Genealogical Terminology and Taxonomies of Controlled Vocabularies

Stephen K. Smith
Family Search International
Building Q Suite 2300
1221 Research Drive
Orem, Utah 84097
smithsk@familysearch.org

Abstract

No one does genealogy without information. This information comes in many types, from many sources, the most authoritative being records written by impersonal scribes, stored in relatively safe and stable locations in government and religious archives. In times past, records have been categorized by the types of information they contain, especially key events, relationships, and residences. In order to facilitate communications between researchers, archivists, genealogical societies, and other interested parties, many different, and sometimes competing vocabularies of genealogical terms have been constructed.

As tools and repositories for researching, recording, and storing family relationships become more available to the interactive genealogical research community, the need for more carefully constructed, openly debated, easily validated, widely supported, and generally agreed upon taxonomies of genealogical terms will likely become more acute. This paper describes the main categories of information sought by genealogists and uses several techniques to think differently about the problem of organizing genealogical terms in a stable hierarchy, including comparing Genealogy to the game of Treasure Hunt, touching on basic scientific taxonomic techniques, and examining the parts of speech for several well-known genealogical terms, all the while avoiding excessive complexity.

Stepping Stones into The Past

The purpose of genealogical research is to make a connection across time to family members and friends, sometimes with the living, but most often with those who are dead. In reconstructing the past, the genealogist, amateur or professional, gathers information and uses it to document the life of the subject of their research to build stepping stones and finding new information about this and other people of interest. Each successive stepping stone is likely to be a little less reliable than the previous one, until eventually, one by one, the branching pathways into the past fail to provide enough information to construct new stepping stones; gaps appear, until finally each path fades and disappears into a forest of unusable information or a desert of no information at all.

Often the most reliable information about ancestors can be found in records of specific events such as births, marriages, and deaths, recorded at the time of the event. Census records are another good source of information; though they sometimes have inaccuracies introduced by fading memories, they have the advantage of being fairly regular and fairly complete. For some types of information, more personal sources provide more accurate facts than official sources, but on average, most progress in genealogical research occurs with the aid of “official” records, which then become the stepping stones to progress in genealogical research.

The Shifting Landscape of Meaning

Making connections across time can be a challenge for those with less experience, or when records are in a different language, or when word usage seems strange compared to modern times. The spellings of a person's name can change as they migrate to new geographical locations, new countries. Places can be renamed to something completely new, or the name of the place can end up in a different location (as with Winchelsea, England, which was swamped by 60 years of storms in the 13th century, and rebuilt 3 miles inland where it stands to this day). The ethnicity and language of an area can be completely replaced in a few generations (as with the Saxon exodus from Transylvania in the 19th century, motivated by a rising Romanian population). Subtle changes in usage, or transplanting of a term to a new context, or variance in record structures, can all contribute to confusion, slowing or halting a genealogical researcher's progress for a time.

In these and similar situations the researcher would benefit from authoritative help, otherwise they can get bogged down in learning the background context for each information source they encounter. Understanding the impact of context on meaning is essential to making full use of available data. The context of the archive, the record, and the record data can all have an impact on the research effort. The amateur genealogist especially does not want to have to rediscover everything for themselves when there are other people with specialized knowledge who could provide templates, heuristics, guides and other hints for interpreting record data.

The Genealogical Community Has Helped

From the beginning of organized efforts to support genealogical research, societies and other organizations have provided guidance for researchers in the form of instruction, access to documents, retrieved records, standardized vocabularies, authoritative lists, computer software, and methods for sharing information. All of these efforts have contributed to improved productivity among genealogical researchers. Standard vocabulary, name, and place lists have sped the adaption of the less experienced to research in unfamiliar territory, while improving communications between participants of all experience levels.

Categorizing Genealogical Information

Genealogical service providers and researchers can benefit from having standardized vocabularies and authoritative lists for the following information types:

Names of People: Hand written documents can be hard to read. The same name can be spelled several different ways depending upon the document or the location. Many immigrants changed their names to fit in better with their new society. In some times and places, name changes have been mandatory (for example, when several decades ago the Indonesian government required, for a time, that those with Chinese names take on Indonesian names). Names have changed for other reasons as well. Name information is largely predetermined and quite extensive. Variance of structure from one language or culture to another may require that this information be represented by a flexible hierarchy.

Personal Information: In addition to name(s) and relevant dates, this category includes country of origin, race, ethnicity, eye-color, language(s), religion(s), occupation(s), hobbies and other physical, behavioral, and cultural facts about the person.

Event Information: Events are occurrences in time that are associated with a relatively short finite period of time and are specified by dates. Among the events examined for genealogical purposes are

births, baptisms, marriages, and deaths. Event data is often accompanied by personal facts and relationship information.

Relationship Information: Family ties are the primary source of new names for research and family tree building. Indicators of relationships can be obtained from oral histories, census and event records, journals, wills, news reports, tax returns, government decrees, family bibles, letters, old Christmas cards, financial statements, and local histories.

