

# Preparing Future Digital Curators<sup>1</sup>

*Paper prepared for 2008 LAUNC-CH Research Forum, Chapel Hill, NC, May 21, 2008*

---

## Part I: A Summary Report on the Digital Curation Curriculum Project

**Christopher (Cal) Lee and Carolyn Hank**

School of Information and Library Science, University of North Carolina at Chapel Hill  
{callee, hcarolyn}@email.unc.edu

### Abstract

Ensuring persistent and meaningful access to digital collections is a challenge – and unprecedented set of opportunities - confronting contemporary institutions of all types and sizes. Work in the areas of digital preservation and access has resulted in a set of strategies, technological approaches, and activities now often called “digital curation.” This evolving area of investigation and practice requires new approaches for professional education and development. An initiative is underway at the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) to develop a graduate-level digital curation curricular framework, course modules, and experiential components. The project, “Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum” (DigCCurr), is a collaboration of SILS and the U.S. National Archives and Records Administration (NARA), guided by an international advisory board. This paper describes the DigCCurr project, with a specific focus on the Carolina Digital Curation Fellowship program.

### Background

In recent years, there has been significant progress in the development of repository architectures; preservation tools and strategies; and approaches for engaging communities that create digital materials. The terms “digital curation” and “data curation” have emerged to represent a set of considerations and activities that reach beyond digital preservation alone. Our working definition of “digital curation” is “the active management and preservation of digital resources over their full life-cycle.” It involves “maintaining and adding value to a trusted body of digital information for current and future use” and is “key to reproducibility and re-use.”<sup>2</sup> Responsible digital curation will require a diverse set of digital curation professionals prepared to work in libraries, archives, museums, data centers, and data-intensive organizations.

While several decades of work have yielded a set of digital curation strategies, technological approaches, and activities, there are relatively few educational

---

<sup>1</sup> This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

<sup>2</sup> JISC, Circular (June 2003a). <http://www.dcc.ac.uk/docs/6-03Circular.pdf>.

opportunities to prepare professionals to work in this area. The “Preserving Access to Our Digital Future: Building an International Digital Curation” (DigCCurr) project, an Institute of Museum and Library Service (IMLS)-funded collaboration between the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH) and the National Archives and Records Administration (NARA) was established in response to this need for dedicated digital curation curriculum at the graduate-level.<sup>3</sup>

### **DigCCurr**

The primary goals of the DigCCurr project are to develop a graduate-level curricular framework, course modules, and experiential components to prepare students for digital curation in various environments. DigCCurr initiatives in support of this goal include the formation of an Advisory Board of experts from Australia, Canada, Italy, the Netherlands, New Zealand, the United Kingdom and the United States. The DigCCurr project also includes two international symposia to engage librarians, archivists, museum professionals, data curators, scholars, other information professionals and the general public in discussions on issues of digital curation and digital curation education. The first symposium, DigCCurr2007, was held April 18-20, 2007 at UNC-CH, attracting nearly 300 participants from ten countries.<sup>4</sup> The second symposium is slated for early April 2009, and will coincide with the culmination of DigCCurr’s three-year grant period. This paper introduces another critical facet of the DigCCurr project – the Carolina Digital Curation Fellowship program.

### **Carolina Digital Curation Fellowship**

The Carolina Digital Curation Fellowship program supports five graduate students interested in research and work in digital curation for two academic years (2007-09). The Carolina Digital Curation Fellows combine coursework with a digital curation practicum assignment in a UNC-CH academic library, archive, data center or service center, leading to a master's degree in Information Science or Library Science. The program's goal is to prepare information professionals to work in the 21st-century environment of trustworthy digital repositories. It offers the Fellows the unique opportunity to interact and collaborate with key international leaders in digital preservation, as well as providing: 1) a practicum assignment in a Carolina data or digital repository; 2) a competitive annual stipend for two years; 3) in-state tuition and health coverage; and 4) mentorship by senior university library, archives, data center or service center administrators. Four practicum settings were chosen for the 2007-08 inaugural year of the Carolina Digital Curation Fellowship program:

- **University Library.** Two Fellows were assigned to the University Library's new Carolina Digital Library and Archives. The Fellows gained valuable, hands-on experience working on a number of established and emerging initiatives, including the award-winning digital publishing initiative, Documenting the American South (DocSouth), and the University's developing institutional repository program.

