# Using OAI-PMH and METS for exporting metadata and digital objects between repositories

Jonathan Bell and Stuart Lewis

Authors: Jonathan Bell is the Repository Bridge Project Officer, Information Services, University of Wales Aberystwyth, UK. E-mail: jon.bell@aber.ac.uk
Stuart Lewis is the Web Applications Developer, Information Services, University of Wales Aberystwyth, UK. E-mail: stuart.lewis@aber.ac.uk

## *Abstract*

### Purpose

To examine the relationship between deposit of electronic theses in institutional and archival repositories. Specifically the paper considers the automated export of theses for deposit in the archival repository in continuation of the existing arrangement in Wales for paper-based theses.

### Design/methodology/approach

We present a description of software that makes use of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as the first stage in the automatic import and ingest of items between institutional and archival repositories. The implications of this approach on the management of the institutional repository are also considered.

### Findings

We show that OAI-PMH is a useful approach to harvesting the metadata for items to be imported into an archival repository. This reduces the difficulty of maintenance of the import and export software components albeit at the possible expense of necessitating certain requirements on the management of the institutional repository.

### Research implications/limitations

The research shows that institutions can make use of OAI-PMH as a part of an automated export/import process, encouraging the preservation of multiple copies of digital items for increased safety of the content.

### Practical implications

The software has been developed and is being tested. It is proving capable of performing the required harvesting but the relative imprecision of searching in OAI-PMH has implications for the management of the exporting repository. These are discussed.

### Originality/value

We present a description and discussion of novel software components that enable the use of OAI-PMH as the first stage in the export and import of digital items between repositories, independently (as far as is practicable) of the software used by the repositories themselves.

Article type: Technical paper

Keywords: Institutional repositories; OAI-PMH; METS, E-theses; Wales; Item export; Metadata export

Word length: 4,676

## 1. Introduction and background

The University of Wales Aberystwyth (UWA) has been investigating the electronic deposit of theses using an institutional repository (IR) for preservation and access to e-theses, as well as other research output. Arising from this there has been an investigation of how e-theses can be exported to a primarily archival repository at the National Library of Wales (NLW). This paper describes and discusses the results of investigations into the automated export of theses and their ingest into the NLW's repository. While most of the development work has been carried out at the UWA, the University of Wales Swansea (UWS) has also installed an institutional repository to support further testing of the resulting software. This work was carried out with funding from the UK's Joint Information Systems Committee (JISC) for the project known as the Repository Bridge (http://www.jisc.ac.uk/index.cfm?name=project_repository_bridge). The project has resulted in the writing of software and the establishment of processes to support the electronic deposit of theses and their export to an archival repository in such a way as to make use of established standards and protocols.

There is a long standing arrangement between the universities in Wales and the NLW that any thesis resulting in the award of a research degree should have a copy deposited in the NLW. In addition, dissertations from students studying taught masters courses that merit the award of a distinction or those deemed to be of Welsh interest are also deposited in the NLW. The aim of the research described in this paper is a continuation of this existing arrangement through its application to e-theses together with, as far as possible, the automation of the export, import and ingest of these into the NLW archive.

The UWA and the UWS use DSpace (http://www.dspace.org/), an open source repository management system, as described in Smith (2003). DSpace was chosen primarily as it provides a usable solution as it stands, providing as it does a built-in workflow system for managing a repository. In contrast, the NLW uses Fedora (http://www.fedora.info/) as the 'back end' of its electronic archiving system. Fedora is also open source software and provides more sophisticated support for long-term preservation of digital materials. Its greater flexibility in managing different types of digital content (such as video and sound recordings), supported by the association of one or more 'disseminators' with the original item, is described in Staples (2003). While it was not originally a part of the project's aim, the Repository Bridge team is also co-operating with another JISC-funded project – Electronic Theses Online Service (EThOS - http://www.ethos.ac.uk/). One of aims of EThOS is the establishment of a UK-wide database of theses. The intention is that the Repository Bridge will allow the NLW to act as a hub for depositing Welsh theses in the EThOS database of theses.

