Authority Control in the Context of Bibliographic Control in the Electronic Environment

Michael Gorman
California State University, Fresno

There is a sense in which authority control and bibliographic control are coterminous — two sides of the same coin. At the very least, bibliographic control is literally impossible without authority control. Cataloguing cannot exist without standardized access points and authority control is the mechanism by which we achieve the necessary degree of standardization. Cataloguing deals with order, logic, objectivity, precise denotation, and consistency, and must have mechanisms to ensure these attributes. The same name, title, or subject should always have the same denotation (in natural language or the artificial languages of classification) each time it occurs in a bibliographic record, no matter how many times that may be. Unless there are careful records of each authorized denotation, the variants of that denotation, and citations of precedents and the rules on which that denotation is based (*id est*, authority control), the desired and necessary standardization cannot be achieved.

Let us begin by looking at the fundamentals of cataloguing. A catalogue record consists of three parts:

- An access point
- A bibliographic description
- A location or (nowadays) the document itself[i]

The access point leads the user to the record; the description enables the user to decide whether the items desired is the one sought; and the location takes the user to the desired document. This is a simple and profound formulation and is the basis of all cataloguing. Even the wretched Dublin Core, of which more later, contains access points, descriptive elements, and locational information. Each element of a catalogue entry is standardized. The description and the location are presented in standardized form (or they could not be understood). The standards that govern the description (principally, the International Standard Bibliographic Description standards) and the local standards for location designations are not part of authority control. Therefore, authority control and the related vocabulary control are concerned with access points and their standardization. The access point has two basic functions. It enables the catalogue user to find the record and it groups together records sharing a common characteristic.[ii] In order to carry out the first function it must be standardized (obviously the user should always find *Il gattopardo* under **Tomasi di Lampedusa, Giuseppe [Gattopardo]**, and not sometimes under that access point and sometimes under **Lampedusa, Giuseppe Tomasi di** and/or translations and variations of the title). This is what is known as vocabulary control—the presentation of every name (personal or corporate), uniform title, series, and subject denotation in a

single, standardized form. The reason why we have descriptive cataloguing rules is that any cataloguer following those rules should, theoretically, achieve the same, standardized result in each case.

Authority records

Vocabulary control is vital to authority control but it is only the first, if most important, aspect of authority work. Robert Burger discusses the concept of the authority record—the vehicle that contains the results of authority work.[iii] In summary, he writes that the role of the authority record has five components:

- To record the standardized form of each access point
- To ensure the gathering together of all records with the same access point
- To enable standardized catalogue records
- To document the decisions taken, and sources of, the access point
- To record all forms of the access point other than the one chosen as the normative form (*id est*, forms from which reference is made)

To which I would add

- To record precedents and other uses of the standardized access point for the guidance of cataloguers.

In many card and other pre-OPAC catalogues, the authority record existed only implicitly—in the evidence of the catalogue entries themselves. Online catalogues demand the explicit formulation of authority records, linked to catalogue entries and containing at least the elements demanded by the roles listed above. That is, they must contain:

- The standardized access point
- All forms from which *See* reference is made
- A connection (a *See also* reference) to all linked authority records
- The sources from which the standardized access point has been derived
- Lists of precedents and other uses of the standardized form.

In addition, in developed machine records, the authority record will itself be linked to all the bibliographic records to which it pertains.[iv] The database made by assembling all the authority records used in a catalogue is called an authority file or, looked at another way, a thesaurus.

From whence does the content of authority records come?

The content of authority records (the authoritative form, variant forms, links, and notes of various kinds) is obviously of the greatest importance. In cases in which there are variants, there is always a reason for choosing one form over the others and, crucially, one source of information over the others. The primary agent in such choices is the code of cataloguing rules in force in the area in which the cataloguing is done. Because we

have no global cataloguing code (though the *Anglo-American cataloguing rules, second edition*—AACR2—has a global reach), global lists of subject headings, or global classification systems, cataloguers in different areas may reach completely different conclusions, even when they are proceeding from exactly the same evidence. That evidence is a mixture of the objective (the evidence presented in the materials themselves and in reference sources) and the subjective (the cataloguer's interpretation of the cataloguing rules or subject matter of the document being catalogued).

Here are some of the sources that have to be taken into consideration in constructing catalogue records:

- Existing national and local authority files
- The applicable cataloguing code, subject heading list, etc.
- The document being catalogued
- Reference sources (using the broadest definition of the term—any source providing useful data).

Each of these has to be weighed against the others and, even within each category, some sources have more authority than others. No source can be regarded as always dominant. For example, the evidence presented in the document itself may be superseded by the evidence found in reference sources when the cataloguing code's rules tell the cataloguer to do so. Again, a national authority file may be more authoritative in one case, but a local authority file more authoritative (because of special local knowledge) in another. Within the document itself, information found in one part may conflict with information found in another. Lastly, there is an obvious hierarchy of reference sources. Some publishers produce works of higher quality than others, and most printed sources are more authoritative than most electronic sources. The result of all this is the need for the cataloguer to be able to negotiate these ambiguities by exercising skill, good judgement, and the fruits of experience. The *skill* lies in knowledge of the type of material being catalogued; knowledge of the rules that govern cataloguing; knowledge of the interpretations of those rules in the past; and knowledge of applicable reference sources and their strengths and weaknesses. The good *judgement* lies in the ability to weigh all these factors and to decide based on the spirit of the rules when the letter of those rules is ambiguous. The *fruits of experience* lie in the cumulation of knowledge of rules, policies, and precedents gained from cataloguing many materials over the years. Given these three attributes, a cataloguer can produce records that are truly authoritative and that will benefit cataloguers and library users across the world.

