features of ePrintsUQ was the provision of download statistic for authors' works. Statistics provided included the following counts

- Top 50 authors
- Top 50 papers
- Statistics by author name (these gave the download counts, per individual paper, for every item an author had deposited in ePrintsUQ. The display provided two separate tallies – counts of the Abstract View, and counts of the full download of any attached full text files).

Given that author names are often entered into the database in varying ways, (e.g. Pender, M. P., Pender, Michael P, and so on,) download statistics were aggregated by means of the AuthorID field in ePrintsUQ. AuthorIDs were added to all authors in records when records were vetted prior to being moved from the submission buffer into the open ePrintsUQ archive.

Given that this feature was so popular, ePrintsUQ material could not be migrated until Fez statistics functionality was at least as good as that in ePrintsUQ.

Fez developers integrated into the Fez software the statistics work Arthur Sale had developed for the ePrint repository at the University of Tasmania. These statistics gave us top 50 authors/papers counts as well as some additional functionality such as country display (which was not available in ePrintsUQ). While authors in UQ eSpace do not have the dedicated look up page of 'By Author' statistics that ePrintsUQ offered, if they wish to see the counts for specific items, these counts (abstract view/full download) are displayed as part of an individual item's

display. If demand is sufficient, we may offer a look-up facility to aggregate all the statistics for a specific author as in ePrintsUQ, but for now, we are happy that we can at least supply the download-per-item data.

Some additional functionality in Fez/UQ eSpace was added to facilitate the AuthorID-style single instances of author names. The author lookup table is tied into the University of Queensland's staff database so that a single version of an author name can be added to UQ eSpace records. Records migrated from ePrintsUQ generally brought Author IDs successfully across, but some records required editing as the ID had been assigned to the wrong author. Institutions that have not used AuthorID in ePrints repositories would not have had that problem.

Statistics began to accumulate for all migrated material as soon as it appeared in UQ eSpace. However, the statistical counts from ePrintsUQ were not transferred across and added to these counts. They will eventually be added in as staff were anxious not to lose the quite enormous counts they had accumulated in ePrintsUQ.

2.4. ePrintsUQ usernames

No functionality had to be added into Fez/UQ eSpace to allow users to log in. In fact, the move to Fez simplified log in for us. Log in to UQ eSpace is via the University of Queensland's LDAP system (which includes students as well as staff). Any UQ staff or student can log in via the LDAP, and since log ins bring user attributes with them, these log ins are thus tied to the rights users have to create or edit records within specific communities and collections. The ePrintsUQ log in was by ePrintsUQ username and password, which, in some cases, was quite distinct from that of the LDAP since users registering with ePrintsUQ could create whatever username they wanted. There was a small portion of ePrintsUQ users who created usernames and passwords quite different from their LDAP logins. These user IDs needed to be edited in UQ eSpace to match the LDAP log in structure. Where an ePrintsUQ user was no longer at UQ, e.g. a completed postgraduate student, a newly created UQ eSpace username and password was assigned to that user. (Fez includes the functionality to add users and groups in addition to LDAP or other look up systems.) Around 200 user IDs had to be edited/upgraded post-migration. This was simply a matter of editing the username and password details of the migrated user records and then, in some cases, deleting the old ones.

2.5. Service providers
ePrintsUQ was an OAI-PMH-compatible repository. Accordingly, this functionality had to be programmed into Fez before migration. It was also important to create a service provider for Picture Australia which needs to harvest image collections deposited in UQ eSpace. Both these providers are now available in Fez/UQ eSpace.

2.6. ePrints copyright information
A lot of copyright information was gathered in ePrintsUQ to comply with publisher policies on uploading author versions or other full text files. It was held in a Notes field in ePrintsUQ (the Additional Information field). This information was transferred into a Notes field in the Fez record and displays accordingly. It is not imperative in UQ eSpace that a file be attached to any citation. Accordingly, many citations where the copyright situation is unclear can be published as 'citation only' entries until the copyright situation is resolved. It is envisaged that copyright information such as publisher copyright statement requirements be included in Fez as a look up table with data able to be selected and added automatically from such a table.

2.7. ePrints document types
It was important that UQ eSpace have a content model for all migrating content. This was not an issue since Fez had more content models than ePrintsUQ.

