GETTING THE CROWD INTO OBITUARIES

How a Unique Partnership Combined the World's Largest Obituary Index with the Utah's Largest Historic Newspaper Database

BY:

John Herbert
Jeremy Myntti
Alan Witkowski
J. Willard Marriott Library
University of Utah

John Alexander
FamilySearch.org
The Church Of Jesus Christ of Latter Day Saints

GETTING THE CROWD INTO OBITUARIES

The Utah Digital Newspapers (UDN) and FamilySearch are joining forces to create an innovative obituary index. UDN contains 282,000 obituaries in its extensive database of historic Utah newspapers. UDN's headlines are manually keyed (double-keyed and reconciled), and are nearly letter-perfect. However, the article text is created from raw optical character recognition software, which is often less than fully accurate.

FamilySearch has the world's largest obituary index. Each entry has dozens of structured metadata fields, all accurately keyed and properly tagged. Some examples of these tagged fields are deceased name, death date, place of death, and parents, children, and siblings of the deceased. This obituary index is text only; it contains no images at all. In addition, and very importantly, FamilySearch has a tremendous volunteer corps available for use in many different genealogical projects.

This partnership will use this crowd of volunteers to key-capture structured metadata for every obituary in the UDN database. This general process is described below.

FamilySearch will use the API for UDN's database and extract the PDF images and metadata of every UDN obituary. They will display this information for their crowd, and using it, their volunteers will key in the contents of the obituary into the FamilySearch obituary index form. After saving the newly created obituary index, FamilySearch will provide this metadata to the Marriott Library in spreadsheet form. The Marriott Library will then update the UDN obituary items with the new metadata. Finally, both institutions will cross-link to the other's website.

Utah Digital Newspapers

The Utah Digital Newspapers program is administered by the J. Willard Marriott Library at the University of Utah (USA). From its modest beginning in 2002, the program has flourished. Now in its twelfth year, UDN contains 1.5 million pages of content and is recognized as a national, even international, leader in newspaper digitization. It is accessed via a fully accessible, free website, which can be seen at http://digitalnewspapers.org. It remains the first hit in the Google search engine for a search on "digital newspapers."

As of January, 2014, the Utah Digital Newspapers holds 100 distinct newspapers titles, ranging from the very first newspaper issue published in the Utah Territory, the Deseret News in June, 1850, to the Garfield County News published in March, 2005. It holds content from 27 of the 29 counties in the state of Utah. Of these 100 titles, we hold the first issue (volume one, number one) of 43.

Some other statistics related to the size of the UDN collection are:

- Number of titles 100

- Number of counties 27 of 29[1]
- Number of newspaper issues 154,409
- Number of pages 1,561,091
- Number of articles 18,178,920
- Number of individual collections in the database 260
- Total number of objects in the database 19,894,420

A Brief History

UDN began in early 2002 with a $93,000 grant from the Utah State Library that purchased server hardware and provided for the first digitization of historical Utah newspapers content. In December 2002, after several months of experimenting with digitization processes, UDN's initial website was launched with 30,000 pages of content, which was comprised of 10,000 pages from each of three titles. Word quickly spread throughout the Utah library community of this unique new resource. The Library recognized immediately that this concept had great potential and that it needed to expand its content and capability.

Later that same month, the initial $93,000 grant was followed by a second, much more substantial grant from the Utah State Library for $278,00. This funding provided for a full-time project director and 106,000 pages of content to be digitized, effectively tripling the size of the database.

When the Library received a $470,000 National Leadership Grant from the Institute for Museum and Library Services (IMLS), a U.S. federal agency, in September 2003, it was a watershed event. With this large infusion of support to fund efforts over a two-year time span, UDN was transformed from a project to a program and emerged into the national spotlight as a leader in newspaper digitization. During the term of the IMLS grant, another 278,000 pages of content were digitized and the database grew to nearly 500,000 pages.

As the IMLS grant wound down in 2005, the National Endowment for the Humanities (NEH), in collaboration with the Library of Congress (LC), launched its National Digital Newspapers Program (NDNP). The University of Utah was one of six institutions awarded a grant in the first test-bed phase of the program from 2005-2007. The Library subsequently received additional two-year awards in both 2007 and 2009, bringing its NEH funding to a grand total of $863,000 to digitize 380,000 pages of content.

