The Home Stretch: Developing Automated Solutions for Legacy Container List Data at the Cuban Heritage Collection, University of Miami Libraries

Natalie Baur, Archivist for the Cuban Heritage Collection, University of Miami Libraries
Lyn MacCorkle, Digital Repository Librarian, University of Miami Libraries
Sevika Singh, Graduate Assistant, University of Miami Libraries
Archival Practice, Volume 3

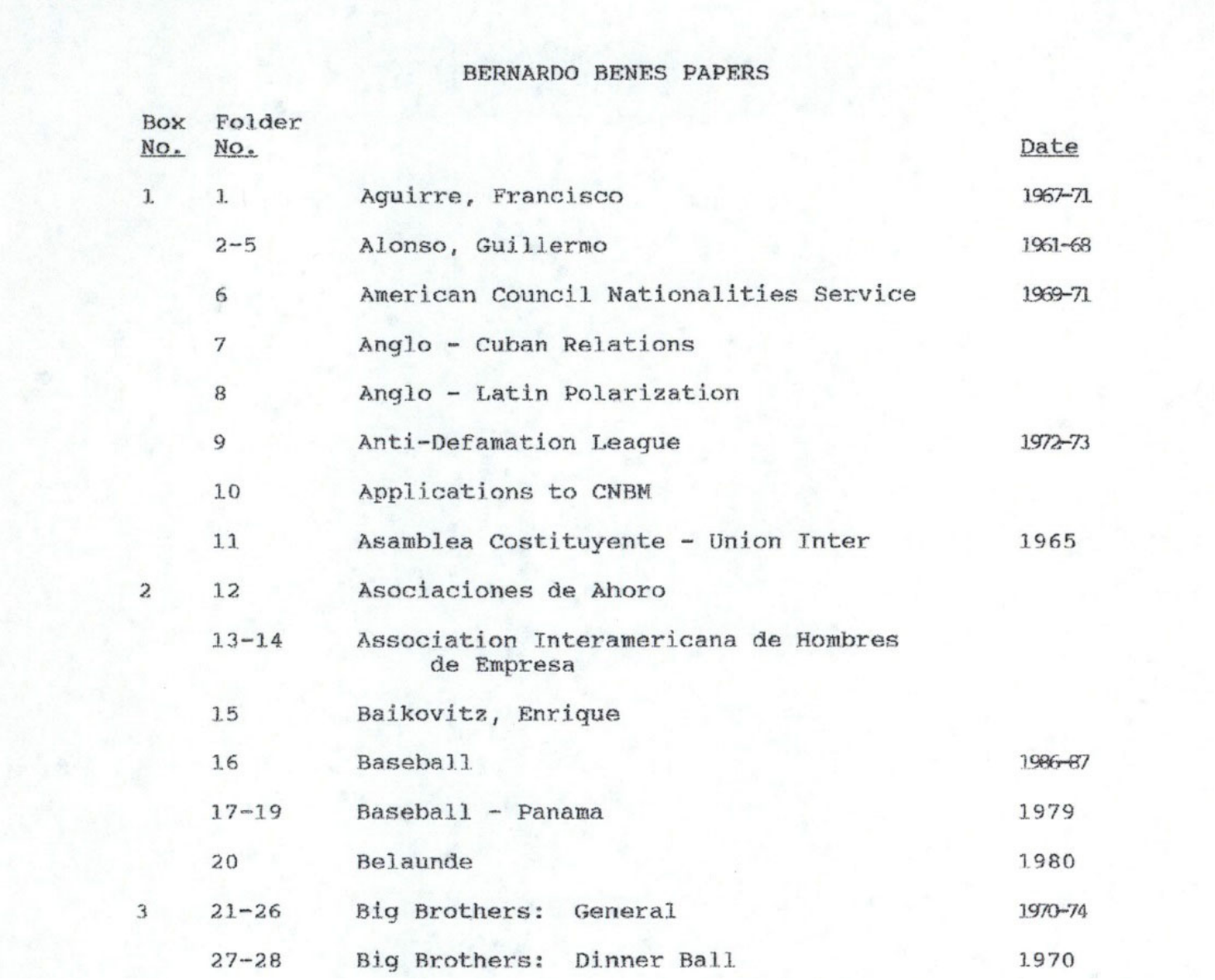
Abstract

Legacy finding aids in formats other than standardized EAD-XML encoding are problematic for archivists wishing to provide deep access to and interoperability for their archival description and metadata. This case study will detail the project development, implementation, and outcomes for the legacy PDF container list conversion project as implemented at Cuban Heritage Collection (CHC) at the University of Miami, which seeks to provide innovative automation solutions based on the tenant of iterative processing and description of archival collections.

Introduction

Legacy finding aids in formats other than standardized EAD-XML encoding are problematic for archivists wishing to provide deep access to and interoperability of their archival description and metadata. Archivists in the last decade have written a variety of articles on streamlining workflows for converting legacy finding aids in typewritten, MS Word, PDF, HTML and even poorly executed EAD-XML formats to standardized and valid EAD-XML markup. We are now beginning to see the power of having archival metadata in an interoperable format that allows for remixing data. Archivists and metadata enthusiasts are finding new and innovative ways to remix archival metadata and description including creating Wikipedia pages from finding aids, as in the case of the University of Miami's 2013 Remixing Archival Metadata Project.[1] Having the creator's biographical notes in EAD-XML format makes creating personal, family, and corporate names in the EAC-CPF standard (Encoded Archival Context: Corporations, People and Families) more readily attainable. Finally, having clean, validated EAD-XML finding aids is invaluable for contributing local archival metadata to statewide, regional, and international digital library cooperatives and metadata aggregators such as Digital Public Library of America, ArchivesGrid, World Digital Library, Wikimedia, and Wikipedia. EAD-XML encoded finding aids are essential for playing well with XML-based GLAM (Galleries, Libraries, Archives and Museums) standards and linked open data initiatives, data visualization applications and digital humanities projects.

