Watch out for the potholes: The bumpy road of digitizing a historical collection

Tony Chu, PhD, MLIS '14

Center of Alcohol Studies
Rutgers, The State University of New Jersey

The collection for digitization

The major task of the digitization project the author involves at the Rutgers Center of Alcohol Studies library is for digitizing the historical collection of Ralph G. Connor Alcohol Research Reference Files (CARRF). The Ralph G. Connor Alcohol Research Reference Files is an archive of instruments (questionnaires, interview schedules, surveys, etc) that have been used in studies of substance abuse. The primary focus of this collection is on various aspects of drinking behavior and alcoholism.

The society for Study Problems founded this collection, which is now named in memory of Dr. Ralph G. Connor, in 1960. The Files were developed and maintained by Dr. Connor at Eastern Washington College, and following his death they were transferred to the instruments to the Center of Alcohol Studies at Rutgers University. The Center has continued to add new instruments to the collection, and periodically issues new updates to the descriptive inventory providing brief information on the materials.

The tasks of digitization and metadata construction

When building a digital library, the digitization of a historical collection can be done by (1) digitizing and indexing each of the print documents one-by-one, (2) digitizing all the print documents, importing them into the new system, and then indexing each of them, or (3) digitizing all the print documents, importing the existing metadata from the old system, and matching the two parts in the new system. The third option is often considered because the metadata is typically available as an online public access catalog (OPAC). For efficiency, this option is also prefered by reducing the operational time.

Issues could arise with the third option. Using the digitization of our library's historical collection as a case study, we documented two such issues. First, digitizing the print documents, our staff had included the archive ID and title of each document as part of the filename, which helped with general identification, but did so with a lack of delimiters necessary to parse the information with data mining and statistical software such as Statistical Package for the Social Sciences (SPSS). For matching other information with the file ID number, this caused problems when attempting to separate the ID number from the document title.

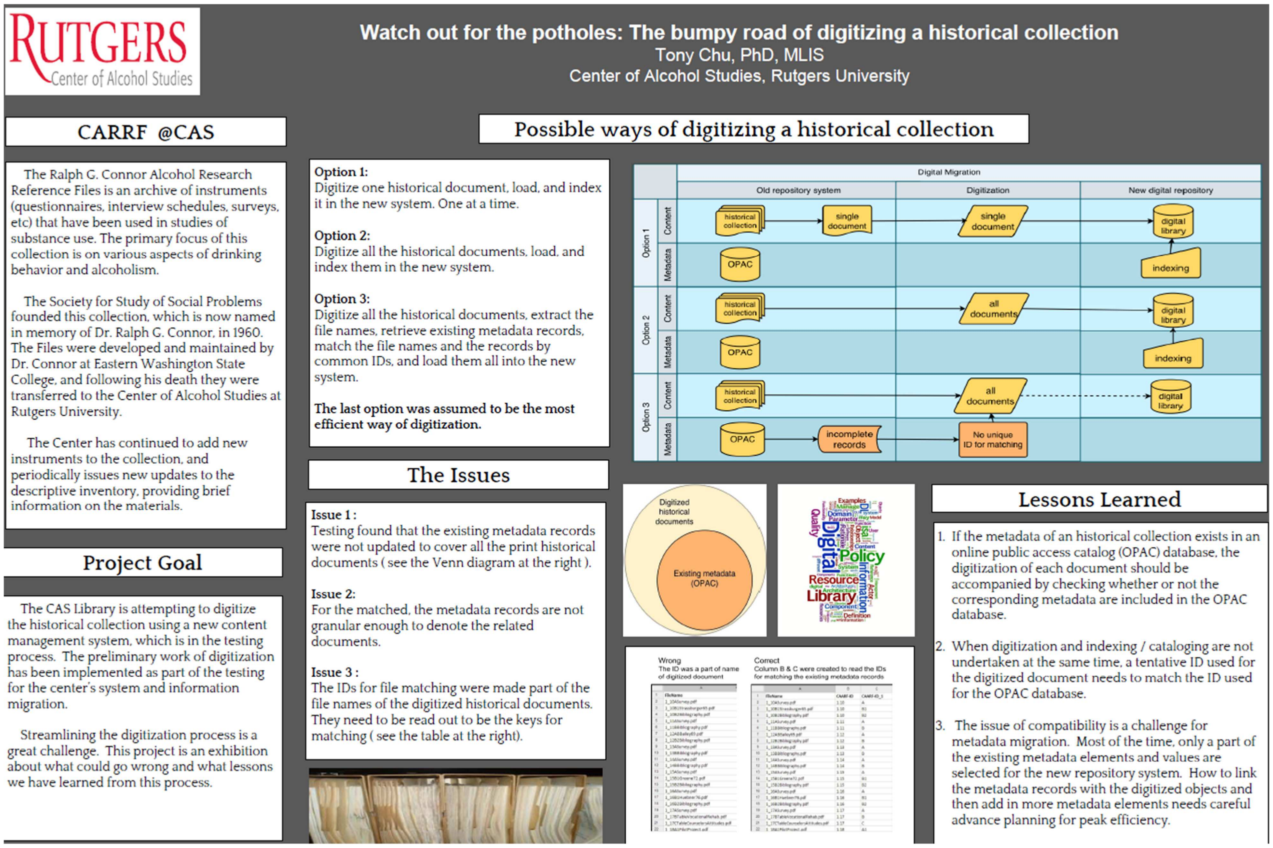
The second problem was about matching the digitized content with the metadata. The metadata of the online public access catalog were extracted to be included in an Excel file. Each record/row was identified by a unique linking ID number, namely the CARRF id number which refers to the print document archived. On the other hand, the file names of the digitized documents were saved as another Excel file. In the latter, the linking ID's were manually read, normalized, separated from the file names, and recorded in new columns. For this part, because a physical archive folder may contain several documents which all link to the same metadata record in the old system, an ID could relate to various types of digitized documents. In other words, the existing metadata are not granular enough to specify and describe all the documents of the historical collection. The records of these two files did not appear to simply match each other on one-to-one basis.

