Andrew J. M. Smith
Emporia State University
asmith37@emporia.edu

https://doi.org/10.18485/mks_dh_skn.2021.1.ch3

COMMERCIAL AND NOT-FOR-PROFIT DIGITIZATION OF HISTORICAL AND CULTURAL RECORDS FOR GENEALOGICAL STUDY: PROBLEMS AND OPPORTUNITIES

Summary: The study of family history and the desire to discover one's roots is an area that is expanding rapidly, especially in those cultures that have become more fragmented and have lost their immediate connections to family and to their past. The rise of popular television programming that models the genealogical process and combines social history with the discovery of the family history of personalities from the entertainment and sports industries has further served to popularize the search for roots and family stories.

In the context of this great increase in desire for information about family members and coupled with the rise in the use of the internet and the availability of hardware and software tools that are easily used by the general public, there has been an explosion in the amount of information that has been digitized by both commercial and not-for-profit entities. However, different providers have used different models of digitization and transcription, with commercial companies relying more on the rapid processing of large numbers of resources, primarily using OCR and full-text searching while not-for-profit entities rely more on human transcription and indexing by volunteers and the careful processing of much smaller data sets.

It is becoming more common for the provision of the same data or parts of the same data sets by different entities using different digitization and indexing processes. Some data is provided by multiple organizations, but specific data known to exist may not be easily found in all versions of the data set. There is also a corresponding rush to differentiate commercial information providers with unique data sets. The use of restrictive licensing agreements and limited distribution rights leads to a situation where although information may be available and may have been digitized, the genealogist must be prepared to search multiple commercial resources (at a cost) to identify an information source and then embark on discovery of information that may not be well digitized, may not be well transcribed, or may not be well indexed. Again, information sources may be partly digitized, or may only be abstracted or indexed only, with no access to original source documents in digital form.

There are many lessons to be learned from the current situation for collections of historical records that have not yet been digitized but would be helpful in the study of individuals' family histories.

Keywords: digitization, cultural heritage, genealogy, indexing, full text searching, manual transcription, OCR.

The study of family history and the desire to discover one's roots is an area that is expanding rapidly, especially in those cultures that have become more fragmented and have lost their immediate connections to family and to their past. The rise of popular television programming that models the genealogical process, albeit in an extremely simplified form, and combines social history with the discovery of the family history of personalities from the entertainment and sports industries has further served to popularize the search for roots and family stories. (For examples see Who do you think you are (British Broadcasting Corporation, 2019); Finding your roots (Public Broadcasting Service, 2018.)

In the context of this great increase in desire for information about family members and coupled with the rise in the use of the internet and the availability of hardware and software tools that are easily used by the general public, there has been an explosion in the amount of information that has been digitized by commercial, government, and not-for-profit entities (Ancestry, 2019; Find My Past, 2019; Free UK Genealogy, n.d.; Intellectual Reserve, Inc., 2019; My Heritage, 2019; and National Records of Scotland, n.d. c) However, different providers have used different models of digitization and transcription, with commercial companies relying more on the rapid processing of large numbers of resources, primarily using Optical Character Recognition (OCR) and full-text searching while not-for-profit entities rely more on human transcription and indexing by volunteers and the careful processing of much smaller data sets (Free UK Genealogy, n.d.; Intellectual Reserve, Inc., 2019.)

It is becoming more common for the provision of the same data or parts of the same data sets by different entities using different digitization and indexing processes. Some data is provided by multiple organizations, but specific data known to exist may not be easily found in all versions of the data set. There is also a corresponding rush to differentiate commercial information providers with unique data sets. (see Why choose Find My Past? FindMyPast, 2019.) The use of restrictive licensing agreements and limited distribution rights leads to a situation where although information may be available and may have been digitized, the genealogist must be prepared to search multiple commercial resources (at a cost) to identify an information source and then embark on discovery of information that may not be well digitized, may not be well transcribed, or may not be well indexed. Again, information sources

may be partly digitized, or may only be abstracted or indexed, with no access to original source documents in digital form.

There are many lessons to be learned from the current situation for collections of historical records that have not yet been digitized but would be helpful in the study of individuals' family histories, or indeed for other research purposes.

Source documents

The first thing to be considered is the source documents that are used for digitization. The optimal source is the original document. This ought to give the digitizer the best opportunity to create a clean, digital image of the document as the basis for further processing, including transcription and indexing. However, it is not uncommon for digitization to take place from secondary sources, from microfilm, for example (Intellectual Reserve, Inc., 2019.) Newspapers are a common example of this second-generation digitization. Material that has been converted to a microfilm can easily and quickly be digitized with little human intervention beyond mounting the microfilm on a reader. However, in the majority of cases, the original microfilming of newspapers was conducted on bound volumes of newsprint. The best option would have been to disbind each volume and film each issue lying completely flat, but the push to microfilm, originally for preservation and access, meant that speed and expediency were valued over more time consuming and expensive methods that may have resulted in better digital images.

