code{4}lib JOURNAL

Issue 8, 2009-11-23

ISSN 1940-5758

Automatic Preparation of ETD Material from the Internet Archive for the DSpace Repository Platform

A big challenge associated with getting an institutional repository off the ground is getting content into it. This article will look at how to use digitization services at the Internet Archive alongside software utilities that the author developed to automate the harvesting of scanned dissertations and associated Dublin Core XML files to create an ETD Portal using the DSpace platform. The end result is a metadata-rich, full-text collection of theses that can be constructed for little out of pocket cost.

By Tim Ribaric

Introduction

The Library where I work recently embarked upon creating an institutional repository. The ultimate goal of the repository is to be the authoritative, archival source of information produced by my institution, Brock University. During the selection process to choose a software platform, many different products, both open source and proprietary, were evaluated. After a thorough search the selection came down to the DSpace platform [1].

From what we saw, DSpace offered the best set of features and being an open source product allowed us to add extra functionality to it. The DSpace platform has a large body of users which has created an extensive pool of expertise to draw from. The only drawback is that the official documentation is a bit lacking and doesn't shed light on everything that DSpace is capable of. The project I embarked upon was not to modify the platform itself but to find a way to populate it with institutional data. I was searching for an automated way to ingest a collection of dissertations created by the University.

The Start of the ETD portal

Like most institutions with graduate programs, my library possessed a collection of printed and bound theses that dated back to the inception of the school. The plan was to create an ETD (Electronic Theses and Dissertations) portal within DSpace. The total list of graduate dissertations numbered somewhere around 1400.

This proposed ETD portal created two distinct challenges: first how to digitize this vast quantity of material, and second how to ingest this information into the fledgling DSpace instance. The answer to the first question turned out to be twofold. Some theses would be shipped out to the Internet Archive for digitization while some were to be digitized in house.

The Internet Archive [2] site summarizes its aims quite succinctly:

> The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library, with the purpose of offering permanent access for researchers, historians, and scholars to historical collections that exist in digital format. [3]

While best known for hosting a variety of digital content, the Internet Archive Foundation also provides a digitization service. Paper materials can be sent to the offices of the Archive (at the University of Toronto in our case) to be digitized. This service delivers a PDF document with full OCR mark up and information about the material itself in the form of metadata. The material is then hosted on the Internet Archive site, presumably indefinitely, to be accessed by anyone. The digitization service offered by the Internet Archive solved half of the problem. The end result was well-crafted digital representations of our theses. Unfortunately they were made available in a way that wasn't the most useful for us. That is to say they existed in an online form that was not locally hosted or integrated into our already existing repository. The next step in the process was to locally load these theses into the DSpace site. This boiled down to two complementary processes, each of which were automated with a web application.

IA_Scraper

The first utility I created, IA_Scraper [4], was an intelligent RSS feed monitor that would (with help from the PHP cURL class) monitor our collection of newly digitized theses and download the required materials. Every item posted on the Archive is displayed with a fairly predictable layout. The URL follows a succinct predetermined pattern and for each digitized item there are a base set of files that are created.

An example record of one of our theses on the Archive, with associated XML and MARC record files, can found at http://ia331419.us.archive.org/1/items/descartesmeditat00yurkuoft/:

Figure 1. Directory of a scanned thesis at Archive.org

The extension of the file names show exactly what is bundled for each item. In my particular case I was looking for the PDF file and the XML Dublin Core file.

Furthermore, each distinct set of material found on the Archive is organized in a unique 'collection'. In my case all the items from Brock University were sorted under http://www.archive.org/details/BrockUniversity. By browsing the collection unique to my institution and by selectively searching for the word 'Thesis' in the description of the item it was possible to put together a list of our digitized theses. In addition, the Internet Archive offers an RSS feed that reports the items as they get added to particular collections. Putting these two features together I was able to monitor our institute-specific RSS feed and fetch the theses as they were added. The end result is a collection of XML and PDF files of the digital versions of our theses.

DS_Ingestor

The DS_Ingestor [5] completes the final steps of converting the Internet Archive data into something that is DSpace ready. DSpace creates a standardized directory and file structure for each imported item. More information about this formatting can be found in the DSpace manual [6]. Simply pointing the DS_Ingestor at the data gathered from the IA_Scraper will process titles into ready to ingest numbered sets (as defined in a config file). The most challenging part of the DS_Ingestor process was modifying the XML produced by the Internet Archive into the format that DSpace expects.

Example of the differing XML syntax

The Internet Archive provides metadata for its items in a variety of different formats. I used the OCR full-colour PDF as our repository object and the Dublin Core XML file associated with the item. Dublin Core was chosen because it has been adopted widely and it is natively supported by DSpace.