Throughout all these years, the Library had very good success raising in-state funds from various institutions to digitize local newspapers. It worked with academic libraries, public libraries, newspapers themselves, historical societies, and other cultural heritage institutions. The largest of these projects brought funding of $631,000 from the Utah Department of Heritage and Arts to digitize 250,000 pages of the Salt Lake Telegram and several other smaller titles. This content contains fifty years (1902-1952) of a major Salt Lake City daily newspaper, and represents 16% of the entire UDN database.

[1] Two counties, Daggett and Wayne, are not represented because the UDN program has not been able to identify a substantive newspaper collection in either locale to digitize.

Operating Principles

During the course of the digitization program, UDN has followed six simple operating principles, all of which were designed to improve the patron experience. First, from the very beginning the Utah Digital Newspapers project focused on achieving a broad statewide scope and representation. Especially in the early years, UDN resisted the temptation to digitize large metropolitan titles. In fact, it consciously pursued the opposite goal, exclusively targeting smaller, rural weekly titles instead. This allowed UDN to generate demand across the entire state while at the same time expanding its chronological coverage with weekly, rather than daily, papers.

Second, after selecting a title to be digitized, UDN's strategy is to scan materials beginning with the earliest dates and then to progress forward in time as far as the available funding will allow. These tactics enable the program to digitize the set of materials that are most likely to have the greatest need for preservation and will be the most in demand by users.

Third, whenever possible, UDN uses modern technology to capture images of original hard copies of newspapers rather than scanning worn and dated microfilmed images. This technique generates digital images worthy of the 21st-century. High-resolution imaging, in turn, contributes to higher accuracy for optical-character-recognition (OCR) software processes, which in turn provides more accurate search results for users.

Fourth, UDN's processing protocols include providing images and metadata for each newspaper article. All OCR text is attached to its article image so that the full article image may be included in search results. This allows users to quickly view and understand the context of hits returned from database searches. Most other digital newspaper programs in the U.S. do not segment page images into their individual articles because of the significant additional cost to do so.[2] Moreover, a much more complex database structure is required to manage information that is article-based. UDN, however, believes strongly that article-level metadata provides a much more rewarding patron experience and is well worth the additional cost. Furthermore, up to this point in time, the UDN database has been able to satisfactorily handle the more complex newspaper issue structure.

Fifth, to further enhance search accuracy, the UDN digitization service provider manually keys in article headlines. In fact, they are double-keyed and verified, which means that two different people key each headline and any discrepancies are reconciled. This process insures nearly 100% accuracy of headline text. Again, this extra processing is more expensive, but the UDN program believes that the corresponding improvement in its patrons' search accuracy justifies the extra expense.

Sixth, to stay in touch with patrons and receive their feedback, UDN offers a simple survey on its website asking users about their use of UDN. The survey has run continuously since 2005 and has collected over 1,500 patron responses. Among the many things learned from the survey are:

- 84% of users gave an overall rating of "good" or "excellent" for their user experience
- 79% will return soon
- 74% will tell others about UDN

[2] It should be noted that article-level segmentation is much more prevalent outside the United States.

- 66% rate search accuracy as “good” or “excellent”
- 65% find new sources for their research
- 63% are more knowledgeable about their own family history
- The most often asked-for improvement is simply “more content!”

Article-Level Segmentation

There has been a long-running debate in the digital newspaper field, especially within the NDNP community, about the costs and benefits of capturing metadata and presenting images for pages or for articles. The simple solution to organizing newspaper metadata is to follow the page-level specifications set forth by the NEH and LC in their National Digital Newspaper Program (NDNP). For better or worse, these specifications are rapidly becoming the industry standard, and any viable digitization processor should be able to deliver newspaper files in this format.

Page-level metadata is of course much easier and less expensive to produce. In a large, national program like the NDNP, overall costs per page for processing are a serious consideration. Page-level items also require a less robust data model and less technically capable database to house the data. Finally, with the increased functionality of image clipping software, it is much easier now-a-days for software to “clip” an article image from a page image on the fly as the reader requests to view an article online. So presenting an article image for viewing can be done even though only a page image is housed in the database.

All that notwithstanding, in Utah we do indeed segment the individual articles on every page and capture the text and other metadata at the article level. The headlines and sub-headings for each article are manually keyed in. In fact, they are double-keyed and verified, which insures search accuracy for headlines to nearly 100%. We also classify each article by its type, which includes news, advertisements, mastheads, obituaries, and birth and wedding announcements. These last three types are very important to genealogists, our largest patron group. In general, they allow for narrower searches, which is increasingly important as these text-heavy, digital newspaper collections grow.

