Exporting Metadata and Content from DSpace

A report from the 2016-2017 LoboVault migration at the University of New Mexico

By Amy E. Winter, MPA
Program Specialist, Digital Initiatives and Scholarly Communications
University Libraries and Learning Sciences, University of New Mexico

Abstract

The University of New Mexico initiated a digital institutional repository in approximately 2009 using the DSpace open-source repository software. By 2015, faced with deteriorating performance from the DSpace installation, library administration decided, rather than performing an extensive upgrade to DSpace, to purchase a subscription to bepress's hosted product, Digital Commons, and migrate all existing DSpace content – more than 23,000 documents – to the new platform. The migration was completed between June of 2016 and June of 2017. Because I could not find much explicit documentation on obtaining and processing data exports from DSpace (version 5.1), this article explains the available export options, my decision-making process, and the broad workflow steps I used to complete the migration. Since bepress provides extensive documentation of their batch uploading process, the importing of content into Digital Commons is not described here.

Project Description

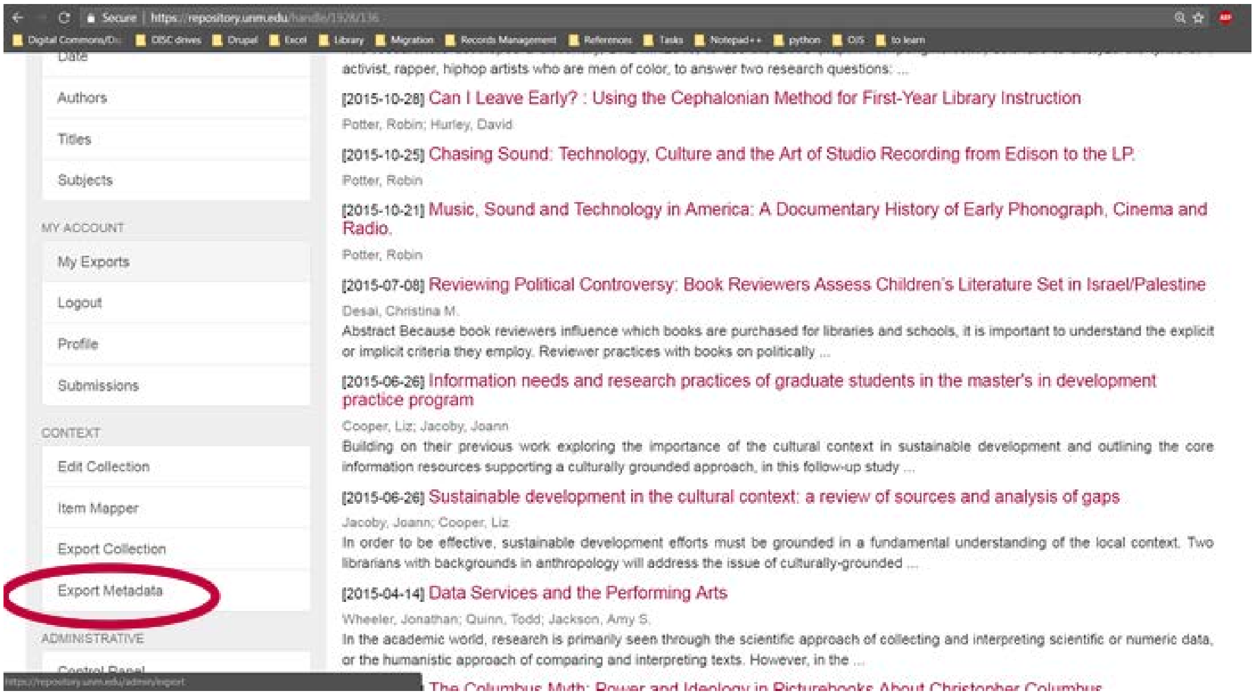
At the beginning of the project, I evaluated the content exported via two different options in DSpace. In most cases collections were individually exported, except where items were to be organized differently in the new repository. I wrote a Python processor that walks through the exported document folder structure and converts the XML metadata and the bitstream (file) URL into CSV. I then processed and cleaned the CSV files in Excel, in preparation for batch import into Digital Commons.

Obtaining Metadata from DSpace

UNM's installation of DSpace provides two simple, front-end interface based methods for exporting data.[1] (Your results may vary depending on the specific options selected for your installation.)

