Repurposing of metadata from a spreadsheet format to individual XML files for ingestion into DSpace

Jessica Davies

Student of the University of Cape Town
Department of CS,UCT,Rondebosch, 7701, RSA
(+27) 0847292793
dvsjes004@myuct.ac.za

Abstract

Whilst, storing intellectual content digitally has become a very important every day aspect of our lives, many people wish for a more effective and efficient way to manage this process. This paper looks at an example of this at UCT, creating a tool which improves the workflow of uploading digital assets into an intellectual repository. This tool allows users to input metadata in the simplest form (a spreadsheet), and converts it to individual XML files for each record. The spreadsheet is designed to make the entering of the data easier and the converter application successfully creates metadata files that can be ingested into DSpace (an example of an intellectual repository).

Categories and Subject Descriptors

H.3.2 [**Information Storage and Retrieval**]: File organisation; H.3.7 [**Digital Libraries**]: Collection; H.4.1 [**Office Automation**]: Workflow management; 5.4.2 [**Office Automation**]: Spreadsheets

General Terms

Metadata repurposing, ingestion into DSpace

Keywords

Metadata, XML files, spreadsheets, DSpace

Introduction

Information for the past few years has been moving from the physical form to the virtual form. However, it is no longer just necessary to be able to store all physical information virtually; one also has to store it in such a manner that ensures that the digital asset keeps up with technology and does not become obsolete (Ross, 2007). For academic institutions, institutional repositories are a solution to this. An institutional repository is an online archive of intellectual output of an institution which provides open access (Giesecke, 2011). An example of this is

DSpace, which is open source software available to anyone. DSpace allows digital assets (text, images, moving images, mpegs and data sets) to be archived and metadata containing information about these digital assets allows them to be easily browsed and searched for (DSpace Foundation (2011).

Metadata forms an integral part of digital objects by facilitating discovery, use, management & preservation of digital content. This means that the manner in which the metadata is entered and stored is very important. Repository software tools, like DSpace, ingest digital content and metadata in a predefined, well structured packaged format. Spreadsheets are among the most popular data management programs, however, they are not well suited for working with structured data. Specifically, DSpace's default metadata schema is Dublin-Core (DC) (DSpace Foundation, 2011) which contains 15 elements all of which are optional and can be repeated (Kurtz, 2010). DSpace's default SIP (Submission Information Packages) format is METS (Metadata Encoding and Transmission Standard) but other methods are used, such as the Simple Archive format for importing which was used in this project.

The goal of this research is to improve the design of a spreadsheet so as to easily create structured data and to create an application that will convert the metadata to individual XML files. The spreadsheet was improved through formatting and VBA code and the application was implemented in Java and converts a text version of an Excel spreadsheet into these files. The XML files are in a format that is compatible for DSpace ingestion. Sai Deng (2010) illustrates that such an application improves the work flow and ascertains that Excel is a useful source of metadata input. This in turn ensures the digital preservation of UCT's intellectual output.

XML files are ideal to represent structured data yet a lot of people using the system will find creating individual XML files tedious or beyond their scope of knowledge. Hence spreadsheets are the better option for entering the metadata. This process to convert spreadsheets for ingest into DSpace has been attempted before, most noticeably during the Google Summer of Code Project 2008 where Blooma created an application to allow for batch imports into DSpace (Sai Deng, 2010) which is run as an Add-on to DSpace. However, it was indicated in a survey that Blooma did that "majority of users (62%) developed their own customized program for the preparation of submission information packages" which is what this research intends to do (Blooma et. al. 2008).

Methodology

On conducting this research an Excel spreadsheet was used to capture the metadata (see Figure 1 for a screenshot of the spreadsheet). It was formatted so that each DC element has its own column and repeated elements lie on a separate row. Each row corresponds to one record but one record can have numerous rows depending on how many times an element has been repeated. VBA code was used to ensure that the elements have the correct entry ID number corresponding to their record. VBA code was also used to normalise the data; for example, the commas had to be replaced with a "pseudo comma" as commas are used as delimiters when the spreadsheet is saved in csv format. A comma within the element's text would create the impression of an extra column. The spreadsheet's aim is to be simple and easy to understand so as to avoid the extra effort of creating individual files for each record.

	A	B	C	D	E	F	G
1	**Dspace Metadata Spreadsheet**						
2	User's Title	Entry ID Number	contributor	coverage	creator	date	description
3		0	If there are multiple entries for an element, please enter each item of the same element in the row beneath				
4	Example 1	1	x	x	Hussein Suleman	2002/11/27	the view that greets you as you emerge the tunnel under the freeway - WOW - no the mountain isn't that close - it just that way in 2-D
5	x	1	x	x	x	x	x
6	x	1	x	x	x	x	x
7							
8							
9							
10							

Figure 1: Spreadsheet to capture metadata

The effectiveness of the design of the spreadsheet was tested using user performance studies via a questionnaire. The participants filled in a set of sample data given to them in a text file, entered it into the spreadsheet and then answered the questionnaire based on their experience. The questionnaire accessed the participants' experience with metadata and DSpace and required them to rate design factors as well as efficiency.