Segmenting pages into separate articles has several other advantages: articles are presented in search results, article images are presented for viewing, and OCR is improved because there are more consistent fonts within an article and hyphenated words are more easily conjoined. We find these reasons compelling enough to justify the additional expense of segmentation.

FamilySearch

Since 1894, FamilySearch (formerly the Genealogical Society of Utah) has gathered genealogical records to enable family history research. From 1938-2005 over 2.5 million rolls of microfilm were gathered and stored in the Granite Mountain Records Vault in Little Cottonwood Canyon. As FamilySearch embraced the digital age, they began the process of scanning their historic microfilm collection and replaced microfilm cameras with digital cameras. Currently FamilySearch adds 10 terabytes of data every day to its collection. In its entirety, the digital collection is just over 30 petabytes.

FamilySearch Indexing

Like most information publishers, one of the primary challenges that faces FamilySearch is the searchability and accessibility of an ever expanding digital collection. The problem is not new, and many of the resources that we have used to track and retrieve our microfilm, such as our catalog, are still being used today. While our catalog can help patrons identify record type, locality, and time periods, it still requires a significant amount of time to find information about a specific individual. FamilySearch has been addressing this problem for several decades by having vital information indexed. This has typically been done by capturing the names, relationships, and vital dates such as birth, marriage, or death. This level of access allows for much quicker and accurate retrieval of historic documents and has become the standard of publication throughout the genealogical industry. Because of the amount of time and resource required to create indexes, FamilySearch is able to acquire records much faster than they can be indexed. Therefore, most of the records in the FamilySearch collection have yet to be indexed.

In 2006 FamilySearch introduced FamilySearch Indexing (FSI) in an effort to increase the rate of indexing. FSI is a standalone application available to the general public that allows volunteers to download images and create fielded indexes that are then uploaded back to the FamilySearch databases. Record projects are currently done in 9 different languages and there are volunteers from over 115 countries. Since 2006, 1.1 billion records[3] have been created by a total of 900,000 individual volunteers. This development has dramatically increased the rate of indexing and allowed FamilySearch to make more records accessible. There are currently 3.7 billion indexed records available on FamilySearch.org, and the number is growing all of the time thanks to the work of volunteer indexers. FamilySearch is currently developing a new web-based tool for FSI that will debut later in 2014.

A constant concern from working with volunteers is the quality and consistency of the data that is created. FamilySearch has a very high standard of quality that is insured in part by our FSI process. In addition to maintaining controlled vocabularies and authorities, FamilySearch attempts to review each record after it is submitted by volunteers. Each image is keyed by two separate individuals, called an A and B key respectively. The FSI system compares the A and B key. If the two versions are identical, then the system accepts the data and it is published. If there is any disagreement between the two, then the record goes to an arbitrator to make a final decision. Arbitrators are experienced indexers who correct errors or make judgment calls when ambiguity exists. This process has allowed us to mitigate most of our quality challenges resulting in some of the highest quality indexes in the industry.

FamilySearch and Newspapers

Historically, FamilySearch has acquired very little newspaper content for several different reasons. The volume of content is daunting, especially considering the global coverage that FamilySearch is trying to achieve. In addition, the genealogically valuable articles are peppered

[3] See https://familysearch.org/indexing/ for current statistics

throughout newspapers and can be difficult to isolate. Even the advent of digital developments such as OCR did not solve the problem as the FamilySearch database uses fielded indexes for search and retrieval. For many years the cost of publishing newspaper content outweighed the benefit, however recent developments have made newspapers a priority within the reach of FamilySearch.

FamilySearch has been enabling genealogical research for many years by providing information about vital information and family relationships. For many users, knowing the name, date of birth, spouse, and children is all that is wanted for research. However, for others and especially beginners, discovering additional details and life stories creates a richer experience. Using this logic, FamilySearch has recently added photos and stories as part of their collection. One of the richest sources for vibrant family detail and also valuable genealogical information are newspapers. Article types like obituaries are especially valuable and rich. Because of this value, FamilySearch has made a strategic decision to explore and develop a way to make newspapers searchable and available to patrons.

One of the many ways that FamilySearch tries to overcome technical- and scope-related challenges is by seeking and cultivating partners within the industry. Through partnerships, FamilySearch is able combine the work of collecting and publishing newspapers content with large volunteer workforce. These types of partnerships will support both newspaper publishers and the genealogical community by creating a new resource that will enable access and searchability within newspapers.

