Batch Import into Dspace: Digital Microfilm Instance

Xiaoxuan Tang
Systems and Electronic Resources Librarian
Randolph C. Watson Library
Kilgore College
Kilgore, Texas. USA
atang@kilgore.edu

ABSTRACT: *The value of microfilm in preserving history is huge and cannot be measured by money. Kilgore Daily News is such a newspaper which records and inscribes important history in Kilgore City since 1930s. However, some factors, including microfilm life limit, usage abrasion, and accidental temperature change are shortening the life of microfilm. In addition, some characteristics of microfilm itself, such as needing support of microfilm reader, not allowed charging out, difficult to retrieve, block the usage and access of them to the public. Under these situations, it is a great idea to digitize microfilm and make them accessible to the public online. After consideration, DSpace is chosen to manage and publicize the content of digital microfilm.*

Keywords: Digital Archiving, Microfilm, Digitization

Received: 1 September 2011, Revised 6 November 2011, Accepted 14 November 2011

1. Introduction

It is compulsory and obligatory to digitize these microfilms and make them public. There is a trend every year more and more people come by to retrieve the content in microfilm and write biography about their family and grandparents. However, the severe abrasion of microfilm in reader day by day already makes some of them vague and illegible. Furthermore, from microfilm preservation theory, the life span of microfilm is shortened under high temperature and humid environment[1]. From another side, under the library policy, these microfilms have the usage limit which prohibits them to check out. However, many old patrons in local or neighbor cities cannot drive a long distance here to browse them. And, some other patrons are too busy to come within library working hour. Under this situation, it is absolutely required to digitize these microfilms and make them accessible online.

Digitizing is the prerequisite to make microfilm online. A microfilm scanning machine is in use to transfer pictures from microfilm to computer and convert them to digital files. In order to make the converted digital files embrace high quality, a detailed inspection during and after transfer is necessary and mandatory. For the purpose of making it accessible to public, and easily to print, Acrobat PDF file format is adopted. Scanned pictures are embedded into PDF file and decrease the risk they are tapered.

2. Choose the system to publicize microfilm

With digitizing in process, it is time to compare and determine what kind of content management system is appropriate to use. It is a painful decision of which content management system we should pick up. Obviously, lightweight CMS cannot undertake the responsibilities of digital microfilm publication, for the reason of that they cannot support heavy load of data retrieval. So, heavyweight repository systems are merely in consideration. Besides, budget is another problem need to consider. Two of them,

Fedora-Common and DSpace, are compared in order to pick up the appropriate one.

DSpace is an open-source software package that provides management tools for digital assets[2], which makes use of Dublin core as data storage format. The biggest advantages of DSpace are open standards and mature[3]. The function is easy to adjust, and the incoming contents are easy to add in any time. The structure of Dspace allow configuration be made in scalable level.

Fedora provides a conceptual framework and foundation upon which some customized high-efficiency repository system can build.[4] It means it is not an out of the box product, but a flexible framework which you can build any content management or repository system you want above them. If the academic units or universities have strong programming team and abilities, Fedora is a fantastic platform to fulfill the grand design of them.

However, until now, we don't have enough programming staff on this project. So, we decide to take the out of the box choice, DSpace.

3. The prerequisite of Batch process

Importing digital microfilm into DSpace in batch is totally different from the traditional way to add record into system. The usual way is to add records into system one by one through the DSpace Administration module. However, it is not a good choice to employ this way when you face the needs to import tons of digital items. Hence, the better way to handle them is to import in batch. For the purpose of making the batch process smooth, first of all, the name of files which represent digital items should be tailored and re-organized. Second step, the metadata generated from microfilm should be hammered and shaped in XML format to well match each of digital items. Third step, each digital item file and metadata file, should be put into individual child folder, and be differentiated by their folder name. Fourth step, DSpace script is employed to import the well-organized metadata and digital items.

