MDR, Vol. 40, pp. 115–121, September 2011 • Copyright © by Walter de Gruyter • Berlin • Boston. DOI 10.1515/mdr.2011.015

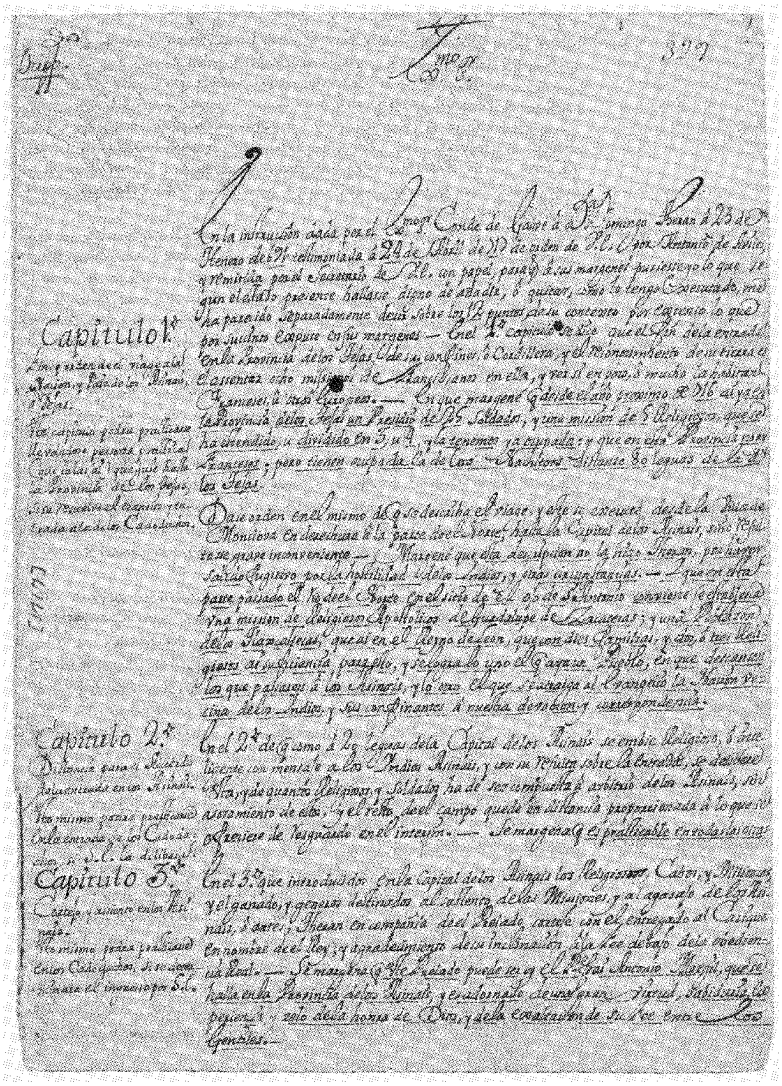
Decades to Digitization: A Pilot Project to Bring the Bexar Archives Online

Zach Vowell

Zach Vowell (zvowell@austin.utexas.edu) is the Digital Archivist at the Dolph Briscoe Center for American History, University of Texas at Austin.

In the summer of 2009, the Dolph Briscoe Center for American History (DBCAH) at the University of Texas at Austin received funding from the Texas State Library and Archives Commission's TexTreasures grant program (itself sponsored by IMLS) in order to digitize a small portion of the Bexar Archives.[1] This archival collection, which consists of more than 250,000 manuscript pages and dates 1717–1836, documents the administrative activities of the Spanish and Mexican governments that ruled over the state of Texas y Coahuila before the Texas Revolution in 1836. The course of administrative busi-

Figure 1. Rebolledo's comments on Gálvez instructions to Domingo Terán, with suggestions for the founding of settlements at Béxar, La Bahía, and among the Asinais and Cadodaches, circa 1717. Bexar Archives, Briscoe Center for American History, University of Texas at Austin; e_bx_000002_001.



ness records in minute detail the military, civil, and political life of this historic jurisdiction. Indeed, the Bexar Archives represent an essential source for any scholar interested in the history of the southwest borderlands.

