

Voyager To VITAL

Table of Contents

1 Introduction.....	1
2 Background information.....	3
3 Downloading the Python Scripts.....	4
3.1 Script dependencies	4
4 Converting MARC to MARC21.....	5
5 Split XML Into Archive.....	6
5.1 Name of the script.....	6
5.2 Location of script.....	6
5.3 Purpose of script.....	6
5.4 Parameters/Arguments.....	6
5.5 Syntax.....	7
5.6 Example.....	7
6 Clean MARC metadata.....	8
6.1 Name of the script.....	8
6.2 Location of script.....	8
6.3 Purpose of script.....	8
6.4 Parameters/Arguments.....	8
6.5 Syntax.....	8
6.6 Example.....	8
7 Obtain files for archive.....	9
7.1 Name of the script.....	9
7.2 Location of script.....	9
7.3 Purpose of script.....	9
7.4 Parameters/Arguments.....	9
7.5 Syntax.....	9
7.6 Example.....	10
7.6.1 Using Authorization.....	10
7.6.2 Without Authorization.....	10
8 Wrap namespace around file.....	11
8.1 Name of the script.....	11
8.2 Location of script.....	11
8.3 Purpose of script.....	11
8.4 Parameters/Arguments.....	11
8.5 Syntax.....	11
9 PDF to full text.....	12
9.1 Name of the script.....	12
9.2 Location of script.....	12

9.3 Purpose of script.....	12
9.4 Parameters/Arguments.....	12
9.5 Syntax.....	12
9.6 Example.....	12
10 Remove XML Node.....	13
10.1 Name of the script.....	13
10.2 Location of script.....	13
10.3 Purpose of script.....	13
10.4 Parameters/Arguments.....	13
10.5 Syntax.....	13
10.6 Example.....	13
11 XSL Transform (marc.xml to dublin_core.xml).....	14
11.1 Name of the script.....	14
11.2 Location of script.....	14
11.3 Purpose of script.....	14
11.4 Parameters/Arguments.....	14
11.5 Syntax.....	14
11.6 Example.....	15
12 Archive To FOXML.....	16
12.1 Name of the script.....	16
12.2 Location of script.....	16
12.3 Purpose of script.....	16
12.4 Additional Information.....	16
12.5 Parameters/Arguments.....	17
12.6 Syntax.....	17
12.7 Example using external web server to host non XML data streams.....	18
12.8 Example using Python simple server to host non XML data streams.....	18
13 Insert XMLNS XSI.....	19
13.1 Name of the script.....	19
13.2 Location of script.....	19
13.3 Purpose of script.....	19
13.4 Parameters/Arguments.....	19
13.5 Syntax.....	19
13.6 Example.....	19
14 Preparing FOXML data for Fedora Ingest.....	20
15 Preparing non XML data for Fedora ingest.....	20
15.1 Using the Python Web Server	20
15.2 Using an Existing Web Server.....	20
16 Fedora Ingest for VITAL 3.....	21
16.1 Ingest procedure.....	21

1 Introduction

Author

Bron Dye, RUBRIC Technical Officer

Tim McCallum, RUBRIC Technical Officer

Purpose

This technical report outlines how Voyager library system XML can be migrated for ingest into VITAL/Fedora as FOXML objects.

Audience

RUBRIC Project Partners and other users of Fedora repositories

Requirements

XML file (output from Voyager library system)

Python 2.4 installed on the system to run the harvest

The libxml2 library, including the Python bindings, installed on the system to run the harvest

An instance of VITAL for ingest

A working installation of marc-edit 5.0

References

Official VITAL website at VTLS

<http://www.vtls.com/Products/vital.shtml>

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 libxml2 website

<http://xmlsoft.org/python.html>

Official py.test tool and library website

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

Official Subversion website

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

Marc-edit site .exe file

http://oregonstate.edu/~reaset/marcedit/software/development/MarcEdit50_Setup.exe

Notes

The scripts have been developed on a Linux based system. Python is a cross platform programming language and therefore the scripts should also run under Microsoft Windows, and the OSX operating systems. However this has not been tested.