# 2. The design of the import system

## *2.1 Requirements*

The main requirement of the project was to develop a ‘bridge’ that enables the automatic export of an item (specifically a thesis) from the repository in which it was originally deposited into some other repository. As far as possible, this is to be carried out without human intervention so a thesis deposited in a university’s IR will also be deposited in the NLW. It will be seen from the description of the existing arrangement that the process cannot be automated in all cases as it is impossible to automate the selection of Welsh-interest taught masters dissertations. Therefore it is a requirement that some method of allowing a user to trigger the export of an item be incorporated alongside the automatic export. This is also useful in cases where automatic export has failed. It is also a requirement that adequate logs of activity are generated and administrators are warned of any failures.

Both DSpace and Fedora are open source, so it is feasible to incorporate any changes felt necessary in their code. However, this is undesirable as it is likely to lead to a maintenance overhead as the changes are incorporated anew into successive versions of the software. Because of this, it was decided that as much use should be made as possible of facilities for metadata harvesting and export and import that are already incorporated into the repository tools.

In addition to the requirements of the software, the use of the bridge places, or might place, requirements on the structure and management of the repositories themselves. The obvious case here is the need to ensure that on deposit of an item for exporting, sufficient metadata is present for both the institutional and archival repository. The depositor also needs to be made aware that the item will also be deposited in the archival repository and agree to that deposit. One disadvantage with the approach taken is that it tends to increase the effect of the bridge on the design and management of the institutional repository.

## *2.2 Method*

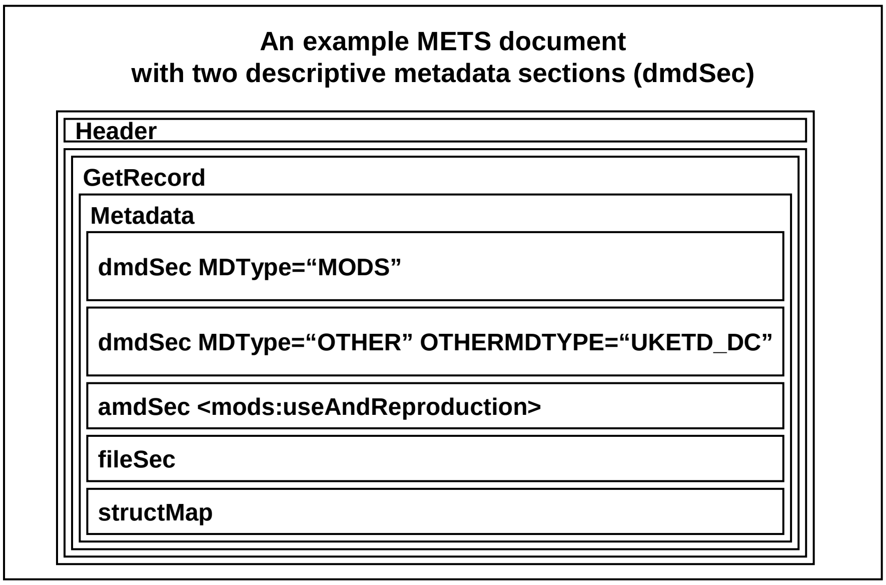
Given that the bridge is to connect repositories using different software there are two possible approaches. One is to get one of the repositories (probably the exporting one) to pass on items in the ‘language spoken’ by the other. This was the approach that initially we expected to adopt. The alternative is to attempt to use a *lingua franca* that both repositories will support. Fortunately there are such *linguae francae* available for different, but related, tasks. Specifically, DSpace and Fedora both support the Metadata Encoding and Transmission Standard (METS - http://www.loc.gov/standards/mets/), so this was a good starting point.

While METS is useful for actually transmitting the required metadata for each item to be exported, there is also the need to identify the items to be exported. As DSpace supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH - http://www.openarchives.org/) as a data provider, this can be used. The OAI-PMH defines a mechanism for harvesting records containing metadata from repositories. As it makes use of the open standards of Hypertext Transport Protocol (HTTP) and the Extensible Markup Language (XML) it allows metadata to be harvested over the Web. Although simple Dublin Core is specified as a metadata standard to enable a basic level of interoperability across all repositories, the metadata can be in any format agreed by some community of users. Metadata can be gathered in one database and services provided based on this collection of ‘aggregated’ metadata. At present, OAI-PMH is only a *de facto* standard but support for it is widespread. There is a history of OAI-PMH in Carpenter (2003) and an overview of the protocol in Lagoze (2004).