4. Metadata preparation and process on digital item

DSpace use Dublin core as metadata format of digital items. A decision must be made on how many fields are required to demonstrate the content of digital items. From now on, our library has no budget to purchase professional OCR to convert contents in microfilm to full text. In addition, during digitizing, historical newspaper is more difficult to generate full text compared with nowadays documents. Some problems, such as deteriorated and vague originals, unusual font, faded printing, shaded background, skewed text, and curved lines, definitely makes text identification accuracy low.[5] Besides, until now, we have no enough employees and time to read microfilm and extract metadata one by one. So, a relatively succinct metadata design is adopted like below.

Full metadata record

DC Field	Value
dc.contributor.author	Kilgore Daily News
dc.date.accessioned	2011-07-02T15:45:01Z
dc.date.available	2011-07-02T15:45:01Z
dc.date.created	04/27/2011
dc.date.issued	07/01/1931
dc.identifier.uri	http://ezproxy.wildcatter.kilgore.edu:2048/login?url=http://library.kilgore.edu:8081/jspui/handle/123456789/111
dc.language.iso	en
dc.publisher	Kilgore News Herald
dc.subject	Newspaper
dc.title	Kilgore Daily News 07/01/1931
dc.type	Digital Microfilm
Appears in Collections:	Kilgore News Herald

Figure 1. Dublin Core fields of Kilgore Daily News digital microfilm collection

The design of digital microfilm file name is significant and determined by patron usage mode and habit. Patrons mainly search for date to retrieve the newspaper, so combining date with newspaper title is a good choice for metadata design. All files name are tailored in this format, KNH_mm_dd_yyyy_pdf (mm is month, dd is day, and yyyy is year), in that the minimum unit of digital newspaper is day.

Then, under different operation systems, the process of file name is a little tricky. A command under Linux system, "ls >filename.txt", will generate filename list without problem. Under windows operation system, a command like "dir > filename.txt" will create filename list, but mixed with the fields of date and file size.

In order to offset the disturbance from unrelated fields, some additional process is required in windows. For instance, by using Excel, it is easy to merely preserve filename field and discard unrelated ones. By using auto number, formulae, and auto format function of Excel, the metadata, such as author, issue date, subject, title, and publisher, can be generated easily.

Metadata of digital microfilm will use XML as data format. DSpace put XML format as the metadata container. The normal process to build metadata by hand is as below (see figure 2) through copy and paste function. However, it is very painful and troublesome to embed thousands of digital microfilm metadata to XML file manually. Hence, we need to figure out a way to make this embedding process automatic. The first step, Excel is used to pre-format data. The second step, a customized program will be designed to help building SIP (Submission Information Package).