In the spring of 2013, the University of Miami Libraries elected to migrate to the newly created ArchivesSpace archival management system, and archivists from each of the Library's three collections were tasked with preparing existing data for migration. One of the most challenging issues that confronted the archivists with the legacy data in Archon was developing a solution for converting finding aids with links to folder-level and item-level container lists in PDF format to EAD-XML encoded container lists so that they could be ingested into Archon and directly migrated to ArchivesSpace when the migration took place. Archivists were particularly interested in using the migration to do this work because of the rich metadata hidden away in PDF folder- and item-level container lists that could be made searchable. Because the PDFs were only accessible at the time through a link from Archon, all of the keywords - including names, dates, and locations - were not accessible from the Archon keyword search interface. Also, based on usage statistics from Aeon, many of the most frequently consulted collections (in this case, those held at the Cuban Heritage Collection) had legacy PDF container lists, and by ingesting the extant, previously "hidden" data into the EAD-XML encoded finding aids in Archon, the Libraries would be able to provide users with enhanced access to these high-priority collections. Finally, because of the high quantity of legacy PDF finding aids in Archon, which numbered from one page to hundreds of pages in length, the team determined that if an automated approach was possible, it would result in huge time savings and lighten the load on staff and student labor resources.



BERNARDO BENES PAPERS

Box No.	Folder No.		Date
1	1	Aguirre, Francisco	1967-71
	2-5	Alonso, Guillermo	1961-68
	6	American Council Nationalities Service	1969-71
	7	Anglo - Cuban Relations	
	8	Anglo - Latin Polarization	
	9	Anti-Defamation League	1972-73
	10	Applications to CNBM	
	11	Asamblea Costituyente - Union Inter	1965
2	12	Asociaciones de Ahoro	
	13-14	Association Interamericana de Hombres de Empresa	
	15	Baikovitz, Enrique	
	16	Baseball	1986-87
	17-19	Baseball - Panama	1979
	20	Belaunde	1980
3	21-26	Big Brothers: General	1970-74
	27-28	Big Brothers: Dinner Ball	1970

Fig. 1: Example of a scanned typewritten legacy container list in PDF form. CHC0216 Bernardo Benes papers, Cuban Heritage Collection, University of Miami Libraries.

