Preservation of the Texas Agricultural Experiment Station Bulletin in the Digital Repository

ROBERT B. MCGEACHIN

Texas A&M University Libraries, College Station, Texas, USA

To assist agricultural librarians in their new role as digital preservation and distribution specialists, this article documents the basic procedures for scanning and digitizing print agricultural serial publications and submitting them to a DSpace digital repository, through a case study of a project at Texas A&M University Libraries to digitize Texas Agricultural Experiment Station Bulletins. It is hoped that a dispersed network of similar agricultural materials in all the various land grant university digital repositories could be crawled to harvest the metadata records and make them accessible in a central user-friendly digital library for agriculture.

KEYTERMS agricultural information, archival preservation, digital repository, digitization, open access, Texas A&M University

---

Received 19 November 2009; accepted 22 January 2010.

This project, "Digitization of Texas Agricultural Agency Publications in Support of Development of the National Digital Library for Agriculture," was funded, in part, by USDA, ARS, National Agricultural Library Specific Cooperative Agreement 58-8021-4-195 and by the Texas A&M University Libraries.

Address correspondence to Dr. Robert McGeachin, Associate Professor/Agriculture and Life Sciences Librarian, Texas A&M University Libraries, 5000 TAMU, College Station, TX 77843-5000, USA. E-mail: r-mcgeach@library.tamu.edu

"*This is an Author's Accepted Manuscript of an article whose final and definitive form, the Version of Record, has been published in the Journal of Agricultural & Food Information 11:2*

INTRODUCTION

In the United States of America all state land grant universities and their affiliated experiment stations and extension services produced serial monographic publications to report on their agricultural research efforts in the form of an "experiment station bulletin" and then translated that research into popular consumer-oriented publication series such as bulletins, leaflets, and/or circulars. Most of the experiment stations began publication of a "Bulletin" in 1888, after their creation by the Hatch Act of 1887 (Anonymous, 1887; McInnis, Scott, & Gulley, 1888). Most of these bulletins were printed on acid-based paper that is now yellowed and brittle. Their valuable historical and scientific intellectual content needs to be preserved. Many bulletins were microfilmed in a national preservation project in the 1980s, but only historians find this to be a tolerable form of access. The general public and scientists are now used to electronic desktop access and often act as if non-digital sources do not exist. The content of the experiment station bulletins is not only of historical value; with the growing interest in topics such as organic gardening, use of heirloom varieties, and low fossil fuel inputs, the content of many pre-World War II bulletins is once again relevant for today's farmers and consumers.

Digital repositories have been developing at many academic and scholarly institutions and libraries over the last decade. Their mission is to both preserve intellectual content electronically and make it accessible to a worldwide audience. But, of course, a library repository must have copyright clearance to distribute material to the world in its digital repository. Since the experiment station bulletins are government publications that were published in the public domain, there are no copyright problems with making them accessible. Even if they were originally copyrighted, those published in or before 1923 are now in the public

domain. Experiment station bulletins are thus a logical starting point for land grant university libraries seeking content for their digital repositories.

This paper describes Texas A&M University Libraries' project to digitize the Texas Agricultural Experiment Station Bulletins and make them accessible in a DSpace digital repository. Step-by-step instructions for scanning and digitizing are given to provide guidance to other librarians who wish to preserve and distribute their own state's agricultural serials in digital repositories.

DIGITAL PRESERVATION REQUIREMENTS

For the sake of preservation, a digital repository should archive the highest resolution copies of original materials that it feasibly can. For print materials like the experiment station bulletins, an image of each page should be captured as an archival Tagged Image File Format (TIFF) or JPEG 2000 and each high-resolution page image file stored in the repository. The minimum resolution for text-only page images is 600 dots per inch (dpi) as a 1-bit bitonal image (National Agricultural Library, 2009). For pages with black and white illustrations or black and white half tone photographs, the minimum resolution is 400 dpi as an 8-bit greyscale image. Pages with color should be digitized at, at least, 400 dpi as a 24-bit color image. One can, of course, produce higher resolution page images than this, which may be desirable, as you can always produce lower resolution derivatives, but you cannot ever get any higher resolution than what is already captured, unless you go back and re-digitize the print again at a higher resolution. Therefore, it is best to do the highest resolution you can the first time, as a second chance may not ever come again.