2.8. Ability to bulk associate (copy/move) objects with other collections
This was programmed into Fez to facilitate the movement of a single collection into many different collections. See below.

2.9. Succession control
The eprints.org software had the facility to link different versions of a document by means of document IDs (the eprints.org 'succession' feature). When a linked item in ePrintsUQ was displayed in Abstract View, any information about other versions in the database, including links to those versions, was provided. This feature was not available in Fez before migration, and was a necessary prerequisite of the migration since it was valuable information that could not be lost. The issue was solved by using an element, 'IsDerivationOf', in the Fedora RELS-EXT. RELS-EXT has around twenty elements of which, at the time, Fez was using only two (now three). The process worked as follows: the ePrintsUQ collection was ingested into Fez/Fedora as a single collection. Within Fez, a look-up table linked the newly created Fez/Fedora PID to the old ePrintsUQ ID. The ePrintsUQ MySQL database was then queried to identify the ePrintsUQ ID of any items with links to other versions. Where items did have versions, the Fez look-up table identified the Fedora PID of any alternative versions. Using the RELS-EXT element

'IsDerivationOf', we were then able to link items together within UQ eSpace. Accordingly, when an item displays, links to alternative versions are provided, and the PID of the linked item appears in the 'Succeeds' field of the UQ eSpace record.

While Fedora 'version controls' all repository content, management of this is not accessible in the current Fez software release (it will be included in later Fez software releases). The 'version control' currently available in Fez/UQ eSpace simply mimics the 'succession' feature of ePrintsUQ so that succession information emanating from ePrintsUQ was not lost on migration. Succession information can also be added to new records created directly within the UQ eSpace system by means of entries in the 'Succeeds' field in the UQ eSpace record.

3. ePrintsUQ functionality not yet offered in Fez/UQ eSpace

3.1. Cloning

The ability to 'clone' records in ePrintsUQ was a popular feature. This feature is still not yet available in UQ eSpace but will be included in future development plans. If a user now wants to copy a record, this has to be done via cut and paste rather than through a simple 'clone' function.

3.2. Saving data

The ability to 'save' work entered into a record before submission was a popular feature in ePrintsUQ. A user could part-populate an item and return later to finish the record creation before submitting it to an editor for final vetting and approval.. This feature is not currently available in UQ eSpace. If a user does not 'submit' a record in which data has been entered, the record will eventually 'time out' and all data will be lost. However, as workflows in UQ eSpace are customisable, it is envisaged that a new workflow of Save will be made available for users who need this function.

3.3. Search functionality

ePrintsUQ offered three searches :

- ***Basic*** or default search – searched the title, keyword and abstract fields (and *only* those fields)
- ***Quick*** – searched the author, title, keyword and abstract fields.
- ***Advanced*** – searched any or all of a combination of all the fields in the record. You could also limit by publication type, status (published/unpublished/in press), refereed/non-refereed, and you could limit to a specific year or a range of years.

UQ eSpace will eventually offer two searches – basic and advanced.

UQ eSpace Basic search looks for matches in the author, title, keyword and abstract fields (an amalgamation of ePrintsUQ's basic and quick searches). The choice of fields can be customised via Fez's Manage Search Keys function.

UQ eSpace Advanced search offers the ePrintsUQ-style advanced search where you can specify data to be sought in specific fields, e.g. author, keyword, and where you can combine fields to make a narrower search. Advanced search is currently switched off in UQ eSpace, as certain functions were not working properly, but it will be restored in a couple of weeks once the full programming for the search facility is finished. Basic search is working well, but the default search is a Boolean OR rather than the more commonly used and expected Boolean AND. This will be changed. The choice of fields to be included in Advanced Search can also be customised to suit local needs.

ePrintsUQ offered the following **Browse** views - by subject, year, author and latest additions. UQ eSpace does the same.

4. Extra functionality in Fez from that offered in ePrintsUQ

Fez did offer some additional functionality that has been used successfully post-migration. These additions are as follows:

4.1. Bulk move

ePrintsUQ was a single collection. The emphasis was on an individual's entries rather than on the output of the university as a whole or that of a particular school or centre. UQ eSpace is based on a number of collections (and has no upper limit). This structure mirrors that of the university, based as it is on schools, centres and institutes. Material migrated from ePrintsUQ had to be assigned to specific collections since create and edit rights on objects are tied to the roles specific user log ins have. These are, in turn, governed by the attributes (including their organisational unit) people bring with them when they log in to the system via the LDAP.

