Digitization in the Real World
Lessons Learned from Small and Medium-Sized Digitization Projects

Edited by
Kwong Bor Ng & Jason Kucsma

Metropolitan New York Library Council
Digitization on a Dime: How a Small Library and a Big Team of Volunteers Digitized 15,000 Obituaries in Just Over a Year

Elizabeth Goldman (Kingston Frontenac Public Library)

Abstract
In 2006 and 2007, Chelsea District Library, a small public library in Chelsea, Michigan, digitized a collection of 15,000 obituaries on a small budget by using a staff of nearly 50 volunteers and open source software. The author describes the research and planning that led up to the project; unique aspects of the staffing and technology for the project; and the resulting database, which contributed to the library being named “Best Small Library in America” for 2008 by Library Journal. The database continues to be updated, expanded, and improved, and the use of volunteers as the primary workforce has had long term rewards for the library. The chapter demonstrates the accessibility of digitization projects to libraries, even those without pre-existing expertise, large staffs, or big budgets.

Keywords: Database, Digitization, Genealogy, Obituaries, Open source, Volunteers.

Introduction and background
Especially in small towns, the public library may serve multiple roles as library, museum, archives, and community center. In many cases, libraries accept donations of rare or unique historical material from
families in the area with little thought for long-term maintenance. Technological advances in the last few decades have given libraries new options for preserving local history collections and making them more accessible through digitization. While even the smallest libraries have staff educated in reference and circulation procedures, however, relevant training in archival methods and technology is rare, making the prospect of a digitization project overwhelming. At the same time, small public libraries often have little or no funding for such projects.

Chelsea District Library is a single-branch public library serving 14,000 people in southeast Michigan, about 50 miles west of Detroit. The library was established in 1932 by the local Women’s Club and, in 1999, became a district library system serving both the town of Chelsea (population 5,000) and surrounding townships. The earliest settlements in the Chelsea area date back to the 1830s and many local families can trace their roots back to the town’s founders, resulting in a rich history. From its earliest days, the library served as a repository for local historical and genealogical material, housing a local history room on the cramped third floor of the McKune House, its location from 1959-2000. In 2006, after extensive renovations and additions to the McKune House, the library moved from temporary quarters back to its historic home on Main Street in Chelsea, leading to renewed interest in the local history collection.

A collection of about 50,000 index cards known as the Family History Index made up a major component of the local history material. A retired lawyer and amateur genealogist named Harold Jones started the collection as a hobby, clipping obituaries from the local Chelsea Standard newspaper and other sources and pasting them onto 4 x 6 index cards, along with cross references that allow women to be located by maiden name. Upon his death in 1987, Jones’ family donated the collection to the library, where it received extensive use by local and visiting genealogists. Library staff and volunteers completed a project from 2000-2002 to clean up and update the collection, since then volunteers have continued adding new clippings.

As part of a planning process leading up to an election to fund an expansion of the building and the staff, the library identified local
history as a priority for Chelsea area taxpayers and the Family History Index as a prime candidate for digitization work. At the time, the professional librarian staff consisted of the director, three department heads (adult services, youth services, technology services), and a part-time librarian, leaving few resources to focus on a project of this scope. In the spring of 2005, the library replaced the departing part-time librarian with a full-time librarian (the author), adding additional duties of managing the library’s website and digitizing the Family History Index.

This paper will describe how a librarian and a team of four dozen volunteers completed the digitization of 15,000 records from the Family History Index in just over a year, resulting in a highly usable database that helped Chelsea District Library earn its distinction of “Best Small Library in America” for 2008 from Library Journal and the Bill and Melinda Gates Foundation. The first section will describe preparation for the project, including research in archival standards and digitization techniques, as well as the recruitment and training of the volunteer workforce. Section two will describe the decision-making process that went into the choice of open source software in order to create a user-friendly, free database of the records on a limited budget, as well as the work of developing and testing the database itself. Section three will offer an overview of the workflow for staff and volunteers as they did data entry, scanning, and proofreading of the records. Finally, the paper will describe the resulting database, current upkeep and expansion, and how the project served as a model for additional digitization work.

**Project preparation**

**Research**

Initial research focused on archival and digitization standards. No one on the staff had a specific background in either archives or genealogy that would fit the requirements of the project, so research started from a very basic level. Research was conducted largely online and in books, as well as by speaking with archivists and libraries that had done similar projects. Internal research played a role, too:
understanding how the Family History Index had historically been used and its importance to the community; exploring the scope, size, and fragility of the collection; and agreeing on project goals.