Given the size of the Bexar Archives, DBCAH recognized that a pilot project would help develop our workflows and web interface, and prepare the way for a more ambitious undertaking. In addition, DBCAH decided to scan the microfilm copies of the original documents, in order to reduce both the handling of the originals and scanning costs. What follows is an account of this pilot project and its results.

Background

Before delving into the specifics of project, it may help the reader understand why DBCAH undertook such an ambitious project. The importance of the collection alone does not explain the project's urgency – DBCAH holds many historically valuable collections and completes only a few digitization projects each year. Why did DBCAH need to digitize this particular collection now?

The project's timing can be traced back over 100 years, when the Bexar County Court of Commissioners transferred the Bexar Archives to the University of Texas in 1899. As a stipulation to this transfer, the University agreed to make the collection more accessible, and to translate the original documents. The translation effort began in 1934, and has continued with little interruption to the present day. While performing this requirement of the transfer, the translators wrote introductory text for each volume of translations, and appended an index of terms and a glossary.

The University Archives staff began calendaring the collection in card form not long after acquiring it in 1899, arranging the collection into 3 series, describing every document in the collection, and assigning document numbers to each document. In 1938, Bexar Archives translator J. Villasana Haggard revised the card calendar and converted it to typescript form.

As translators continued their work into the 1960s, microfilm technology gained wide acceptance in the library and archives communities. In 1967 the National Historical Publications and Records Commission (NHPRC) began funding a project that would allow the University to microfilm the entire Bexar Archives. Access quickly expanded once the University created microfilm copies and translation volumes, and the Bexar County Clerk's office and several other institutions around the country began adding the copies to their collections, enabling researchers to consult the Bexar Archives without traveling to Texas. In addition, the public could purchase copies of the microfilm for personal use.

The microfilming project triggered an update to the Bexar Archives Calendar. The new Calendar replaced the original Calendar's document numbers with microfilm frame numbers, corrected errors in the original Calendar, and incorporated a shorthand identification scheme that would help researchers know quickly what type of administrative document they were viewing. The new Calendar totaled 37-volume work further subdivided by microfilm rolls.

The translations, the Calendar, the indices, and the glossaries all provided rich resources through which researchers could better access the Bexar Archives. But as of 2009, the resources were somewhat disconnected from each other. The Bexar Archives Online project proposed to bring all these resources together with the original documents, and provide researchers both a way to find the information they seek and information on how to access the original documents at DBCAH, if necessary.

Prominent project features

Recognizing that translations are open to interpretation, DBCAH wanted to give researchers the ability to compare a translation with the original Spanish text. Consequently, DBCAH envisioned side-by-side presentation of the originals with their translations.

DBCAH also wanted to provide researchers with the capability to search the English-language translations' text. Such full-text search capabilities would require optical character recognition (OCR) software and some maintenance of the OCR output. Furthermore, DBCAH would require the design of a search engine that could search both the translation text and the metadata that would be associated with each document.

As for metadata, the project would re-purpose the Calendar's extensive metadata about each document. And staff planned to record the metadata *before* we scanned the original documents and translations.

DBCAH's typical workflow follows a reverse course, where metadata is recorded after scanning has been done. Due to the idiosyncrasies of our cataloging software (Extensis Portfolio), which required a digital asset to exist before cataloging could be done, DBCAH needed to think critically about software and workflow.

DBCAH intended to import the metadata and digital documents into the Center's existing database and web framework. This web-based database, the Digital Media Repository (DMR), is composed of several PHP templates that receive content from a MySQL database. The flexibility offered by this custom-configured system would allow us to craft a distinct interface required by the new features, while maintaining the existing database structure which had served so well for past digitization projects.

Specifications

The project's scanning specifications followed a more typical path. The *BCR's CDP Digital Imaging Best Practices, Version 2.0*, dated June 2008, presents clear, concise guidelines for digitizing a wide variety of archival content. This document had proved useful in past digitization projects, and so DBCAH adopted the standard for the Bexar Archives Online project as well.