The installation of the Python programming language and the libxml2 library, including Python bindings is outside the scope of this technical report. Many Linux distributions,

such as Ubuntu, will have these already installed.

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.

Interest was expressed in being able to migrate exported files produced from local library systems into other repositories, such as VITAL or DSpace. This technical report, and the associated Python scripts, comprise the strategy to achieve this goal.

The Python scripts create an archive or directory, similar in structure to a DSpace Archive. Within this directory are created item directories each with a temporary xml file storing dublin core metadata, all files relevant to the xml item and a file listing all relevant files attached to the exported item.

The Python scripts have been developed using a unit testing approach using the testing framework provided by the `py.test` tool and library. More information about the library is available at the website listed in the references section of this technical report. These scripts have been developed using Python version 2.4, and may work with earlier versions. However this has not been tested.

The Python scripts are modular in nature and use functionality provided by modules that have been used in other migration strategies. It is anticipated that this type of architecture will allow modification and customisation as required.

3 Downloading the Python Scripts

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:

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

3.1 Script dependencies

The structure of the toolkit is important; file dependencies are relative to the scripts used. These dependencies include

dspace_archive directory

A directory containing Python modules that provide utilities for the creation of dspace archive objects.

foxml_class directory

A directory containing Python modules that provide utilities for the creation of FOXML objects.

utils directory

A directory containing Python modules that provide general utilities almost all scripts within the migration toolkit.

4 Converting MARC to MARC21

The following sections of this technical report outline the procedure converting the Voyager output file to MARC XML and then using the Python scripts to extract the XML objects from a larger XML file.

It is assumed that you have Python and the Libxml2 library, including the Python bindings, already installed.

1. Download and open marc edit.
2. Select marc maker
3. Set the input file as the Voyager file
4. Set the output file
5. Click on marc to marc21 in the marc functions
6. Click execute to create the MARC XML
7. Open MARC XML in a text editor.
8. Remove the marc namespaces from it, Replace `<marc:` with `<` and replace `</marc:` with `</`

5 Split XML Into Archive

5.1 Name of the script

split_xml_into_archive.py

5.2 Location of script

migration_toolkit/split_xml_into_archive.py

5.3 Purpose of script

To convert a large xml file (containing multiple records) into a dspace archive format. The dspace archive format consists of one top level directory that houses multiple subdirectories. Each subdirectory represents one record from the large xml file. After this script has been executed you will find two files with in each subdirectory. A temp XML File containing metadata for one individual record (extracted from the large xml file) and a basic text file called contents. This file lists the contents of the subdirectory in which it is contained.

5.4 Parameters/Arguments

dataFileName

Path to the large input xml file that will be split.

archiveName

Name of the dspaceArchive that will be created (this will create the dspaceArchive in current directory)

OR

Full path to location where you would prefer the dspaceArchive to be created, must specify name of the dspaceArchive after full path when using this option (directory does not have to exist it will be created when script is run)

xpath

An xpath to match the root node for each individual record in the large xml file. For example `//*[local-name()='record']`

tempXmlFileName

Filename for the temp XML file within each subdirectory containing metadata for each record.

templateFile

The filepath to a wrapper file if needed. This parameter can be used if it is necessary to wrap the temp XML File file located within the subdirectory with extra data, for example extra namespace.

5.5 Syntax

```
python split_xml_into_archive.py dataFileName archiveName xpath  
tempXmlFileName templateFile
```

5.6 Example

```
python split_xml_into_archive.py large_xml_file.xml dspaceArchive  
"//record" temp.xml False
```

6 Clean MARC metadata

6.1 Name of the script

clean_marc_metadata.py

6.2 Location of script

migration_toolkit/clean_marc_metadata.py

6.3 Purpose of script

The **clean_marc_metadata.py** script is used to check if date and title metadata have extra trailing characters. For example: A date field may have a trailing “.” or “]”. A title field may have a trailing “:”. The script iterates through the dspaceArchive subdirectories and cleans the temp XML file contained within. This script is also used as a tool to insert additional marc elements to the marc metadata. The marc elements to be inserted are hard coded into the script, they are inserted into the marc based on the record type. See parameters/Arguments below for more info on specifying record types.

