Packaging Orphaned Collections for Long Term Preservation

Joel Da Costa
UCT Computer Science
University of Cape Town
Private Bag X3, 7701
Rondebosch, Cape Town
+27 21 650 2663
joeldacosta@gmail.com

ABSTRACT
This report gives an introduction and background to the area and necessity of archiving online collections. The goal of the research was to develop a tool able to archive metadata and data from OAI-PMH compliant repositories effectively. After running tests it was seen that while the metadata harvesting and archiving was effective, the data harvesting needed more work due to the brute force nature of link extraction.

General Terms
Management, Documentation, Experimentation

Keywords
OAI-PMH, METS, Archive, Render, Metadata, Package, Collection.

1. INTRODUCTION
The project tackled the problem of maintaining access to orphaned collections. Over time, due to loss of interest, resources, and the like, many online collections are neglected and end up no longer available. The proposed system that was developed is a tool that makes use of common standards such as the Metadata Encoding Transmission Standard (METS), Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) and XML in order to try to create a solution to the above problem.

The expected outcome was a tool that would be able to collect and preserve contents from OAI-PMH compliant repositories and effectively maintain accessibility to the data and metadata. The tool was not planned to be an all-round solution, but rather is currently limited to PDF files. The rest of the report will cover: the background to the problem; the design of the solution implemented; Evaluation (Experiments & Results) and the Conclusions drawn from the experiments.

2. BACKGROUND
While physical deterioration is always problematic with any form of preservation, this paper is mainly concerned with 'Digital Obsolescence'. With the proliferation of new technology and digital formats, the loss of data due to outdated hardware and software is fast becoming an unignorable problem. Thus, digital preservation is starting to have a more widespread influence as a typical practice, but unfortunately lacks a one-size-fits-all solution to the problem. It typically requires a large amount of effort, time and money to implement it effectively, which is not always considered to be a profitable process.

Naturally the lack of standards, protocols and formats in the digital realm means there is no guarantee that in several years any particular file is going to be usable. At the same time, with the ever increasing pace of technological progress, it seems that digital preservation is being compromised from two sides.

Digital preservation typically consists of two main processes: 'Retrieval' whereby the necessary digital files are retrieved and placed into error free storage; and 'Interpretation/Rendering' where the files that have been stored are decoded from their various formats and made usable and available for human access [1].

There are several strategies to mitigate digital deterioration: Refreshing, the copying of data to the same medium again; Migration, the transferring of data to a new format; Replication, copy the data to more than one system; Emulation, recreating the same functionality of a system that is no longer available; and Metadata attachment that involves including information regarding the file format [2]. This solution focuses on a combination of replication and metadata attachment.

3. DESIGN OF SOLUTION
The tool is made up of two primary functionalities: Packaging and Rendering.

3.1 Packaging
The packaging function allows the user to connect to an online repository via the base URL. After connecting, the tool implements the OAI-PMH protocol in order to query the repository, which can be based on the variables From (date), Until (date), Set and Metadata format. Once the query parameters are selected, the user can browse the Metadata identifiers returned in the table and, should they choose to, they can package the query results.

The packaging will take the metadata returned by the repository, as well as download the actual PDF data files associated with and specified in the metadata. The downloading of files uses a relatively brute force method. The address specified (in the field chosen by the user) is scanned for links to PDF files, which are then downloaded, or. in the case that it is a direct link to the PDF file, simply downloaded. If more than one file is downloaded, all of them are still packaged and associated with the same metadata file.

Once this is done, all the metadata, data files, as well as any auxiliary metadata (regarding Technical, Rights, Source and Digital Provenance specifications) are used to create a single METS file. [3]

The attachment of auxiliary metadata allows, to some extent, a mitigation of digital obsolescence, as mentioned above.

3.2 Rendering

The rendering function takes the METS file created by packaging as an input, and makes the metadata and data available to users. The primary and auxiliary metadata are displayed in tables and trees allowing the user to browse through them. The export function will recreate the metadata as well as the PDF files in a specified location.

Initially, the plan was to export the data files as navigable HTML, which would probably be more sensible with, for instance, JPEG files. But, given the nature of PDF it seems more appropriate to just export the files as they are and allow the user to decide how to make them more widely accessible.