---

<sup>3</sup> DigCCurr project website: <http://www.ils.unc.edu/digccurr/index.html>

<sup>4</sup> DigCCurr2007: An International Symposium on Digital Curation website: <http://www.ils.unc.edu/digccurr2007/>

- **Information Technology Services (ITS) Teaching and Learning division.**  
The Fellow assigned to the Teaching and Learning division worked on several projects with the overall intent of applying principles of digital curation to practical teaching and learning applications. Particular areas of focus included open source course management systems, outreach and engagement in regard to emerging technologies, and development of enhanced virtual learning environments
- **The Odum Institute for Research in Social Science.** This practicum setting provided the assigned Fellow with opportunities to investigate data life cycle characteristics in the context of a well-established and successful social science data repository. The Fellow contributed to initiatives exploring Data Documentation Initiative (DDI) compliance and IQSS Dataverse Network usability, clarity and functionality.
- **ibiblio.org.** [ibiblio.org](http://ibiblio.org), a conservancy of freely available information and home to one of the largest "collections of collections" on the Internet, provided several unique opportunities for its assigned Fellow. The Fellow worked on the improvement of collection browse functionalities through the identification and tagging of archival sites, and locating and registering unregistered sites. Additionally, the Fellow contributed to the development of automated collections' index registration processes at the point of collection creation, and investigated Open Archives Initiative (OAI) compliance capabilities through application of Dublin Core metadata elements.

#### Facilitating the Fellowship Program

Several activities have facilitated the Fellows digital curation educational experience. A one-credit seminar exclusively for the Fellows was designed and administered by the DigCurr project team in Fall 2007. The objective was to provide Fellows an introductory understanding of digital curation, including historical context and key principles. This seminar was held approximately every other week. Additionally, we prepared and shared with the students a detailed syllabus supplement, listing resources for terminology and important websites listing key projects, organizations, and electronic mailing lists. Further, the seminar provided opportunities for Fellows to discuss their unique practicum experiences and share new discoveries with their “fellow Fellows” and academic advisors in a supportive and collaborative group setting.

Since the Fellows are expected to provide 20 hours of service at their practicum settings during the Fall and Spring semesters, a Carolina Digital Curation Fellowship Practicum Handbook (2008-09) was created to provide guidance on the Practicum aspect of the Digital Curation Fellowships, and was distributed to Fellows, the Practicum site supervisors, referred to as DigCCurr Partners, and Fellows’ academic advisors. Contents included: 1) introduction to DigCCurr project and Fellowships; 2) identification of roles and responsibilities, including a) Fellows’ requirements; b) Fellows’ recommendations; c) DigCCurr Partners’ recommendations; and d) SILS Academic Advisors’ requirements; 3) listing of required forms (Learning Agreement and Evaluations); 4) Calendar of important dates; and 5) contact information.

Fellows, in consultation with their Practicum site supervisors and their academic advisors, completed a Carolina Digital Curation Fellowship Practicum Agreement form.

The form is essentially a learning contract between Fellows and their DigCCurr Partners, describing goals and objectives for their annual Practicum assignment. If changes in goals or objectives transpire over the course of the two-semester practicum, Fellows are expected to submit a revised Contract, following consultation with their respective practicum supervisors.

Further, Fellows and their Practicum site supervisors had two opportunities to evaluate performance and work toward the goals set forth in the learning contract. Fellows and DigCCurr Partners, respectively, were each provided with an interim evaluation, due at the end of Fall semester, and a year-end evaluation, completed at the close of the Spring semester.