For import, there is no need for Fedora to support OAI-PMH, though it does so, and this is useful for re-exporting items to EThOS. The relationship with EThOS caused some complication to the METS to be exported. The NLW and EThOS have adopted different standards for descriptive metadata. The NLW uses the Metadata Object Description Schema (MODS - http://www.loc.gov/standards/mods/) as does the DSpace support for METS. However, EThOS has adopted its own qualified Dublin Core (qDC) metadata set for electronic theses and dissertations (ETD), the UKETD metadata set. This was derived from an earlier metadata set for ETD described in Copeland et al. (2005). As DSpace uses  qDC internally, it was decided to add a set of  qDC metadata to the exported METS alongside the MODS. This seemed simpler than converting the  qDC to MODS and then converting it back for export to EThOS. As METS supports the presence of multiple descriptive metadata sections, the adopted approach is readily handled. The structure of the resulting METS document is outlined in Figure 1.

Take in Figure 1

Figure 1.  An example of a METS document


An example METS document
with two descriptive metadata sections (dmdSec)
Header
GetRecord
Metadata
dmdSec MDType=“MODS”
dmdSec MDType=“OTHER” OTHERMDTYPE=“UKETD_DC”
amdSec <mods:useAndReproduction>
fileSec
structMap

It can be seen that the METS includes:

- two descriptive metadata sections (dmdSec), one of which  uses MODS and the other UKETD;
- an administrative metadata section (amdSec) that holds a copy of the deposit  agreement;
- a fileSec that describes the individual files that make up the actual content of the thesis;

- a structure map (structMap) describes the relationships between these files.

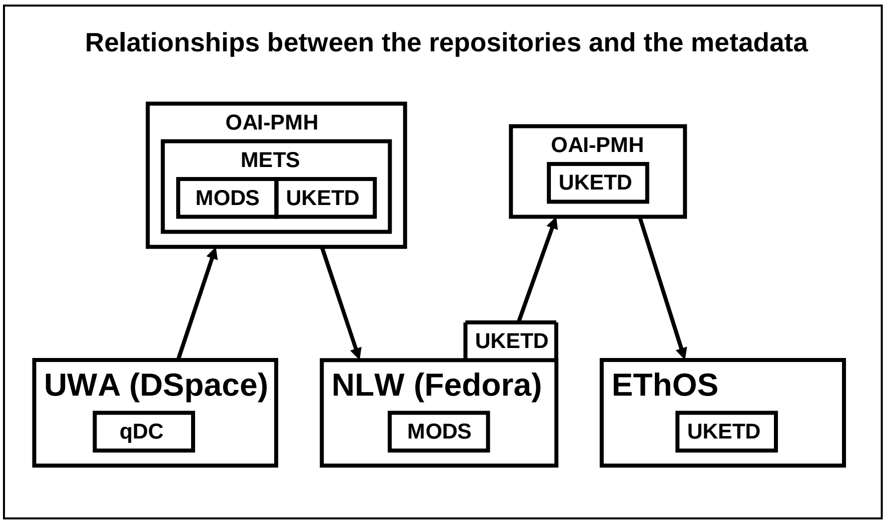
The use of METS and OAI-PMH allows the metadata importer to be a simple program that is external to both repositories. This minimises the changes required to the repository software itself.

## 3. The harvester in use

The process works by the importer (installed at NLW) sending OAI-PMH requests to each contributing institution in turn. First, a list of new theses is obtained by sending an OAI-PMH ‘ListIdentifiers’ request, with the date of the previous thesis import operation as the earliest date of interest (the OAI-PMH ‘from’ date argument). As this request can be qualified using a ‘set’ argument (which DSpace equates to collection), this operation is repeated for each collection in an institution’s repository that is known to contain theses. The repetition of ‘ListIdentifiers’ results in the production of a list of identifiers of all theses that are yet to be imported by the NLW. The import is carried out by calling the OAI-PMH ‘GetRecord’ request with each identifier in turn. The resulting METS files are saved in the NLW client’s filestore, ready for ingest into Fedora. The METS metadata includes identifiers for each of the files of the item itself so, during ingest of the METS, Fedora automatically downloads the content (the thesis) using these file identifiers. This approach works whether a thesis content is in one file (say a PDF) or several (such as a sequence of Microsoft Word documents, one for each chapter). The import software will also support cases where there is no content available, so only the metadata will be exported. The need for this feature is questionable but a requirement of EThOS is that it will support import of metadata alone, if the thesis itself is not to be made available. The NLW is then in a position to forward the metadata to EThOS, using a simpler METS format as there is now no need for the MODS descriptive metadata. The relationship between repositories and metadata is illustrated in Figure 2.