At the time, in mid-2005, quite a few libraries had launched efforts to put obituary indexes online for use by genealogists and historical researchers. The vast majority of these projects resulted in static webpages or simple databases that provided access to citations but not complete text. This early wave of digitization projects represented an important first step on the path toward full electronic access by giving researchers more complete information about the contents of a library’s collection. However, those who found an obituary citation online still had to contact the library to obtain more detailed information or the complete text of the obituary.

Chelsea District Library was lucky enough to have a collection that included full-text clippings of obituaries on a significant portion of the cards in the Family History Index. Because of this and advances in technology, one of the library’s primary goals was to make complete text available freely online, meaning genealogists and researchers had at least the possibility of meeting their research needs without leaving their computer or contacting the library at all. To further this goal, the library sought to develop a database that was free, simple to use, and contained relatively small image files accessible even by those with dial-up internet connections.

Important sources, although they sadly have not been updated in recent years, were the book and accompanying website, “Moving Theory into Practice,” and the Making of Modern Michigan project. Moving Theory into Practice: Digital Imaging for Libraries and Archives by Anne R. Kenney and Oya Y. Rieger (Mountain View, CA: Research Libraries Group, 2000) provided a good grounding in what to consider when planning a digitization project, as well as minimum standards for use and long-term preservation. An online tutorial housed at Cornell University Library (2010) offered step-by-step instructions for taking a project successfully through to completion. The Making of Modern Michigan was an IMLS-funded joint effort by the Michigan State University Library, the Library of Michigan, the
Michigan Library Consortium, and others (Michigan State University Libraries, 2005). Although its heyday had already passed by the time Chelsea embarked project, the website provided helpful background information on the structure of digitization projects and, especially, scanning equipment that had been vetted. The scanner models in the Making of Modern Michigan are no longer produced and libraries may have moved to higher minimum DPIs for scanning, but the advice offered at both sites remains sound.

Two more recent resources, available freely online, are BCR’s CDP Digital Imaging Best Practices, from the Bibliographical Center for Research in Colorado (Collaborative Digitization Program, 2008) and NISO’s IMLS-funded A Framework of Guidance for Building Good Digital Collections (National Information Standards Organization, 2007), which within the framework includes links to many other resources on more specific topics.

**Project scope**

Armed with this information, Chelsea District Library took a look at the Family History Index. While the cards had been stored away from the light in filing drawers, they also had been used heavily by patrons over the years, resulting in some wear and tear. In addition, neither the index cards nor the glue were acid free, and newsprint is one of the most acidic papers. The oldest cards and the oldest clippings dated back to the 1950s, resulting in some that were in very delicate condition. Along with the unique and irreplaceable nature of the collection, this delicate condition led the library director to decide that the cards should be scanned manually in-house rather than sent to a contractor who would likely feed them into a scanner, possibly resulting in damage.

This decision meant significant labor for library staff members and volunteers. In May 2005, the library applied for a grant from the State of Michigan to fund the digitization, which would have allowed for the hiring of contractors, but the application was turned down. Luckily, the library already had a well-established volunteer services program. Lacking any funding beyond $5,000 committed by the library, the project manager developed a volunteer-driven plan and
made the choice to seek an open source solution for the database, resulting in savings on the equipment side. An RFP was opened to competitive bidders in the fall of 2005, with selection of a contractor and design and testing of the database completed by early 2006.

During this time, the library also made decisions about the scope of the project. While the collection itself was estimated to consist of more than 50,000 cards, closer to 25,000 obituaries were represented, due to Jones’ system of cross-referencing women by maiden name. A database eliminated the need for this. The remaining set of obituaries came from a number of sources:

- gravestone transcriptions from Chelsea’s three cemeteries and several others in neighboring areas
- notes culled from early histories of the area, as well as scrapbooks and other material in the local history collection
- notes from death notices published in the local newspaper, The Chelsea Standard, taken from microfilmed versions of the paper, dating from about 1887-1950
- complete obituaries from The Chelsea Standard, clipped and pasted on cards, dating from about 1950 to the present
- complete obituaries from the newspapers in two nearby cities, Ann Arbor and Jackson, clipped and pasted on cards, dating in the 1970s and 80s.

Of this material, the first three could be included in the project without further consideration of copyright, as the donation of the Family History Index to the library included rights to copying the material. The Chelsea Standard, a weekly publication owned by Heritage Newspapers, supported the project from the beginning, granting copyright release for material originally printed in its pages.