### **Informing the Curriculum**

The educational objectives for the Carolina Digital Curation Fellows have been informed by the ongoing development of a Matrix of Topics for Digital Curation Curriculum. The draft matrix is organized into six dimensions (factors, topics, issues): 1) type of resource; 2) functions and skills; 3) professional, disciplinary or institutional/organizational context; 4) mandates, values, and principles; 5) prerequisite knowledge; and 6) transition points in information continuum. The matrix is intended to serve as a tool of organizing, comparing and planning for digital curation curriculum content. Educational objectives have been further informed by the development of High Level Categories of Digital Curation Functions, identifying twenty-four high-level function categories (e.g. production, transfers, ingest), and three meta-level functions to be applied to any of the high-level functions (e.g., evaluation and audit).<sup>5</sup> Ongoing feedback from the Fellows is a valuable data source in this process. Their in-class and practical experiences are helping us to determine what aspects of our Matrix may require further revision, and even more importantly, what aspects should receive more or less attention in our courses, in order to best prepare digital curation professionals.

---

## **Part II: Four Perspectives on Applying Academic Understanding in a Practice Setting**

**John Blythe, Lisa Gregory, Samantha Guss, and Jennifer Mantooh**  
Carolina Digital Curation Fellows, School of Information and Library Science,  
University of North Carolina at Chapel Hill  
{blythej, gregoryl, jmmantoo}@email.unc.edu; sag231@gmail.com;

### **Abstract**

This paper presents four perspectives on the application of digital curation principles in real-world settings, presented by the Carolina Digital Curation Fellows. The Carolina Digital Curation Fellowship program supports five graduate students interested

---

<sup>5</sup> For a more specific understanding of these two curriculum products, see DigCCurr Progress Report: Development of a Graduate-Level Digital Curation Curriculum, a poster presented at ALISE 2008 (Philadelphia, PA): <http://www.ils.unc.edu/digccurr/digccurr-alise2008-poster-v01.pdf>

in research and work in data and digital curation for two academic years (2007-09). Funded through a grant from the Institute of Museum and Library Services (IMLS), the Carolina Digital Curation Fellows combine coursework with a digital curation practicum assignment in an academic library, archive, or data center at the University of North Carolina at Chapel Hill.

Presented are reflective accounts of specific digital curation challenges from the perspectives of four of the Fellows, within the context of their Fellowship settings. John Blythe writes on the challenges associated with inconsistent file naming conventions from his experience building a directory within the dark archive for the Carolina Library and Digital Archive's (CDLA) Documenting the American South (DocSouth) publishing initiative. Lisa Gregory has been working on image scanning projects for the CDLA's Digital Production Center (DPC), and writes on the intersection of the image needs and expectations of project managers, production center staff and librarians, and the importance for all stakeholders to have a common, basic understanding of imaging best practices and standards. Samantha Guss works at the Odum Institute for Research in Social Science's Data Archive, one of the oldest and largest collections of social science datasets in the United States. She writes on her project to standardize the archive's metadata to prepare it for migration to a Data Documentation Initiative (DDI) compliant system and how such activities necessitate the need for good ingest metadata practices and education of data creators. Lastly, Jennifer Mantooth writes on her experience attempting to identify archival value in defunct WebSpaces hosted at her practicum site, *ibiblio.org*. Because *ibiblio.org*'s implementation of openness extends to contributor access and maintenance of their own WebSpaces, it leads to a unique situation of how to archive abandoned and often context-less files, and the question of whether these WebSpaces should even be considered archival in the first place.

---

## **Current Status of DocSouth CD Migration: A Report from the Carolina Digital Library and Archives**

John Blythe

### **Background**

During Fall 2007, I migrated 44,787 files (903 GB) from 1642 CDs to the DocSouth section of the dark archives. The migration process took 100 hours. Currently the directory structure in the dark archive mimics the structure of the DocSouth CD library. Top level folders match the names of the CDs in the library (i.e., Disc 1 in the dark archive is an exact copy of CD1 in the CD library). The plan is to restructure the dark archive directory to match the structure of the DocSouth database.

During the fall semester I also evaluated and recommended a software application for automatic metadata extraction from the image files. The recommended software is the open-source application, JSTOR/Harvard Object Validation Environment (JHove), which can extract technical metadata from these archive image files.<sup>6</sup> Batch metadata extraction has yet to begin. The dark archive directory needs to be rebuilt first.

