January 2011

Name Disambiguation – Learning From More User-Friendly Models

Bob Thomas
Western Washington University, bob.thomas@wwu.edu

Follow this and additional works at: http://cedar.wwu.edu/library_facpubs
Part of the Library and Information Science Commons

Recommended Citation
http://cedar.wwu.edu/library_facpubs/21

This Article is brought to you for free and open access by the Western Libraries and the Learning Commons at Western CEDAR. It has been accepted for inclusion in Western Libraries Faculty & Staff Publications by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.
Name Disambiguation – Why do We Continue to Ignore Patrons' Needs Learning From More User-Friendly Models

BOB THOMAS
Western Libraries, Western Washington University, Bellingham, Washington, USA

Library catalogs do not provide catalog users with the assistance they need to easily and confidently select the person they are interested in. Examples are provided of Web services that do a better job of helping information seekers differentiate the person they are seeking from those with similar names. Some of the reasons for this failure in library catalogs are examined. This article then looks at how much information is necessary to help users disambiguate names, how that information could be captured and shared, and some ways the information could be displayed in library catalogs.

KEYWORDS: authority control, cataloging, cataloging standards, catalog indexing, catalog display, catalog design, library catalogs, OPACs, discovery platforms, discovery services

Address correspondence to Bob Thomas, Integrated Systems Librarian and Head of Cataloging, Western Washington University, 516 High Street MS 9103, Bellingham, WA 98225-9103. E-mail: Bob.Thomas@wwu.edu

INTRODUCTION

When a patron searches for a name in a library catalog, for example “Winston Churchill”, one of two things will happen in most catalogs. In a traditional OPAC, if the name is searched as a name or subject heading, the patron is shown a list of various Winston Churchill’s. The patron is then supposed to pick the specific Winston Churchill they are interested in. Yet the only information provided is the name, sometimes a middle name, and a birth and/or death date. In the Library of Congress OPAC, a search for “Winston Churchill” results in a list of eleven Winston Churchill’s (six Winston Churchill authorized headings, one additional name containing Winston Churchill, and four see references). If the same search is made as a keyword search, or in one of the newer search interfaces released over the last several years, patrons are usually presented with a set of resources by or about a wide variety of Winston Churchill’s, and often a list of some of the names on either the left or right of the screen that can be used to further refine their search. In the WorldCat.org search interface, the names are listed as shown in figure 1 – without the differentiating fuller forms or dates!
In a discussion on PCCLIST, Antony Franks pointed out that part of the problem is "the confusion of what makes a unique heading versus what identifies a heading as a specific person." While the heading Churchill, Winston, 1874-1965 is a unique heading within the LC Names Authority File, it does not clearly indicate which specific person the heading represents. On the NGC4Lib list, Ed Jones recognized that "the suggested string – the preferred access point used by the library – might itself be ambiguous to the searcher." The lack of information users need to easily and confidently select the person they are seeking, at the point it would be most useful, is one factor that prevents library catalogs from successfully competing in an environment where there are many similar tools that are more helpful.

NAME DISAMBIGUATION AND AUTHORITY CONTROL

Researchers have developed a variety of automated approaches that attempt to automatically group resources into clusters, with each cluster related to a separate person. So far, these approaches have split the works of one individual into many separate clusters, included resources created by multiple people in a single cluster, or often both. In library catalogs, instead of automated clustering, separate groups of resources related to individual names have traditionally been created using the authority control process. This process encompasses three principal steps. First is the creation of unique strings, or authorized headings, for each separate name, e.g., Churchill, Winston, 1874-1965 and Churchill, Winston, 1871-1947. Second, the appropriate authorized heading is selected and recorded in the bibliographic record for each resource. Alternatively, instead of using unique headings, other unique identifiers such as the International Standard Name Identifier (ISNI) or the Open Researcher and Contributor ID (ORCID) could be recorded in the bibliographic record. The catalog software could then obtain the display text from the authority record indicated by the unique numeric identifier. Lastly, a catalog user must be given the ability to select the resources related to a single person. This is...
normally accomplished by offering a list of names and allowing the user to select the name they are seeking, which allows the catalog system to display the resources linked to the name selected. This process has been very effective in allowing catalogs to provide both precision (the ability to retrieve only the items created by a particular individual) and recall (the ability to retrieve all the items for that individual).