Before ePrintsUQ material could be reassigned after migration, new communities and collections had to be set up to reflect the university's structure. Security also had to be set on all such collections.

The creation of a bulk move workflow greatly facilitated the reassignment of materials to these new collections after the migration. Records could be marked and then moved en masse to their new homes. Where items were moved to the wrong collection in error, these could be easily reassigned by another bulk move. The bulk move workflow runs in the background so did not preclude other work while items were moving.

Migrating an eprints collection to a Fez repository does not necessarily involve re-allocating the migrated materials to new collections. While the Fez community/collections structure facilitates object security inheritance and control, the choice of whether to split up the eprints.org collection or leave it as a single entity would need to be an institutional decision.

At UQ, in addition to the facilitation of object security, we decided to use a community structure based on university organisational units so that academics would feel a sense of 'ownership' about their collection. In time, we plan to personalise collections with icons and style sheets that mirror what currently appears on the Web pages of specific organisational units throughout the University.

4.2. New roles

ePrintsUQ had only a single role for most users – that of creator. Editorial control was managed centrally, though editor roles could have been rolled out more widely if required. However in UQ eSpace there are a number of roles that a user can have –

- Lister (can list objects, collections and communities)
- Viewer (can list and view objects, collections and communities)
- Creator (can list, view and create objects in collections and communities)
- Editor (can list, view, create and edit objects in collections and communities)
- Approver (can list, view, create, edit and approve, i.e. publish objects in collections and communities)
- Community administrator (can list, view, create, edit, approve, (i.e. publish) and delete records within the community and its collections)

This gives us much greater flexibility for collection management. Different collections can set different levels of security, and changes to security are very easily managed.

4.3. New record types

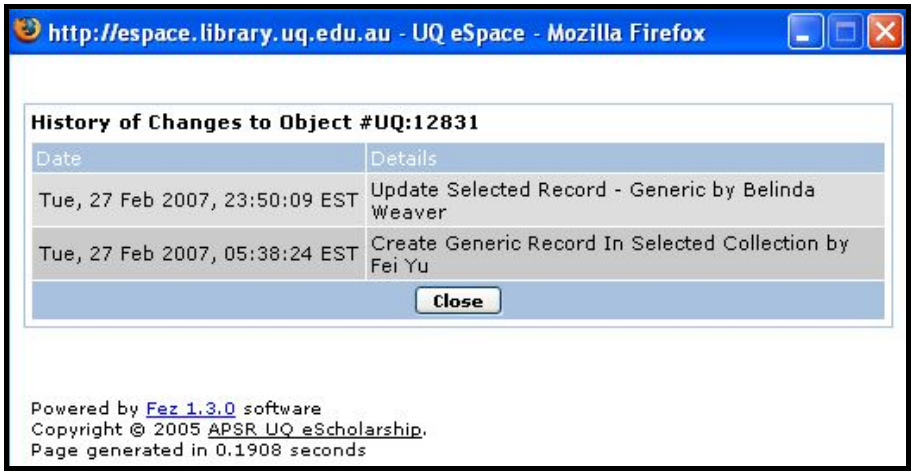
We have added new content models such as image, audio and generic document to the list of content models imported from ePrintsUQ. We are looking to add more, such as a content model specifically for video. We have customised the content models we already have in open collections to include extra reporting data for research assessment trial collections. The point here is that Fez can have as many content models as an institution requires.

4.4. Preservation metadata and object history logging

One of Fez's strengths is its support for the PREMIS model. Fez provides audit trail data for Actors and Events. Objects in repositories will only remain usable over the long term if information about the item and its format is stored with the object, and if a system for warning of format obsolescence is also built in. Fez offers object history logging for all deposited items. Any change to an object, such as a create, edit or delete, is fully recorded (thus providing the audit trail). The Actor in the Event is also recorded as part of object history logging so that all 'events' can be traced back to the individuals carrying them out. Preservation metadata (in XML format) is automatically generated for all objects on ingest. Fez

has also integrated the work of the Automatic Obsolescence Notification System (AONS) although this has not yet been released with a Fez distribution, pending further testing, documentation and packaging.