---

<sup>6</sup> For the JHove project website, see <http://hul.harvard.edu/jhove/>

## Current Status

During Spring 2008, I worked with programmers to rebuild the DocSouth directory in the dark archive. The process was slow, with obstacles encountered along the way. In short, many file names in the dark archive do not match file names in the database. This has made automated matching and renaming difficult. A DocSouth programmer wrote and ran a script for matching. The results of this first attempt at auto-matching were:

- 16,847 file names matched and assigned appropriate item IDs from the database;
- 24,280 file names did not match item IDs in the database; and
- 1,257 file names matched more than one item ID.

The auto-matching process had included 9,848 files related to Dickens' material. These CDs, which are included in the DocSouth CD library, were migrated to the dark archive but, as one might expect, are not included in the DocSouth database. After removing these files from the total of unmatched file names, we were left with about 14,800 file names that did not match.

My next step was to begin working through the list of unmatched file names and file names with multiple matches to determine the reasons for the matching problems. In particular I wanted to see if any patterns could be discerned. My hope was that knowledge of patterns could help with a second attempt at auto-matching.

### Instances of "No Match"

Unfortunately I was able to discover only one pattern that proved at all helpful. Almost every folder has a file named *small.jpg*. These files are thumbnails of larger images used in DocSouth, often thumbnails of a portion of the title page, cover or verso. These files were not included in the DocSouth database. Originally I attempted to determine the larger file for which *small.jpg* was a thumbnail and then assigned the appropriate item ID to the thumbnail. Based on discussions with DocSouth staff and administrators, it was decided that all *small.jpg* files can be removed from the dark archive.

Other perceived patterns did not prove particularly helpful in developing a second approach to auto-matching. In some cases the dark archive includes several files of the same image, each at a different resolution. For instance, Disc 65a contains a folder named *confnational*. That folder, in turn, contains images named *confissue2\_at\_75.jpg* and *confissue2\_at\_150.jpg*. Unfortunately, the database does not contain any iteration of this file name. There is no *confissue.jpg*, *confissue2\_at\_75.jpg* or *confissue2\_at\_150.jpg*. In addition to the suffixes described above, others include *-thumb*, *-1*, *-50*, *-75*, *-150*, *\_100*, *\_150*. Programming staff made a second attempt at auto-matching by stripping the suffixes from the file names. Unfortunately, only 85 new matches were made as a result of this process.

In many cases, the file names for images on CD and in the dark archive are slight variations of the file names associated with Item IDs in the DocSouth database. In other cases, files were named differently in the database and in the DocSouth website than they were on CD and, consequently, in the dark archive. In working my way down the list manually, I had some success making matches by following the method described below.

1. Establish Parent\_ID of folder in dark archive by searching ITEM table in DocSouth database.

2. Using Parent\_ID, search ITEM\_CHILD table to determine the Item\_ID's that correspond with Parent\_ID.
3. Use Item\_ID of child to search ILLUSTRATION\_ITEM table for the specific file name associated with that ID.
4. *Does file name in database match file name in dark archive? If not, go to next step*
5. Look on DocSouth website at image corresponding to file name.
6. Look in dark archive at images associated with same collection.
7. *Do images match but have different file name? If yes, then go to step 8*
8. Assign image in dark archive appropriate item ID.

Following is an example of the process described above for a specific disc in the DocSouth CD library:

1. Disc 16 contains a folder named Uncle Johnson, the Pilgrim of Six Score Years. A search of the ITEM table shows that the Parent\_ID for this group of images is 39.
2. Searching 39 in the ITEM\_CHILD table shows that Items 53968 and 53969 correspond to this Parent\_Item.
3. Searching 53968 in the ILLUSTRATION\_ITEM table shows that the ID corresponds with a file named *fostecv.jpg*.
4. The file on CD and in the dark archive is *unclejcv.jpg*
5. Found image with name *fostecv.jpg* on DocSouth
6. Found image in dark archive with name *unclejcv.jpg*.
7. Images match.
8. Assign the item ID 53968 to the dark archive image named *unclejcv.jpg*.