6.4 Parameters/Arguments

archiveName

Name of the archive containing the xml files to be cleaned

tempFileName

Filename of the xml file to be cleaned

recordType

The type of record Ethesis(use E) or Brunner (use B) Library Publications (use L) Working Papers (use W) The script will add marc metadata elements that are hard coded into the python script based on the type of record entered here.

6.5 Syntax

```
python clean_marc_metadata.py archiveName tempFileName  
recordType
```

6.6 Example

```
python clean_marc_metadata.py dspaceArchive temp.xml E
```

7 Obtain files for archive

7.1 Name of the script

obtain_files_for_archive.py

7.2 Location of script

migration_toolkit/obtain_files_for_archive.py

7.3 Purpose of script

The obtain_files_for_archive.py script is a Python script that searches the temp.xml files looking for URL links to external pdf files. Once found, the files are then downloaded by the script and put into the dspaceArchive. Use optional protocol, username and password arguments when files to be harvested require authentication.

7.4 Parameters/Arguments

archiveName

name of the archive to be accessed

tempXmlFile

name of the file in each archive to be accessed

fileType

file extension of external file to be accessed and downloaded

protocol

protocol used to access file(optional)

username

username required to access file(optional)

password

password required to access file(optional)

7.5 Syntax

```
python obtain_files_for_archive.py archiveName tempXmlFile fileType fileType username password
```

7.6 Example

7.6.1 Using Authorization

```
python obtain_files_for_archive.py dspaceArchive temp.xml pdf  
http:// username password
```

7.6.2 Without Authorization

```
python obtain_files_for_archive.py dspaceArchive temp.xml pdf
```

8 Wrap namespace around file

8.1 Name of the script

wrap_namespace_around_file.py

8.2 Location of script

migration_toolkit/wrap_namespace_around_file.py

8.3 Purpose of script

Large master files are sometimes supplied for migrations, the master file generally has one namespace that wraps the entire file containing multiple records. Once the master file is split into single records (eg temp file produced during split_xml_into_archive.py) each single temp file is stored without a namespace declaration. This

wrap_namespace_around_file.py script is used to iterate through the dspaceArchive and wrap a namespace declaration around each single temp file, it then saves the result as a new file (see newFile parameter below)

8.4 Parameters/Arguments

tempXmlFile

Name of the file inside the archive that needs wrapping

wrapperFile

The full path to the wrapper file (eg wrapper_file_for_marc_file)

archiveName

Name of the archive (eg voyagerArchive)

newFile

The name of the output file ie: marc.xml

8.5 Syntax

```
python wrap_namespace_around_file.py tempXmlFile wrapperFile
archiveName newFile
```

```
python wrap_namespace_around_file.py temp.xml
wrapper_file_for_marc_file dspaceArchive marc.xml
```

9 PDF to full text

9.1 Name of the script

pdf_to_full_text.py

9.2 Location of script

migration_toolkit/pdf_to_full_text.py

9.3 Purpose of script

The **pdf_to_full_text.py** script is the Python script that iterates through a dspaceArchive and converts the harvested pdf files to **fulltext**.

9.4 Parameters/Arguments

ArchiveName

Name of the dspaceArchive that contains the pdf files

9.5 Syntax

```
python pdf_to_full_text.py ArchiveName
```

9.6 Example

```
python pdf_to_full_text.py dspaceArchive
```

10 Remove XML Node

10.1 Name of the script

remove_xml_node.py

10.2 Location of script

migration_toolkit/remove_xml_node.py

10.3 Purpose of script

The **remove_xml_node.py** script is the Python script that iterates through a dspace archive and removes unwanted nodes and their contents from a file (specified as an argument)

10.4 Parameters/Arguments

archiveName

Name of the archive to be accessed

fileName

Name of the file from which the node is to be removed

xpath

The xpath to locate the node to be removed

10.5 Syntax

```
python remove_xml_node.py archiveName fileName xpath
```

10.6 Example

```
python remove_xml_node.py dspaceArchive marc.xml //*[local-name()='datafield'][@tag='856']
```

11 XSL Transform (marc.xml to dublin_core.xml)

11.1 Name of the script

xsl_transform.py

11.2 Location of script

migration_toolkit/xsl_transform.py

11.3 Purpose of script

The **xsl_transform.py** script is a Python script that iterates through a dspace archive. It carries out an XSL transformations on one XML file per item within the archive. A new XML file is created as a result of each transformation. This file is stored along side the original XML file in the item.