Once high-resolution single page images of a document have been archived, a lower resolution derivative document of the entire work is created to serve as the main access for viewing on the Internet. Currently, this is usually a Portable Document Format (PDF) file that is a fraction of the total size of the original individual page images and can be more readily transmitted over the Internet. In this case, each bulletin is preserved as an item in the digital repository with accompanying descriptive bibliographic metadata, all of the original page image files, and a viewing copy of the whole bulletin as a PDF document.

DIGITIZATION METHODS USED IN THE TEXAS A&M UNIVERSTIY PROJECT

The Texas A&M University Libraries' project uses Dell Optiplex GX620 computers with an Intel Pentium 4 CPU at 3.20 GHz, with 1.00 GB of RAM using the Windows XP operating system. The scanners are the Plustek OpticBook 3600, which is a book edge scanner that can scan the entire surface of the glass, including up to 6 mm from the front outer edge.

The project is using bound copies from the Library circulating collection, because it is our most complete set of this serial. Each bulletin is examined for physical defects that may require finding another physical copy that is more complete from the Texas A&M University Archive's collection or from microfilm. A list of bulletins to obtain from another source was created when gaps were found in the bound collection. A spreadsheet was created using Microsoft Office Excel 2007 for initially recording the metadata for each bulletin. The metadata include the following Dublin Core elements (Dublin Core Metadata Initiative, 2008): creator (author), title, relation.isPartOf (series title and number), publisher, date.issued (date of publication), relation.isPartOf (table of contents), subject (keywords), subject.NALT (National

Agricultural Library Thesaurus terms), description (notes about the physical item, e.g., number of pages), language (English_US), type (article), and coverage.spatial (Texas).

The metadata were derived from either direct examination of the bulletin and hand entry of the data into the spreadsheet or gleaned electronically from exporting records as flat text files from the AGRICOLA database. The records in AGRICOLA originally came from Texas A&M University cataloging of the 1980s national project to microfilm experiment station and Extension Service bulletins and related agricultural serials. The electronic text was easily copied and pasted into the spreadsheet records but only covered through 1982, the upper end of the microfilming project. For each bulletin, the subject content was examined and appropriate National Agricultural Library Thesaurus terms were assigned (National Agricultural Library, 2002).

A coherent file-naming scheme was devised to make the content of each file recognizable from the file name. The TIFF file for each page of a bulletin is "b" for bulletin, followed by a four-digit number for the bulletin number, a space, and a four-digit number for the specific page number. So the first few pages of bulletin #1 are b0001 0001.tif, b0001 0002.tif, b0001 0003.tif, etc. There needed to be four digits in both the bulletin number and page number parts of the file name, because there are instances in both of up to four digits in the series. Page numbers reach four digits because, at one point in the series, pages were numbered consecutively for several years across many bulletins. The complete derivative PDF file of the whole bulletin is designated by "Bull" followed by four digits of the bulletin number, for example, Bull0001.pdf, Bull0002.pdf, Bull0003.pdf, … up to Bull6192.pdf. The files are stored on a local external hard drive until uploaded into the repository. They are arranged in folders designated by the year of publication, and each bulletin's files are in a subfolder named with the bulletin number, for

example, folder 1888 with subfolders 1, 2, 3, 4, 5; folder 1889 with subfolders 6, 7, 8; folder 1890 with subfolders 9, 10, 11, 12, 13; etc. Starting in the 1950s, updated and substantially revised editions of some of the bulletins were produced with the same bulletin number but in a different year of publication. Storing the different editions in their respective year's folders keeps the identical file and folder names segregated.