Programming staff tried to incorporate this process into the 2<sup>nd</sup> attempt at auto-matching. Unfortunately, this second attempt was minimally successful in reducing the number of unmatched files. Just 550 files were matched. It was theorized that one of the reasons for the low number of matches is that there are slight variations between the names assigned to the folders in the dark archive and the names listed in the *sort\_by* field of the *Item* table in the DocSouth database. In some cases the names varied because a comma was used in one and not the other. In other cases the wording was different. It is speculated that a script can be written that would strip punctuation from the name fields, but it is unclear how much such a step will help in reducing the size of unmatched files.

There are other reasons for the difficulties in auto-matching. In some cases there was a typo in either the filename in the DocSouth database or in the dark archive. Typos might include an extra space, a mistyped letter or an extra letter. For instance, the Circular to the City Council folder in the dark archive (in Disc 2) contains a file *circlrtp.jpg*. In the database, that file is named *circltp.jpg*. Such slight variations made auto-matching impossible.

In a slight different version of the problem described above, file names on Discs 29 and 30 (the History of Louisiana Negro Baptists) were one number off from those in the DocSouth database. For instance, a file named *hicks22.jpg* in the dark archive (and on CD) was named *hicks23.jpg* in the database. This was the case for 44 *jpgs* and 44 *tiffs*.

This inconsistency could only be confirmed by opening each image in the dark archive and determining whether it matched an image on the website. Then I had to check the name of the image on the website and verify that the same file name was in the database.

There is yet another reason that some files in the dark archive do not show up in the DocSouth database. Some images were scanned and archived on CD, but then not used on the DocSouth website. Consequently, these were never assigned an Item\_ID.

### Instances of Multiple Matches

There are numerous cases where images from different publications share the same file name. Consequently, auto-matching produced a list of item IDs matching the file name. In some cases, the appropriate ID could be determined by choosing the number that fell in sequence with other IDs from the publication. For instance, the file *taylorp.jpg* (from Reminiscences of My Life in Camp with the 33<sup>rd</sup> and found on Disc 16) matched items 49333, 49934 and 51474. Because other images from the same publication had such Item numbers as 51478, 51476 and 51473, I safely assumed that the appropriate item ID was 51474. This process was done manually. It may be possible to automate this task following further investigation by programming staff.

### **The Future**

Manual matching is tedious and time consuming. It requires that multiple windows be open on the desktop – the spreadsheet listing all files and their status, the DocSouth website, the DocSouth database (which I viewed and worked with via Navicat), a Word document for listing discrepancies and problems, and the digital archive (dark archive) folder. Future attempts at auto-matching should explore the possibilities described above, including:

- Stripping punctuation from folder names in the dark archive and *sort\_by* names in DocSouth; and
- Seeking to reduce the multiple item ID possibilities by creating an automatic process that can evaluate Item IDs in terms of their numerical proximity to each other.

Because there was wide variation in file naming and how data was entered into the database, auto matching may still do little to reduce the number of unmatched files. If this is the case, then it may be necessary to hire a student to work his/her way down the list using the 8-step method described earlier in this document. Once matching has concluded there are several steps that should occur. Several folders should be deleted from the dark archive. I was told to migrate into the dark archive all CDs from the DocSouth library. It turns out that in a few cases CDs are duplicates of each other. In cases of duplication, one disc (folder in the dark archive ) should be deleted.

The next step should be restructuring of the directory in the DocSouth section of the dark archive. This is a task that programmers will need to carry out. Following restructuring, extraction of technical metadata should occur. I recommend the use of JHOVE for this task since it can be scripted to perform batch metadata.

Finally, proper management of digital files in archives requires the creation of checksums. A checksum should be created when the file is moved in to the dark archive, then the checksum should be verified on a periodic basis. Additionally, the checksum should be verified any time the file is edited or moved. An MD5 hash is the standard

algorithm used for this function. Thus far, no checksums have been created for the DocSouth images in the dark archive.