Take in Figure 2

Figure 2. Relationships between the repositories and the metadata



The UWA stores its metadata as qDC and this is converted to MODS and UKETD Dublin Core for export. The NLW keeps the MODS for its own use and simply relays the UKETD metadata on to EThOS, which also uses the file identifiers in the metadata to download the actual item.

The importer keeps a list of institutions and sets (collections) to harvest from in a file that also keeps contact information of the institutional repository administrators for e-mail alerts and logging. These entries can be changed by the administrators using a simple Web interface, so an institution's administrator can add or remove collections that should be harvested from, and change or correct contact details. The Web page also allows users to view logs of the interactions between the import tool and their institution. Changes to the sets for harvesting are assisted by use of OAI-PMH, as the 'ListSets' request is used to allow the Web interface to show a 'pick' list of available sets in the IR.

This approach minimises the internal changes to the DSpace code, though additions were found to be necessary (at least to version 1.3.2). Apart from the addition of the UKETD Dublin Core descriptive metadata, changes were made to support the export of metadata of items that contain more than one content file. This entailed providing a correct structure map.

Beside the relatively minor changes to DSpace itself, a program to carry out the import was written. This program follows the sequence of operations described above, allowing the ingest program to recover the METS files later and incorporate them into Fedora. The intention is that this program will be run periodically, maybe as often as every night. This was felt preferable to the idea of triggering export on submission of a thesis because of the danger of the exporter attempting to export when the importer is unable to accept the item. The periodic harvest approach simply means that if an institution is unable to respond to a request from the importer, any theses that are not imported will wait until the next run of the harvester. Each set has its own date of last harvest and this is not updated when the harvest fails. Having the exporting institutions drive the process would also result in the need for similar export code at each institution, instead of the one copy of the importer at the NLW. This would increase the difficulty of maintenance and of adapting the import for other repository software. How often the harvest should run is still to be decided. This would obviously affect the delay between deposit of a thesis and its export to the NLW and EThOS.

Because not all the theses for export are susceptible to automatic selection, the bridge needs a way of triggering the export of a thesis or dissertation manually. As all items for export are placed in specific collections to allow the harvest to be restricted by set, this problem has been solved simply by allowing the depositor of a dissertation that is to be exported to place the item in a target collection. This was felt to be less troublesome for the depositor than requiring the trigger of any specific operation, as it is simply part of the deposit process. This means that the depositor can trigger such an export without accessing the importer itself. This highlights the fact that the operation of the bridge has effects on the design and management of the IR.

## 4. Discussion

Using OAI-PMH as the starting point for our harvesting has advantages and disadvantages. There is a trade-off between easing the maintenance of the software and easing maintenance of the repository itself. The most important advantage of making use of existing standards, such as OAI-PMH, is that support for these exists in the repository software. This applies to us in the case of OAI-PMH itself, merely requiring the addition of files to support the correct metadata formats. METS and MODS are supported by DSpace, though as noted earlier it was found necessary to improve the METS support. The EThOS project team has written the code necessary to support export of its preferred UKETD metadata set by OAI-PMH, and we were able to modify this so as to enable its use for generating the UKETD descriptive metadata section in our exported METS. Our slightly changed version of this code was adopted by EThOS as a result of the close co-operation between the two projects. The benefit of reducing the project input into the repository software’s code is, of course, that there is no need to keep changing our code to keep pace with new versions of the software. For example, now that EThOS has adopted our version of the UKETD file, any changes to its metadata set, resulting in changes to that file, will be incorporated into our export on installing the updated version of the file. No other changes are required. Similarly, as all interactions with the repository itself are through OAI-PMH, we can be confident that new versions of DSpace will support these export operations. The use of established standards and the resulting minimising of additions to the repository software also eases installation into other repositories. The importer itself is installed in the importing institution (the NLW) so each exporting institution only needs the additional files to support the OAI-PMH export, essentially the file to support the UKETD metadata set and a file to incorporate it into the METS. The necessary changes were made to UWS’s DSpace in a matter of minutes.