A Microsoft Office Excel 2007 spreadsheet was created to keep track of the workflow and what had already been accomplished among the various workers so that the next item and pages to be scanned or processed could be readily identified. As each bulletin is scanned, the range of page file names, the worker's initials, and the date are recorded As each PDF file is created, that file name, the worker's initials, and the date are recorded As each set of files is submitted to the repository, the worker's initials and the date are recorded. As each bulletin item submitted to the repository is "approved" and "published," the collection manager records the permanent repository item number.

The scanning process begins with turning on the power first to the external hard drive and OpticBook 3600 scanner before booting the computer, so that these external devices are recognized and loaded into operation properly. To begin scanning, the BookPilot software for the scanner is invoked by pressing on the corresponding button on the scanner. The following settings in the software are designated:

- Scan settings for greyscale set to 400 dpi for all text pages
- Scan settings for greyscale set to 600 dpi for pages with illustrations or photographs
- Scan settings for color set to 600 dpi for pages with any amount of color

- Set to rotate 180^{o} on even scans (all the bulletins begin page numbering with the front cover as page 1, which is placed right side up on the scanner)
- Select file folder path on the external hard drive to save page images
- Select file type as TIF
- Enter the base file name – b#### (the scanning software will add the _####.tif page number portion sequentially)

To begin the first page scan, place the bulletin right side up on the glass of the scanner ~~to~~ in the upper right corner, with the left side of the volume hanging over the scanner's book edge. In the scanning software, click on the preview scan button. Once the results of the preview scan are displayed, adjust the scanning margins so that the entire content portion of the page is included, but the margins stay just within the edges of the page. This prevents extra dark edges from being included in the scan. In bound volumes, adjust the pressure on the spine of the volume to make the page sit as flat as possible on the glass to reduce the amount of stray light entering the scanning surface. This stray light will produce a grey to black band along the page gutter, depending on the intensity of the light. In very tightly bound volumes, the volume cannot be completely pressed flush with the glass and some amount of grey banding cannot be avoided. This must be accepted as a compromise to the alternative of unbinding the volume to obtain completely flat page scans. If multiple copies exist, it might be justifiable and desirable to disbind. The text will still be legible even in a partially greyed band. Once the page is properly aligned and the scanning area defined, the page is actually scanned at the delineated resolution by pressing either the grey or color scan buttons. For the next even numbered page, the volume is turned upside down and the left-hand side placed on the scanner, and the process is repeated.

In this instance, once the scan is completed, the BookPilot software inverts the page image back to right-side up before it is saved. If there are any totally blank pages interspersed in the bulletin, a dummy page with the phrase, "This page is blank in the original bulletin" is scanned instead to keep the scanning sequence in numerical page and rotation order and to let the reader know it is not an error. Any completely blank pages at the end of a bulletin are not scanned and not included in the metadata description of page count.

Next a derivative PDF file of the complete bulletin is created with Adobe Acrobat Pro software. First, a base document is created by selecting all the TIFF files in numerical page order with the "Create PDF from multiple files" selection. In the select file window, one browses to the proper folder on the external hard drive and selects all the files. When all are in order on the "create PDF" window, the process is started. The software stitches together the individual page images into one document. Then optical character recognition (OCR) is performed on the PDF using the Adobe Acrobat Pro software, and it processes through the entire file and produces a text version of the document that corresponds to the text images in it. This makes the resulting PDF file fully searchable by the reader. If the repository software is capable of it, then, in addition to indexing the metadata associated with the bulletin, it also indexes the OCRed text of the entire PDF of the bulletin, greatly increasing the searchability of the bulletin. The bulletin's PDF file is then saved to the folder, along with the corresponding TIFF image files, and all of this is noted in the workflow spreadsheet as completed.

SUBMISSION TO THE REPOSITORY

The Texas A&M University Libraries' digital repository uses the DSpace open source software and the DSpace Manakin user interface that allows for customization of the look and feel of the

user interface (Digital Initiatives, 2005; DSpace Foundation, 2002). DSpace repositories are made up of "communities" that contain "collections" that contain "items," which are digital files of almost any type, and text metadata describing the content of those files. There is a standard DSpace submission interface that consists of five stages for submitting an item, including entry of the metadata and uploading of the file or files that comprise the item. To be able to submit to a collection, one must have an account within the repository with rights to submit. There is also a higher-level collection administrator who can submit and then review, approve, and thereby publish submissions to make them accessible. The collection administrator can also do post-publication editing of item metadata, if necessary.