---

## **The Intersection of Image Needs and Expectations: A Look from the Carolina Digital Library and Archive's Digital Production Center**

Lisa Gregory

### **Introduction**

As a fellow, I have had the opportunity to work in the newly formed Carolina Digital Library and Archives (CDLA). The CDLA was formed to encompass library digital initiatives, from the well-established Documenting the American South (DocSouth) to newer endeavors like a partnership with the Internet Archive. My particular position is in the Digital Production Center (DPC) of the CDLA. Staff in the Production Center use a variety of scanners (including flatbed scanners, film scanners and a Zeutschel book copier) to digitize source materials both for patrons and for library digital collections. This last year, I worked primarily with large format materials, specifically maps. We image most maps by placing them within a vacuum table and then scanning them using a BetterLight digital scanning back. My workspace looks a lot like a traditional photographic studio, except that on the camera mount there is a scanning back on the rear standard where the film holder would be, and I do all of my post-capture work on a MacIntosh, not in a darkroom. For those interested in technical details, the standard image is captured at 100% of its original size, 300pixels per inch. We generally use the .TIF file format, and file sizes range in the hundreds of megabytes.

Most of the materials I have been digitizing are maps for two different projects under development in the CDLA. One of the interesting challenges I've encountered in my position has been related to image standards and quality control. Before I explain this further, I should give you an idea of the workflow for these particular projects. First, one of the project managers selects maps from the library's map collection. If fragile, the maps might be examined by conservation staff before heading to the DPC. Maps ready to be imaged are brought to the DPC in batches of approximately twenty or thirty. I image the maps, creating both a preservation copy for our dark archive and a display copy that will be used for the respective project. Minimal technical metadata is also recorded.

### **Negotiating Expectations with Operations**

The intersection of standards and department-specific expectations has, at times, been an interesting issue to navigate. As the DPC operates independent of any particular project, we have adopted certain imaging standards based on industry standards, our own technological capabilities, and our department's resources. Our monitors and printers have been color calibrated. We strive to adhere to best practices regarding our workspaces and tools. In addition, it has been reinforced to me that our first priority is to serve our staff's needs. What happens, then, when staff members have different expectations of image standards than those we've established in the production center?

Let me give a specific example. Recently, we had the opportunity to upgrade our digital scanning back and lights. This occurred around two-thirds of the way through one of the projects I had mentioned before, a project that involves scanning several hundred

Sanborn maps of North Carolina towns. These maps are all very similar in color, size and format, which served to highlight the fact that the images produced after we upgraded our equipment were visibly different from those before the upgrade. Compound this with the fact that staff members in other parts of the library were viewing these maps on monitors whose screens had not been calibrated. Though the DPC has improved our technology and is better able to approximate the original objects, we now had a problem of consistency between departments and between the scans done before and after the technology upgrade.

### **Recommendation**

This situation led to staff members trying to put into words fine details that different people perceived differently in images, and discussions about the importance of monitor calibration. On a personal level, it highlighted for me the tension between producing an image that accurately reflects the original and fulfilling patron and staff demands – these aren't always the same ends. Despite technological sophistication, deciding what to do with the raw scan sitting in front of you on your computer monitor is never a black and white decision. Post-processing choices range along a continuum. On a more practical level, this experience has made me realize that in any extensive digitization project involving different stakeholders, an afternoon of basic image processing education for all, held early in the project, would give everyone a common vocabulary. It would alert project managers, or student workers, or librarians to the DPC's, or the like at other institutions, own standards and allow them to express their own expectations.

When multiple people compare a digital image to the original under different lights, with different technology, through different eyes, it underscores the often subjective and sometimes infuriating nature of trying to be faithful to the original material while serving human needs. As I expect many are discovering every day, technology and standards cannot take the place of human collaboration.