The weak link in this process is expecting catalog users to be able to tell which Winston Churchill represents the novelist, and which the prime minister, based solely on the information contained in these two headings. One of the principle reasons that library catalogs do not better support users in this task is that the name headings used in library catalogs were designed for dictionary catalogs – a paper-based format that severely restricted the amount of information that could be displayed for each entry. Phonebook-like listings on paper that predated even card catalogs. As Rochkind pointed out, “there’s a reason Cutter’s opus is called Rules for a Dictionary Catalog.” In a dictionary catalog, the limitations of the format prevented the display of information beyond the simple unique headings. Even though libraries moved beyond the dictionary catalog in the early twentieth century, and into the electronic catalog in the late twentieth century, our catalog and authority standards do not yet take full advantage of the new flexibility in information display available in electronic catalogs we have chosen to remain slaves to the display restrictions that are now a century old. Libraries have so far failed to take advantage of the modern catalog environment to provide users with better support in disambiguating names.

NAME DISAMBIGUATION ON THE WEB

Non-library information search tools on the web are often faced with the same issue of helping users select the desired name from a list of similar names. A search for Chris Brown in The Internet Movie Database (IMDb) takes the user directly to a name disambiguation page (see figure 2) containing all Chris Brown’s in the IMDb database plus similar and variant names. IMDb provides three additional data points about each name to assist the user in selecting the name they are seeking – a work the person was involved in, the date of the work, and their role in that work. Although IMDb provides some information to assist the user in disambiguating similar names, this automated approach isn’t as helpful when a person has performed multiple roles in the movie industry and the single work displayed is not a role or work well known to the searcher.
Wikipedia uses a very different approach to address name disambiguation. A search for Winston Churchill in Wikipedia takes the searcher to the entry for the name the Wikipedia software determined was most likely to be the one desired, while at the same time offering a link leading to a list of all similar names. This disambiguation page (figure 3) not only groups similar names into clusters, in this case People, Ships, Schools, and Other uses, it also provides a manually created descriptive phrase for each name in the listing. Unlike IMDb, the Wikipedia descriptive phrases can include multiple roles, multiple well-known works, a range of dates for when the person was active, etc.
Spool et al. popularized the notion of the scent of information, derived from information foraging theory. Information scent refers to how well the terms used in navigation options match a user’s need and how well they predict what the user will find if that option is selected. For instance, if a person is seeking information about a saxophone player named Jones, and the information about one of the Jones’ offered as an option contains the phrase saxophone player, the user’s confidence would be very high that selecting that particular name would likely lead to resources by the person sought. Spool dubbed these key terms trigger words. His team found that users tend to scan web pages for words or phrases they associate with the content they’re seeking in an attempt to pick up the scent. If they spot one of the terms they’re looking for it triggers an action by the user – they attempt to follow the trail marked by the trigger word(s). Spool’s team found that if the link labels are missing the clues the user needs to move forward, i.e., if the link doesn’t give off enough “scent”, users are significantly less likely to find the resource they’re seeking.
One approach to providing information about a name in a way that would include a large number of trigger words is to create a page of information about each name. WorldCat.org has taken this approach with its WorldCat identities pages. From within the display of a bibliographic record, the user is offered a link to “Find more information about” each name associated with the resource. Selecting this link for Winston Churchill (the prime minister) sends the user to a WorldCat Identities page containing a wealth of information about Winston Churchill – roles, related subject headings, publication timeline, most widely held works by Winston Churchill and \textit{about} Winston Churchill, related identities and associated subjects. Informational pages such as WorldCat identities pages are excellent resources for catalog users wishing to know more about a person. The information gathered on a WorldCat identities page, however, is too much for name disambiguation -- it cannot be displayed in a list of names in a way that is helpful in easily selecting which name among similar names represents the person being sought.