After consulting the *BCR/CDP* document, DBCAH set the master files' specifications as such:

Original Documents (microfilm)
File format: TIFF
Bit depth: 8-bit
Colorspace: Grayscale
Resolution: 600 dpi

Translations (typescripts)
File format: TIFF
Colorspace: Bitonal
Resolution: 400 dpi

The scanning vendor, Neubus, produced two derivatives from each translation master file. First, the vendor derived OCR output from the translation TIFFs, resulting in plain text files encoded in UTF-8 characters. Furthermore, Neubus generated PDF derivatives of the translation typescripts, and, by placing the OCR text in the background of the PDFs, made the PDFs searchable.

Once the Center received the master TIFFs, we created 3 types of image file derivatives, in accordance with the requirements of the DMR's display mechanisms:

"Large" image
File format: JPEG
Height (or longest dimension): 500 pixels
Resolution: 72 dpi

Thumbnail image
File format: JPEG
Height (or longest dimension): 120 pixels
Resolution: 72 dpi

Zoom image
File format: JPEG
Width (or shortest dimension): 50 inches
Resolution: 72 dpi

The zoom images, however, would need more work. Specifically, the Zoomify Converter application converted the zoom image JPEG into hundreds of different "tile" images. The Bexar Archives Online site then could display these "tiles," via the Zoomify Viewer web application, in a way that displays each document in crisp detail.

Metadata creation – Phase 1

Descriptive metadata had been previously created during the original microfilming project, and encapsulated in the Bexar Archives Calendar. This work contains metadata regarding each document's description, author, date, place of composition, document type, and other information. DBCAH merely needed to transcribe the metadata as it existed in typescript form into the DMR database where it could be integrated into the Bexar Archives Online interface.

However, two hurdles immediately presented themselves. First, the project began with a project intern and the Calendar, but no digital documents to which they could correspond. This first hurdle proved a more conceptual obstacle, related to file naming. For the DMR, each document page requires a unique identifier. Staff could estimate the number of pages that the vendor would ultimately deliver, but we could not proceed with creating a unique database record for each page until they had been

scanned, processed, and uniquely identified. Furthermore, one feature of Bexar Archives Online would require direct links between each original document and its translation. These links need unique identifiers to enable the desired feature, and so staff recorded shorthand information about each document that would enable them to record the direct links once the unique identifiers had been established.

Secondly, DBCAH staff typically uses Extensis Portfolio desktop software to describe digitized items, but Portfolio requires the digital items to exist before one can describe them. This more practical dilemma upset our normal workflow. If staff could not rely on the Portfolio interface and functionality to transcribe the metadata, what could we use until the digital documents arrived at the Center's receiving-room door?

The solution ultimately resolved both concerns. Staff turned to .CSV-formatted spreadsheets, via Microsoft Excel. A spreadsheet provided the flexibility of not having to commit to a specific number of unique records (i.e., it allowed us to estimate the number of pages that the vendor would deliver), and it also provided a format through which the metadata could be imported into Portfolio and the DMR.

So, staff created several spreadsheets (eight for each microfilm roll, and 125 for each translation volume) and parsed the Calendar metadata into 22 qualified Dublin Core fields. The fields were based on the Center's internal document *Metadata Schema & Style Guide, Version 1.1*, which align with the fields in the DMR database. While staff awaited the digital documents from the vendor, staff entered the metadata into these spreadsheets.

Digital file processing

Not long after staff finished transcribing the metadata into spreadsheets, the vendor delivered the digital documents on May 18. Staff now confronted three primary tasks that would convert the master TIFF files to a format suitable for the functionality DBCAH envisioned for Bexar Archives Online.

The original microfilm project filmed 2 pages per microfilm frame. To be exact, the first page of each document occupied its own microfilm frame, but subsequent pages of that document would be arranged 2 pages per frame, until the next document. Not only would staff need to split each 2-page TIFF into 2 separate TIFFs, we would most likely have to parse out the 1-page TIFFs in any sort of automated processing.

Our solution lay in Photoshop. Staff created 2 action scripts within Photoshop, one of which would crop the left side of each file and saved as a new file, while the other would crop the right side. The idea being that staff would perform *both* of these cropping functions on each vendor-produced TIFF. The left and right cropped files would be combined in a single directory, where they would await file renaming and thus assume the role of archival masters. The vendor-produced TIFFs would be discarded at the project's end.