4. EVALUATION

4.1 Experiments

The evaluation for this tool is largely Proof of Concept, in that it must be able to achieve its purpose, as explained above, effectively and without error. In order to demonstrate this, random queries were used on several OAI-PMH compliant repositories and packaged. The repositories were chosen from the Open Archives Inventory of Data Providers [4]. The archives were then exported and their contents examined to determine correctness.

4.2 Tests & Results

Please see Table 1.

5. CONCLUSTIONS

When analysing the results it becomes clear that the tool is far from perfect. While the metadata harvesting runs error free and effectively, the data harvesting lacks efficiency due to a number of reasons:

(I) The metadata does not specify an address.

(II) The address specified is no longer valid.

(III) The LinkExtractor.java class used to determine the URLs for the data does not detect them.

(IV) The LinkExtractor.java class detects more than one URL for the data and downloads more than is necessary.

Naturally, the problems of (I) and (II) are uncontrollable for all intents and purposes here. The prominent problem is with the link extraction, and it follows that with enough improvement the tool would become sufficiently more effective at archiving online collections.

That being said, the tool performs well in all other areas. There are no apparent problems with the archiving and exporting process, and the tool largely meets the goals set out in the beginning.

6. FUTURE WORK

6.1 Link Extraction

As mentioned above, this is where one of the primary deficiencies of the tool lies and improvement would greatly improve the tool's performance.

6.2 Extend to allow Multiple Package Formats

While METS is sufficient, other formats, such as MPEG7 or Simply CT, could be used effectively to offer a wider range of archiving uses.

6.3 Extend to allow Multiple File Formats

Currently the tool only allows for archiving of PDF files. The work to extend to different formats is not substantial. The problematic part is, once again, the link extraction. For formats such as JPEG, for instance, it would have to be far more intelligent so as not to download every picture on the html page.

6.4 Converting file formats

While not necessary with the PDF formats, if other file formats are included it could be useful to convert them to formats with better expected long-term survivability

6.5 Indexing Archives

It could be particularly useful to Index the data and metadata in larger archives in order to make them more accessible.

7. REFERENCES

[1] UKOLN, June 2008. *An Introduction to Digital Preservation*. [Online]. Available: http://www.ukoln.ac.uk/cultural-heritage/documents/briefing-31. [15 February 2010]

[2] Cornell University Library, 2008. *Digital Preservation Strategies*. [Online] Available:http://icpsr.umich.edu/dpm/dpm-eng/terminology/strategies.htm [15 February 2010]

[3] METS: An Overview & Tutorial', February 27 2009. [Online] Available: http://www.loc.gov/standards/mets/METSOverview.v2.html. [15 February 2010]

[4] Birgit Matthaei, 2003. *Open Archives Forum: OAI Archives – Inventory of Data Providers*. [Online]. Available: http://www.oaforum.org/otherfiles/tv-dp.pdf. [15 February 2010]

Table1. Tests and Results

Repository	From	Until	Set	Metadata Protocol	Number of Results	Number of Metadata Files Exported	Number of Data Files Exported
http://epub.wu-wien.ac.at/dyn/OAI/oaicgi.pl	09/01/24	10/08/25	Theses	oai_dc	0	0	0
http://eiop.or.at/cgi-bin/oaiserv.pl	09/07/30	10/11/00	Euirscas*	oai_dc	103	103	37
http://sammelpunkt.philo.at:8080/perl/oai2	09/03/20	10/07/06	Subject = Philosophie: Geschichte der Philosophie: f) 19.Jahrhundert*	uketd_dc	43	43	18
http://217.21.43.5/cgi-bin/oai	09/01/11	10/07/26	Economics*	oai_dc	0	0	0
http://eprints.biblio.unitn.it/perl/oai2	09/10/01	10/01/18	Published	oai_dc	48	48	52
http://eprints.cs.vt.edu:8000/perl/oai2	09/12/02	10/08/31	Software Engineering*	oai_dc	5	5	4
http://philsci-archive.pitt.edu/perl/oai2	09/12/22	10/05/19	All	oai_dc	29	29	24
http://phy043.tours.inra.fr:8080/perl/oai2	09/10/09	10/10/28	Status = In Press	oai_dc	0	0	0

* These set fields were changed to 'All Sets' after the randomly chosen one yielded no results in an effort to gain more meaningful tests.