Colleagues were consulted for advice on how to automate the process of extracting and adding the PDF container data to Archon, and the Digital Repository Librarian suggested scraping the text from the PDF container list and creating Perl scripts to generate a correctly formatted spreadsheet file. After editing, this file could be uploaded to Archon using the software's CSV "collection content import" feature. This case study will detail the project development, implementation, and outcomes for the legacy PDF container list conversion project as implemented at Cuban Heritage Collection (CHC) at the University of Miami.

Project Development

Nearly all of the Cuban Heritage Collection's 450-plus archival collections have, at least, a DACS-compliant, collection-level record. The occurrence of legacy PDF folder- and item-level container lists in those collection-level or series-level finding aids was 10-15% of total finding aids in Archon. Before moving to Archon, the special collections at the University of Miami used their websites to host both HTML finding aids and scanned PDFs of typewritten finding aids, and many of the HTML finding aids had folder- and item-level container lists. When migrating to Archon, archivists and digital librarians opted to convert the HTML container lists to PDF documents, providing a link to the PDF container list in the Administrative Information field in Archon.

In the earlier work of converting HTML finding aids to PDF documents, formatting in the PDF container lists was fairly standardized, which allowed for relative ease in automating the text scraping and subsequent Perl script routines for parsing the data into the Archon ingest spreadsheet template. Most, but not all, of the PDF container lists scanned from

legacy typewritten finding aids were also in a format standardized enough to be run through the same automated procedure. One basic Perl routine was created that required 14 minor edits to handle the all the different conventions used in the legacy PDF container lists, and a standard Perl structure was created to parse the data. About 20 customized script selectors were created, and were deployed depending on the precise format of the finding aid in order to capture the data correctly.

The Archon "collection content ingest" feature only allows import of collections with one sub-series level. For the finding aids with sub-subseries, a valid EAD-XML container was created and imported into Archon using the EAD-XML collection import feature. Creation of the EAD-XML container list required these steps:

1. Generate - using Perl - the container list spreadsheet from the scraped text; separate fields for Series, Subseries, Sub-subseries, Box, Folder, Title, Date
2. Edit the spreadsheet as necessary
3. Create a 'Level' and 'Container' field in the spreadsheet. In the 'Level' field the proper ‹C› component level was added, and in the 'Container' field series, subseries or file was added as appropriate
4. Save the marked up spreadsheet as an Excel 2003 (.xls) file
5. Import the .xls file to Oxygen and save as an XML document (checked the option to create XML element names from the 1st row headings), then export the XML file
6. Use PHP SimpleXML to parse the exported file and create a fully formed and valid EAD-XML container list that will replace the Archon legacy PDF container list

Project Implementation

Project development consisted of three phases:

1. Identification and prioritization of finding aids with legacy PDF container
2. Conversion of the PDF container lists to spreadsheets for editing
3. Exporting the spreadsheet to CSV format for upload to Archon (those with sub-subseries had to be converted to EAD XML before upload and required a different editing process)

For the first phase of the project, the archivist consulted reading room statistics in Aeon to identify collections with legacy PDF container lists that were most utilized and had the richest container list metadata that would benefit from being keyword searchable. Of that first critical group, thirty finding aids were identified as being prime candidates for the first exploratory phase of the project, including finding aids for core CHC collections, processed at least at the folder-level. These collections included the Gastón Baquero papers, Fulgencio Batista papers, and INTAR Theater records. These collections are heavily used and the container lists contained rich metadata such as personal names in correspondence files and personal papers series relating to other people and corporate bodies with collections held at the CHC. Opening this rich metadata to keyword search and linked relationship data exposes the networks of people, places, time, and events that co-exist in the CHC's holdings.