Option 1. Metadata Export

The first method is to export the metadata using the "Export Metadata" option in the "Context" section of the sidebar of the community or collection you would like to export. This option is available to logged-in administrators with appropriate permissions.


Secure | https://repository.unm.edu/handle/1928/136
Digital Commons/D... DISC drives Drupal Excel Library Migration Records Management References Tasks Notepad++ python OJS to learn
Date
Authors
Titles
Subjects
MY ACCOUNT
My Exports
Logout
Profile
Submissions
CONTEXT
Edit Collection
Item Mapper
Export Collection
Export Metadata
ADMINISTRATIVE
Control Panel
https://repository.unm.edu/admin/export

activist, rapper, hiphop artists who are men of color, to answer two research questions: ...

[2015-10-28] Can I Leave Early? : Using the Cephalonian Method for First-Year Library Instruction
Potter, Robin; Hurley, David

[2015-10-25] Chasing Sound: Technology, Culture and the Art of Studio Recording from Edison to the LP.
Potter, Robin

[2015-10-21] Music, Sound and Technology in America: A Documentary History of Early Phonograph, Cinema and Radio.
Potter, Robin

[2015-07-08] Reviewing Political Controversy: Book Reviewers Assess Children's Literature Set in Israel/Palestine
Desai, Christina M.
Abstract Because book reviewers influence which books are purchased for libraries and schools, it is important to understand the explicit or implicit criteria they employ. Reviewer practices with books on politically ...

[2015-06-26] Information needs and research practices of graduate students in the master's in development practice program
Cooper, Liz; Jacoby, Joann
Building on their previous work exploring the importance of the cultural context in sustainable development and outlining the core information resources supporting a culturally grounded approach, in this follow-up study ...

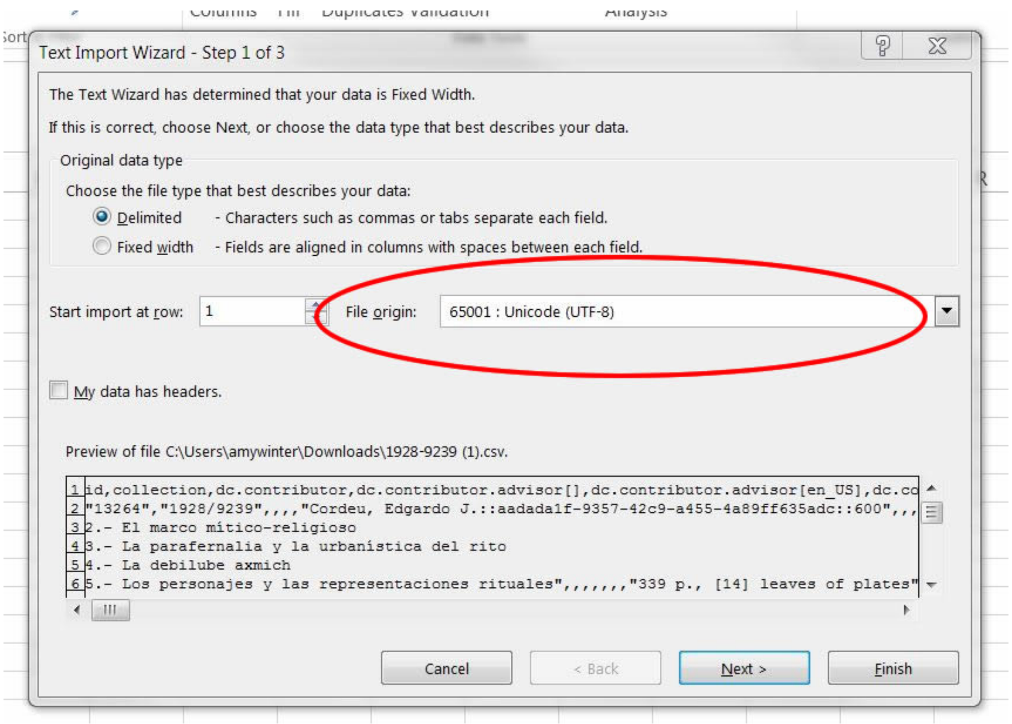
[2015-06-26] Sustainable development in the cultural context: a review of sources and analysis of gaps
Jacoby, Joann; Cooper, Liz
In order to be effective, sustainable development efforts must be grounded in a fundamental understanding of the local context. Two librarians with backgrounds in anthropology will address the issue of culturally-grounded ...

