Databases, Search Engines, and the Genealogical Proof Standard

David Ouimette, CG, CGL
david.ouimette@familysearch.org

In keeping with genealogical standards, what should family historians understand and what methodologies should they employ to best leverage the billions of digitized and indexed records available online?

How to Approach Genealogical Websites

"**13. Source-based content**. Research plans…list databases, finding aids, indexes, search engines, and other mechanisms for accessing sources." (Board for Certification of Genealogists (BCG), *Genealogy Standards*, pp. 12-13)

Think of each large genealogical website as a conduit to tens of thousands of archives. You ought to know what the websites offer, their strengths, and their limitations. Online sources and research tools merit careful study and consideration, given their ever-increasing value for genealogical research. Understanding the processes of record selection, negotiations, digital acquisition, and online publication, as well as the functionality of search templates, search engines (on genealogy websites), and search results, will empower you in your online research.

From Archival Records to Online Family History

The large genealogical websites follow most of these steps to make records available:

1. **Targeting** – prioritize countries, archives, and records
2. **Archival agreements** – negotiate digital preservation agreements
3. **Camera capture** – digitally preserve original manuscripts and microfilms
4. **Cataloging** – describe the contents of record collections
5. **Arrangement** – organize images for online browsing
6. **Indexing** – abstract genealogical details from records
7. **Standardization** – apply name, date, and place standards to improve findability
8. **Publishing** – make images and indexes available online
9. **Search results** – interpret user input, apply match criteria, and display results
10. **Guidance** – offer record hints, contextual help, wiki articles, and training videos

Each step introduces benefits and raises questions that may influence your research. Specific genealogy standards apply to each step of this process.

1. Targeting

"**17. Extent**. …Thorough research attempts to gather all reliable information potentially relevant to the research question…." (BCG, *Genealogy Standards*, p. 14)

Each genealogical website prioritizes countries, archives, and records to pursue. For instance, FamilySearch considers these factors:

- **Current demand**: provide historical records from ancestral countries of highest demand
- **Future demand**: provide recent records where researchers live, anticipating the future
- **Coverage**: target records offering broad population coverage
- **Record value**: target records that help identify individuals uniquely within families
- **Discovery**: attract more people to family history via engaging records
- **Perishability**: prioritize top-tier records at high risk of destruction, even if demand is low
- **Archival support**: leverage archival resources and windows of opportunity
- **Partner opportunities**: accelerate efforts with partners

Targeting invariably focuses on a subset of the localities, archives, and records that a researcher may need. Most genealogical research requires a combination of online and "offline" sources.

2. Archival Agreements

"19. Data-collection scope. Genealogists attempt to collect all information potentially relevant to the questions they investigate." (BCG, *Genealogy Standards*, p. 16)

Archives seek to preserve and provide access to records. Record-digitization agreements aim to support both archival goals. Genealogists should understand that agreements with archives might limit publication rights by restricting access based on timeframe, record type, audience, or specific record details. An archive might require redaction of sensitive information of value to the genealogist. Once authorities sign a digitization agreement, camera capture begins.

3. Camera Capture

"34. Agents. Genealogists may use agents to find and examine sources, make images from them, and provide the images directly to the genealogist…." (BCG, *Genealogy Standards*, p. 21)

Ideally, the camera operator inputs bibliographic metadata for each digitized item as supplied by the archival staff. Archivists might know every nuance of every record or have entire rooms of uncatalogued documents. Most archives fit somewhere in between. Also, the camera operator should image the volume cover and spine, providing additional reference details.

Typically, digital cameras capture high-quality TIFF originals. Subsequent online publication produces lossy JPG derivatives. Usually, these are grayscale images, although over time, more

archives are requiring color. This choice often impacts your source analysis, as grayscale images may obscure what was clearly visible in the original.

4. Cataloging

"**5. Citation elements**. Complete citations use a standard format to describe at least four facets of each cited source: Who…, What…, When…, Where." (BCG, *Genealogy Standards*, p. 7)

Camera operators should enter complete and accurate citation details. This does not always occur, as the archive may supply scant information or the camera operator neglects or does not discern certain details. This may obscure the nature of the records and their provenance.

Even when adequate information is provided for each item, the cataloger still needs to supply collection-wide summaries of the records. For example, when describing the date range of a record collection, does the year of the earliest record or the first year of substantial record coverage dictate the start date for the collection? Catalogers make these decisions which in turn directly impact researchers. Question the degree of population coverage of each source.

5. Arrangement

"35. Source analysis. As they examine potentially relevant sources, genealogists appraise each source's … internal consistency – how parts of the source agree or disagree with each other." (BCG, *Genealogy Standards*, pp. 21-22)

When images arrive from an archive, they must be organized for online browsing. Ideally, the website adheres to the archival standard known as "*respect des fonds*," a principle requiring an archive—and by extension, a genealogical website—to group records per their provenance and preserve their sequence. Records from the same archive or *fond* should remain together.