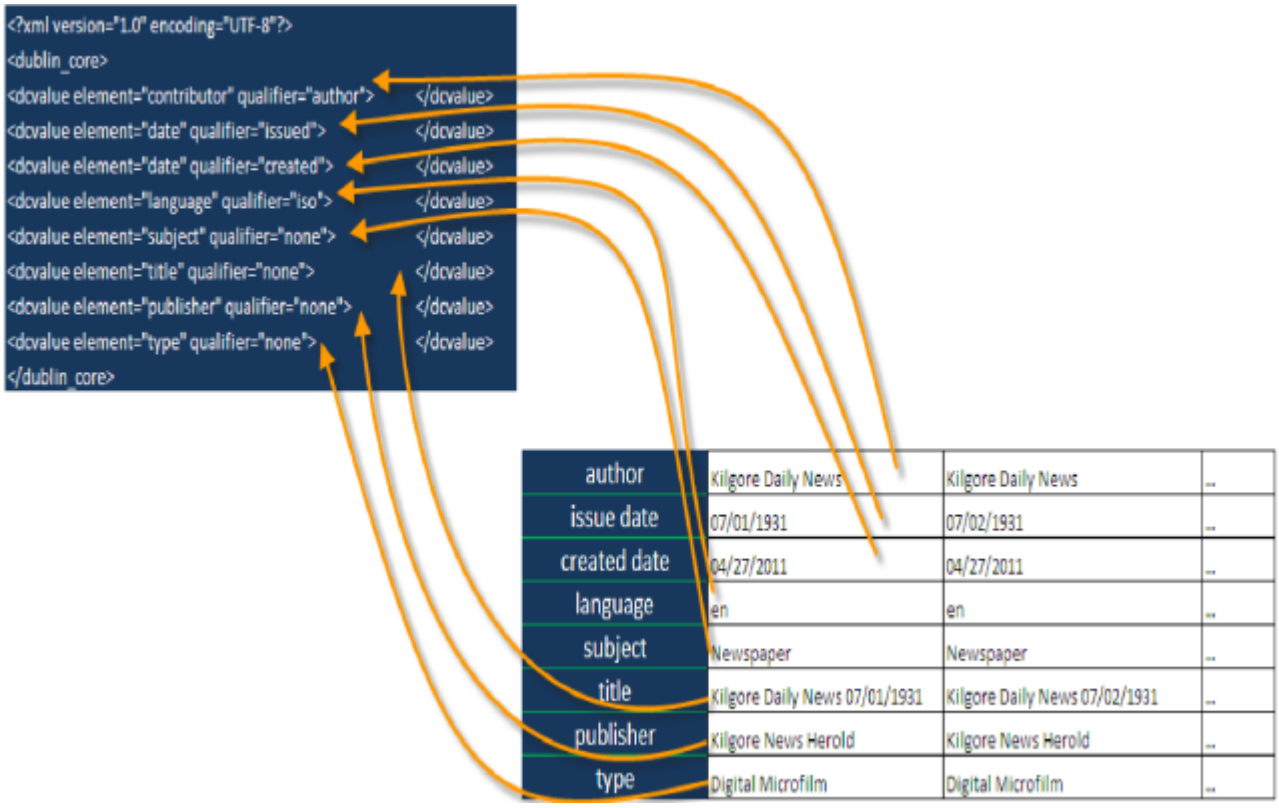

Figure 2. Embedding metadata in XML

Using Excel formula to embed metadata into XML file is highly efficient. An excel template is created, and formulae are pre-injected to make the process automatic (http://library.kilgore.edu/libraryapplication/DSpace-excel.xlsx). By using Excel and concatenate functions, all of the microfilm metadata are embedded into XML profile. Thereafter, all of the XML metadata are exported as one accumulated file, for preparation of usage by DSpace system lately.

The next step is to prepare the SIP (Submission Information Package) (see figure 3). The item concept in Dspace is so special,

that we have to pay attention to that. The "*item*" in Dspace may contain and enclose several digital files in one container, like a bucket.[6] Each metadata file and corresponding digital items, as a unit, should be put into individual child folders. All of the units will be differentiated by their folder name, and authoritative under mother folder. Until now, accumulated XML metadata file is produced, and digital microfilm files are created. Because Kilgore Daily Newspaper is in unit and measurement of day each "*item*" merely requires matching only one digital file. However, if you like, you can make every "*item*" matches or accommodates more than 2 digital files. Hereafter, each child folder, which represents container or bucket, will at least encompass or enclose at least three files, including "*contents*" (pointer), Dublin_core.xml(metadata), and real digital microfilm file (see Figure 3).