Working with the Digital Repository Librarian and her graduate student assistant, the CHC Archivist started work on the thirty finding aids identified for the first pilot phase of the conversion project. The Digital Repositories Librarian and her graduate assistant developed Perl routines based on the standardized format of the legacy PDF container lists to parse the scraped text. They developed the following workflow for scraping the text and parsing it into a text file with semi-colon delimiters that could be imported into Excel and, in the case of sub-subseries, transform the container list to a validated EAD-XML container list.

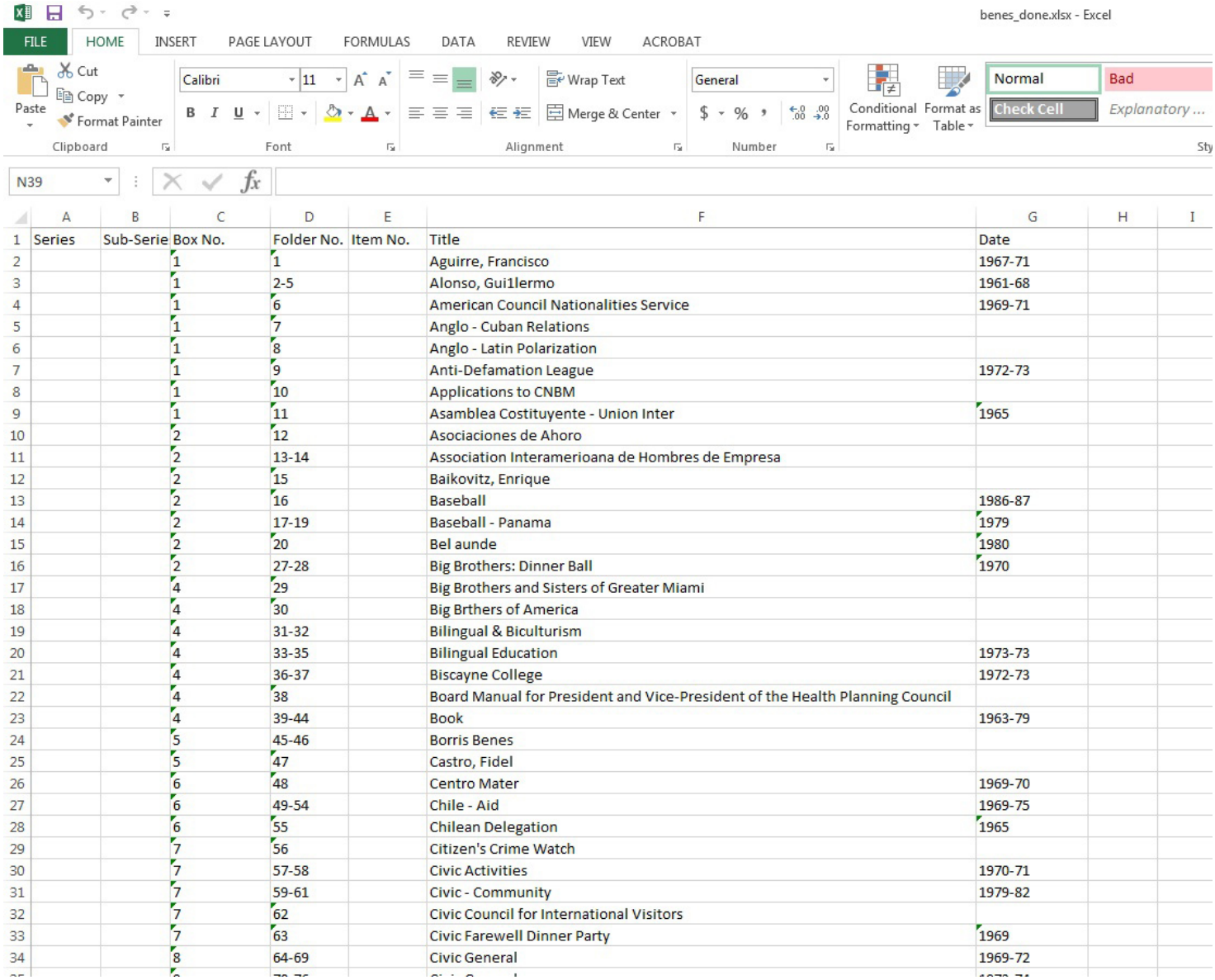
1. Highlight and copy text
2. Use a Perl script to extract; format; add series number, title, subseries, box numbers, folder numbers, folder titles; and parse dates. Separate elements with a semi-colon delimiter and save as a text file (fig. 2)
3. Import semi-colon delimited text file into Excel with the following headings: Series, Subseries, Sub-subseries, Box Number, Folder Number, Folder Title, and Date