In creating descriptions that help people differentiate names, the descriptive phrase needs to be long enough to include the trigger words most likely to be recognized as relevant, but short enough to be displayed at the point of need. They do not need to be full or even brief biographical sketches. Seldom would a user be faced with a group of people with similar names where they had to differentiate between two prime ministers, or two writers on plastic surgery. Normally, a group of similar names will contain people in diverse fields, so a description containing just a few well chosen trigger words would often be sufficient to differentiate between Jim Clark the Formula One world champion and Jim Clark the founder of Netscape. Two cases, however, argue for descriptions that are longer than just a few words. The first are those cases where there are multiple people with similar or identical names who were active in the same field, such as Johann Strauss, 1804–1849 (Austrian composer, famous for popularizing the waltz, best known work is \textit{Radetzky March}, father of the composer known as “The Waltz King”), and Johann Strauss, 1825–1899 (Austrian composer of dance music and operettas, known as “The Waltz King”). The other case that would seem to require longer descriptive phrases is when one individual has been active in multiple fields. An example would be David Levine, the music director of the Met, a baritone, and a conductor. In researching the balance of “long enough but not too long,” Spool found that the best labels, at least for web links, have 7-12 words, and that if the link labels get too long they become “noisy,” masking the trigger words.

\section*{Adding Descriptive Phrases for Names to Catalogs}

Ockerbloom, in describing the concept-oriented catalog, states that the “catalog should work with the user by providing an easily understood context for their search. Users should be given a clear idea of what concepts they’re looking at (e.g., by clearly naming and describing them), and what resources are associated with those concepts” (emphasis mine\textit{added}). Two of the basic design rules he proposes give further clarity: use clear names for your concepts, and provide useful information about concepts when people come across them.
If library authority records contained short descriptive phrases designed to help users differentiate between similar names, developers of OPACs and other discovery tools could display this information in a variety of different ways. First, this type of information could be displayed simply as additional information following each name in a traditional name browse display. A number of recently introduced discovery tools, however, take the user from a search directly to a related set of resources. Often these interfaces include a navigation pane on the side of the screen allowing users to refine the search by selecting among a list of options, such as the creators of the resources in the original result set. In this type of display, one method of displaying the descriptive phrases would be as tool tips, i.e., text displayed whenever the cursor hovers over a link. A third method for systems designed to actively guide a user through the search process could use the descriptive phrases to formulate queries that allowed the user to indicate whether a system suggested route was the direction the user wished to pursue, such as, “Did you mean Teddy Roosevelt, the former President of the United States?” Finally, OPACs could offer users the option of going to a special name disambiguation page such as those used within Wikipedia.

NAME DISAMBIGUATION IN CATALOGING PRINCIPLES

The 1961 Paris Principles, in describing the collocating function of the catalog, confirmed the need to select a single form of a name for use as the authorized name heading. The Principles did not address the capture of additional information to assist in differentiating among similar names. Guidelines for Authority and Reference Entries, published in 1984, permitted the recording of non-cataloging data concerning the author’s identity and sources used, but was silent on the direct use of the information by the users of the catalog. While that report was being written, Authorities: a MARC Format had been prepared which, while also recognizing the need to capture information related to the author’s identify and sources used, relegated this information to a 670 Note, a field intended for use only by catalogers and therefore not formatted for viewing by catalog users.