Once staff had finished splitting the TIFFs, we could proceed to giving each TIFF a unique identifier (and in the process, prepare the way for completing our metadata). Having not planned for batch file-naming software in our project budget, DBCAH was forced to scour the Internet for free software compatible with Mac OSX. Fortunately, the developers at Name Mangler continue to offer that software's free predecessor, File List, on their website.[2]

With this software, DBCAH proceeded to consider a suitable file-naming convention. DBCAH decided to adopt one comprised of 3 parts: 1) a project/collection-based prefix, 2) a document number, and 3) a page number. Following the tradition of previous Briscoe Center digitization projects, part 1) would be **e_bx_**, where the **e_** part denotes the digital object as a digital surrogate and the **bx_** identifies the object to be a part of the Bexar Archives. Part 2) would consist of a 4-digit number, and would be assigned serially to each document, regardless of whether the document was an original or a translation. Part 3) would consist of a 2-digit number, and would be assigned in the sequential order of pages within each document. The first page of the first document, then, would be **e_bx_0001_01**.

Staff also needed to make one slight adjustment to the translation TIFFs. The vendor delivered the bitonal translation TIFFs in a curious, custom pixel aspect ratio. When Photoshop opened these TIFFs, they looked squished in at the sides. To resolve this issue, staff created an action script that would convert each translation TIFF to a square pixel aspect ratio.

The final major processing task involved creating 3 types of image files that would be suitable for web presentation: 1) the "large" JPEG (which measures

500 pixels on the long side), 2) the thumbnail JPEG (120 pixels on the long side), and 3) a collection of "tile" JPEGs that would allow a user to view each page through the Flash-enabled Zoomify Viewer application. In addition to these image files, staff created another set of derivatives for the translation documents: multi-page PDF files. Whereas each image file represented only one page of a document, these searchable PDFs, which could be downloaded directly from Bexar Archives Online, combined every page of a given document.

All told the processing of the vendor-produced files netted 31,285 TIFF files (each corresponding to a document page), totaling 572 GB. With a "large" and "thumbnail" JPEG created for every TIFF, staff created an additional 62,570 JPEGs, and then 1,881 translation PDF documents, and approximately 17,000 OCR text files derived from the translation pages. Factoring in the literally millions of "tile" JPEGs, it is not hard to imagine the stress put on our local network connections during file transfers from our working space to the web server.

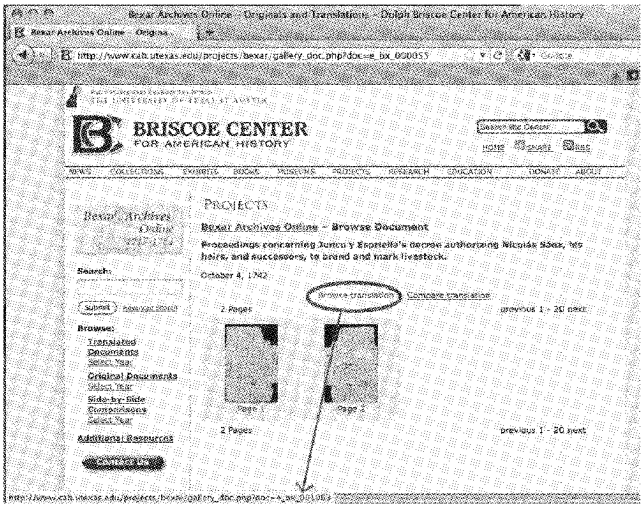
Metadata – Phase II

When we last left metadata, the project intern had entered preliminary descriptive metadata into .CSV spreadsheets. Now that the digital documents had been received and processed, staff began completing metadata for each document, a process vital to making Bexar Archives Online work the way DBCAH intended it to work.

Recall that the .CSV spreadsheets held no information about file names, because the digital files did not yet exist. The first step in this second phase of metadata creation involved matching our newly processed digital files with the rows and rows of metadata in the spreadsheets. With digital files in hand, staff could correct any errors (too many pages, or not enough pages), and precisely associate descriptive metadata with each digital representation of the original document pages. Furthermore, staff furnished each page, both originals and translations, with a unique identifier.