The data manipulating software application SPSS was used for reading and matching these two Excel files by the ID's. Before the match/file merger was applied, the duplicates of the ID's in the file that included the file names of the digitized documents were removed. The result of matching showed that the metadata from our existing database system had not been updated to cover the entire collection.

Possible resolutions

The preliminary steps of linking the file names of the digitized documents and the metadata in the existing system have been taken. For a simple one-to-one merger, a decision was made to only link the file names of the digitized surveys with the metadata. This merged table is being loaded into the new content management system and eventually will be linked with the digital objects of the surveys.

For those archived documents which were not indexed and covered by the existing metadata, workflows need to be created to manually index the digitized documents.



RUTGERS Center of Alcohol Studies

Watch out for the potholes: The bumpy road of digitizing a historical collection

Tony Chu, PhD, MLIS

Center of Alcohol Studies, Rutgers University

CARRF @CAS

The Ralph G. Connor Alcohol Research Reference Files is an archive of instruments (questionnaires, interview schedules, surveys, etc) that have been used in studies of substance use. The primary focus of this collection is on various aspects of drinking behavior and alcoholism.

The Society for Study of Social Problems founded this collection, which is now named in memory of Dr. Ralph G. Connor, in 1960. The Files were developed and maintained by Dr. Connor at Eastern Washington State College, and following his death they were transferred to the Center of Alcohol Studies at Rutgers University.

The Center has continued to add new instruments to the collection, and periodically issues new updates to the descriptive inventory, providing brief information on the materials.

Project Goal

The CAS Library is attempting to digitize the historical collection using a new content management system, which is in the testing process. The preliminary work of digitization has been implemented as part of the testing for the center's system and information migration.

Streamlining the digitization process is a great challenge. This project is an exhibition about what could go wrong and what lessons we have learned from this process.

Possible ways of digitizing a historical collection

Option 1:
Digitize one historical document, load, and index it in the new system. One at a time.

Option 2:
Digitize all the historical documents, load, and index them in the new system.

Option 3:
Digitize all the historical documents, extract the file names, retrieve existing metadata records, match the file names and the records by common IDs, and load them all into the new system.

The last option was assumed to be the most efficient way of digitization.

The Issues

Issue 1 :
Testing found that the existing metadata records were not updated to cover all the print historical documents (see the Venn diagram at the right).

Issue 2:
For the matched, the metadata records are not granular enough to denote the related documents.

Issue 3 :
The IDs for file matching were made part of the file names of the digitized historical documents. They need to be read out to be the keys for matching (see the table at the right).

Wrong
The ID was a part of name of digitized document

	A
1	FileName
2	1_10ASurvey.pdf
3	1_10B1Strassburger65.pdf
4	1_10B2Bibliography.pdf
5	1_11ASurvey.pdf
6	1_11BBibliography.pdf
7	1_12ABBailey69.pdf
8	1_12BBibliography.pdf
9	1_13ASurvey.pdf
10	1_13BBibliography.pdf
11	1_14ASurvey.pdf
12	1_14BBibliography.pdf
13	1_15ASurvey.pdf
14	1_15B1Greene72.pdf
15	1_15B2Bibliography.pdf
16	1_16ASurvey.pdf
17	1_16B1Huebner76.pdf
18	1_16B2Bibliography.pdf
19	1_17ASurvey.pdf
20	1_17BTableVocationalRehab.pdf
21	1_17CTableCounselorsAttitudes.pdf
22	1_18A1PilotProject.pdf

Correct
Column B & C were created to read the IDs for matching the existing metadata records

	A	B	C
1	FileName	CARRF-ID	CARRF-ID_1
2	1_10ASurvey.pdf	1.10	A
3	1_10B1Strassburger65.pdf	1.10	B1
4	1_10B2Bibliography.pdf	1.10	B2
5	1_11ASurvey.pdf	1.11	A
6	1_11BBibliography.pdf	1.11	B
7	1_12ABailey69.pdf	1.12	A
8	1_12BBibliography.pdf	1.12	B
9	1_13ASurvey.pdf	1.13	A
10	1_13BBibliography.pdf	1.13	B
11	1_14ASurvey.pdf	1.14	A
12	1_14BBibliography.pdf	1.14	B
13	1_15ASurvey.pdf	1.15	A
14	1_15B1Greene72.pdf	1.15	B1
15	1_15B2Bibliography.pdf	1.15	B2
16	1_16ASurvey.pdf	1.16	A
17	1_16B1Huebner76.pdf	1.16	B1
18	1_16B2Bibliography.pdf	1.16	B2
19	1_17ASurvey.pdf	1.17	A
20	1_17BTableVocationalRehab.pdf	1.17	B
21	1_17CTableCounselorsAttitudes.pdf	1.17	C
22	1_18A1PilotProject.pdf	1.18	A1

Lessons Learned

1. If the metadata of an historical collection exists in an online public access catalog (OPAC) database, the digitization of each document should be accompanied by checking whether or not the corresponding metadata are included in the OPAC database.
2. When digitization and indexing / cataloging are not undertaken at the same time, a tentative ID used for the digitized document needs to match the ID used for the OPAC database.
3. The issue of compatibility is a challenge for metadata migration. Most of the time, only a part of the existing metadata elements and values are selected for the new repository system. How to link the metadata records with the digitized objects and then add in more metadata elements needs careful advance planning for peak efficiency.