Unfortunately, the publisher of both the Ann Arbor News and Jackson Citizen-Patriot declined to grant copyright permission. The library considered this a minor setback, as obituaries from those two newspapers covered a span of only about 20 years. The impetus for collecting from newspapers in the neighboring cities was that some Chelsea area residents chose to publish obituaries only in these publications. The number of citizens who fit this scenario, however,
was greatly outweighed by the number of obituaries of people who had no connection to the library’s primary service area at all. The library considered entering data from these obituaries but refraining from scanning them, which would not have violated copyright law, but chose instead to leave the cards for a potential future project and focus instead on truly local residents.

**Staffing**

At this project’s initiation, Chelsea District Library had about twenty employees. The project could not be completed by paid staff, and the small budget precluded the hiring of contract labor.

The library had a well-established and strong volunteer program, including a tradition of volunteers working with the Family History Index. Started as a volunteer effort in 1932, the library had always had strong support from volunteers, and genealogy in general is a topic that draws volunteers. Nevertheless, bringing in volunteers double the size of the library’s own staff would not have been possible without a coordinator, who had developed procedures, documentation, and processes for intake, training, and evaluation.

Building on this strong foundation, the library advertised through its newsletter, the local newspaper, word-of-mouth, and presentations to organizations such as the county genealogical society. The Family History Index was a well-used collection and one of the best sources for obituaries for the area, so the library was able to draw in volunteers who did not live in the Chelsea district itself. Volunteers filed standard application forms, which covered basic contact information, times available for work, and special skills. The initial group of about three dozen volunteers received training at one of two sessions set up in early February 2006, after which additional volunteers received training one-on-one or in small groups as they signed on. Later, existing volunteers would train new recruits. Over the course of the project, nearly 50 volunteers contributed to various aspects of the project.
**Database development**

In considering technology options the library’s predominant constraints related to both funding and expertise. In 2005, Chelsea District Library contracted most of its technology services to the library cooperative of which it was part, including website hosting. The library itself at the time had only six public computers and about a dozen non-networked staff computers, with the single on-site server running the public computer time management system. No one on staff had knowledge of server administration, leaving staff nervous about hosting a server for the digitization project but also open to any of a number of configurations.

Open source software has seen increasing adoption by public libraries in recent years. While often referred to as “free,” open source software is monetarily free only in the sense that to obtain a copy of the code requires no exchange of funds. In the truer sense, “free” refers to the user’s freedom to view and adapt the software, generally with an agreement to then share improvements with the larger community. Chelsea District Library initially considered both proprietary and open source options for this project, as well as both in-house and contract solutions. While the librarian hired as project manager had extensive experience with Microsoft Access, the licensing costs for the accompanying Microsoft SQL server allowing multiple simultaneous users made that option prohibitive. An open source database based on PHP and MySQL appeared to be a more realistic option, with the drawback that no one on staff had the requisite familiarity with these programming languages. The library decided to solicit requests for proposals to get a better sense of its options. Replies to the RFP highlighted the range of options: from a $40,000 proposal that involved proprietary software and taking cards offsite for more efficient scanning to a $1,500 proposal based on open source software and leaving data entry and scanning purely up to the library. The library selected the able services of a programmer who worked at a nearby library and who recommended the purchase of a server and quickly designed and built a PHP/MySQL database meeting specifications. While open source was not the initial goal, that such
software ended up forming the basis for the database contributed greatly to the library’s ability to produce a high-functioning, easy-to-use database on a limited budget.

Once the library selected the underlying software, the real work of database design began. Priorities were a simple interface on both the administrative and public ends; completely web-based access for both data entry and retrieval; and the ability to attach multiple images to each record. The library pictured a database that would be simple and fast for access by users all over the world and one that could be expanded to meet larger goals for the local history collection. Starting with attached obituary images, the library envisioned eventually allowing researchers to submit their own photos, family trees, marriage licenses, or other material that could supplement the library’s own collection to tell the broader story of each person represented in the database. Flexibility for growth of both the size and scope of the collection was important.

These ambitious plans remained in the future. In the near term, the library had to balance providing extensive access to the obituaries with completing the work within a relatively brief time frame. The poor quality of the newsprint and cards meant doing optical character recognition (OCR) was not realistic. Thus, while researchers would be able to view an image of the complete obituary, searching would be limited to data entered by volunteers. This meant maximizing the number of access points was ideal; at the same time, too many access points could slow work to a crawl. In the end, after consultation with genealogical researchers, the library chose the following fields:

- first, middle, last, and maiden names of subject in separate fields
- first and last names in one field for: mother, father, spouse(s), children
- date and place (city, state) of birth and death
- cemetery and funeral home
- metadata covering obituary source and date and source of digital record
The database allowed for multiple spouses and children. It also included a notes field which, at the early stages, was left blank but proved to be invaluable for later expansion.