Preservation metadata in Fez also allows objects to contain more dates. Apart from the bibliographic publication date, ePrintsUQ had only one date – the date the item was deposited (more often called the creation date). With object history logging, UQ eSpace can report more than just the date of creation. It can list the dates of all changes to a record.



4.5. Internal search links

UQ eSpace offers more inward linking of items. Within ePrintsUQ, users could link to other items by means of subject terms or author names. Within UQ eSpace, they can link via author and subject but also by other fields such as keywords.

5. Post-migration tasks

5.1 Bulk move

This task, of breaking up the monolithic eprints collection and moving items into specific collections, took one staff member at a remote location almost two full days to complete. The items moved numbered around 3,700. The process was somewhat slow given that there was very

little ‘provenance’ information within most ePrintsUQ records. Staff lists had to be consulted, and in some cases, items were moved based on RFCD codes. No problems have yet been reported by academics of items being in the wrong place.

5.2 Editing of user records

Editing of former ePrintsUQ user records had to be done to make these usernames comply with the LDAP system. The number to be edited was around 200, and editing was minimal, usually just involving a change of password.

5.3 Editing of some keyword fields in records

In ePrintsUQ, keywords and keyword phrases were entered in a single box string separated by semi-colons. Fez provides boxes for the entry of keywords and keyword phrases. Where ePrintsUQ keyword phrases included a comma (e.g. names in reverse order such as Lawson, Henry), UQ eSpace read the two names as separate keywords and placed each name in a separate box. When these records are found, they are edited, but since all the words are searchable in any case, this division of names or phrases is not seen as a major problem.

5.4 Problems encountered

Several migration trials were done before the full migration occurred. In total, six test migrations were done before the final successful migration in December 2006. Each test migration revealed problems that had to be fixed.

In the initial migration test, the migration failed after ingesting only one-tenth of the records. For that initial test, we tried to use the eprints.org export script – the export command line tool `export_xml` perl script which will generate an XML file of the contents of an ePrints “archive”. However, in addition to falling over, the export file lacked important information. Only the DC core fields were transferring across to UQ eSpace – other data and, more importantly, the attached full text files – were not transferring. The XML file contains metadata for each record in an eprints.org repository but does not contain the file attachments or URLs to the full text files (e.g. the PDFs, PostScript or HMTL files). These file links are however available in the eprints.org OAI-PMH service provider.

We ended up directly querying the ePrints MySQL database from Fez to import objects into Fez. Institutions using later versions of eprints.org software might not encounter these problems, or might seek to overcome them, if they do encounter them, in a different way.

Our process worked as follows. When Fez finds an ePrints XML export file during batch import of a selected directory, it will create a Fez+Fedora object for each ePrints record and do an OAI-PMH `getRecord` look-up to the ePrints server to get the URL links for each file attachment, download them, and add them to the new Fez object. Fez will also read the document type of each ePrints record and match those against Fez document types automatically. If no document type match is found, an item is created of the type “Generic Document”.

More detail on setting up the Fez **config.inc.php** for an eprints.org repository batch import is included in ePrints section of the Fez Wiki (http://dev-repo.library.uq.edu.au/wiki/index.php/EPrints).

Finally, the migration trials highlighted the issue of faulty ePrintsUQ records. The trials highlighted some filename issues that had gone undetected in ePrintsUQ. Some of the links to PDFs in ePrintsUQ were not functioning. In two cases, users had uploaded files that included the # character in a filename. ePrintsUQ refused to open these files. In other problem cases, filenames that included spaces in the file names did not transfer successfully to Fez which cannot accept filenames with spaces in filenames. Before the final migration, all problem filenames were edited directly within the ePrintsUQ MySQL database and they then transferred correctly.

5.5 Acknowledgements

Thanks to the eScholarshipUQ team at the University of Queensland Library – Christiaan Kortekaas, Andrew Bennett, Matthew Smith, and Lachlan Kuhn – for their assistance in compiling this document. Thanks also to Kingsley Gurney for his assistance in managing ePrintsUQ since 2002.

Belinda Weaver

6 March, 2007