Below is the Dublin Core metadata that I retrieved from the Internet Archive for one of our digitized theses located at: http://www.archive.org/details/descartesmeditat00yurkuoft. (Updated, see correction note).

```
<?xml version="1.0" ?>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>Descartes' Meditations : can the idea of God be derived from a meditation on the will and substance? /</dc:title>
  <dc:creator>Yurkewich-Liddell, KellyAnn.</dc:creator>
  <dc:type>text</dc:type>
  <dc:publisher>St. Catharines, Ont. : Brock University, Department of Philosophy,</dc:publisher>
  <dc:date>2004.</dc:date>
  <dc:language>eng</dc:language>
  <dc:description>Thesis (M.A.)--Brock University, 2004.</dc:description>
  <dc:description>Includes bibliographical references (l. 94-98)</dc:description>
  <dc:subject>Descartes, Ren&amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;eacute;, 1596-1650.</dc:subject>
  <dc:subject>God</dc:subject>
</oai_dc:dc>
```

The XML native to the Internet Archive is written against the openarchives.org namespace with metadata elements enclosed within Dublin Core colon-delimited tags.

DSpace, however, requires things to be formatted a bit differently. Here is the same information from our example record after it has been converted in the DSpace format by DS_Ingestor.

Continuing with our previous example now found at: http://dr.library.brocku.ca/handle/10464/2298

```
<dublin_core>
  <dcvalue element="title" qualifier="none">Descartes' Meditations : can the idea of God be derived from a meditation on the will and substance? /</dcvalue>
  <dcvalue element="contributor" qualifier="author">Yurkewich-Liddell, KellyAnn.</dcvalue>
  <dcvalue element="type" qualifier="none">text</dcvalue>
  <dcvalue element="publisher" qualifier="none">St. Catharines, Ont. : Brock University, Department of Philosophy,</dcvalue>
  <dcvalue element="date" qualifier="none">2004.</dcvalue>
  <dcvalue element="language" qualifier="iso">eng</dcvalue>
  <dcvalue element="description" qualifier="none">Thesis (M.A.)--Brock University, 2004.</dcvalue>
  <dcvalue element="description" qualifier="none">Includes bibliographical references (l. 94-98)</dcvalue>
  <dcvalue element="subject" qualifier="none">Descartes, Ren&amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;eacute;, 1596-1650.</dcvalue>
  <dcvalue element="subject" qualifier="none">God</dcvalue>
</dublin_core>
```

The DSpace XML itself is a nested set of 'dcvalue' tags with a series of attributes, one for element, and another for qualifier. The basic metadata like 'Author' and 'Title' is listed, but there is also a 'Subject' tag, which is the LC subject headings that have been assigned to the Thesis when it was being digitized. By incorporating this information into the DSpace record of each thesis another facet of categorization and search is added to the ETD Portal.

The completed workflow

Figure 2. Sample thesis at Archive.org

The complete workflow for our example would look something like this:

1. The thesis is digitized and made available on the Archive at url http://www.archive.org/details/descartesmeditat00yurkuoft
2. IA_Scraper downloads the required information by examining an RSS feed and saves it locally as follows:
   ```
   ia_scraper_data_dir/descartesmeditat00yurkuoft/
              descartesmeditat00yurkuoft.pdf
              descartesmeditat00yurkuoft_dc.xml
   ```
3. DS_Ingestor runs and adds the files to a bulk directory with other titles:
   ```
   ds_ready_dir/batch_2009-09-14-01-19/
     ...
     descartesmeditat00yurkuoft/
          contents
          dublin_core.xml
          descartesmeditat00yurkuoft.pdf
     ...
   ```
4. DS_Ingestor reports back that data is ready along with the final command line instruction that needs to be entered that commits the data to the repository. (This will vary depending on DSpace configuration.) For this example it might end up being (on a Unix style system):
   ```
   /dspace/bin/import --add --eperson=person@repository.edu
     --collection=1111/1000
     --source=/var/www/dsingestor/ds_suitable/batch_2009-09-14-01-19/
     --mapfile=/home/dspace/ingest_files/batch_2009-09-14-01-19.map.ingest
   ```

Briefly looking at the components:

`/dspace/bin/import` – the location of the DSpace binaries directory where the import command is located.

`--add` – the switch to tell the importer you are adding items

`--eperson=person@repository.edu` – the email address of the person that has permission to deposit items into the repository

`--collection=1111/1000` – is a combination of your Handle id [7] and the id of the collection you are adding your content to

`--source=/var/www/dsingestor/ds_suitable/batch_2009-09-14-01-19/` – where the information downloaded by IA_Scraper and processed by DS_Ingestor is held.