Depending on the location of an issue within the bound volume, material may be obscured in the gutter of the binding, or the curve of the page next the gutter may cause the image to be out of focus at those edges, or for the lines of text to be misaligned. Materials that were legible to the human eye in original form then become illegible in the microfilm copy. When this copy is used as the basis for digitization, then the digital copy also contains areas of illegibility. Although software can make an attempt to restore this, success may be limited and information becomes inaccessible. From the family history perspective, although the illegible areas of each page may not be large, because of the structure of newspaper columns, and the tendency for announcements of births, marriages and deaths to appear in the columns closest to the margins, this can mean that such family information is disproportionately affected by legibility issues.

Even if it were decided to digitize from original documents rather than microfilm, in many cases the original materials may not be available or may not be robust enough to be handled for digitizing. (Newspaper is especially vulnerable as it was never designed to be a permanent record, and especially in the 19th and early 20th centuries, much of the newsprint used was highly acidic, leading to rapid disintegration of the source material.)

Process of digitization

The actual process of digitization can also pose problems. Digitizing materials that are in a secondary form may be smooth and rapid, the limitations coming from the shortcomings of the secondary source. Digitization of primary sources can be more challenging. How is it accomplished? Are materials removed to a specialized digitization facility where the best possible images are created or are the materials scanned in situ, simply doing the best that can be achieved under often adverse circumstances? The answer to this question is often driven by the purpose of the digitization and who is performing it. The tendency is for large scale commercial genealogy companies to concentrate on speed and numbers, and digitization frequently takes place where the records are. A single person will be dispatched with basic digitization equipment and will set up a temporary digitization station and will simply work through the materials the company wants to acquire. Manipulation is by hand – placing materials on a scanner or turning pages of a book for a camera. A common requirement for this type of set up is to digitize 2000 pages per day, which equates to 250 pages per hour, 4 pages per minute or 15 seconds per page. (Freelance Digitizer, personal communication, April, 2014.) While this does have the advantage of digitizing large amounts of material in a short space of time, the emphasis is on speed and quantity, rather than completeness or quality. The digitizer is not responsible for checking that each page of a volume has been digitized, nor that all the images are of a high quality. If a page is missing from the final digital version, there is no way for the reader to know if the page exists and was simply omitted from the digitization, or if the page is missing from the original.

In contrast, material that is digitized by not-for-profit groups or by the original data producers, tends to be digitized more slowly, with a greater emphasis on accuracy and completeness, as the goal is to provide the best possible digital versions of a smaller amount of material to make this available, and the digitization is designed to both preserve the originals and make the

information more widely available, without the added impetus of generating profit from a unique data set.

Digitization as preservation

Note that digitization as a form of preservation is mostly concerned with limiting the use and handling of original documents. Digital files themselves can be fragile, and are subject to data loss, to format obsolescence, as well as to the failure of storage media, whether in the cloud or on site. Although digitization seems to be common now, standards and procedures are still evolving, and material from the same collection may be digitized at different resolution quality, using different software, and with different hardware compatibility. Digital files also require a great deal of curation and management and can be just as expensive to maintain as paper-based originals, if not more so, as costs are recurring and often not obvious to the casual observer (such as power, digital storage, software and hardware maintenance, and personnel.) The actual creation of the digital file may be the least expensive part of the process. The time and costs of preparing the digital images for publication and their maintenance thereafter may far outweigh the initial digitization expense.

OCR vs human transcription

Another major difference between the large commercial genealogy companies and other organizations is the method used for transcribing the records. Again, the goals of speed and quantity mean that commercial companies rely heavily on optical character recognition (OCR) technology to process large numbers of document pages. Although high degrees of accuracy can be obtained, OCR success is not 100% reliable and can be affected by whether the original document is printed, typed, hand-written, or some combination of these. The language of the document also has an effect on accuracy of OCR, as well as language usage. OCR software often does not handle abbreviations or shorthand notations well, nor is it good at coping with slanting lines of writing, or material that is not evenly spaced (frequently the case with hand-written forms and documents). Human transcription is able to see solutions in a way that OCR software is not, and can decipher difficult handwriting, expand abbreviations (or record them accurately), piece together mismatched lines of text across pages, and other transcription tasks that are

currently beyond the capabilities of OCR, although taking much more time to do so.