Next, staff had to convert the shorthand dc.Relation metadata into metadata that represented the file identifiers. For instance, metadata recorded shorthand as **18 (25-36)** (where 18 referred to the translation volume, and 25-36 referred to the page range) had to be converted to **e_bx_2567**. The PHP code

Figure 2. Web page displaying the pages of a Bexar Archives document. The highlighted "Browse translation" link leads the user to the translation through the translated document's file identifier prefix, e_bx_001963.



within Bexar Archives Online required us to record the document prefix only, which is why the value within dc.Relation omits any page number suffixes.

As figure 2 illustrates, one can see that the URL embedded in the "Browse translation" link refers only to the document prefix. The illustration example is e_bx_1963. When the link is clicked, every page in document e_bx_1963 is displayed for the user.

Web presentation

All along, this project's goal was a web-based presentation of a small portion of the Bexar Archives. Furthermore, DBCAH wanted to avoid merely dumping the digital documents into one space on our site – rather, DBCAH had high hopes of harnessing all of the work previous generations had accomplished to provide greater access, and convert those access tools into web-based functionalities.

While the features themselves appear intuitive and self-explanatory, we dare highlight a few of the features possible only within the online environment, but which nevertheless derive from paper-based sources:

- The full-text search capability took advantage of the tireless work of Bexar Archives translators over the past 70 years.

Figure 3. Display window for a typical DMR record.

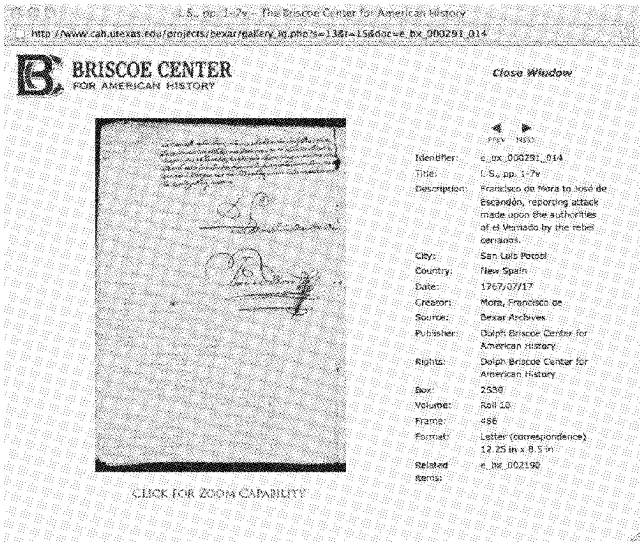
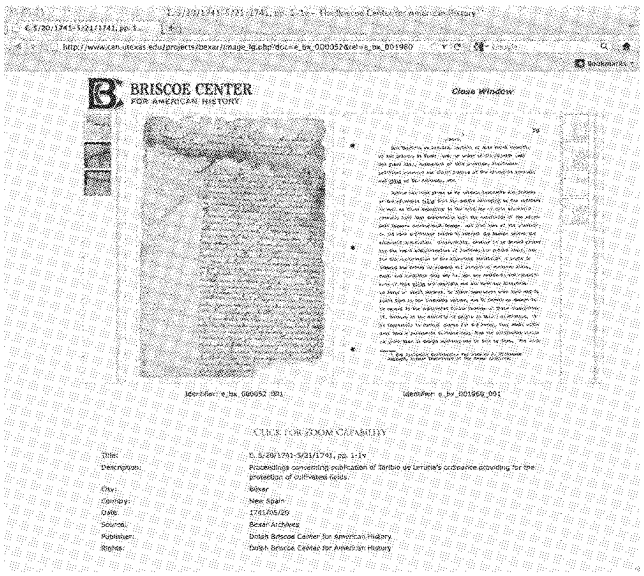


Figure 4. Side-by-side display window, which allows users to select the two pages to compare through the scrolling page selection side-bars.



- The "document description" search owes its existence to the Bexar Archives Calendar, originally a paper typescript.
- The browse feature was made possible by the date information found in the Calendar (and sometimes corrected by the translations).
- The side-by-side comparison tool was largely made possible by the meticulous descriptive work of the Calendar compilers and the Bexar Archives translators, both of which groups carefully noted the dates, page lengths (in manuscript and typescript forms), and document synopses, thus allowing us to definitively match each original document to a translation document.