[2015-04-14] Data Services and the Performing Arts
Wheeler, Jonathan; Quinn, Todd; Jackson, Amy S.
In the academic world, research is primarily seen through the scientific approach of collecting and interpreting scientific or numeric data, or the humanistic approach of comparing and interpreting texts. However, in the ...

The Columbus Myth: Power and Ideology in Picturebooks About Christopher Columbus

Export metadata menu option

This provides ~~a~~ metadata output in csv format and can be helpful as a snapshot of the contents of the collection. Be sure to import this file into Excel using the wizard in the Data tab, and in Step 1 of the wizard, choose as the File origin: "65001: Unicode UTF-8," to avoid problems with character set mismatches.



Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Fixed Width.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

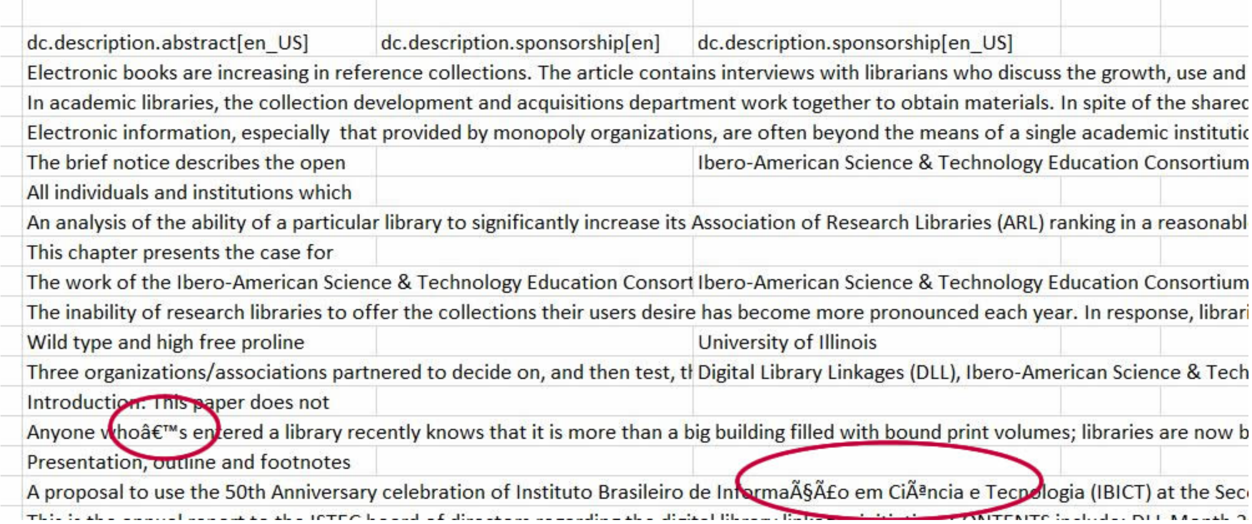
Start import at row: 1 File origin: 65001 : Unicode (UTF-8)

□ My data has headers.

Preview of file C:\Users\amywinter\Downloads\1928-9239 (1).csv.

```
id,collection,dc.contributor,dc.contributor.advisor[],dc.contributor.advisor[en_US],dc.co
"13264","1928/9239",,,,"Cordeu, Edgardo J.::aadada1f-9357-42c9-a455-4a89ff635adc::600",,,
2.- El marco mítico-religioso
3.- La parafernalia y la urbanística del rito
4.- La debilube axmich
5.- Los personajes y las representaciones rituales",,,,,,,"339 p., [14] leaves of plates"
```

If you open the export with Windows or other encoding in Excel, you may see errors with special characters:



dc.description.abstract[en_US]	dc.description.sponsorship[en]	dc.description.sponsorship[en_US]
Electronic books are increasing in reference collections. The article contains interviews with librarians who discuss the growth, use and		
In academic libraries, the collection development and acquisitions department work together to obtain materials. In spite of the sharec		
Electronic information, especially that provided by monopoly organizations, are often beyond the means of a single academic institutic		
The brief notice describes the open		Ibero-American Science & Technology Education Consortium
All individuals and institutions which		
An analysis of the ability of a particular library to significantly increase its Association of Research Libraries (ARL) ranking in a reasonabl		
This chapter presents the case for		
The work of the Ibero-American Science & Technology Education Consort		Ibero-American Science & Technology Education Consortium
The inability of research libraries to offer the collections their users desire has become more pronounced each year. In response, librari		
Wild type and high free proline		University of Illinois
Three organizations/associations partnered to decide on, and then test, tl		Digital Library Linkages (DLL), Ibero-American Science & Tech
Introduction. This paper does not		
Anyone whoâ€™s entered a library recently knows that it is more than a big building filled with bound print volumes; libraries are now b		
Presentation, outline and footnotes		
A proposal to use the 50th Anniversary celebration of Instituto Brasileiro de InformaÃ§Ã£o em CiÃªncia e Tecnologia (IBICT) at the Sec		
This is the annual report to the ISTEC board of directors regarding the digital library linkages initiative. CONTENTS include: DLL Month 2		

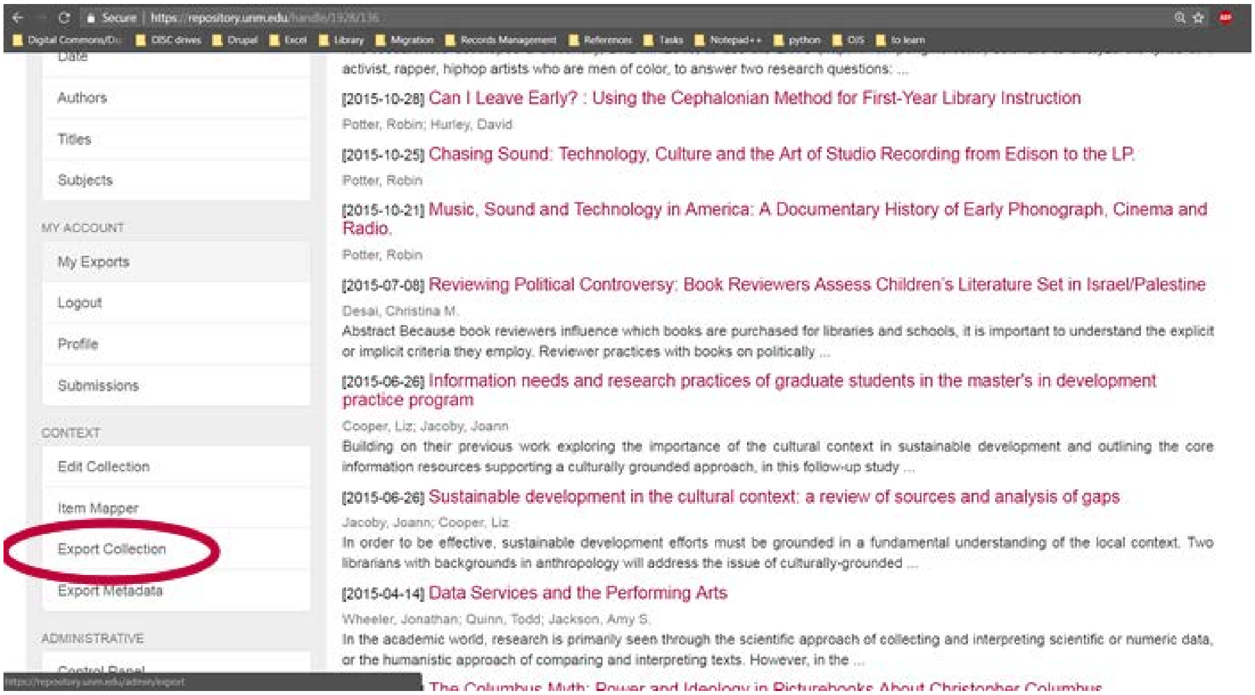
Character set interpretation and field duplication problems in DSpace metadata export

The major drawbacks to the metadata export, at least for UNM's installation, is that it included a number of fields we didn’t want to migrate; it didn’t contain some fields that we wanted (for example, provenance); and in some cases contained multiple fields for the same data.

Manually combining duplicate columns (in the above example, dc.description.sponsorship[en]” and “dc.description.sponsorship[en_US]”) for large collections becomes tedious. Google’s OpenRefine may better automate the cleaning of this export, but because I already had extensive Excel experience[2], I chose not to spend the time learning another application. For these reasons, for the most part I did not use this option to export metadata, preferring the export obtained in Option 2.