One of the objectives of the IFLA Functional Requirements and Numbering of Authority Records (FRANAR) Working Group, established in April 1999, was “to clarify the underlying concepts [of authority data] to provide a basis for refining and improving on current practice in the future.” Two of the user tasks listed by the working group seemed to indirectly acknowledge that the catalog user needs sufficient information to determine which name is the one being sought:

- Identify: Identify an entity (i.e., to confirm that an entity represented corresponds to the entity sought, to distinguish between two or more entities with similar characteristics).
- Contextualize: Place a person, corporate body, work, etc. in context; clarify the relationship between two or more persons, corporate bodies, works, etc.; or clarify the relationship between a person, corporate body, etc. and a name by which the person, corporate body, etc. is known.
FRANAR, in the definition of the *contextualize* user task, appears to acknowledge that knowing a name is not enough to determine which person that name represents. One of the functions identified by the working group was that “information in an authority file [provides] information that can be used in distinguishing one person, corporate body or work from another.” Again left unanswered was who this information is for — catalogers, patrons, or both.

Functional Requirements for Authority Data (FRAD)\(^1\) published in 2009, repeats much of the information on user tasks described in FRANAR, but expands on the types of information about a person that should be captured in an authority file such as gender, place of birth/death, country, place of residence, affiliation, address, language used, field of activity, profession/occupation, and biography/history. Left unaddressed is when and in what form this type of information might be made available to users of the catalog.

It is the author’s contention that concise phrases useful in disambiguating names need to be crafted by people—that an automated algorithm would not be capable of selecting the few most important attributes related to a person that allow disambiguation. Since disambiguation phrases are related to the group 2 entity, and not the works produced, the appropriate mechanism to store and share these phrases is in the authority record, along with similar attributes such as biographic information. The FRAD introduction states that the FRAD report fulfills the first of three terms of reference for the FRANAR Working Group, which was “to define functional requirements of authority records . . .”. Therefore, the author proposes that the list of FRAD\(^1\) attributes be expanded to include brief disambiguation phrases.

According to the *Statement of International Cataloguing Principles*,\(^2\) the highest principle of the construction of cataloging codes should be for the convenience of the users of the catalogs. Yet, at least when recording authority data, there is a marked lack of emphasis on capturing and formatting authority information specifically intended for display to and in support of users of the catalog. As the FRAD model of authority information is revised and expanded, the model needs to clearly reflect an emphasis on recording information specifically designed to improve the user experience.

**CONCLUSION**

A number of new web resources have shown that a descriptive phrase, displayed at the point of need, would allow patrons/users to easily select the person they are seeking from a list of similar names. Library catalogs would be more user friendly if they added name disambiguation phrases to any list of names. Catalogers need to continue recording much of the descriptive information about people, corporate bodies and titles, recorded by catalogers in authority records could be used to assist users if it was formatted for use by both users and catalogers, instead of being but stop recording the information in notes intended for only catalogers. Our existing authority records could easily accommodate a short descriptive phrase field formatted for display to catalog users. FRAD recognizes a basic user task is to *contextualize*—to place a name in context and to clarify the relationship between two or more names. The author recommends
that FRAD be revised to include an attribute for brief disambiguation phrases. Additional research is needed to determine if it is possible to develop a set of principles that could be used to create effective disambiguation phrases.

When catalogers NACO contributors were given the freedom to add death dates to authority records, and later when OCLC initiated the Expert Community Experiment allowing catalogers the broader cataloging community to update more records in WorldCat, the cataloging community showed that a large number of people making small individual contributions, can add up to a huge community-wide contribution in very little time. The cataloging NACO community is fully capable of adding descriptive disambiguation phrases to authority records in the same way – through small individual contributions.

It is essential that we constantly return to Rosenberg’s statement that “the design of any … information system should be the system’s ease of use … If an organization desires to have a high quality of information used, it must make ease of use of primary importance.” If library catalogs are to survive in today’s rich, web-based information environment, I believe they must improve their ease of use and offer services which are both valued by information seekers and which cannot be replicated in the open web environment dominated by the major search engines. Due to our rich authority structure, the library community could quickly begin capturing and displaying the information users need to easily disambiguate names. There is no reason the user of a library catalog should be faced with multiple Winston Churchill’s and not be shown which one is the one he or she is seeking in a way that makes sense to them.
NOTES