While the ability to process huge numbers of records is a distinct advantage of OCR, the point is that for each genealogical researcher, there may only be one part of a document or record set that is relevant to their research. If this is missing, illegible, or mis-transcribed, it does not matter how much other material is available, correctly transcribed, the data source has failed to provide the required information.

Dispersed data sets

A major problem in the genealogical world is that of dispersed data sets. This has come about because of the patchwork of cooperative and licensing agreements that existed from the beginnings of the modern era of genealogical research, as organizations that did not yet have the capability to digitize and process their own records formed alliances with other entities, both commercial and non-commercial who had the necessary expertise and resources. For example, the Church of Jesus Christ of Latter Day Saints (LDS) pursues genealogical research as part of their belief system, and has gathered a large collection of records from all over the world. They have high quality digital images of civil registration vital records (births, marriages, and deaths) in Scotland from the beginning of registration in 1855 to 1875, as well as the years 1881 and 1891. These images, which are higher quality than those digital images available through the Scottish Government's own genealogy research website, ScotlandsPeople (National Records of Scotland, n.d. c) are, however, only available online to non-church members in one of their Family History Centers. Vital records are available from the LDS's website for other years, but only as partial record transcriptions – and these partial transcriptions are also available on some commercial sites. Full records are available only from the Scottish Government website.

Although data from the Scottish portions of the UK census are available through multiple providers, in differing formats, for all censuses from 1841 to 1901, the material from the 1911 census is only available from the Scottish Government, as they have reached a stage of digital maturity where they are able to digitize, transcribe and index data by themselves, with no external help, and are also therefore able to reap the monetary rewards of being the sole provider of this one data set. This exclusivity will also apply to the planned release of the 1921 census (National Records of Scotland, n.d. a).

Indexing

The quality of the transcription has a major effect on the retrieval of the information. Commercial companies tend to rely more heavily on whole word searching and retrieval within their data sets, or on more limited index fields (the larger size of their data sets and the desire for speed of response dictate how many fields are indexed and how many may be cross-searched at one time.) By contrast, non-commercial entities with human-transcribed digital resources can provide stronger indexes, based on more accurate transcription, and on a larger number of index fields. For example, the FreeCen project in the United Kingdom (Free UK Genealogy, 2018) provides high quality transcriptions of census records that may be searched not just by name, approximate age and location, but also by occupation, an invaluable tool when trying to distinguish among many people with the same name, and where age may not be a particularly accurate identifying factor, due to data gathering practices including rounding up or down ages to the nearest five years, or to the fact that age was not such an important factor to many people, or that they simply did not really know their exact age, and reported it differently in different censuses. Social convention also led to deliberate misstatements of age, as in wives decreasing their age to comply with cultural ideas about wives being younger than their husbands, or husbands increasing their age for the same purpose.

Again, for the 1881 Scotland census, a version transcribed by the LDS is available from them, also through every commercial genealogy provider, partially from the FreeCen project, and from the Scottish Government, although different search algorithms on each site may give different results for the same search parameters.

To give one example, a broad search for the name "Robert Henderson" on the ScotlandsPeople website shows there are 689 matched records in the index of the 1881 Scottish census prepared by ScotlandsPeople but only 682 matched records in the index for the same census prepared by LDS. If you limit the search to the city of Glasgow, the Scotland's People index offers 108 matches, while the LDS index offers 29 (National Records of Scotland, n.d. b). The same broad search conducted on FindMyPast.com results in 771 matches allowing for first name variants or 725 records with an exact match to the first name, while the search in Glasgow yields 113 exact matches or 119 allowing for first name variants (FindMyPast, n.d.). Similarly for the broad search, Ancestry,com claims 840 matches with first name variants and 687 matching the first name exactly while for the comparable Glasgow search, 27 and 22

matches respectively (Ancestry, n.d.). These results are presented in Table 1.

(Note that while the structure of the database search engine in Scotland's People does allow for the use of wild cards within the search fields, so that user-suggested variants can be accommodated within the search, it does not offer the same search capability of predetermined name variants provided by FindMyPast and Ancestry, so the results of variant name searches are not provided for the ScotlandsPeople or LDS indexes.)

Table 1.
Search Results for Robert Henderson in the 1881 Scottish Census

Indexer	Scotland		Glasgow	
	Exact name	First name variants	Exact name	First name variants
ScotlandsPeople	689	-	108	-
LDS	682	-	29	-
FindMyPast	725	771	113	119
Ancestry	687	840	22	27

While it may be easy to suggest reasons for these discrepancies, such as different matching algorithms for the broad searches or varying definitions of local areas for the more targeted searches (and even these definitions may change over time, so one indexing problem is whether to index based on geography at the time the data was created or current definitions), the end result for the user is that the desired information may not be retrieved. Again, the point to remember for the genealogist is not the total number of available records but whether the one required record is retrievable.