If a website chooses to combine records from separate archives for ease of use online, the original context of each record might not be readily discernable. The researcher should examine the digital images and bibliographic details to discover how the online arrangement of records corresponds to original bound volumes or loose papers as physically archived.

6. Indexing

"23. Reading handwriting. Genealogists correctly read all legible handwriting in materials they consult." (BCG, *Genealogy Standards*, p. 17)

"30. Abstracts. Abstracts omit redundant, repetitive, and formulaic wording in the abstracted record…." (BCG, *Genealogy Standards*, p. 19)

Large genealogical websites obtain genealogical abstracts, or indexes, from various sources, including archives, governments, partner agreements, volunteer-based indexing projects,

genealogical societies, optical character recognition software, handwriting recognition software, and commercial keying vendors. The resulting databases are essentially genealogical abstracts or transcriptions of either original manuscripts or derivative works.

Each genealogical website uses its own criteria when making decisions about fields to index, quality standards, template design, software tools, name standardization, gazetteers, and other aspects of the indexing process. For instance, a for-profit genealogical website naturally seeks to reduce indexing costs to improve the return on investment; this directly impacts the number of fields indexed, quality thresholds, and the degree of utility of the resulting database.

7. Standardization

"35. Source analysis. As they examine potentially relevant sources, genealogists appraise each source's likely accuracy, integrity, and completeness." (BCG, *Genealogy Standards*, p. 21)

Prior to publication, genealogical websites attempt to increase the usefulness of raw indexes and make it easier to find relevant records. Pre-publication operations add calculated or derived information to indexes. For example, the raw index might contain an age field from which a derived birth year is calculated. Personal names might be compared with names common to the locality, and place names might be compared with contemporary gazetteers. Additional fields add this information with the intent of enhancing findability.

8. Publishing

Genealogical websites publish online images and indexes in accordance with their contractual rights, legal rights, and business models. Your access to certain records, images, or website features may depend upon your subscription, the country you live in, where you access the website, or other parameters dictated by the website and its legal agreements. Also, some genealogical websites allow user-submitted corrections while some produce record hints. Remember that indexes are primarily finding aids, guides to richer details found in original records and their digitized surrogates.

9. Search Results

"10. Effective research questions. Questions underlying research plans concern aspects of identity, relationship, events, and situations. The questions are sufficiently broad… [and] sufficiently focused…." (BCG, *Genealogy Standards*, p. 11)

The questions you "ask" a genealogical website should be tailored to coax the best results. Target the right content—relevant databases, record collections, or indexes—and leverage the power of search templates and nuances of the underlying search engine to obtain information.

How can you tame a genealogical website and cause it to yield the results you seek?

- ***Learn the tools: study how search results relate to your search input***
    - What results do you see and not see? Why, and what can you do about it?
    - Learn how each control on the search template works
    - Relevancy ranking and thresholding (tuned to balance precision and recall)
    - Exact match (providing better control of search results)
- ***Stop surfing: drill down to specific localities or sources per your research plan***
    - Best to search "on location" rather than from the stratosphere
    - Search templates for individual databases offer better access to records
- ***Be a power user: go beyond basic searches to unlock the potential of each database***
    - Ancestry: switch from locality to keyword searches
    - Switch from relevancy ranking to exact match
    - Use wildcards extensively (see below)
- ***Rephrase you question until the website gives a good answer***
    - Search for individuals and search for families (these are done differently)
    - Expand one parameter while shrinking another (test multiple cul-de-sacs)
    - Expand the circle of people, places, date ranges, and records you research
- ***Be creative with names of people and places***
    - Test the limits of wildcards (e.g., '*' and '?' characters) on each website
    - Gather all the spelling variations you can and construct wildcard templates
    - Learn the jurisdictional hierarchies and study relevant maps and gazetteers
- ***Search the same records on multiple websites***
    - A shared database will yield different results on different websites
    - The same source indexed twice yields significantly different results
    - Use the unique functionality of each website and correlate what you learn
- ***Be thorough before declaring someone absent from the records***
    - Scour the images, not just the indexes (negative evidence vs. negative search)

"**40. Evidence mining**. Genealogists obtain evidence from … sets of information items…. Evidence mining requires attention to detail." (BCG, *Genealogy Standards*, p. 22)

You can mine evidence across surnames, localities, date ranges, or entire databases to inform your research. For instance, counting the number of people in a census by age produces an age-distribution curve with small spikes for ages ending in '5' and large spikes for ages ending in '0'. This shows why you should be suspicious of ages that were likely rounded.

10. Guidance

"**82. Development goals**. Genealogists improve and update their … knowledge of genealogically useful materials and contexts." (BCG, *Genealogy Standards*, p. 43)

Avail yourself of the wide variety of research guidance offered on each genealogical website. FamilySearch Wiki articles, Ancestry Academy videos, contextual help, and other guidance may help you develop expertise with digitized records, search engines, and individual databases so that you may best leverage online resources in your family history research.