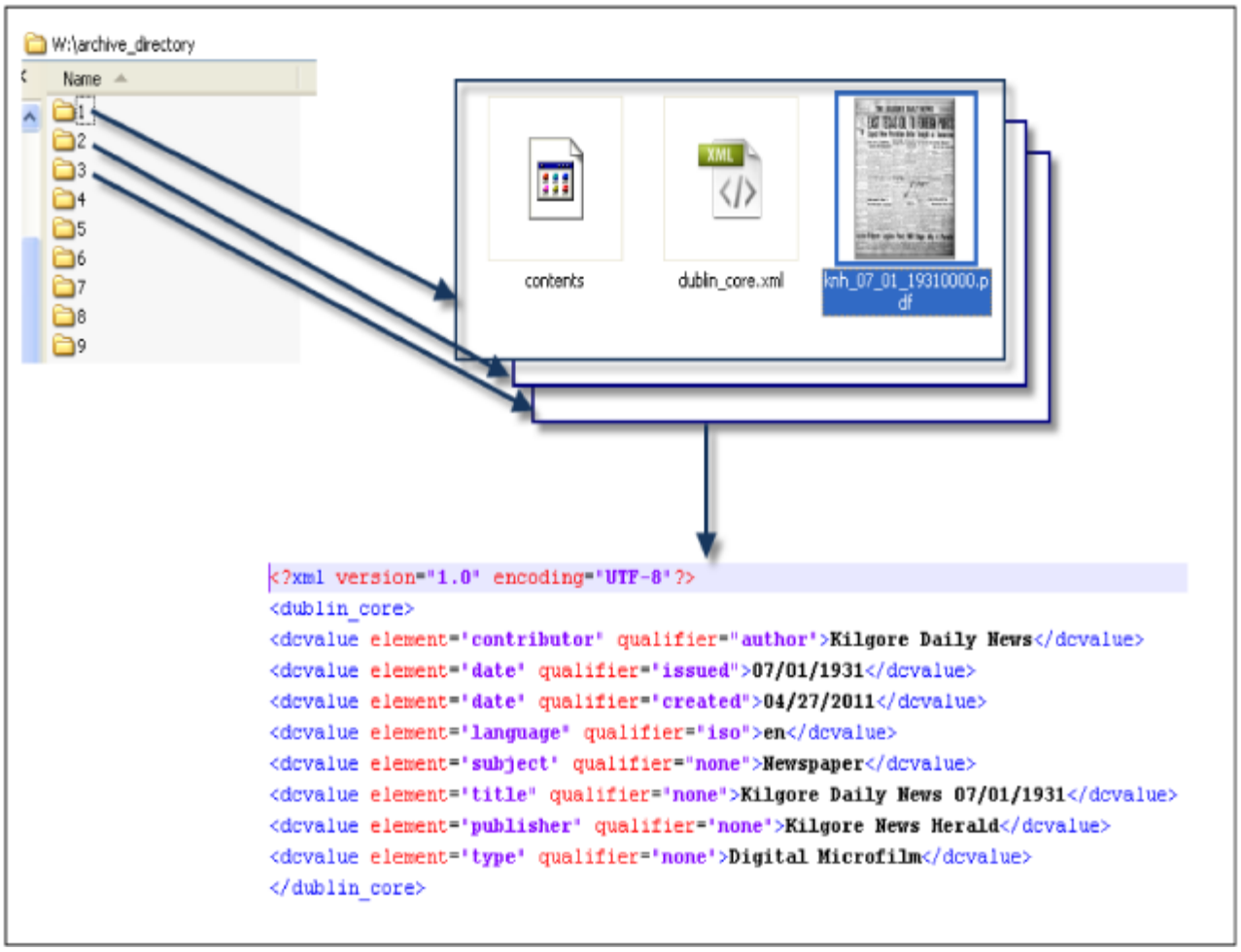

Figure 3. SIP (Submission Information Package) preparation

Now, we meet the problem on a very detailed layer on how to generate the file named "*contents*" in each child folder, how to tailor metadata against each "*item*" and save them into each child folder, and how to move each digital microfilm file to appropriate folder. Obviously, if facing thousands of digital files, it is not wise to do it manually.