4. If there is a sub-subseries, the list cannot be imported into Archon. Component levels, and container types have to be added to the spreadsheet. Save the Excel file as 2003 version to in order to generate an XML version in Oxygen. Then parse the exported file with PHP SimpleXML.

```
#c:\perl\bin\perl.exe
# OPEN INPUT FILE
open (IN,"benes.txt");
# OUTPUT
open (OUT, ">benes1.txt") ;
while (<IN>) {
next if (/^$/); # next if line empty
next if ($_ =~ /^Cuban/);
#next if ($_ =~ /^Series/);
#next if ($_ =~ /^Subseries/);
next if ($_ =~ /^file/);
$_ =~ s/\s+/ /mg;   # collapse multiple spaces to one space

$text = $_;
#capture series number
if ($text =~ /^(Series\s)(\w{1,3})(\W\s)(\w.+$)/){
$sn = $2;
$st = $4;

print OUT "$sn;;;;$st;\r";
}

#capture sub series number
if ($text =~ /^(Subseries\s)(\w{1})\W\s\w.+$/){
$ssn = $2;
}

#capture box number
if ($text =~ /^(\d{1,3})\s(\d{1,3}|\d\d\-\d\d\d|\d{1,3}\-\d{1,3})\s(\D.+$|\d\d\d\d.+$|\d\d\d\d\s.+$|\d{4}

$bn = $1;
$fnx = $2;
$ftx = $3;

$ftx =~ s/n.d./Undated/;
$ftx =~ s/View Digital Object//;

print OUT "$sn;$ssn;$bn;$fnx;$ftx\r";
}

#capture Folder # and Folder Title
```