Names of Places: The same place name can exist in several different contexts. Dekalb is a city in Illinois and a county in several southern states including Alabama and Georgia. Some place names have fallen out of use. Others can simultaneously refer to a village, a parish, and a district, all in the same vicinity. This is another category where the values in the standardized lists are historically predetermined. Identifying the correct values is more a process of discovery and organization than interpretation and synthesis. Largely hierarchical, but occasionally peer-to-peer, the structures and values used to identify the relationships of geographical entities must allow for the representation of a variety of ad hoc, bottom up, and imposed, top down confederacies, so that all locations can be uniquely identified in all time periods and all forms the place existed.

Residence Information: This is a special case of the place-name category, being a place where an ancestor lived. Knowing where a person was born, lived, and/or died enables a researcher to dig deeper into the past and find new people to research. Sometimes visiting the one-time residence of an ancestor can lead to meeting living relatives, such as distant cousins. Seeing the places where past family members lived out their lives can increase interest in genealogy and spur people on to overcome the difficulties of research in the effort to connect with family.

Record Types and Subtypes: These categories are not genealogical information but metadata about documents where genealogical information, can be found, something good to know when looking for reliable data, especially if good indexes or waypoints into their contents exist. Organizing this type of data into useful lists involves both discovery and synthesis, and like place identifiers and person's names, also likely requires a flexible hierarchical structure to represent all of the kinds of lists that are needed to meet all of the requirements of archives, planners, analysts, systems builders, consultants, and end users.

Enabling the Game of Genealogy

Looking at Genealogy as a game reveals it to be very similar to a Treasure Hunt with fairly simple rules. Except in this version of Treasure Hunt, the clues may not exist at all, and the treasure, information about ancestors, can be difficult if not impossible to find. If we want to encourage people to play this game and search out their kindred dead, we will want to make the clues as understandable as possible. We do not have complete control over the situation because the most fundamental clues were not written in our time. However, we can provide the best tools possible for deciphering the clues, storing the progress of researchers, and facilitating sharing. Records of various types, often created to record specific events in the community, can provide fairly accurate personal and historical data about each research subject so long as the data is accurately interpreted.

Among the tools being provided for deciphering the game clues are standardized and authoritative lists of people's names covering an increasing number of languages and place identifiers covering a growing number of locales. Within Family Search there is a single team dedicated to providing programmatic access to name and place data that services can eventually make available to

end users and partners. While such lists may still have errors, they are errors that can be detected and corrected by comparing the list against the historical data. As a result, the tools for deciphering the clues in the Treasure Hunt can improve over time.

Other standardized lists based on authoritative information, covering terminology for the categories of personal, event, relationship, residence, and record type data are also being researched and constructed. These lists are sometimes referred to as Controlled Vocabularies. The terms these lists contain give genealogical service providers the language they need to characterize and catalog collections of records and they can be used by researchers and indexers to decipher and interpret individual records.

Record types and subtypes are often combined with events to form a hierarchical taxonomy, which, given the many implementations available, are apparently easy enough to create. Often it is easier to create a new list rather than search out an existing list. Sometimes existing lists are readily available but either their contents or organization do not meet the current need, prompting the creation of a new list. Competition between service providers can also stimulate the proliferation of vocabularies, including online glossaries.

While much of the criteria for determining the structure and content of genealogical terminologies are determined by the research context (such as the country, the archive, the collection, and the documents,) many of the decisions about the structure and content of various controlled vocabulary lists are driven by a desire to meet an immediate need, an approach that is often detrimental to improved quality, expanded capability, and the building of sustainable common ground. The resources spent in repeated duplications of effort are significant, but may be justified at times.

To the extent a list is subjectively derived, it is neither verifiable nor falsifiable. In the absence of objective criteria for evaluating vocabularies, no one can say a particular list is good or bad, adequate or inadequate. Without a means of obtaining feedback from interested parties about specific vocabularies, without a way for users to give feedback to list suppliers, the proliferation of vocabulary lists will most likely continue, and systematic improvement in the controlled vocabulary space will less likely occur.

Normal Science, Normal Genealogy

In his highly influential book, *The Structure of Scientific Revolutions* (3rd ed. 1962, 1970, 1996, p. 200), Thomas Kuhn argued that what he called "Normal Science" proceeded using existing models as a foundation for research until a crisis of incongruence between the model and empirical data caused some scientists to seek a theory that was more coherent with observed reality. History shows that whenever multiple theories have existed at the same time, each was supported by a faction, and every faction sought to refine their own model and question other models until one prevailed.

According to Kuhn, "The practice of Normal Science depends on the ability … to group objects and situations into similarity sets which are primitive in the sense that the grouping is done" more or less intuitively. In other words, a usable model produces fairly predictable results of categorization when those who understand the model do the categorization. He also notes that good models are always communicable by exemplars to neophytes in ways that result in effective participation with the model.