In order to overcome the difficulty, a small program/procedure is compiled to help fulfilling this boring task in acceleration and solve the problem automatically (see figure 4). The coding language of the program is Visual Foxpro, which is easy to understand and effortless to transfer to Visual Basic. The code has been compiled to EXE file and easily to run at any computers with Windows system. The EXE file is available in http://library.kilgore.edu/libraryapplication/DSpace_SIP_Prepare.zip. The source code of this program is available in Appendix 1. Before using this program, we not only need to download and extract it to a folder, but also require moving the metadata XML file, filelist file, and digital microfilm PDF files into that folder. After that, this folder becomes the "*Default Work Folder*". If clicking "*Clean Child Directory*", any folders under the "*Default work Folder*" will be cleansed. Then, we tell this program the location of XML metadata file, filelist filename, and how many items we would like to add. After clicking on "*Run*" button, all of the digital files, description file, and related metadata XML will be processed,

assembled, and moved into appropriate folders. Until now SIPs are created and prepared to be imported in Dspace. Hereafter, an out-of-box script from DSpace will be employed to import all of data from mother and child folder, into DSpace.

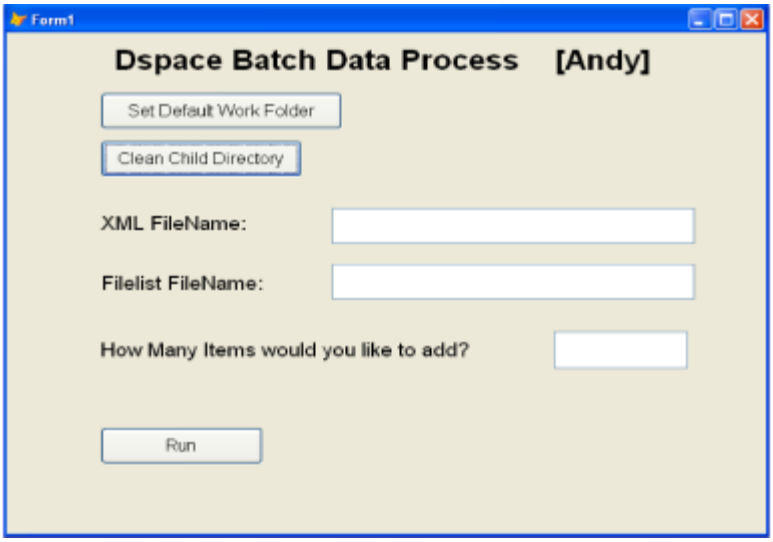

Figure 4. DSpace Batch Data Process Procedure

5. Import data into DSpace

DSpace provides the script of batch importation (see figure 5) for usage. First of all, we copy the item folders, including mother folders and child folders, into an archive directory in DSpace server, to form SIP. Afterwards, by using script like below[7], all of digital microfilm files and related metadata are imported into this collection in DSpace.

"*DSpace--add--eperson = somebody@kilgore.edu--collection* = 123456789/2*--source=c:\archive_directory--mapfile =c:\maps_archive*\06202011.*txt*"

Command used:	*[dspace]* /bin/dspace import
Java class:	org.dspace.app.itemimport.ItemImport
Arguments short and (long) forms:	Description
-a or --add	Add items to DSpace ‡
-r or --replace	Replace items listed in mapfile ‡
-d or --delete	Delete items listed in mapfile ‡
-s or --source	Source of the items (directory)
-c or --collection	Destination Collection by their Handle or database ID
-m or --mapfile	Where the mapfile for items can be found (name and directory)
-e or --eperson	Email of eperson doing the importing
-w or --workflow	Send submission through collection' workflow
-n or --notify	Kicks off the email alerting of the item(s) has(have) been imported
-t or --test	Test run—do not actually import items
-p or --template	Apply the collection template
-R or --resume	Resume a failed import (Used on Add only)
-h or --help	Command help

Figure 5. Batch import command description

After the batch import of digital microfilm is completed, the final result of importation is demonstrated as below. (See figure 6)

Figure 6. Kilgore College digital microfilm repository

6. Conclusion

DSpace is an out of the box institutional repository software with excellent design. It can handle large scale of digital microfilms in server with good performance. However, metadata preparation before batch import still depends on programming ability of patron or some third-party software. If DSpace can develop some easy-used online tools to simplify the metadata preparation work, it will save patron a lot of time and decrease heavy workload.