The application to convert the text version of the excel spreadsheet (CSV file) into XML files is written in Java. It reads the CSV file, manipulates the data so as to store it in a series of lists and then writes each record in its own XML file containing DC elements (see Figure 2 for a sample).

To ingest the file into DSpace, the Simple Archive format was used. This involves creating an archive hard drive, which is a directory full of items with a subdirectory per item. Each item directory contains the metadata data file named dublin_core.xml, the file being ingested (the digital asset) and a textfile named contents, which contains one line consisting of the digital asset's file name.

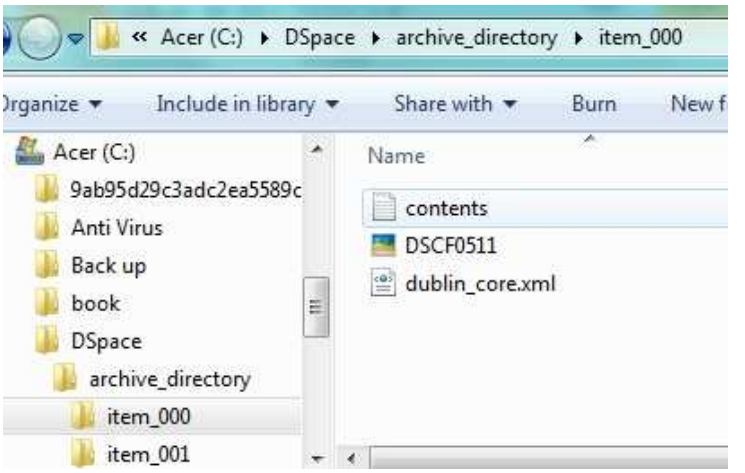

Figure 3: Example of the archive directory

```
- <dublin_core>
    <dcvalue element="creator" qualifier="none">Hussein Suleman</dcvalue>
    <dcvalue element="date" qualifier="none">2002/11/27</dcvalue>
    <dcvalue element="description" qualifier="none">the view that greets you as you emerge
      from the tunnel under the freeway - WOW - and, no the mountain isn't that close - it
      just looks that way in 2-D</dcvalue>
    <dcvalue element="format" qualifier="none">image/ jpeg</dcvalue>
    <dcvalue element="identifier"
      qualifier="none">http://www.husseinsspace.com/pictures/200230uct/02ust1.jpg</dcvalue>
    <dcvalue element="language" qualifier="none">en-us</dcvalue>
    <dcvalue element="publisher" qualifier="none">Hussein Suleman</dcvalue>
    <dcvalue element="relation" qualifier="none">http://www.husseinsspace.com</dcvalue>
    <dcvalue element="relation" qualifier="none">http://www.uct.ac.za</dcvalue>
    <dcvalue element="rights" qualifier="none">unrestricted</dcvalue>
    <dcvalue element="subject" qualifier="none">Visit to UCT</dcvalue>
    <dcvalue element="subject" qualifier="none">UCT</dcvalue>
    <dcvalue element="subject" qualifier="none">2002 travels</dcvalue>
    <dcvalue element="title" qualifier="none">02uct1</dcvalue>
    <dcvalue element="type" qualifier="none">image</dcvalue>
  </dublin_core>
```

Figure 2: Metadata using Dublin_Core XML file example

DSpace is then accessed through the command the command prompt and this line of code is used: `import -a -e dvsjes004@myuct.ac.za -c 123456789/6 -s C:\DSpace\archive_directory -m mapfile`

`Import` indicates importing into DSpace, `-a` indicates adding data and `-c 123456789/6` is the reference to the collection where the digital asset is being imported to in DSPace. `-s C:\DSpace\archive_directory` is the location of the archive file on the drive and `-m mapfile` is a file created which stores the mapping of item directories to item handles.

Results

The questionnaires were answered by 4 of the 11 people to whom they were sent (from the postgraduate lab at UCT), all of whom possess expert computer experience. All were familiar with the concept of metadata, 2 out of 4 have had experience with DSpace and those who did not have worked with other metadata managing systems (e.g. Green Stone, SA-NETD Mashup, EPrints, Fedora and SimplyCT). Each participant managed to fill in all the information in the spreadsheet correctly and 50% thought that this was a quicker process than filling out a web form for each record.

The questionnaire allowed for comments and some noteworthy ones were “it is difficult to see which line is associated with which item”, “The columns are well labelled, and the system allows flexibility to allow multiple entries for each object” and “the floating “Done” button seems out of place”.

The application successfully managed to create individual XML files for each record that were in DC schema. Ingestion into DSpace using the Simple Archive format was also successful.

Discussion

The results of the questionnaire can be divided up into different sections: usability, effectiveness and potential problems. The usability questions were rated on a scale of 1 to 5 (1 being very poor and 5 being very good) and three aspects were assessed (understandabilty, ease of navigation and aesthetics). Figure 4 shows the average rating for each aspect.