11.4 Parameters/Arguments

InputFile

Filename of the XML file in the archive to be converted.

XslFilePath

File path to the stylesheet used for the XSL transformation

OutputFile

Filename of new XML file that will be generated as part of conversion process

ArchiveName

Name of the archive to be accessed

RemoveInputFile

Remove the original XML file file? - set to False

11.5 Syntax

```
python xsl_transform.py InputFile  
XslFilePath OutputFile ArchiveName RemoveInputFile
```

11.6 Example

```
python xsl_transform.py marc.xml  
xsl/marc_dc.xsl dublin_core.xml dspaceArchive False
```

12 Archive To FOXML

12.1 Name of the script

archive_to_foxml.py (if you are building FOXML for VITAL 3.X.X system)

archive_to_foxml_vital_2.py (if you are building FOXML for VITAL 2.X.X system)

12.2 Location of script

migration_toolkit/archive_to_foxml.py

migration_toolkit/archive_to_foxml_vital_2.py

12.3 Purpose of script

The **archive_to_foxml.py** script is the Python script that iterates through a dspace archive. The script builds a FOXML file for each item in the dspace archive using the MARC, Dublin Core and MODS metadata contained within each item. All FOXML files created during this process are stored in a single directory (specified as an argument)

12.4 Additional Information

It is important at this step to ascertain whether you have any non XML data streams as part of your migration.

How do I tell if I have non XML data streams?

There are a couple of ways to determine this. Firstly, the script `obtain_files_for_archive.py` is designed to fetch non XML data streams. If you have run this script then you would have non XML data streams as part of your ingest. Secondly, you can list the contents of a few items in your dspace archive and see if there are any non XML data streams present such as PDF files or full text files.

If non XML data streams exist in your archive (PDF, full text etc), you are required to provide a URL where these data streams can be served during ingest. You can serve these files on an independent web server or you can use a Python simple server on the same machine that VITAL is running on.

If you will be using the python simple server as part of your ingest into Fedora then use `http://localhost:8000` for the `URLforNonXmlDataStreams` argument below. If you will be using an existing independent web server during the Fedora ingest then enter the full URL to where the non XML data streams will be served during ingest.

If no pdf files or fulltext datastreams exist, set `URLforNonXmlDataStreams` argument to `FALSE`

12.5 Parameters/Arguments

archiveName

Full path to name of the archive to be accessed (write down the value that you put for this argument, it is needed during the ingest into VITAL/Fedora)

startNum

Starting number for the Fedora PID increment

PIDPrefix

Name of the Fedora PID

outputDirectory

Name of the directory for storing the FOXML objects

labelPrefix

Prefix to be added to title for reference. Eg Imported Item:

foxmlObjectState

Set this to Active (A), Inactive(I) or Deleted (D).

MARCFileName

Name of the MARC xml file contained in the Archive.

MARCDataStreamState

Set this to Active (A), Inactive(I) or Deleted (D).

DCFileName

Name of the Dublin Core file contained in the Archive

DCDataStreamState

Set this to Active (A), Inactive(I) or Deleted (D).

MODSFileName

Name of the Dublin Core file contained in the Archive. Set to False if MODS file is not present.