Online OCR function is another function we need very much. Maybe DSpace can develop some plug-in to solve this problem, or we can complete it ourselves in future.

```
*****************************************************************
* Author: Andy Tang
* Compiled Date: 06/20/2011
* Copyright: Open to Public!
*****************************************************************
CLOSE All
CLEAR ALL
CLEAR

ii=1
FOR ii=1 TO 106
```

```
&&Clear the child directory and avoide the contradiction.
bb=STR(ii)
   IF DIRECTORY (bb)=.T.
      loFSO  =  Createobject("Scripting.FileSystemObject")
       loFSO.DeleteFolder(&bb)
   ENDIF

ENDFOR

*****************************************************************
* The beginning to process the xml folder and xml file
*****************************************************************

&&book3_1.xml is accumulated metadata file name
stringa=FILETOSTR( "book3_1.xml")

&&Create a array with 106 position. If you have more data need to import, change this.
DIMENSION arrayt(106)
isolator=LEN("</dublin_core>")
jj=1
FOR jj=1 TO 106
                              &&arrayt(jj)=STR(jj)+" "+"victory"
                              &&? arrayt(jj)
         IF jj=1
                              arrayt(jj)=SUBSTR(stringa,1, ATc("</dublin_core>", stringa,1)+isolator)
  ELSE
         arrayt(jj)=SUBSTR(stringa, ATc("</dublin_core>", stringa,jj-1)+isolator+2,ATc("</dublin_core>", stringa,jj)+isolator-
(ATc("</dublin_core>", stringa,jj-1)+isolator))
         ENDIF
ENDFOR

mm=1

FOR jj=1 TO 106
   MKDIR ALLTRIM(STR(jj))
   CD ALLTRIM(STR(jj))
            IF FILE('dublin_core.xml')

            ELSE
               filehandle=fcreate('dublin_core.xml')
               =FWRITE(filehandle,arrayt(jj))
              =FCLOSE(filehandle)
            ENDIF

   cd..
endfor

*****************************************************************
* The beginning to process the filename list
*****************************************************************

&&filename.txt is the name list of all digital microfilm files in this folder.
stringb=FILETOSTR("filename.txt")
```

References

[1] Robek, Mary, F., Gerald, Brown, F., David, O. Stephens. (1995). Information and records management: document-based information systems. New York: GLENCOE/McGraw-Hill.

[2] Kurtz, Mary. (2010). Dublin Core, DSpace, A Brief Analysis of Three University Repositories. Information Technology & Libraries 29 (1) 40-46. Academic Search Complete, EBSCOhost (accessed July 11, 2011).

[3] An Introduction to DSpace, Stuart Lewis & Chris Yates, http://cadair.aber.ac.uk/DSpace/bitstream/handle/2160/617/ (accessed July 1, 2011)

[4] Getting Started with Fedora,Fedora Development Team, https://wiki.duraspace.org/display/FCR30/Getting+Started+with+Fedora (accessed June 20, 2011)

[5] Microfilm, Paper and OCR: Issues in Newspaper Digitization, Kenning Arlitsch and John Herber, http://digitalnewspapers.org/public/pdf/MicroFilmArticle.pdf (accessed June 22, 2011)

[6] DSpace Batch Import Format, Texas Digital Library, http://www.tdl.org/wp-content/uploads/2009/04/DSpaceBatchImportFormat.pdf (accessed June 23, 2011)

[7] DSpace Manual, DSpace Development Team, http://www.DSpace.org/1_7_1Documentation/DSpace-Manual.pdf (accessed June 27, 2011)