Option 2. Collection/Community Export

The second option for exporting data from DSpace is to use the "Export Collection" or "Export Community" feature, also available in the "Context" section of the sidebar.



activist, rapper, hiphop artists who are men of color, to answer two research questions: ...

[2015-10-28] Can I Leave Early? : Using the Cephalonian Method for First-Year Library Instruction

Potter, Robin; Hurley, David

[2015-10-25] Chasing Sound: Technology, Culture and the Art of Studio Recording from Edison to the LP.

Potter, Robin

[2015-10-21] Music, Sound and Technology in America: A Documentary History of Early Phonograph, Cinema and Radio.

Potter, Robin

[2015-07-08] Reviewing Political Controversy: Book Reviewers Assess Children's Literature Set in Israel/Palestine

Desai, Christina M.

Abstract Because book reviewers influence which books are purchased for libraries and schools, it is important to understand the explicit or implicit criteria they employ. Reviewer practices with books on politically ...

[2015-06-26] Information needs and research practices of graduate students in the master's in development practice program

Cooper, Liz; Jacoby, Joann

Building on their previous work exploring the importance of the cultural context in sustainable development and outlining the core information resources supporting a culturally grounded approach, in this follow-up study ...

[2015-06-26] Sustainable development in the cultural context: a review of sources and analysis of gaps

Jacoby, Joann; Cooper, Liz

In order to be effective, sustainable development efforts must be grounded in a fundamental understanding of the local context. Two librarians with backgrounds in anthropology will address the issue of culturally-grounded ...

[2015-04-14] Data Services and the Performing Arts

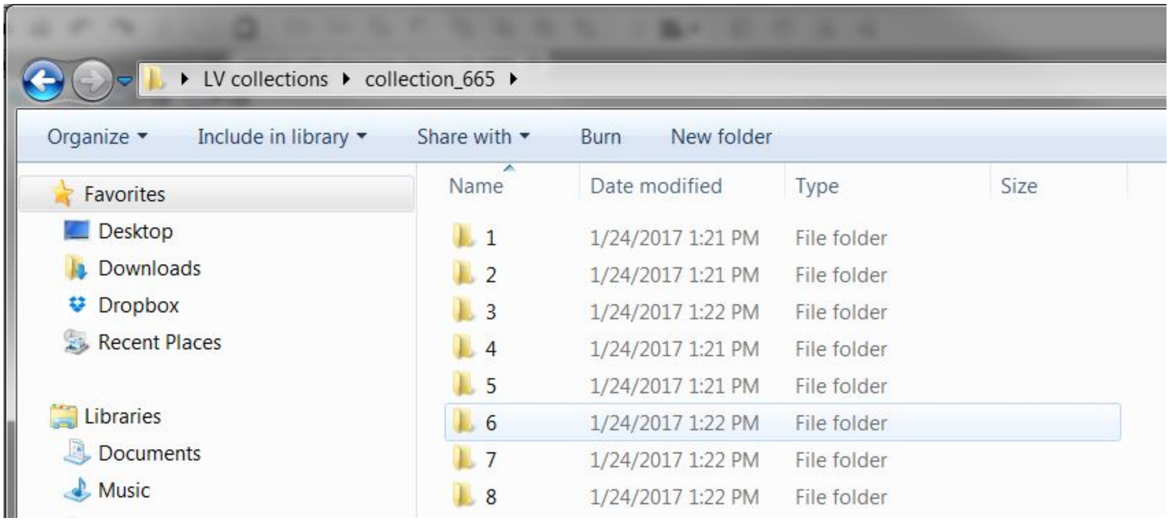
Wheeler, Jonathan; Quinn, Todd; Jackson, Amy S.

In the academic world, research is primarily seen through the scientific approach of collecting and interpreting scientific or numeric data, or the humanistic approach of comparing and interpreting texts. However, in the ...

