

# AANRO Research System to FEZ

## 1 About this document

### Author

Peter Sefton, Tim McCallum & Bron Dye

### Purpose

This document follows the steps to ingest a set of AANRO records into FEZ.

### Audience

AANRO repository implementers, anyone else looking for sample code to add data to Fez

### Requirements

AANRO data

Python 2.4 installed on the system to run the harvest, with the Cheetah templating system installed.

Fedora 2.2 installed and running.

Optionally, an instance of FEZ for ingest.

### References

Fez wiki:

[http://dev-repo.library.uq.edu.au/wiki/index.php/Main\\_Page](http://dev-repo.library.uq.edu.au/wiki/index.php/Main_Page)

Documentation on the FOXML (Fedora Object XML) specification:

<http://www.fedora.info/download/2.1.1/userdocs/digitalobjects/introFOXML.html>

Official Python website:

<http://www.python.org/>

Official py.test tool and library website:

<http://codespeak.net/py/current/doc/test.html>

Official Subversion website:

<http://subversion.tigris.org/>

### Notes

The scripts have been developed on OS X and Linux based system. Python is a cross platform programming language and therefore the scripts should also run under Microsoft Windows operating systems. but we have not tried.

Installing the Python programming language, Fedora and Cheetah is outside the scope of this technical report.

## 2 Background Information

A component of the work undertaken at RUBRIC-Central is the development of various data migration strategies. These strategies are designed to assist RUBRIC Project Partners to migrate data into, and out of, various systems. The data migrations specifically target the three institutional repository solutions under consideration as part of the project: Dspace, Fez and VITAL.

This document covers a script to take a propriety database dump of the AANRO data and it is off little direct use to others, but may serve as an example.

This data migration differs from most of the others we have published at RUBRIC so far. in that is not modular, does not use the DSpace archive format is an intermediate stage, and makes no use of XSLT. For performance reasons we wrote this a stand-alone script that goes straight from the data files to FOXML records for ingest into Fedora.

There are a few reasons for the change all to do with performance:

Linux only allows 32000 files in one directory so our simple Dspace archive class would have needed to be rewritten.

1. The libxslt library for Python has memory leaks so the script fails on large data sets.
2. A multi-step process is ideal for small (up to ten thousand or so records but is unmanageable when dealing with two hundred thousand.
3. We wanted to try using Cheetah templates as the basis for data migration as they seemed to offer an easy-to-use alternative to XSLT.

### 3 The Python Scripts

The data migration uses a single script to process the text-based data dump into FOXML records you can ingest into Fedora.

#### **aanro\_dump\_to\_foxml.py**

The main script. This takes .dmp files from the AANRO database.

This script is accompanied by a minimalist py.test test-script: we recommend adding new test cases before you change the script.

The result is an output directory which contains FOXML to allow it to work with large numbers of files the output is broken into a number of directories (0..n) with up to 10,000 records in each.

NOTE: The script can currently deal with the AANRO publications data, publications archive, and the research archive but not the research data it uses a different format which will require more development.

#### **aanro\_foxml\_template.tmpl, aanroFoxmlTemplate\_research.tmpl**

Cheetah templates to transform AANRO data to FOXML, including both MODS and Dublin Core versions of the data. They deal with publication and research project respectively.

## 4 Download the Files

All of the data migration scripts, and associated code libraries, modules and files, are made available via a publicly accessible website.

If you have the subversion client installed you can download the Python scripts, test files, and other files used during development. The URL that you will need to check out is as follows:

Type this:

```
svn co https://rubric-  
central.usq.edu.au/svn/Public/code/migration_toolkit
```

## 5 Run the Python script

1. Change into the AANRO directory

```
cd AANRO
```

2. Compile the templates using cheetah

(do this again after any changes you make to the .tmpl files)

```
cheetah compile *.tmpl
```

### 5.1 Name of the script

aanro\_dump\_to\_foxml.py

### 5.2 Parameters/Arguments

#### Input (-i)

Full path to the AANRO data file to be transformed (.dmp) file

#### Output (-o)

Name of directory that transformed files will be sent to (This directory will be created if it does not exist)

#### Template (-t)

Name of template file to be used in transformation

### 5.3 Syntax

```
python aanro_dump_to_foxml.py -i Input -o Output -t Template
```

### 5.4 Example Research Data

```
python aanro_dump_to_foxml.py -i research.dmp -o output_directory  
-t aanroFoxmlTemplate_research
```

### 5.5 Example Publication Data

```
python aanro_dump_to_foxml.py -i publication.dmp -o  
output_directory -t aanroFoxmlTemplate
```

## 6 Ingest the data into Fedora

```
sudo /usr/local/fedora-2.2/client/bin/fedora-ingest.sh d  
[output_dir] foxml1.0 O localhost:8080 fedoraAdmin fedoraAdmin  
http ""
```

<http://www.rubric.edu.au/>

Follow the latest instructions for Fez to index the resulting data into the Fez indexes.