---

## **Metadata Issues at a Large Social Science Data Repository: A Snapshot of Digital Curator Needs in the Practice Setting**

Samantha Guss

### **Background**

I went to work at the Howard W. Odum Institute for Research in Social Science (Odum) data archive not knowing much of anything about social science data, digital archives, or digital curation at all. After a year of exploring the technical and managerial aspects of the archive, I have a much better understanding of these things and, more importantly, where they all intersect.

One of Odum's biggest recent undertakings has been participation in the Data Preservation Alliance for the Social Sciences (Data-PASS) project, which makes interoperability, backups, and increased data security possible across many of the country's largest social science data archives, including Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan and the Harvard-

MIT Data Center.<sup>7</sup> Under Data-PASS, Odum will offer all of the other archives' data to its users and Odum's data will be stored and made available at other sites as well. Part of this process involves complying with the Data Documentation Initiative (DDI), an XML standard for the social science community to easily share data. Eventually, all of the studies archived at each of these digital repositories will be interoperable and DDI-compliant, with a logical and easy user interface.

### **Metadata Investigation**

Before any of this was possible, the individual studies held by Odum needed a little housecleaning. While other graduate assistants worked to make sense of the studies – rearranging variables according to their accompanying codebooks and uploading the new studies to a new database system – I was set to the task of standardizing the studies' metadata records. Looking at all of Odum's study metadata in a table made it immediately clear that different fields are not used the same way by all study creators. Even if a field name's meaning was unmistakable – place of creation, for instance – no standard thesaurus or structure was used to enter the data. For example, some creators used only a state, some spelled out the state and some abbreviated it, and some abbreviated city names while some used terms like "RTP" instead of spelling out "Research Triangle Park." While these might be understandable to a human reading through one study at a time, inconsistencies like these presented a larger problem for machine readability (searching) and making sense of a large number of studies at one time. Searching for a study performed in New York City might not turn up ones entered as "NY, NY" or "NYC." To fix this, I went about running scripts and hand-correcting the records, as well as writing up rules for standard data entry in the future.

### **Lessons Learned**

As a result of this long and often tedious process, and through consultation with the other graduate assistants, I learned a few things about data curation best (or better!) practices. Namely, incorporating "clean up" into the ingest process and helping the data creators do a better job of metadata creation in the first place. First of all, the best metadata is generated by those who know and understand the study the best – the researcher/creator. However, the creator doesn't always understand the importance or benefits of good metadata and may not understand how to fit their study into the archive's system. It isn't their job to know that, but it would be beneficial for everyone if we, the data archivists, would teach and help them with it. It's only a few extra keystrokes and no extra work for a data creator to enter "New York, NY" instead of "NYC," and would save many hours later for the study's curators. Good study metadata also benefits the creator – it helps with recall later and decreases the amount of time spent devising one's own organization system. Instruction in data management seems simple yet is not offered widely or uniformly as far as I know. Therefore, a future project at Odum (possibly part of my own Fellowship activities for next year) will be to incorporate this instruction in the popular short courses on statistical packages that the Institute already offers.

Of course, it is improbable that even the best instruction program could resolve every issue of standardization in study metadata, and I recognize that there will always

---

<sup>7</sup> Data-PASS project website: <http://www.icpsr.umich.edu/DATAPASS/>

need to be some work done by the archive to incorporate new studies into the interoperable system. Therefore, a data "cleanup," like I did across the whole archive, should be a part of the ingest process – each incoming study will be cleaned up as it is received to avoid large-scale problems in the future. If things are done consistently, they become much easier to manage. Even if things need to be changed later for whatever reason, a script can be used to change many fields at once instead of having a person go through one at a time and sort through a wide range of problems.

My time at the Odum Institute's Data Archive so far has taught me a lot about the procedures for curating usable, logical, and trustworthy data sets, as well as how these procedures can be improved. My work over the past year has exposed me to some of the challenges of digital data curation, but also inspired some exciting ideas about how to solve them in the future.