MODSDataStreamState

Set this to Active (A), Inactive(I) or Deleted (D). or False if MODS file is not present

URLforNonXmlDataStreams

Set this to the full URL where non XML data streams can be served during ingest

Note: Remove any trailing slashes (created by tab completion) from arguments before executing script

12.6 Syntax

```
python archive_to_foxml.py archiveName startNum PIDPrefix
outputDirectory labelPrefix foxmlObjectState MARCFileName
```

```
MARCDataStreamState DCFileName DCDataStreamState MODSFileName  
MODSObjectState URLforNonXmlDataStreams
```

12.7 Example using external web server to host non XML data streams

```
python archive_to_foxml.py dspaceArchive 0 vital foxml_items  
Imported_Items A marc.xml A dublin_core.xml A False False  
http://servername/directoryname
```

12.8 Example using Python simple server to host non XML data streams

```
python archive_to_foxml.py dspaceArchive 0 vital foxml_items  
Imported_Items A marc.xml A dublin_core.xml A False False  
http://localhost:8000
```

13 Insert XMLNS XSI

13.1 Name of the script

`insert_xmlns_xsi.py`

13.2 Location of script

`migration_toolkit/insert_xmlns_xsi.py`

13.3 Purpose of script

This script iterates through FOXML items in a single directory and sets the correct namespace declaration for the MARC XML section of the FOXML

13.4 Parameters/Arguments

outputDirectory

The name of the directory where the FOXML files are located

13.5 Syntax

```
python insert_xmlns_xsi.py outputDirectory
```

13.6 Example

```
python insert_xmlns_xsi.py foxml_items
```

14 Preparing FOXML data for Fedora Ingest

1. Log into the machine that VITAL/Fedora is running on as the dbadmin user.
2. Copy the directory containing all of the FOXML items created during this migration to the /home/dbadmin directory on the VITAL/Fedora server, ensure that dbadmin has ownership of this directory after it has been copied.

15 Preparing non XML data for Fedora ingest

15.1 Using the Python Web Server

To use the simple HTTP server that comes with Python follow these steps:

1. Log into the machine that VITAL/Fedora is running on
2. Start a new terminal session (or place the & after the command to make the server run in the background)
3. Copy the dspace archive directory (created during this migration) to the /home/dbadmin directory on the VITAL server. Make sure that the name of the dspace archive directory is not changed in any way during the copying process.
4. Execute the following command from within the /home/dbadmin directory

```
python -c "import SimpleHTTPServer;SimpleHTTPServer.test()"&
```

This command will invoke Python and start the SimpleHTTPServer. Allowing Fedora to fetch the items (this is done based on a URL in the FOXML. This URL was constructed during the archive_to_foxml.py script).

Please note that the SimpleHTTPServer will not be able to service requests other than those from the local machine, and therefore this process will only work when the data streams are on the same server as the VITAL repository.

15.2 Using an Existing Web Server

1. Copy the dspace archive directory (created during this migration) to the directory specified as the URLForNonXmlDataStream. This was an argument in the archive_to_foxml.py script. Go to this directory on the webserver using a web browser and make sure that it is resolving and files are being served.

16 Fedora Ingest for VITAL 3

16.1 Ingest procedure

To ingest the items into VITAL complete the following procedure:

1. Log into the VITAL server as the dbadmin user
2. Navigate to the following directory on the server

```
/opt/vtls/vital/applications/fedora/client/bin
```

1. Ensure the FEDORA_HOME and JAVA_HOME shell variables exist. If they do not exist, sample commands are outlined below

```
export FEDORA_HOME=/opt/vtls/vital/applications/fedora
export JAVA_HOME=/opt/vtls/java
```

1. Invoke the following command to start the fedora-ingest utility, where outputDirectory is the directory containing all of the FOXML items and password is the fedoraAdmin password

Note: This may take some time to complete

```
./fedora-ingest.sh d outputDirectory foxml1.0 O localhost:8080
fedoraAdmin password http ""
```

note the single O before the work localhost in the above command is the capital letter O not zero

Further information on the Fedora ingest utilities is available at the following URL:

<http://www.fedora.info/download/2.1.1/userdocs/client/cmd-line/index.html#ingest>

Once the ingest is complete, check the XML log file, as specified by the output of the program, for any errors

If the new objects are to be made available via the VITAL portal, ensure sufficient time has elapsed to allow the VITAL indexer to become aware of the additional objects