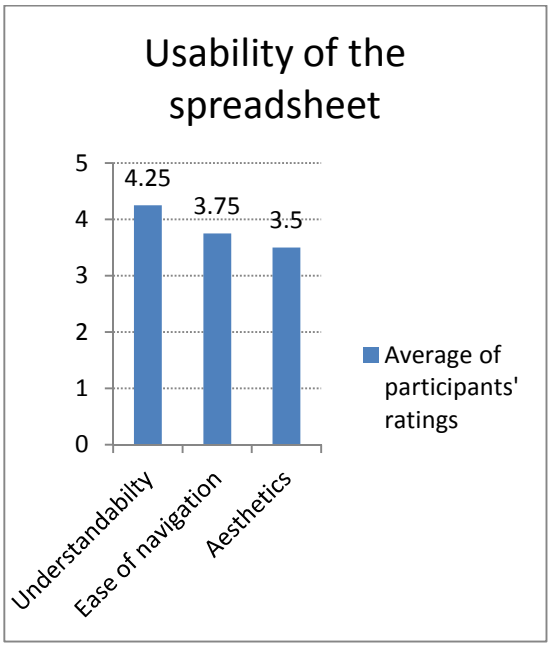

Figure 4: Bar chart indicating the usability of the spreadsheet

Whereas the usability was assessed through the questionnaire, the effectiveness of the spreadsheet was judged on how complete the returned spreadsheets were. One spreadsheet had all the metadata entered correctly but none of the metadata entered related to a record number. Consequently, the converter application would be unable to create the correct XML files. This was because the participant had not clicked the “Done” button after every record but only after filling in the information for all records.

Whilst the overall feedback was positive and most participants thought this method an efficient one for dealing with metadata, some issues were raised. The participants pointed out that using code in a spreadsheet application can be problematic as the macros might not be transferrable across different applications or operating systems. Moreover, one of the participants actually disabled all VBA code because he/she was not sure of the source of the active elements of the code. The participant then explained this was due to security reasons. This can be avoided by a warning of active code being implemented. The error mentioned in the previous paragraph is attributed to the ambiguity of the "Done" button, as it is not clear when to click it. Lastly, it was suggested that a more effective way to deal with empty fields should be found.

After following the method for ingesting into DSpace described earlier, the digital asset along with the metadata was successfully uploaded into DSpace. This indicates that the XML files created by the converter application are in a format suitable for DSpace ingestion.

This program is an example of metadata repurposing which Foulonneau and Cole (2005) describe as when metadata is converted into a format for use in a different application than originally created in which is exactly what the application does. It takes metadata in a spreadsheet and changes it into an acceptable format for storage on DSpace.

Conclusion

This research indicated that a spreadsheet can be used as an effective tool for storing structured data. However, this may not be the case for very large amounts of metadata and other options, e.g. databases, would have to be considered. One also has to look into the possibilities of subfields within the elements and how those would be dealt with within the current system. Hyperlinks within the spreadsheet are a possible solution. Despite these considerations, the metadata entered into the spreadsheets was successfully converted into XML files and furthermore, these files were successfully ingested into DSpace. The converter application can also be expanded to include code that ingests the files into DSpace without the user having to do that personally. Thus the spreadsheet and converter can be used by UCT as tool in using DSpace as an intellectual repository and help preserve UCT's intellectual content.

References

Blooma, M. J., Kurian, J. C., & Lewis, S. 2008. *An add-on to facilitate the existing DSpace Batch Import Procedure*. Google summer of code projects 2008. Retrieved from http://wiki.dspace.org/index.php/Google Summer of Code
2008 Batch Import Program can be downloaded at: http://code.google.com/p/dspace-gsoc/downloads/list

DSpace Foundation. 2011. *DSpace 1.8.1 Manual*. Retrieved from http://www.dspace.org/

DSpace Foundation. 2011. *About DSpace.* Retrieved from http://www.dspace.org/introducing

Foulonneau, M., & Cole, T. W. (2005). *Strategies for reprocessing aggregated metadata.* In 9th European Conference on Digital Libraries, ECDL 2005, September 18–23, 2005, Vienna, Austria. *Lecture Notes in Computer Science*. Heidelberg, Germany: Springer-Verlag.

Giesecke, J. 2011. Institutional Repositories: Keys to Success. *Journal of library administration.* 51(5):529-542.

Kurtz, M. 2010. Dublin Core, DSpace, and a Brief Analysis of Three University

Repositories. *Information technology & libraries.* 29(1):40-46.

S. Ross (2007), *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, Keynote Address at the 11th European Conference on Digital Libraries (ECDL),
Budapest (17 September 2007)

Sai Deng. 2010. Optimizing Workflow through Metadata Repurposing and Batch Processing. *Journal of Library Metadata*. 10:4, 219-237