The principal disadvantage of the use of OAI-PMH is its relatively undiscriminating approach to harvesting. As harvest can only be limited by set and date, and as set is equivalent to collection with DSpace, then if harvesting for export is to include all theses and only theses, it is necessary to ensure that theses for export are placed in the appropriate sets and that these sets contain nothing but theses. At UWA, the approach taken (initially) is to have special collections for export. There are separate collections for doctoral and research masters theses and an additional collection for those taught masters dissertations that are to be exported. These are the collections listed as sets to harvest from in the harvest software's configuration file.

Fortunately, DSpace has the facility to allow an item to be held in one collection and 'mapped into' some other collection so that it appears in both collections. This means that the apparent constraint that the Repository Bridge places on the design of the IR is less severe than it appears. Items mapped into the export collection are harvested just as though they were really in that collection. If an academic department wished to have its theses in its own collection, they could be deposited there and mapped across to the export collection. In addition to this need for one or more collections for items to export, this places some extra work in the submission (deposit) process, as whoever submits the finished thesis, or whoever checks the submission is complete, must ensure that it is mapped across to the export collection unless, of course, it was deposited directly into that collection. This is less of a problem for repositories using the EPrints open source software (http://www.eprints.org/) as that software places theses in their own set.

Some thought has been given to using the metadata format to restrict the harvest. We believe that it should be possible to make the UKETD metadata format applicable only to those items which are of type thesis (or dissertation). This would then allow the harvester's 'ListIdentifiers' request to be applied to the whole repository, restricted only by date. Efforts to implement this have not been successful, however. This approach would have the disadvantage that if the bridge were to be used in future for export of items that were not theses, some other metadata format would have to be used and this would almost certainly be a standard METS and so would still need to be constrained by set, with specific export collections to support this. An alternative approach would be to harvest all the metadata from a repository and post filter it to extract the required items for import. This might lead to problems with scaling, as all records, not just all identifiers, would need to be harvested so that the field used for post filtering was available. Another possible difficulty in the case of the Repository Bridge is the need for having some field that distinguishes those masters dissertations that are to be harvested. Having a separate type, say, for "Welsh interest dissertations" seems to complicate the deposit process and increase the danger of incorrect selection of that field. This would, of course, result in errors in the selection of items for import.

The alternative to using OAI-PMH would be to construct a harvester that directly queried the repository, searching for all items added since the previous export and that were theses. The drawback of constructing such a query is that it is sensitive to the repository software and /or the underlying database, so might need considerable changes in response to any changes in the repository software itself.

## 5. Submission of e-theses

Naturally, for the Repository Bridge to be of any use, it is necessary to secure the deposit of electronic copies of theses in IRs. In Wales, the rules for submission of a thesis and award of the degree specify that the thesis is to have copies deposited in the relevant university library and in the NLW. For the bridge to work, then, the submission process needs updating to allow electronic deposit either as well as, or instead of, the current deposit of the paper copies. In general there are three levels at which this could operate:

- voluntary deposit - when a university might invite a candidate to deposit an e-copy of the final version of the thesis alongside the paper copy, taking steps to ensure that the content of the two versions is identical.
- compulsory deposit (as a condition of the award of the degree ) of an e-copy of the final version which would also exist as a paper copy.
- full electronic submission (that is, for examination) which could introduce the possibility of writing a thesis in a non-linear structure, arranging the work rather like a Web site, for example.

At UWA we are at the first stage with proposed changes to the regulations for submission of research theses to allow (but not enforce) the deposit of an e-copy of a thesis alongside the paper copies. Candidates are expected to sign a declaration to the effect that the e-copy is identical in content to the final, corrected paper copy, as is the case at Cranfield University (Bevan, 2005). They must also declare that they have taken suitable steps to obtain copyright clearance for the deposit of any third party copyright material in an open access repository and indemnify the UWA against claims regarding third party copyright. These conditions also cover the deposit of the thesis in other repositories, such as at NLW or in EThOS. Issues relating to the intellectual property and licensing of e-theses are discussed further in Jones and Andrew (2005) and Andrew (2004).

## 6. Conclusion and future work

The Repository Bridge software has been implemented along with the required changes to DSpace itself. It is undergoing testing at the time of writing (late April 2006) and appears to meet the specified requirements.