The database would be accessed via the library’s website by users through either a basic or advanced search. Upon visiting the Family History Index Online, users see a basic search screen for the name field, which searches all name-related fields. This is often enough to get users to the obituary they need.

Users also have access to an advanced search screen, which searches first, last, and maiden name as separate fields. It also allows access via date and place of birth or death, cemetery, and funeral home, making it useful to those who may not be searching for a specific person but for more general historical information. Data typed into any of the fields on the basic or advanced search screens will also search the notes field, which may contain additional data from various sources.

The designer brought another feature to the database that would set the *Family History Index Online* apart from similar projects at other libraries, a hyperlinking feature that makes Chelsea’s project unique in allowing researchers to jump from record to record, following the obituaries of family members represented in the database. If the parent, child, or spouse of an obituary subject is represented in the database, his or her name will appear as an active link. Clicking on this link takes the user to that person’s obituary record. In this way, researchers may discover family connections they did not know existed and be better able to visualize how families relate to each other. Rather than noting the names of relatives, backing out to the initial search screen, and starting a new search, database users can simply hop from one relevant record to the next.

Once the database design was complete, both staff and an initial group of volunteers participated in testing. Volunteers tested for ease of use from the administrative and public perspectives as well as for how the database would meet the needs of genealogy researchers. The library was lucky to have a number of experienced genealogical researchers among its volunteer corps. They provided invaluable
feedback throughout the project. The database designer and project manager worked together to tweak the database in late 2005 and early 2006, leading up to training of volunteers and beginning of data entry work.

**Timeline**

- April 2005 – Chelsea District Library creates a librarian position with duties including digitization of the Family History Index
- October 2005 – Database development begins
- January 2006 – Database development and testing completed
- February 2006 – Volunteer training and data entry work begins
- May 2006 – Scanning of obituary cards begins
- October 2006 – Data entry of 15,000 records completed/ library closes to move to new building
- January 2007 – Scanning resumes; proofreading and database updates continue
- June 2007 – Final image attached to database
- October 2007 – Family History Index Online released to public; timed to coincide with Family History Month in Michigan

**Digitization process**

The library had two old PCs available for use by volunteers in a back office, a setup that proved beneficial to the project, as volunteers found they could often focus better on the mundane task of data entry while working in pairs. Because the database was entirely web-based, no software installation was required. Volunteers also used each other as resources to answer questions such as the interpretation of unclear wording or how to enter data in a particular field. Each volunteer had committed to working two hours per week for a period of at least three months, in an attempt to minimize the amount of retraining that would need to be done. Most volunteers stayed much longer, seeing the project through to completion, and many also worked multiple shifts each week. Through the volunteer program, these workers were able to sign up for shifts during all hours the library was open, including evenings and weekends, maximizing the number of people who could be involved.
As the project progressed, it became clear both that some volunteers wanted to participate but lacked interest or ability in data entry and also that volunteer tasks existed beyond what the library had originally envisioned. This led to some refocusing of efforts before a final workflow developed. In particular, a pair of volunteers took on the task of taking cards from the filing cabinet and sorting out those that need digitization. They stored the sorted cards in a box and transported them to the office where volunteers were engaged in data entry. Cards not in use remained in the files, which prominently displayed signs explaining the project. Volunteers doing data entry took the sorted cards and entered them into the database. Cards that brought up questions went into a separate pile for review by the librarian. The rest went into a “completed” file and moved on to step 2, proofreading. Two volunteers with especially good eyes for detail, as well as genealogical research experience, were recruited as proofreaders, tasked with checking every fifth card. While it would have been ideal to have a second set of eyes on every single database record, this was not realistic, and conferral with professional archivists confirmed that a 20 percent rate was more than sufficient.

After proofreading, cards moved on to scanning. Scanning represented another challenge, in part because the library had only one scanner which was also used for other purposes. In addition, while the scanning software that accompanied the purchase HP Scanjet 5500c was relatively simple, not all volunteers felt they had the requisite level of computer skills. As a result, scanning did not begin until about three months after data entry and was handled by a subset of about 10 volunteers. Volunteers scanned cards at 300 dpi and saved them in the archival standard TIF format. Because they were on black-and-white newsprint and newsprint already has a very low resolution, a higher resolution would not provide any benefits. These archival copies of the cards have been retained in separate, backed up files so that the original cards should never require rescanning. The entire database, including these archival images, was set to copy to a tape drive, with the tape changed daily. After one week, tapes were reused for new backup copies, leaving the library with multiple recent backups for added security.
At this point, the cards were refiled by a volunteer, returning access to the public who still relied on the paper file for research. The digital images then underwent additional processing. The TIF images were converted, using Adobe Photoshop Elements into compressed JPEG images with small file sizes. In general, the image of each card posted to the database was no more than 100 Kb. It would download quickly even with dial-up internet connections. Volunteers manually attached these images to each database record, completing the cycle.