---

## **Determining Archival Value in Orphaned WebSpaces: An Investigation within the ibiblio.org Domain**

Jennifer Mantooth

### **Introduction**

ibiblio serves as the “public’s library and digital archive” and offers a collection of creator contributed WebPages and Linux software. As one of the earliest WebSpaces on the Internet, ibiblio prides itself in its long history of offering a free forum for sharing information and empowering contributors. As may be expected, ibiblio’s fifteen-year tenure has yielded many valuable websites, with many of these created by special interest groups that may not have had the funds to communicate elsewhere. Certainly then, for an age where archives are broadening their collecting focus to document both mainstream and under-represented groups, ibiblio offers rich archival fodder. Additionally, because all website materials are housed on ibiblio’s servers, ibiblio staff has complete access to ‘deep web’ materials – something that many other web archiving projects do not. However, despite the benefits of the ibiblio environment, the overarching focus on openness and contributor empowerment somewhat conflicts with the policies of traditional archives.

### **The Situation**

ibiblio's implementation of openness extends to unrestricted contributor access and maintenance of their own WebSpaces, at the moment of initial creation and throughout the life of the site. Although ibiblio does not routinely create and preserve snapshots of their contributors’ websites, it does preserve content and access to abandoned websites. Most ibiblio websites depend heavily on index pages to render meaning to the supporting files. When index pages have been removed either by the contributor or maliciously (due to weak permissions on the account), ibiblio is left with a list of files with little to no context available. As the solitary archivist working part-time at ibiblio and attempting to re-appraise over 6,000 websites, I faced the question of how to deal with abandoned, context-less sites: should these Websites be archived, and if so, how?

In accordance to archival principles, context-less data offers little benefit to users. Although, due to decreasing storage costs, it may be easiest to keep the files as they are and proclaim them archived, an archive's role is to add value to its holdings, not just to maintain access. And so, I determined that I could not just ignore them and, furthermore, the websites deemed 'archive worthy' would need additional attention. That is, either re-creating an index page, or composing a description to enter into the ibiblio index page that would give context to the information presented.

### **The Response**

In an effort to err on the side of caution rather than attempt to gauge archival value, I determined to save everything that was remotely archival and discard only those that provided a compelling reason for removal. This reappraisal revolved around the following questions.

1. Does the website contain content?
2. Is the material accessible?
3. Is the material easy to negotiate?
  - a. Are the file names meaningful?
  - b. Do the files have value on their own?

The first question weeds out abandoned sites that are just empty directories, although it is possible to argue that the shells of websites – empty cgi-bin directories, temp directories, and the like – may offer informational value to future researchers. Many archivists agree that appraisal based on anticipated use is dangerous and that arguably everything has the possibility to provide informational value to some individual. Therefore, informational value is not a good basis for appraisal. The absurdity of creating a special entry in the ibiblio collection index for empty websites reinforced the decision to remove these websites.

The second question attempts to measure the independency of the website. Due to the age of many of the abandoned sites, most of the links to external pages were defunct. As stated earlier, the re-appraisal was based on identifying a compelling reason for removal. While the presence of defunct links was noted in the collection index website description field, the mere presence of defunct links was no cause for removal of the website. However, abandoned websites that contained little content beyond a list of defunct links were removed. A few abandoned websites were no longer renderable due to the fact that the databases had not been migrated from older versions. These websites were set aside for future investigation.

The final question served two roles: it identified personal websites for removal and determined the quality of the contextual description entered into the collection index for the website. Due to the quantity of websites to be re-appraised, website re-appraisal was conducted concurrently with website description. The investigation of file naming conventions and the originality of the information contained in the files aided the description and actually had little to do with re-appraisal except in situations where the website was determined to be personal and was thereby removed due to breach of collection guidelines.

### **Ongoing Issues**

The majority of abandoned websites consisted of .html pages and therefore posed no immediate format issue. However, some database-driven abandoned sites required data migration in order to view the contents. The re-appraisal guidelines for these may necessarily be more stringent than their .html dependent counterparts due to ibiblio's scant resources. These sites would need to be actively maintained and migrated by ibiblio staff and therefore it is likely that only a small list of websites deemed extremely important could actually be cared for. Additionally, the ibiblio non-interference policy currently does not support modifying the existing files within the abandoned website in order to alert potential users of defunct links or outdated facts.