Fig. 2: Example of a Perl that re-formats the text scrapped from a PDF container list to the format required for Archon upload and conversion to EAD-XML for CHC0216 Bernardo Benes papers.

The digital team then sent the Excel files to the archives team for editing. While the Perl routines automated the process for immeasurable time savings, some of the formatting in the original PDF container lists were not compatible with Archon conventions. For example, if a repeating folder title was indicated with one line: Folder 2-7 "Correspondence," that line was parsed and extracted into the CSV file, and the archives team had to add lines 3 through 7 to the spreadsheet so each individual folder had a line in the Archon container list. Other elements that required special editing were the normalization of dates for DACS compliance, changing series numerations from Roman numerals to Arabic numerals, and correcting any small punctuation and spacing errors resulting from anomalies in the original formatting in the PDF documents.

After editing was completed, the spreadsheet was saved in CSV format to upload into Archon using the "collection content import" feature. For a container list that had a sub-subseries, after editing the container lists, the digital team ran the file through a PHP program to generate a valid EAD-XML container list. The archivist then exported the existing collection-level EAD-XML finding aid into Archon, added the container list to the collection-level document, and imported the complete EAD-XML files into Archon using the EAD-XML import feature. (fig. 3)



	A	B	C	D	E	F	G	H	I
1	Series	Sub-Serie	Box No.	Folder No.	Item No.	Title	Date		
2			1	1		Aguirre, Francisco	1967-71		
3			1	2-5		Alonso, Gui1lermo	1961-68		
4			1	6		American Council Nationalities Service	1969-71		
5			1	7		Anglo - Cuban Relations			
6			1	8		Anglo - Latin Polarization			
7			1	9		Anti-Defamation League	1972-73		
8			1	10		Applications to CNBM			
9			1	11		Asamblea Costituyente - Union Inter	1965		
10			2	12		Asociaciones de Ahoro			
11			2	13-14		Association Interamerioana de Hombres de Empresa			
12			2	15		Baikovitz, Enrique			
13			2	16		Baseball	1986-87		
14			2	17-19		Baseball - Panama	1979		
15			2	20		Bel aunde	1980		
16			2	27-28		Big Brothers: Dinner Ball	1970		
17			4	29		Big Brothers and Sisters of Greater Miami			
18			4	30		Big Brthers of America			
19			4	31-32		Bilingual & Biculturism			
20			4	33-35		Bilingual Education	1973-73		
21			4	36-37		Biscayne College	1972-73		
22			4	38		Board Manual for President and Vice-President of the Health Planning Council			
23			4	39-44		Book	1963-79		
24			5	45-46		Borris Benes			
25			5	47		Castro, Fidel			
26			6	48		Centro Mater	1969-70		
27			6	49-54		Chile - Aid	1969-75		
28			6	55		Chilean Delegation	1965		
29			7	56		Citizen's Crime Watch			
30			7	57-58		Civic Activities	1970-71		
31			7	59-61		Civic - Community	1979-82		
32			7	62		Civic Council for International Visitors			
33			7	63		Civic Farewell Dinner Party	1969		
34			8	64-69		Civic General	1969-72		

Fig. 3: Final edited Excel file of the CHC0216 Bernardo Benes papers to be converted to CSV for upload into Archon's "collection content import" feature.

Finally, the archivist was left with legacy typewritten finding aids for which no original electronic version existed, and PDF scans had not been run through an optical character recognition software (OCR). For the five finding aids that were in this format, the CHC staff ran the original typewritten paper finding aids through a document feed scanner and an OCR program which then allowed the team to proceed with the conversion workflow.

Project Outcomes

The first phase of this project involved converting thirty high-use, content-rich legacy PDF finding aids that ranged in size from one page to several hundred pages in length. This phase took three months, working an average of five dedicated hours per week from conception to completion. The second phase of the project is currently underway. A part-time staff member is examining all of the Archon finding aids to identify the remaining legacy PDF finding aids and adding these to the conversion project master spreadsheet. Phase two of the project has about one-third the number of legacy PDFs as the first phase and once the workflow is deployed again, the second and final phase of the CHC project will be completed in one month at five hours per week.

One of the primary benefits of investing in workflow and programming for this project is the amount of time saved by converting the data from a previously unusable format to a format that was discoverable and interoperable. Automating this process saved months of data-entry work that would have been done by a team of student workers. It also saved the time of staff that would have had to implement intensive quality control and workflow management as well as direct supervision. As it stands, one staff member with in-depth experience with spreadsheets and Archon was

dedicated to the project and worked closely with the archivist during the data clean-up and ingest steps. Because the project only required around five hours per week from an experienced staff member, other processing and descriptive work could go on simultaneously during the conversion project.

In conclusion, this case study can be used as a basis for implementing automated legacy finding aid conversion projects at other repositories. The key elements for the success of the pilot project at the CHC included the legacy container lists being in more or less standardized formatting, a participant who will develop extraction and conversion scripts (in this case, PHP and Perl), a project leader with strong project management skills and the ability to work with a cross-departmental team, and staff knowledgeable in standardized and local descriptive practices for the editing phase of the project. The team ultimately saw this as the final step in an iterative descriptive process implemented by earlier archivists and librarians: from paper finding aids scanned to PDF attachments, from HTML to PDF attachments, and finally, those legacy PDF attachments fully integrated into a complete and powerful, interoperable EAD-XML discoverability tool.