`-- mapfile=/home/dspace/ingest_files/batch_2009-09-14-01-19.map.ingest` – Is where the generated mapfile of the input process will be kept. This is handy to keep because it will tell you what the ID number is of every item that was imported. This is also discussed in the DSpace manual [6].

Figure 3. Final DSpace page at Brock University

The end result and the limits of DSpace

The final repository populated with the Internet Archive content ends up being a very useful and attractive product. Since the PDF's have been OCR'ed it is possible to have a full text searchable database of theses. This can easily be accomplished by adding the full text of repository items to the search indexes that DSpace maintains. The specific details can be

found in the DSpace manual [8].

By also doing some refinements in how things are categorized it is possible to have the theses organized by rank and discipline so that easy browsing is facilitated. As previously mentioned, the inclusion of LC Subject Headings allows for even more entry points into the content. Our finalized Dissertations and Theses portal can be found at: http://dr.library.brocku.ca/handle/10464/4

DSpace in general takes some time to get used to. The fact that it runs in a Tomcat environment using a PostgreSQL database is a stark contrast to the traditional LAMP stack that most popular open source applications can be classified under. This difference of infrastructure created a real barrier to learning how to use DSpace effectively. As mentioned previously, official documentation always seemed out-of-date and not at all thorough. However, DSpace is a platform well worth implementing and supporting. In fact, all open source OAI products should be encouraged. These platforms become the basis of huge collaborative initiatives like OAIster [9] that will eventually be crawling all of the world's digital repositories and unifying all the content under one search mechanism.

A brief word on the ETD-MS

As previously mentioned, DSpace out of the box supports the Dublin Core metadata format. Dublin Core is a great schema that is widely applicable to many different publication types but it does not describe some essential pieces of information that pertain to dissertations. That is why the ETD-MS [10] was created. Proponents of other repository platforms point out that DSpace should support this schema. In fact , in order to have Theses Canada regularly harvest your digitized theses, support for the ETD-MS is a requirement [11]. This shouldn't be seen as a drawback to implementing DSpace. Previous versions of DSpace had user contributed software that allowed for the mapping of the DSpace Dublin Core schema to ETD-MS. With luck this mapping will once again be possible in current and future versions of the DSpace platform.

Final words

The IA_Scraper and DS_Ingestor are generalized enough to use for any Internet Archive content, and are not restricted simply to dissertations. Quite recently I have used the packages to download and locally archive a collection of out of copyright material that is of local interest. By reusing the metadata and digitized items created by the Internet Archive I was able to locally archive a collection of about 300 books in a matter of minutes. The combination of digitization services offered by the Internet Archive, the IA_Scraper and DS_Ingestor software, and DSpace platform allows anyone to create a fully functional open source digital repository. The primary benefit is an end product that requires no licensing fees and only costs staff time and a negligible scanning fee to get online.

References

[1] DSpace Homepage. http://dspace.org
[2] Internet Archive Homepage. http://archive.org
[3] Internet Archive: About IA. http://www.archive.org/about/about.php
[4] IA_Scraper. http://elibtronic.ca/software/ia_scraper
[5] DS_Ingestor. http://elibtronic.ca/software/ds_ingestor
[6] DSpace Item Importer and Exporter. http://www.dspace.org/1_5_2Documentation/ch09.html#N13795
[7] The Handle System. http://handle.net
[8] Configuring Lucene Search Indexes. http://www.dspace.org/1_5_2Documentation/ch05.html#N11980
[9] OAIster. http://oaister.org
[10] ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations. http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html
[11] Theses Canada Requirements for Harvesting Electronic Theses and Metadata. http://collectionscanada.ca/thesescanada/s4-262-e.html

About the Author

With a BSc. in Computer Science and a MLIS, Tim Ribaric has been the Digital Services Librarian at Brock University, located in the Niagara Region of Ontario Canada, since 2006. His areas of interest include OpenURL resolvers, the DSpace platform, and finding a way of using Twitter effectively in the Library environment. He blogs at http://elibtronic.ca and can be reached at tribaric (at) brocku.ca.

Correction

November 25th, 2009: The two code blocks in the Example of the differing XML syntax were inadvertantly switched. This error is now corrected and the surrounding paragraphs slightly rewritten to reflect this change.