There are two specific areas for future work, one technical and one concerned with policy. The technical work involves writing code to support IRs in Wales using the EPrints software as it is hoped that the Repository Bridge will be implemented in other Welsh universities. EPrints is OAI-PMH compliant and this should present few difficulties. The ease of adapting the bridge for other software is an advantage of the chosen approach.

The other area is the handling of deposit of embargoed theses. Any thesis might be subject to an embargo for various reasons such as:

- to protect commercially sensitive information;
- to protect the content while it is being prepared for publication as a book;
- to protect the candidate if the work is controversial in nature (such as being concerned with political controversy or animal experimentation).

As items deposited in the repository are expected to be open access, in general, then some way of preventing access to an embargoed thesis is required during the embargo period. The approach adopted at present is the simple one of not accepting deposit of an embargoed thesis until expiry of the embargo. Another approach might be to accept the thesis on disc and keep it in a safe until the embargo has expired. Although it is possible to deposit an item in an IR and restrict access to that item there are possible objections to this approach. One is the fact that the metadata will still be exposed and the other is the danger of legal implications of keeping data in an electronic source and not being prepared to make it available. This might arise from freedom of information legislation, such as the UK Freedom of Information Act 2000 (http://www.opsi.gov.uk/acts/acts2000/20000036.htm).

The work undertaken for the Repository Bridge is also supporting the development of an IR at UWA as well as an investigation of the export of other items, besides theses, to the archival repository at NLW.

As the importer itself is quite self contained and the necessary additions to DSpace easily installed, it is possible that our work could be used as a starting point for other related projects. The code as developed, especially for metadata exporting, is fairly specific to theses (if only because of its use of the UKETD metadata set) but that could be removed and MODS used on its own. For this reason, the code is not presently available to download, but anyone interested is welcome to contact the authors.

To summarise, we have developed a working model of a bridge to allow the export and import of digital materials (metadata and content) between institutional and archival repositories, allowing for the possibility that different repository software might be in use at each end of the bridge. We have accepted the effect that the choice of approach has on the design and management of the IR in the interest of easing the overhead of maintenance of the bridge itself, while being aware that this is a solution that suits our circumstances but might be considered an unacceptable compromise in other instances. The advantages of our chosen approach include ease of maintenance and of its potential for being made available for other institutions, particularly across Wales.

**Editor's Note**

A PowerPoint presentation giving slightly more technical detail than is given in this paper was presented at the DSpace Users Group Meeting, University of Bergen, Norway, April 20-21 2006. Available from: http://dsug2006.uib.no/archive/lewis.ppt

## References (All URLs checked 11th April 2006)

Andrew, T. (2004), *Intellectual Property and Electronic Theses*, JISC Legal Information Service. Available at: http://www.jisclegal.ac.uk/publications/ethesesandrew.htm.

Bevan, S. J. (2005), "Electronic thesis development at Cranfield University", *Program*, Vol. 39 No. 2, pp. 100-111.

Carpenter, L. (2003), *History and Development of OAI-PMH*, UKOLN, Bath. Available at: http://www.oaforum.org/tutorial/english/page2.htm.

Copeland, S., Penman, A., and Milne, R (2005), "Electronic theses: the turning point", *Program*, Vol. 39 No.3, pp 185-197.

Jones, R. and Andrew, T. (2005), "Open access, open source and e-theses: the development of the Edinburgh Research Archive", *Program*, Vol.39 No. 3, pp 198-212.

Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S. (2004), *The Open Archives Initiative Protocol for Metadata Harvesting*, Open Archives Initiative. Available at: http://www.openarchives.org/OAI/openarchivesprotocol.html.

Smith, McK., Barton, M. R., Bass, M. J., McClellan, G., Stuve, D., Branschofsky, M., Harford Walker, J. and Tansley, R. (2003), "DSpace: an open source dynamic digital repository", *D-Lib Magazine,* Vol. 9 No. 1. Available at: http://www.dlib.org/dlib/january03/smith/01smith.html/

Staples, T., Wayland, R. and Payette, S. (2003), "The Fedora Project: an open-source digital object repository management system", *D-Lib Magazine*, Vol. 9 No. 4. Available at: http://www.dlib.org/dlib/april03/staples/04staples.html.