5.1. The need for control

Overview

Much of the task of organizing information involves creating representations for information objects using words and phrases. We know that when we use words, they can mean different things. More importantly, different words or phrases can mean the same thing. A goal of information organization is to bring like things together (e.g., all items written by the same author, all items that deal with the same subject). This bringing together is called collocation. A method we use for collocation is to exert control over the words and phrases and names we use in our representations. In information retrieval, a benefit of collocation is to improve recall: if all like items in the system are retrieved, recall is 100%.

This series of modules explores the problem, identifies some of the issues, and presents possible solutions. In addition, these modules prepare you to develop a name authority control mechanism for your project (section 5) and get you ready to address vocabulary control for subject representation (section 4).

The first thing we need to discuss, however, is the notion of access and access points, and then see how certain access points are handled in special ways to reach the goal of collocation.

Access and access points

Traditionally, library users were expected to search for items in catalogs with one of three pieces of information: author, title, or subject. Regardless of whether they had that information when they went to the library catalog, we trained them to think of searching by one of those three pieces of information. We trained them by providing entry into the catalog only by those three pieces of information.

Some of you may remember when library card catalogs were divided into three separate sections: authors, titles, and subjects. While each card (i.e., the bibliographic record) contained a lot of information, catalogers put at the top of a card either the author's name, the title, or the subject heading. We refer to these three areas of the bibliographic record that we could search as the access points. Authors, titles, and subjects are the three traditional access points provided by library bibliographic instruments. Think of an access point as the information in the bibliographic record you can look at to find the record (or in an electronic database, the area of the record that is searchable).
The problem

An access point contains a data value. In an access point for author, the data values are authors' names. Similarly, the title access point contains data values of titles and the subject access point contains data values for subjects. But given the many variants of these values, how do we decide exactly which values to put in the access point? Let's introduce the problem by looking at names of people.

The first thing you have to remember is that computers are stupid. The vast majority of search engines simply take the string of characters (i.e., letters and words and phrases) you type in and match those characters with strings of characters in the fields in the database record. Remembering this will help you understand the problem.

We'll use the name William Eugene Moen as an example. This person is also known as Dr. William E. Moen, Bill Moen, and W.E. Moen. All these labels are ways of identifying him. When he writes a paper or article for publication, he puts his name (as author) as William E. Moen. A few years ago, he represented himself this way in an article for the National Information Standards Organization (NISO) newsletter Information Standards Quarterly. When the article appeared in print, it listed him as Bill Moen. Here's a challenge: find all papers, articles, book chapters, technical reports, etc., that this person has written in the past 10 years. If you searched under the form of the name he typically uses, you would never find the article he wrote for NISO. The computer has no way of knowing that "William E. Moen" and "Bill Moen" are the same author.

Let's take a case where a person changes his or her name. The former dean of the School of Library and Information Science at Louisiana State University has gone through several name changes, and she has published many articles under each name: Kathleen Heim, Kathleen de la Pena, and Kathleen McCook. Unless you know all the names under which she has published, you would not be able to do a comprehensive literature review or develop a comprehensive bibliography for her.

Finally, there are authors who write under different names (pseudonyms). The classic library school example is Mark Twain and Samuel Clemens. Samuel Langhorne Clemens was the real name of the person who wrote any number of items under various literary names. Those names include Mark Twain, Quintus Curtius Snodgrass, and others. Go to http://iii.library.unt.edu/ and do an author search for "Snodgrass, Quintus" and see what happens. Let's assume that the user knows that Snodgrass is one of Samuel Clemens' pseudonyms. The catalog output can be seen as a helpful way to get the user to all books by the author Samuel Clemens. Go ahead and try the search now using the author name "Clemens, Samuel." What happens?

The problem of different names of people (and organizations) is one aspect of the problem. Another one we'll cover in more detail when we get to subject representation in section 4 of the project deals with using different terms to mean the same thing. For example, let's say we have two books that are both about computers. We could say the subject of the books is computational machinery (an older phrase to refer to computers) or we say the subject is computers. Which would be right?

Both would be right. Some books may be labeled "computational machinery" and other books "computers." As in the case of Moen, Heim, and Twain name problems, users would either have to do searches using both "computational machinery" and "computers" as search terms to make sure they found all books on this topic (i.e., get collocation).
One concept, many instances

The way we have approached and solved the problem introduced above is through something call **authority work**. In our textbook, Taylor (2004, 356) defines authority work as:

The process of determining and maintaining the form of a name, title, or subject concept to be used in creating access points. In the name and title areas, the process includes identifying all variant names or titles and relating the variants to the name or title forms chosen to be access points. In some cases it may also include relating names and/or titles to each other. In the verbal subject area, the process includes identifying and maintaining relationships among terms—relationships such as synonyms, broader terms, narrower terms, and related terms.

For example, in name authority work, we agree to use a certain form of a person's name consistently throughout our representations so user need only to use that form of the author's name as a search term to locate all items written by a single person.

Another way to think about this is in terms of the key concept of **representation**. There was a real person named Samuel Langhorne Clemens who wrote books and essays under that name, under the name Mark Twain, and under other pseudonyms. Remember our job is to bring order to the bibliographic universe, and we do that by showing relationships between entities in the universe. We have the challenge of showing that Samuel Clemens, Mark Twain, and Quintus Curtius Snodgrass are related; that they all refer to the same person. We can call these three entities the different bibliographic identities for Samuel Clemens.

To solve the problem of multiple names, we create a representation of this person and these pseudonyms, where that representation is a standard form of a name that we will use consistently throughout our records. In this case, the agreed upon or authorized form is Twain, Mark, 1835-1910.

This agreement is based on rules in the library community to standardize on the most well-known form of the name as the "authorized name." The names "Samuel Clemens" and "Quintus Curtius Snodgrass" are considered variant forms of the name, and we show the relationship of the variant names to the authorized name in an **authority file**.

The graphic below depicts a model of name authority control. For a person who writes under multiple names, we standardize one form of the name according to specific rules and put that authorized form in an authority file. We then use the authority file as the source of data to enter in bibliographic records. For all items written by the person Samuel Clemens, no matter what pseudonym (or bibliographic identity) he used, we will always show the author as Twain, Mark, 1835-1910.
The concept of authority control is very broad. We can say that authority control is any effort to standardize on a form of a word, phrase, or code to use in the bibliographic records to bring consistency to the data in those records. Inmagic offers the feature of a content validation list, and that is nothing more than bringing control to the data that appear in specific fields in the record. Authority control, whether for names, titles, subjects, or other types of data in the record helps users. Users do not have to search under all possible variants of a name, title, subject, etc., to make sure they have found all items in the database. Using a standard form allows us to achieve the goal of collocation and to improve recall in information retrieval.

Summary

In this module, we introduced some basic concepts related to the topics of access and authority control. This is simply to set the context for subsequent modules that will take a closer look at the importance and issues of authority control as it relates to providing more effective access to the bibliographic records. Remember the big picture here: we are trying to organize the bibliographic universe through representations. Authority work is really about another set of representations. This time the representations take the form of authorized forms of names, titles, subjects, and other data. Through these representations and the mechanism of authority work, we show relationships among entities and help to achieve the goal of collocation and to improve recall in information retrieval.

Cites & sites