In implementing these and other features, staff primarily relied on our existing digital asset management system, the DMR. Staff began with the DMR's basic integration of a MySQL database that stored the metadata and OCR text, a web server that delivered the JPEG and PDF derivatives, and PHP code which pulled all of this data together into a usable interface. As each feature was added on top of the DMR, staff made substantial modifications to the PHP. The side-by-side windows represent the most extreme example of these modifications. Staff started out with the PHP template for a typical DMR record, which presents the "large" image, the Briscoe Center logo, a "Close Window" link, a "Zoom Capability" link, page navigation buttons, and a selection of the metadata running down the right side of the window (see fig. 3). From this foundation, the project's PHP developer was able to create a page where users can select which pages they would like to view side by side (see fig. 4).

Future features are planned for subsequent phases of the Bexar Archives Online project, such as integrating the translation volumes' indexes into the online interface. The flexibility of the PHP/MySQL framework (and the Center's specific configuration of it) will allow features such as the index to be smoothly added to the Bexar Archives Online interface.

Challenges: Unforeseen and otherwise

In general, DBCAH's major project decisions and workflows succeeded. However, it should be reiterated that this was a pilot project after all. Several challenges and unforeseen difficulties presented themselves to us along the way, and this article provides the perfect forum to share the more significant problems.

First, after the project ended, staff learned that the file-naming convention proved insufficient. Several documents exceed 99 pages, pushing the file names, for instance, to **e_bx_0001_135**. Violating the convention's number of digits caused substantial disrup-

tion in the way that Bexar Archives Online interacted with the DMR database. It also became clear that if DBCAH planned to digitize the entire Bexar Archives, there would be well over 9,999 documents. So, in the year following this project, staff undertook to rename the files to **e_bx_000001_001**, which involved not only renaming the TIFFs, but also the web derivatives – JPEGs, PDFs, OCR text files – and revise the PHP code to peel back the workarounds staff had written to smooth over the disruptions caused by the 100-plus page documents. Furthermore, staff needed to rename the actual database records. Completing this naming-convention overhaul felt at times like a high-wire act, and more thoughtful planning would have forestalled such massive revisions.

An entirely unexpected web function arose to surpass another challenge. As DBCAH began thinking through the side-by-side feature, we realized that the text of a given original Spanish page did not fit on one English translation page. No 1-to-1 ratio exists between the originals and translation pages, making side-by-side display infinitely trickier. DBCAH overcame this by providing scrolling sidebars on either side of the side-by-side window, allowing users the ability to pick exactly which 2 pages they wanted to compare.

Time management proved to be challenging as well. Project staff included one paid intern working 10 hours per week, and a volunteer working 5 hours per week for 6 months, and the *preliminary* round of metadata creation took 3 full months to complete. While DBCAH knew that metadata creation could be a time-consuming process, we did not allot 3 months. Since the metadata had been previously created in the Calendar, DBCAH overestimated the value of possessing such "raw" metadata, and underestimated the extent to which staff would need to devote time to transcription of the Calendar descriptions, parsing the descriptions into qualified Dublin Core elements, and correcting mistakes and oversights of the circa 1960s calendar compilers.

Furthermore, it took a full 2 months to process the digital files. In particular, splitting original documents from 2-page frames into individual pages proved much more time-consuming than staff had anticipated (as far as we anticipated it at all).

These two primary time management miscalculations deprived us of time for quality control. Consequently, staff has refined the content and metadata of Bexar Archives Online well past the project's original 12-month time frame, up to and including the present day.

Tried and one step closer to true

It has now been nearly a year since we completed this pilot project, and in the interim staff has been working to add new content (7 more rolls of microfilm, and 62 additional translation volumes) and refine the web interface. DBCAH enlisted the help of a group of graduate students at the University of Texas' School of Information to conduct a usability study on the site. Several of the recommendations that came out of the study have already been implemented.

In all the time since the pilot project ended, staff has used the lessons learned during that process to improve the site as it appears today. And soon enough, DBCAH hopes to have reached the point where we will be ready to digitize the entire run of microfilm rolls (172 rolls in all) and present them all online.

Endnotes

1 Bexar Archives Online, http://www.cah.utexas.edu/projects/bexar/index.php.

2 Name Mangler, http://manytricks.com/namemangler/.