Indexing Death Notices in FamilySearch

During our first phase, we will index the 282,000 death notices that have been identified by UDN. FamilySearch was able to access the CONTENTdm API and download the death notice images and metadata. This data will then be ingested into the FSI and distributed to our indexing volunteers. We are asking the indexer to read the article, and then key in all of the names and vital information within the article. Our final metadata schema will include the following fields:

- Name of Deceased: Given Name(s)
- Name of Deceased: Surname (Last Name)
- Name of Deceased: Title & Terms
- Event Type
- Date of Death: Month
- Date of Death: Day
- Date of Death: Year
- Place of Death: State/Country
- Place of Death: County
- Place of Death: Town

- Age of Death
- Estimated Birth Year
- Date of Birth: Month
- Date of Birth: Day
- Date of Birth: Year
- Birth Place: State/Country
- Birth Place: County
- Birth Place: Town
- Relative's Surname – We are capturing all Additional Relatives
- Relative's Given Name
- Relative's Titles & Terms
- Relative's Relationship to Deceased

This information will be keyed for each article, allowing FamilySearch to provide a high quality and detailed index entry for each death notice. The metadata will be published on FamilySearch as part of our genealogical database. When a user selects an entry, they will see all of the metadata on FamilySearch. If the user wishes to view the article in its entirety, they will be linked directly to the article on the UDN website. The intent of FamilySearch is to drive users and traffic to the resources available at UDN.

Obituaries in UDN

Of the three genealogical article types in UDN, obituaries are the most desired by family historians. They often are historically important and contain a great many facts about the deceased's life and family. In the Utah Digital Newspapers database there are 282,000 obituaries among the 18.2 million articles. We know this because our articles are segmented and classified as described above. Without article-level segmentation, we would know next-to-nothing about our obituaries. But with it, obituaries are:

- Easily identified as "type=death notices"
- Easily found using standard database queries
- More easily read as a stand-alone article
- Additional metadata about the obituary is more easily attached to the item

So this entire obituary metadata initiative between the Marriott Library and FamilySearch would not be possible, or even conceivable, without article-level segmentation.

UDN Metadata Decisions

The data that FamilySearch delivers for ingestion into UDN is in a spreadsheet format, making it easy to manipulate and parse in order to meet UDN requirements. Before the data could be added to UDN, there were several decisions that had to be made regarding the fields that would be used and how the data should be formatted.

The first decisions were regarding the personal names to include in UDN and how to format these names. FamilySearch had multiple columns in the data for each name: given name, surname, and full name (formatted as "GivenName Surname"). Personal names in traditional library data are typically formatted to the NACO (Name Authority Cooperative Program of the Program for Cooperative Cataloging) standard. In order to do this, the two separate fields with the given name and surname were concatenated into one string in the format "Surname, GivenName." An advantage of this format is that surnames could be sorted alphabetically when a list is created or browsed.

Since there are potentially many names in each obituary, a decision had to be made as to which names would be most important to search within UDN and what metadata fields they should be mapped to. It was decided that the following name fields would be included in the UDN metadata scheme for obituaries:

- Deceased Name
- Father of Deceased
- Mother of Deceased
- Spouse of Deceased
- Children of Deceased
- Siblings of Deceased

Other names (e.g. in-laws, grandparents, non-relatives, etc.) that are mentioned in the obituaries are not being included in UDN. These names were left out because the closest relatives (spouse, parents, siblings, children) would be the most useful for researchers. Other names in the obituary could be discovered by retrieving the data on the FamilySearch website.

The second category of metadata decisions that had to be made were with regards to birth and death dates. Dates in UDN are standardized according to the ISO 8601 format, which is YYYY-MM-DD. Birth and death dates returned from FamilySearch were each in four separate fields: year, month, day, and full date formatted as "DD MMM YYYY" (e.g. "29 Apr 1943"), so the data had to be manipulated in order to conform to the ISO standard. Both the birth and death dates were included in UDN since they can be useful for researchers and genealogists. They are useful in disambiguating similar names and also provide options to search for people who were born or died within a particular timeframe.

Additional information about the deceased person that is being included in the UDN metadata scheme includes the age of the deceased at the time of death and the birth and death place. These place names make it possible to search for people from a particular place while the age at the time of death can be useful to distinguish between two people that have the same or similar names. Fields that FamilySearch was able to generate that were not included in UDN include the title of the deceased (e.g. Mr., Mrs., Dr.), gender, and additional relatives.