The original text of this section read:

Below is the Dublin Core metadata that I retrieved from the Internet Archive for one of our digitized theses located at: http://www.archive.org/details/descartesmeditat00yurkuoft.

```
<dublin_core>
  <dcvalue element="title" qualifier="none">Descartes' Meditations : can the idea of God be derived from a meditation on the will and substance? /</dcvalue>
  <dcvalue element="contributor" qualifier="author">Yurkewich-Liddell, KellyAnn.</dcvalue>
  <dcvalue element="type" qualifier="none">text</dcvalue>
  <dcvalue element="publisher" qualifier="none">St. Catharines, Ont. : Brock University, Department of Philosophy,</dcvalue>
  <dcvalue element="date" qualifier="none">2004.</dcvalue>
  <dcvalue element="language" qualifier="iso">eng</dcvalue>
  <dcvalue element="description" qualifier="none">Thesis (M.A.)--Brock University, 2004.</dcvalue>
  <dcvalue element="description" qualifier="none">Includes bibliographical references (l. 94-98)</dcvalue>
  <dcvalue element="subject" qualifier="none">Descartes, Ren&amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;eacute;, 1596-1650.</dcvalue>
  <dcvalue element="subject" qualifier="none">God</dcvalue>
</dublin_core>
```

The XML itself is a nested set of 'dcvalue' tags with a series of attributes, one for element, and another for qualifier. The basic metadata like 'Author' and 'Title' is listed, but there is also a 'Subject' tag, which is the LC subject headings that have been assigned to the Thesis when it was being digitized. By incorporating this information into the DSpace record of each thesis another facet of categorization and search is added to the ETD Portal.

Dspace, however, requires things to be formatted a bit differently. Here is the same information from our example record after it has been converted in the DSpace format by DS_Ingestor.

Continuing with our previous example now found at: http://dr.library.brocku.ca/handle/10464/2298

```
<?xml version="1.0" ?>
<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>Descartes' Meditations : can the idea of God be derived from a meditation on the will and substance? /</dc:title>
  <dc:creator>Yurkewich-Liddell, KellyAnn.</dc:creator>
  <dc:type>text</dc:type>
  <dc:publisher>St. Catharines, Ont. : Brock University, Department of Philosophy,</dc:publisher>
  <dc:date>2004.</dc:date>
  <dc:language>eng</dc:language>
  <dc:description>Thesis (M.A.)--Brock University, 2004.</dc:description>
  <dc:description>Includes bibliographical references (l. 94-98)</dc:description>
  <dc:subject>Descartes, Ren&amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;amp;eacute;, 1596-1650.</dc:subject>
  <dc:subject>God</dc:subject>
</oai_dc:dc>
```

With the inclusion of a name space and some strict XML encoding, DSpace XML is a bit more rigorous than the Internet Archive's format. Nevertheless, every digitized thesis that passed through DS_Ingestor was able to be converted, indicating that the XML from the Archive was well-formed even without the addition of a namespace.

3 Responses to "Automatic Preparation of ETD Material from the Internet Archive for the DSpace Repository Platform"

Please leave a response below, or trackback from your own site.

1. Karen Coyle, 2009-11-24

   Tim,

   I hadn't ever seen the "dcvalue" form of DC, so I was curious… When I go to the IA dc.xml file, I get pretty standard looking DC:

   Descartes' Meditations : can the idea of God be derived from a meditation on the will and substance? /
   Yurkewich-Liddell, KellyAnn.
   text
   St. Catharines, Ont. : Brock University, Department of Philosophy,
   2004.
   eng
   Thesis (M.A.)–Brock University, 2004.
   Includes bibliographical references (l. 94-98)
   Descartes, René, 1596-1650.
   God

   Source: http://ia331419.us.archive.org/1/items/descartesmeditat00yurkuoft/descartesmeditat00yurkuoft_dc.xml

   Has something changed? Did you use a different file?

2. Karen Coyle, 2009-11-24

   Oops, need to escape all of the &lt:

   <?xml version="1.0"?>
   <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/">
   <dc:title>Descartes' Meditations : can the idea of God be derived from a meditation on the will and substance? /</dc:title>
   <dc:creator> Yurkewich-Liddell, KellyAnn. </dc:creator>
   <dc:type>text</dc:type>
   <dc:publisher>St. Catharines, Ont. : Brock University, Department of Philosophy,</dc:publisher>
   <dc:date>2004.</dc:date>
   <dc:language>eng</dc:language>
   <dc:description>Thesis (M.A.)–Brock University, 2004.</dc:description>

   <dc:description>Includes bibliographical references (l. 94-98)</dc:description>
   <dc:subject>Descartes, René, 1596-1650.</dc:subject>
   <dc:subject>God</dc:subject>
   </oai_dc:dc>

3. Tim Ribaric, 2009-11-25

   Hello Karen,
   Thanks for the astute observation. As it turns out there was a typo with the XML code blocks, (they were inverted) and that created the confusion. This has now been corrected and the article reads as it was intended to.