The Columbus Myth: Power and Ideology in Picturebooks About Christopher Columbus

Export collection menu option

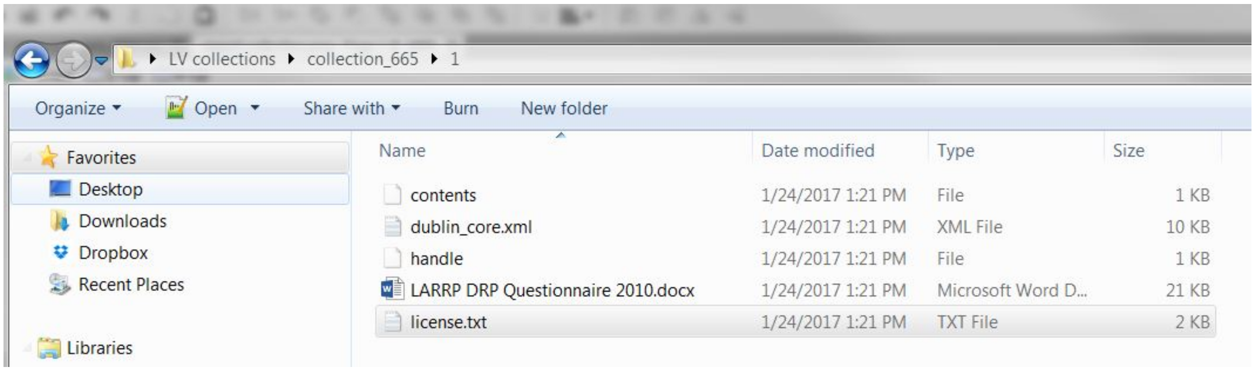
This option generates a zip file containing one folder for each item in the collection. If you export a community, a folder for that community is created, containing a folder for each collection within the community, which contains a folder for each item.



Name	Date modified	Type	Size
1	1/24/2017 1:21 PM	File folder	
2	1/24/2017 1:21 PM	File folder	
3	1/24/2017 1:22 PM	File folder	
4	1/24/2017 1:21 PM	File folder	
5	1/24/2017 1:21 PM	File folder	
6	1/24/2017 1:22 PM	File folder	
7	1/24/2017 1:22 PM	File folder	
8	1/24/2017 1:22 PM	File folder	

DSpace export package folder structure

Each item folder contains a number of files pertaining to the item.



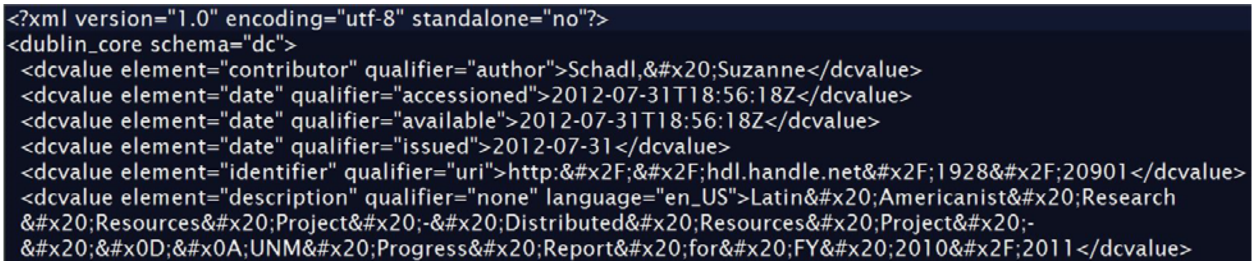
Name	Date modified	Type	Size
contents	1/24/2017 1:21 PM	File	1 KB
dublin_core.xml	1/24/2017 1:21 PM	XML File	10 KB
handle	1/24/2017 1:21 PM	File	1 KB
LARRP DRP Questionnaire 2010.docx	1/24/2017 1:21 PM	Microsoft Word D...	21 KB
license.txt	1/24/2017 1:21 PM	TXT File	2 KB

DSpace export package item folder contents

In this example, the "contents" file contains a reference to the license. The "handle" file contains the unique identifier for the item in the repository. The Word file is the actual document[3], and "license.txt" is the text of the license the user agreed to upon uploading the item. Along with the copy of the document(s), the most relevant file in this batch is "dublin_core.xml" which I will return to in a moment.

The first task for any script is to parse the exported folder structure. There are many ways to do this, but since I know a little Python, I chose to use the os library. In a few simple steps I could walk through the folder structure and extract the content I wanted into a list.

Python also has multiple libraries for parsing XML, but unfortunately the output from DSpace, in the dublin_core.xml file, was not well-formed XML.