Data entry work progressed remarkably quickly, with volunteers putting in close to 2,000 hours between February and October 2006. In October 2006, the project went on hiatus while the library packed up its temporary quarters and moved into a new facility. Volunteers completed data entry for the final card just before the move commenced, adding more than 15,000 records into the database in eight months. Between half and two-thirds of the scanning had also been completed at this point. The move to the new facility caused significant delays in the project due to other priorities for the library’s technology staff. Scanning resumed in February 2007, however, and the final image was attached to the database in June. The library used the next few months to continue proofreading and to test the robustness of the database, releasing it to the public in conjunction with Michigan’s Family History Month in October 2007.

**Results**

Chelsea District Library’s Family History Index digitization project was a major success. Through a combination of creativity, open source software, and volunteer contributions, the library produced a highly usable online database providing full access to more than 15,000 obituaries for less than $5,000, completing work in about 15 months. It demonstrates that even at an institution with little staff and little specific expertise, a project can be developed to meet both the community’s needs and the standards set by the library and archives world. The community involvement and grassroots nature of this project made it truly special for those who participated. It drew attention to the library, increased the volunteer corps, and gave
community members a sense of ownership for a key piece of the library’s collection.

The flexibility with which the database was designed has proven vital to its success. Once done with the initial work of populating the database, the library turned its attention to improvements. One goal was to provide full-text access to all obituaries, even those from newspapers old enough that the only existing copies were on microfilm. In early 2008, the Friends of the Library applied for and received grant funding to help purchase a digital microfilm machine, which volunteers are currently using to scan obituaries from the Chelsea Standard going as far back as copies survive, into the late 1800s. These digital images are then either added to existing database records or used to supplement the database.

There also remained the problem of providing access to obituaries of local residents that were printed in neighboring newspapers. To this end, the library looked to its partnerships with local businesses. Chelsea is represented by two local funeral homes, and directors of both proved willing to provide the library with access to their files. When the funeral homes submit obituaries to any area newspaper, they also send an electronic copy, including photo where available, to the library. These partnerships have allowed the library to enrich the database with full-text access and color photos for more recent obituaries. The text that is sent electronically is pasted into the notes field, and images are higher quality than those ultimately printed in the newspaper. Along with this material, if the final obituary is printed in the Chelsea Standard, the scanned newspaper clipping is attached. The funeral homes have also indicated that they have computerized files going back a number of years, and the library is investigating the possibility of further enhancing the database with this material.

As another extension, the library returned to its initial contractor in 2008 to develop a database on the same platform to house the library’s local history collection, which had been brought out of storage and organized only in 2007. Much of this material was even more valuable and unique than the obituaries, leaving the library with a strong desire to have it made accessible to the public primarily, if
not exclusively, in digital format. Scanning of material and database development have continued in 2009. While working on this and other local history initiatives, including a series of oral history projects, the library has found additional material to enhance the obituary database records of members of Chelsea’s founding and prominent families.

Aside from the primary lesson that ambitious digitization projects are not beyond the reach of even very small libraries, the Family History Index digitization project provided additional lessons that apply to similar projects and beyond:

- The existence of a well-setup volunteer program allows a library to think much bigger than would otherwise be practical. People are out there in all communities who have the time, expertise, and interest to contribute.
- Planning is good but flexibility is essential. Projects may stray somewhat from their original vision or carefully thought out procedures, but that isn’t necessarily bad. Being open to new opportunities and listening to workers and users can ultimately make a project much richer.
- Open source doesn’t have to be terrifying. In fact, open source solutions are generally very stable, as the Family History Index Online has proven to be. Aside from minor software upgrades, the server has provided consistent access to the database for more than two years with virtually no staff intervention required.
- There’s nothing wrong with thinking big, but it doesn’t hurt to be realistic. Ambition leads to projects being even more successful than initially imagined. That said, a realistic assessment of aspects such as which tasks could be handled in-house (project management) and which should be contracted out (database design) prevented later stumbles.
- Partnerships enhance any library activity. In this case, a good working relationship with the local newspaper eased the process of gaining copyright access, and new partnerships with funeral homes have enhanced both the database and the library’s reputation in the larger community. In other situations,
Digitization projects not only preserve the past but provide an opportunity for greater community involvement, partnerships, and identification of the library as a key to the community’s overall health.

References