Future directions

So, where does this leave us? This is a case where we would be better off from the genealogist's point of view if there were cooperation among information providers, and we could benefit from the automated processing of large scale printed data collections that are easily processed by OCR and accessed by full-text searching, as well as having access to collections that require human transcription or correction and multiple human-created indexes. Unfortunately, we are in a place where the provision of digitized historical records has become highly commercialized and where the successful paradigm of record provision is based on uniqueness of collections, quantity

of records (regardless of accuracy or searchability), and supposed ease of use, rather than the librarians' perspective of providing the best information most efficiently.

Opportunity certainly exists for a different model of digitization of historical materials for genealogical research focusing on the quality of the record sets, quality indexing and reliable retrieval. Until a new model is implemented, genealogists will need to search multiple resources and employ sophisticated search strategies in the hope of extracting appropriate information.

References

Ancestry. (2019). *About us.* Retrieved June 12, 2019, from https://www.ancestry.com/corporate/about-ancestry/our-story

British Broadcasting Corporation. (2019). *Who do you think you are?* Retrieved June 12, 2019, from https://www.bbc.co.uk/programmes/b007t575

Find My Past. (2019). Why choose Find My Past? Retrieved June 12, 2019, from https://www.findmypast.com/content/why-choose-findmypast

Free UK Genealogy. (n.d.). *About free UK genealogy*. Retrieved June 12, 2019, from https://www.freeukgenealogy.org.uk/about/

Free UK Genealogy. (2018). *UK census online.* Retrieved June 12, 2019, from https://freecen1.freecen.org.uk/

Intellectual Reserve, Inc. (2019). *About Family Search.* Retrieved June 12, 2019, from https://www.familysearch.org/en/about

My Heritage Ltd. (2019). *About us.* Retrieved June 12, 2019, from https://www.myheritage.com/about-myheritage/

National Records of Scotland. (n.d.). *What records are in the site?* Retrieved June 12, 2019, from https://www.scotlandspeople.gov.uk/what-records-are-in-the-site

Public Broadcasting Service. (2018). *Finding your roots.* Retrieved June 12, 2019, from https://www.pbs.org/weta/finding-your-roots/home/

Ендрју Ц. М. Смит
Емпорија Универзитет
asmith37@emporia.edu

КОМЕРЦИЈАЛНА И НЕПРОФИТНА ДИГИТАЛИЗАЦИЈА ИСТОРИЈСКИХ И КУЛТУРНИХ ЗАПИСА ЗА ГЕНЕАЛОШКУ СТУДИЈУ: ПРОБЛЕМИ И МОГУЋНОСТИ

Сажетак: Проучавање породичне историје и жеља за откривањем нечијих корена подручје је које се брзо шири, посебно у оним културама које су постале уситњеније и изгубиле непосредне везе са породицом и прошлошћу. Пораст популарног телевизијског програма који повезује друштвену историју са откривањем породичне историје додатно је послужио за популаризацију потраге за коренима и породичним причама.

У контексту овог великог пораста интерсеовања за информацијама о члановима породице, упоредо са порастом употребе интернета и доступности хардверских и софтверских алата широј јавности, дошло је до експлозије информација које су дигитализовали комерцијални и непрофитни субјекти. Међутим, различити провајдери користили су различите моделе дигитализације и транскрипције, при чему се комерцијалне компаније више ослањају на брзу обраду великог броја извора, првенствено користећи оптичко препознавање карактера и претрагу целог текста, док се непрофитне организације више ослањају на транскрипцију, индексирање и пажљиву обраду много мањих скупова података које ради човек, а не машине.

Све је чешће да исте податке или делове истих скупова података пружају различити извори, користећи различите процесе дигитализације и индексирања. Неке податке пружа више организација, али одређене податке није лако пронаћи, иако се зна да постоје. Коришћење рестриктивних уговора о лиценцирању и ограничавање права на дистрибуцију доводи до ситуације да, иако информације могу бити доступне и можда су дигитализоване, генеалог мора да буде спреман да претражује више комерцијалних извора (уз наплату) како би идентификовао извор информација, а затим да крене у откривање информација које можда нису добро дигитализоване, можда нису добро транскрибоване или можда нису добро индексиране. Опет, извори информација могу бити делимично дигитализовани или могу бити само индексирани, без приступа оригиналним изворним документима у дигиталном облику.

Из тренутне ситуације може се научити много о колекцијама историјских записа које још нису дигитализоване, а које би биле од помоћи у проучавању породичне историје појединаца.

Кључне речи: дигитализација, културно наслеђе, генеалогија, индексирање, претрага целог текста, ручна транскрипција, оптичко препознавање карактера.