The first step in item submission is to login to the repository. Then at the "submission and workflow tasks" page, one can "start a submission" or, at the collection's main page, one can click on the button to "submit a new item to this collection." On the first page of the submission process, "Initial Questions," there are two questions to tick in boxes: 1) multiple titles, the item has more than one title, e. g., has a translated title; and 2) published, the item has been published or publically distributed before. Ticking either of these appropriately changes the metadata item boxes on the subsequent description screens . The second and third screens, "Describe," are for entry of descriptive metadata information about the item. The information is copied and pasted into the submission form spaces from the Metadata Spreadsheet. The fourth screen, "Upload," is for uploading all the files that are included with one item (bulletin). Each page image TIFF file and the whole bulletin PDF file are uploaded to the repository from the external hard drive one at a time. Depending on the file size, this can take from 30 seconds to a minute or more each. In the Texas A&M University project in which the bulletins range in size from 4 to over 400 pages, the uploading can take from a couple of minutes to several hours, with an average of about 25

minutes. The fifth screen is a "Review" of the metadata and file uploads, with the opportunity to correct any metadata errors or other problems noticed by the submitter. The sixth screen is to grant a non-exclusive right to distribute the item by checking the box below the license and agreeing to the "Submission Copyright Statement" (Texas A&M University Libraries, 2009). In summary, this grants the institution a non-exclusive right to distribute the item and warrants that the submitter has sufficient copyright authority or permission to grant the license. The seventh and final screen, "Submission Complete," confirms completion of the submission and indicates it will be reviewed for inclusion in that particular collection. For this digitization project, the submitters are library student workers.

The collection administrator, a librarian, who supervises the student workers and does quality control checking of the items, would then log in to their repository account and go to the "Submission and Workflow Tasks" page. The collection administrator takes "ownership" of new items submitted and then reviews the items. First a brief check of the displayed metadata is done. At this point, the reviewer can go to the metadata editing screen, if needed, to correct any errors noted. The reviewer checks the list of uploaded page files to make sure none were missed. The whole PDF file of the item is examined to check the quality of the page scanning and for any errors in creation of the PDF file, such as missing pages, incorrect page orientation, or lack of the text being OCRed. The reviewer also checks for page scans that miss part of the page, are badly skewed, or have some other major flaw in quality. Any of these uploading, scanning-quality, or PDF file problems requires rejection of the item from the repository; the problem has to be fixed and the item and its files must be completely resubmitted to the repository. It is vital that items not be approved until completely inspected as, once approved and published, the submission is

assumed to be permanent and only a repository administrator can remove an item with a problem at that point.

The final step in the process of bulletin submission is addition of the National Agricultural Library Thesaurus subject terms to the item metadata. Since this is not a standard metadata item in the default DSpace descriptive metadata input screens, it must be added as a post-submission editing step. While logged in, the collection administrator goes to the item record in the repository; in the left hand content navigation selects "edit this item"; and, at the "Edit Item" screen, selects "Item Metadata." At the top of the metadata editing screen, the administrator selects "dc.subjects.nalt" from the "name" pick list, copies and pastes the NALT subject from the Metadata Spreadsheet into the "value" input box, and clicks on the "add new metadata" button. This is repeated for each NALT subject and saved, when finished, by clicking the "return" button. At this point, any other corrections or additions to the metadata can also be made. The item review, approval, and post-submission addition of the NALT subjects takes the librarian, on average, 10 minutes per bulletin, with 5 minutes spent on the post-submission NALT subject additions.