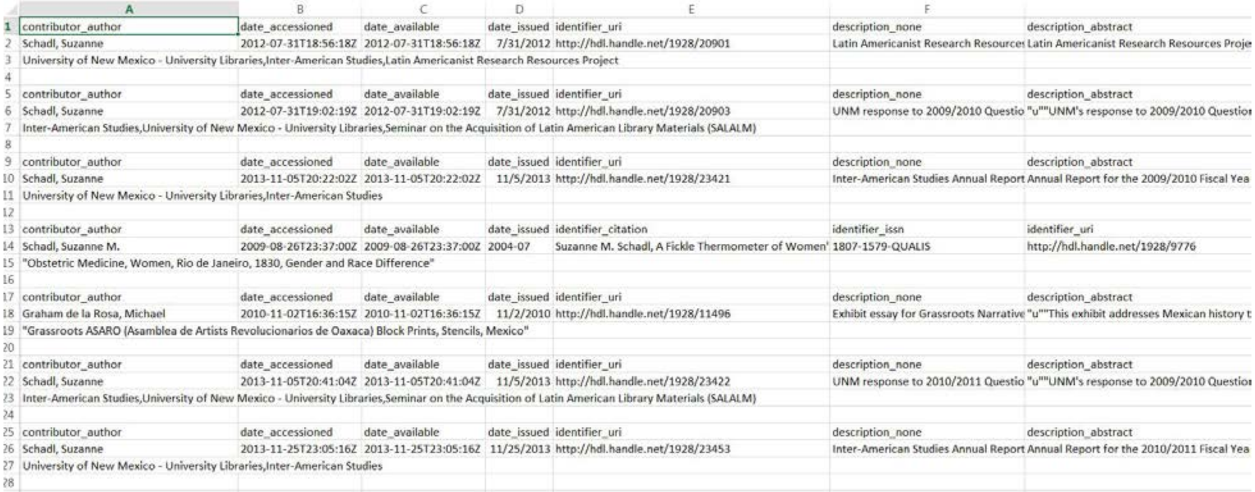
```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<dublin_core schema="dc">
  <dcvalue element="contributor" qualifier="author">Schadl,&#x20;Suzanne</dcvalue>
  <dcvalue element="date" qualifier="accessioned">2012-07-31T18:56:18Z</dcvalue>
  <dcvalue element="date" qualifier="available">2012-07-31T18:56:18Z</dcvalue>
  <dcvalue element="date" qualifier="issued">2012-07-31</dcvalue>
  <dcvalue element="identifier" qualifier="uri">http:&#x2F;&#x2F;hdl.handle.net&#x2F;1928&#x2F;20901</dcvalue>
  <dcvalue element="description" qualifier="none" language="en_US">Latin&#x20;Americanist&#x20;Research
  &#x20;Resources&#x20;Project&#x20;-&#x20;Distributed&#x20;Resources&#x20;Project&#x20;-
  &#x20;&#x0D;&#x0A;UNM&#x20;Progress&#x20;Report&#x20;for&#x20;FY&#x20;2010&#x2F;2011</dcvalue>
```

dublin_core.xml contents

All of the tags are <dcvalue> with elements and qualifiers referring to the DSpace data type and other attributes; my barely intermediate Python and nonexistent XSL skills were insufficient to the task of parsing this accurately.

Instead, I decided to use the Python library BeautifulSoup, designed for parsing HTML tags, and then output two lines of CSV per collection item (using the csv and cStringIO libraries).

The first line of delimited content shows the element and the qualifier from the <dcvalue> tag, and the second line contains the content of that field. This created a sort of header-row, item-row structure so that each piece of data had an identifying label right above it. The CSV output, imported into Excel, looks like this:



	A	B	C	D	E	F	
1	contributor_author	date_accessioned	date_available	date_issued	identifier_uri	description_none	description_abstract
2	Schadl, Suzanne	2012-07-31T18:56:18Z	2012-07-31T18:56:18Z	7/31/2012	http://hdl.handle.net/1928/20901	Latin Americanist Research Resources	Latin Americanist Research Resources Proje
3	University of New Mexico - University Libraries,Inter-American Studies,Latin Americanist Research Resources Project						
4							
5	contributor_author	date_accessioned	date_available	date_issued	identifier_uri	description_none	description_abstract
6	Schadl, Suzanne	2012-07-31T19:02:19Z	2012-07-31T19:02:19Z	7/31/2012	http://hdl.handle.net/1928/20903	UNM response to 2009/2010 Questio	"u""UNM's response to 2009/2010 Questior
7	Inter-American Studies,University of New Mexico - University Libraries,Seminar on the Acquisition of Latin American Library Materials (SALALM)						
8							
9	contributor_author	date_accessioned	date_available	date_issued	identifier_uri	description_none	description_abstract
10	Schadl, Suzanne	2013-11-05T20:22:02Z	2013-11-05T20:22:02Z	11/5/2013	http://hdl.handle.net/1928/23421	Inter-American Studies Annual Report	Annual Report for the 2009/2010 Fiscal Yea
11	University of New Mexico - University Libraries,Inter-American Studies						
12							
13	contributor_author	date_accessioned	date_available	date_issued	identifier_citation	identifier_issn	identifier_uri
14	Schadl, Suzanne M.	2009-08-26T23:37:00Z	2009-08-26T23:37:00Z	2004-07	Suzanne M. Schadl, A Fickle Thermometer of Women'	1807-1579-QUALIS	http://hdl.handle.net/1928/9776
15	"Obstetric Medicine, Women, Rio de Janeiro, 1830, Gender and Race Difference"						
16							
17	contributor_author	date_accessioned	date_available	date_issued	identifier_uri	description_none	description_abstract
18	Graham de la Rosa, Michael	2010-11-02T16:36:15Z	2010-11-02T16:36:15Z	11/2/2010	http://hdl.handle.net/1928/11496	Exhibit essay for Grassroots Narrative	"u""This exhibit addresses Mexican history t
19	"Grassroots ASARO (Asamblea de Artists Revolucionarios de Oaxaca) Block Prints, Stencils, Mexico"						
20							
21	contributor_author	date_accessioned	date_available	date_issued	identifier_uri	description_none	description_abstract
22	Schadl, Suzanne	2013-11-05T20:41:04Z	2013-11-05T20:41:04Z	11/5/2013	http://hdl.handle.net/1928/23422	UNM response to 2010/2011 Questio	"u""UNM's response to 2009/2010 Questior
23	Inter-American Studies,University of New Mexico - University Libraries,Seminar on the Acquisition of Latin American Library Materials (SALALM)						
24							
25	contributor_author	date_accessioned	date_available	date_issued	identifier_uri	description_none	description_abstract
26	Schadl, Suzanne	2013-11-25T23:05:16Z	2013-11-25T23:05:16Z	11/25/2013	http://hdl.handle.net/1928/23453	Inter-American Studies Annual Report	Annual Report for the 2010/2011 Fiscal Yea
27	University of New Mexico - University Libraries,Inter-American Studies						
28							

CSV metadata, written by Python script, imported into Excel

This method leaves a lot of cleanup to do in Excel:

- Cells are not always aligned and occasionally there are multiple columns for the same data type;
- Once cells are aligned correctly, all but the top “header” rows must be removed (a few steps or a macro can be used in Excel to automate this);
- Subject headings and URLs to the documents were handled separately.

With good Excel skills, or facility with OpenRefine or another data cleaning tool, this was an acceptable solution for someone with beginning to intermediate programming skills. Except for the largest and most complex collections, I was usually able to clean 2-3 collections per day, transfer the data and links to the Digital Commons batch upload sheets, and upload them to the new repository.

Code

The code referred to in this article (and an example XML file) is available on GitHub:

https://github.com/amyewinter/LV-scraper

Conclusion

After evaluating the structure and content of the metadata exports available from UNM’s installation of DSpace, I decided to use the Collection/Community export feature, process the exported XML metadata with Python, and then I cleaned the resulting output in Excel to ready it for batch uploading to Digital Commons.

Acknowledgements

Thanks are due to Jon Wheeler, Data Curation Librarian at UNM, for assistance in understanding how to access the data export options in DSpace.

Notes

[1] Jon Wheeler has written a more complex Python script whose output is almost completely ready for upload into Digital Commons. However, it requires the Digital Commons upload spreadsheet to be available before it can process the DSpace export. Generally, I preferred to have a better understanding of the exported metadata content by working with it directly, before building publication structures in Digital Commons.

[2] A good resource for quickly improving basic Excel skills, if needed, is the course “Excel 2013: Shortcuts” from lynda.com, if you have access to it.

[3] The fact that the export folders contain copies of the documents is useful. This provides a local backup of your repository contents (albeit one too cumbersome for use in any but the most extreme circumstances). It is also possible to navigate through the folder structure to quickly access and add supplemental files to migrated repository items once they are created in Digital Commons via batch upload.