While everyday genealogy is seldom performed with the same rigor as Normal Science, success in the genealogical space becomes more routine and predictable when concepts and models can be clearly communicated and the research process

used by experts can eventually be duplicated by non-experts. In fact, this is already true of genealogy in the general sense, but there are still opportunities for refinement in the state of the art.

Among the areas where improvement is needed is in the compilation, organization, and management of the terms and lists that comprise controlled vocabularies. The ability to group terms in similarity sets and hierarchical relations is not easily replicable because the model for relating terms in taxonomies either is not well understood, is not well defined, or is not well communicated to those engaged in the process, and as a process is incapable of generating the same list from the same data if done by different educated authors. It is too dependent upon individual preferences.

The people who have built vocabulary lists in the past have done a fine job on an individual basis. But collectively, there is room for improvement and benefits to be realized from reducing duplication and increasing rigor, especially if the problem is approached incrementally so that existing development projects can move forward in the near term. There are fairly simple mechanisms for classifying and relating genealogical terms, such as taxonomic approaches, that could increase the universality of the resulting classifications among those who are prepared to utilize them.

The Art and Science of Taxonomies

"Taxonomy in its origins is one of the most elementary of disciplines" – EB 21:855

Taxonomy comes from the Greek for taxies (arrangement) and namos (law) and designates a systematic arrangement of the kinds of things, usually by way of a hierarchy of classes and subclasses. "The purpose of taxonomy is to develop a convenient and precise method of classifying knowledge that conforms to fundamental principles." (EB 21:853)

The most famous taxonomy in use today is the one that Swedish botanist Carl Linnaeus developed for biological classification. His solution for the haphazard and confused mess that existed prior to the advent of his system was based on an elegant hierarchy of Kingdom (plant, animal), Phylum, Class, Order, Family, Genus, and Species. This system was improved over time by a participatory process of proposal and critique. The Linnaeus model possessed the characteristic that it could be corrected and expanded with new information without a loss of structural integrity. (EB 21:853)

Criteria for Classifying Terms

In order to achieve the best fit for a term in a specific taxonomy, there are some quick checks that can be used to analyze its attributes.

The values of names of people and places are proper nouns and are used as such to identify a person or place. As mentioned before, residences are a special class of places.

Personal information values can be adjectives, as in "blue" for eye color, or they can be nouns, as in "doctor" for occupation.

Event type names have a common verb form and can pass a test such that the verb form of an event type will make sense when used in a small set of well known sentences that can be constructed for this purpose, and used to test a term for event status. Consider the following sentences:

1) I experienced _____ today.
2) I _____ today.
3) I will be/become _____ today.
4) Today is the day of my _____.

True events can be fit into all of these sentences because they have a verb form. Here are event related words along with the forms that fit each sentence:

Event (1) an event, (2) involved in an event,
(3) in an event, (4) event.
Birth (1) birth, (2) was born, (3) born, (4) birth.
Baptism (1) baptism, (2) was baptized,
(3) baptized, (4) baptism.
Enlistment (1) enlistment, (2) enlisted, (3) enlisted,
(4) enlistment.
Marriage (1) marriage, (2) married, (3) married,
(4) marriage.

An attempt to put relationships into these same questions does not work as well. For example, the sentence, "I experienced a relationship today" makes it sound like it wasn't really a relationship at all, but two ships passing in the night. Relationships only work in these sentences when accompanied by at least a few modifying words. For example, here is the Husband-Wife relationship inserted into the event questions:

Husband-Wife (1) being a husband for the first time (2) was a husband for the first time, (3) a husband, (4) becoming a husband.

Relationships are close to events because they are created by events, but hopefully they endure much longer than an events-length time frame.

Record types are distinguished by the fact that each consists of a noun used as an adjective (adjectival), or an adjective used as such. Noting that the word "records" is a plural noun we can see which record types are designated using an adjective and which are designated using an adjectival.

Cemetery Records (adjectival)
Birth Records (adjectival)
Civil Records (adjective)

By using reasonable criteria for organizing standard genealogical terms, we can avoid obvious taxonomical categorization mistakes roughly equivalent to making tigers and zebras more closely associated than tigers and lions in biology.

Conclusion

There are undoubtedly more exhaustive methods and criteria for classifying terms than what is being presented here. Simplicity is preferred in the near term because it is better to take simple steps than take no steps at all, and these simple measures may be sufficient to increase the precision of term classification schemes over their present state. Such an increase in precision has a chance of accelerating progress toward a more unified future for controlled vocabularies. As these and other simple suggestions are implemented, additional research may be necessary to obtain greater improvements in the organization of genealogical terminology.

Other areas of research that are needed with Controlled Vocabularies include abstract analysis of requirements from many different sources, possible digital representations of the data, development of computer tools to ease list management costs, term and list versioning schemes, translation to and from languages other than English, and distribution of lists to potential users.

References

The Structure of Scientific Revolutions, by Thomas Kuhn, 3rd ed. 1962, 1970, 1996, p. 200

Taxonomy, Encyclopedia Britannica, 1966, pp. 853-855