RESULTS AND DISCUSSION

Currently, 951 bulletins are available in the Texas A&M University Libraries' Digital Repository for worldwide access and usage (Texas Agricultural Experiment Station, 2005 . There are about 1,850 bulletins in total that will be accessible in the Digital Repository when this five-year project is completed. They are discoverable through a number of Open Archives Initiatives (OAI) harvesting search engines, such as Google, AgNIC, and OAIster, in addition to the browsing and searching functions of the Digital Repository's DSpace Manakin interface. In an

item record display in the Digital Repository, the metadata are displayed first, and then a list of the item's page files continues down the page, with the PDF file of the whole bulletin last. The file display is by ASCII character sort order, such that the PDF file that starts with a "B" follows the sequential TIFF page files that all start with a "b."

The Texas Agricultural Experiment Station (TAES) bulletin series has historical significance as a record of agricultural research and rural development in Texas, but it still has scientific and popular consumer significance, too. For example, there is currently an upswing in interest in organic gardening and organic agriculture, in general, and most of the publications prior to 1945 are providing research information on best practices that are essentially "organic." Recently, in response to a reference question about "heirloom" varieties of vegetables and fruits used in Texas, the author was able to refer the patron to three bulletins from the 1890s and 1900s that were available online in our Digital Repository and provided the information. The current problem of bee colony collapse syndrome has created heightened interest in bee diseases, and the TAES Bulletin #116 titled, "The Foul Brood of Bees and the Foul Brood Law," from 1908 was viewed 250 times in July 2009—the second most accessed item in the Digital Repository that month. The bulletin literature covers all fields of Texas agricultural research, including all aspects of animal sciences, veterinary medicine, soil and crop sciences, horticulture and floriculture, forestry, entomology, agricultural economics and marketing, home economics, education and agricultural leadership development, rural sociology, agricultural engineering, and food processing and technology.

The Texas A&M University Libraries' stacks contain partial runs of Agricultural Experiment Station Bulletins from all 50 of the United States of America and Puerto Rico, but it is not our priority to do more than digitize the agricultural serials of Texas. Responsibility for

preserving other bulletins and providing access to them in digital repositories falls on the issuing states and their land grant university libraries and librarians. As each state makes its agricultural serial publications available in OAI metadata harvestable digital repositories, a digital library for agriculture could become a central access point for all of this valuable agricultural information. The digital library for agriculture would harvest metadata from all the state digital repositories and provide a focused, user-friendly central interface to link to all the agricultural publications in their respective digital repositories.

REFERENCES

Anonymous. (1887). *Hatch Act of 1887 establishing agricultural experiment stations*. Retrieved July 8, 2009, from http://www.iahees.iastate.edu/projects/hatch.html

Digital Initiatives, Texas A&M University Libraries. (2005). *Digital initiatives ::Xmlui*. Retrieved August 5, 2009, from http://di.tamu.edu/projects/xmlui

DSpace Foundation. (2002). *DSpace - Home*. Retrieved August 5, 2009, from http://www.dspace.org/

Dublin Core Metadata Initiative. (2008, June 9). *DCMI metadata terms*. Retrieved May 13, 2009, from http://dublincore.org/usage/terms/dc/current-elements/

McInnis, L. L., Scott, T. M., & Gulley, F. A. (1888). *Plan of organization* (Bulletin of the Texas Agricultural Experiment Station, no. 1). College Station, TX: Texas Agricultural Experiment Station.

National Agricultural Library. (2002, May 5). *Agricultural Thesaurus and Glossary home page.* Retrieved August 5, 2009, from http://agclass.nal.usda.gov/agt/agt.shtml

National Agricultural Library. (2009, March 3). *NAL collections: Preservation: Scanning specifications*. Retrieved May 13, 2009, from

http://riley.nal.usda.gov/nal_display/index.php?info_center=8&tax_level=3&tax_subject=158&topic_id=2009&level3_id=6471

Texas A&M University Libraries. (2009). *Submission copyright statement*. Retrieved August 12, 2009, from http://digital.library.tamu.edu/services/scholarly-communication/texas-a-m-repository-policies-and-procedures/submission-copyright-statement

Texas Agricultural Experiment Station. (2005). *Texas Agricultural Experiment Station Bulletin.* Retrieved May 13, 2009, from http://repository.tamu.edu/handle/1969.1/2829