Another new field named “Additional Information” was created which will contain a link back to the original metadata stored on the FamilySearch web server. With this link, all of the metadata that was not included in UDN can easily be viewed by browsing to the FamilySearch site. This field may also be used for other links to additional websites that may provide useful information in the future if other projects such as this are completed in the UDN collections.

Updating the UDN Database (CONTENTdm)

CONTENTdm stores its metadata for a collection in a single text file called desc.all. For example, a single obituary record looks like this:

```
<title>The Roosevelt Standard 1939-07-06 Roosevelt Man Dies Suddenly of Heart Disease</title>
<subjec></subjec>
<descri></descri>
<creato></creato>
<publis>Digitized by: Univ. of Utah</publis>
<contri></contri>
<dateor>1939-07-06</dateor>
<date></date>
<type>death notices</type>
<format>text/PDF</format>
<identi></identi>
<source></source>
<langua>eng</langua>
<relati></relati>
<covera></covera>
<rights>Material in the public domain. No restrictions on use.</rights>
<itemye>1939</itemye>
<itemmo>July</itemmo>
<itemda>06</itemda>
<itempa>Page 1</itempa>
<itemtr></itemtr>
<genre>newspaper</genre>
<dmaccess></dmaccess>
<dmimage></dmimage>
<dmoclcno></dmoclcno>
<find>124520.pdf</find>
<dmcreated>2005-07-28</dmcreated>
<dmmodified>2013-10-25</dmmodified>
<dmrecord>124519</dmrecord>
```

Each line represents a field and its value. Field data can span multiple lines using the CR+LF endline characters. Fields in this file are referenced by their autogenerated nicknames. To insert a new field, the field nickname and its data simply need to be added to the record block. Since there is no official record block delimiter, the first field of the file needs to be noted, in this case <title>.

For adding obituary metadata, we use the Python programming language to read data from the tab-separated spreadsheet from FamilySearch and insert data into the desc.all file. The Python script starts by reading the spreadsheet and building an internal look-up table of each row, with the dmrecord field as the key. Once the spreadsheet is parsed, the script reads in a record block from desc.all and stores it in a temporary string variable. A regular expression search is used to find the dmrecord value for the block. It then checks the dmrecord value against the look-up table to see if new obituary metadata applies to that record. If found, it builds another string based on the values from the look-up table in the same format as the the record block shown earlier. This smaller string is inserted between the <genre> and <dmaccess> fields using a simple string replace. The final string is then written and appended to an output file. This process is then repeated for every record block in the file.

After insertion the metadata xml looks like this:

```
<title>The Roosevelt Standard 1939-07-06 Roosevelt Man Dies Suddenly of Heart Disease</title>
<subjec></subjec>
<descri></descri>
<creato></creato>
<publis>Digitized by: Univ. of Utah</publis>
<contri></contri>
<dateor>1939-07-06</dateor>
<date></date>
<type>death notices</type>
<format>text/PDF</format>
<identi></identi>
<source></source>
<langua>eng</langua>
<relati></relati>
<covera></covera>
<rights>Material in the public domain. No restrictions on use.</rights>
<itemye>1939</itemye>
<itemmo>July</itemmo>
<itemda>06</itemda>
<itempa>Page 1</itempa>
<itemtr></itemtr>
<genre>newspaper</genre>
<deceas>Williams, Roland</deceas>
<deceaa>41</deceaa>
<death>1939-07</death>
<deathp>Vernal, UT</deathp>
<birth>1896-06-22</birth>
<birthp>Gentile Valley, ID</birthp>
<deceab></deceab>
<deceac>Williams, Barbara</deceac>
<decead>Williams</decead>
<deceae>Williams, Robert H; Williams, Dale</deceae>
<deceaf></deceaf>
<dmaccess></dmaccess>
```

```
<dmimage></dmimage>
<dmoclcno></dmoclcno>
<find>124520.pdf</find>
<dmcreated>2005-07-28</dmcreated>
<dmmodified>2013-10-25</dmmodified>
<dmrecord>124519</dmrecord>
```

After the new desc.all file has been generated, it is uploaded to the UDN server and indexed by CONTENTdm. A separate Bash script is used to automate the backup and index all 260 collections.

Schematic Diagram of the Data Flow for the UDN/FamilySearch Obituary Project