Metadata and authority control

Metadata—literally "data about data" (a definition that would include real cataloguing if taken literally)—arose from the desire of non-librarians to improve the retrievability of Web pages and other Internet documents. The basic concept of metadata is that one can achieve a sufficiency of recall and precision (see below for a discussion of these criteria) in searching databases without the time-consuming and expensive processes of standardized cataloguing. In other words, something between the free-text searching of

search engines (which is quick cheap, and ineffective) and full cataloguing (which is sometimes slow, labor-intensive, expensive, and highly effective). Like all such efforts to split the difference, metadata ends up being neither one thing nor the other and, consequently, has failed to show success on any scale, which is the touchstone by which all indexing and retrieval systems must be judged. Any system can be effective if the database is small. The real test is how the system handles databases in the millions. Catalogues, even vast global catalogues such as the OCLC database, have been shown to be effective. Search engines, even the supposedly advanced systems such as Google, are demonstrably ineffective in dealing with vast databases.

After many papers and numerous conferences (a process in which renegade librarians joined), a quasi-standard promoted by OCLC and called the Dublin Core emerged as the shining example of metadata and what it could achieve. The Dublin Core (DC) consists of 15 denotations, each of which has a more or less exact equivalent in the MARC record. As any true cataloguer knows, MARC contains far more than 15 fields and sub-fields, in addition to the information contained in coded fixed fields. In addition, there are MARC formats for a variety of different kinds of publication, from books and serials to electronic resources, which adds to the variety of denotations. Those who advocate metadata and, implicitly or explicitly, believe that the whole range of bibliographic data can be contained in 15 categories ignore the fact that the MARC formats are not the result of whimsy and the baroque impulses of cataloguers but have evolved to meet the real characteristics of complex documents of all kinds. What we have is a simplistic (in many ways naïve) short list of categories that is expected to substitute for cataloguing when put in the hands of non-cataloguers.

The literature of metadata is littered with references to "MARC cataloguing," an ignorant phrase that betrays the hollowness of the metadata concept.[v] MARC, as any cataloguer knows, is a framework standard for holding bibliographic data. It does not dictate the content of its fields—leaving that to content standards such as AACR2, LCSH, etc. People who talk of "MARC cataloguing" clearly think of cataloguing as being a matter of identifying the elements of a bibliographic record without specifying the content of those elements. It is, therefore, clear that those people do not understand what cataloguing is all about. The most important thing about bibliographic control is the *content* and the controlled nature of that content, not the denotations of that content. So, when all the tumult and the shouting are over and the metadata captains and kings have spoken, we are left with the absurd proposition that a 15 field subset of the MARC record, with no specification of how those fields are to be filled by non-cataloguers, is some kind of substitute for real cataloguing. The fact that metadata and the Dublin Core have been discussed *ad nauseam* for about five years with very few people pointing out this obvious flaw in the argument is reminiscent the story of the little boy and the naked Emperor. In this case, however, the Emperor keeps strolling around *sans* clothes (controlled content), at least so far.

Authority control and the content of bibliographic records

Even if one leaves aside the limited number and nature of the categories proposed for the Dublin Core and other metadata schemes, they lack the concepts of controlled

vocabularies and authority work—the means by which controlled vocabularies are implemented and maintained. Given the complex structures of bibliographic records and the need to standardize their content, it is evident that the Dublin Core cannot succeed in databases of any size. Random subject, name, title, and series denotations that are not subject to any kind of standardization —vocabulary control—will lead to progressively inchoate results as databases grow and, when a Dublin Core database is of a sufficient size, the results will be no more satisfactory than those using free text searching on the Web.

Precision and recall

All retrieval systems depend on two crucial measurements—*precision* and *recall*. In a perfectly efficient system, all records retrieved would relate exactly to the search terms (100% precision) and all relevant records would be retrieved (100% recall). To take a simple example, both measurements would be perfect if a user approached a catalogue searching for works by Oscar Wilde and found all the works by that author the library possessed and no other works. In real life, a library will possess a number of works by Oscar Wilde that are not retrieved by a search in the catalogue (poems in anthologies, essays in collections by many writers), but one might expect to find all the books, plays, collected letters and poetry collections by Wilde. One might also expect the precision to be high in that a search for Wilde in a well-organized library catalogue will yield no, or very few, materials unrelated to that author. Compare that to the results of free-text searching using search engines. Even a simple author search for someone with a relatively uncommon name will yield aberrant results. For example, for the purposes of this paper, I did a search on "Michael Gorman" on Google. It yielded "about 7710" results. Three in the first 10 (supposedly the most relevant) related to me. The other references were to a philosopher of that name in Washington, DC; a historian at Stanford; an Irish folk musician; and a consulting engineer in Denver, Colorado. The remaining 7700 entries are in no discernable order and some do not even relate to anyone called Michael Gorman. This is what results from the absence of authority control. Were each entry on Google to be catalogued, it would be assigned standardized name, title, and subject access points, so that the more than 7700 entries would appear in a rational order—each entry relating to each Michael Gorman being grouped together and differentiated from the entries relating to other Michael Gormans. In other words, the searcher would have a reasonable chance of identifying those entries relating to the Michael Gorman she or he is searching for (precision) and identifying all the entries with that characteristic (recall).

Two things are obvious. The system with authority control is clearly superior to the system without. In fact, the latter can hardly be said to be a system as the results of the search are almost completely useless. What is a searcher to do with thousands of records in no order and with no differentiation? The second obvious thing is that supplying the vocabulary and authority control necessary to make searching with the Google system would be prohibitively time-consuming and expensive. These two factors are at the core of the dilemma concerning the "cataloguing of the Web" and bringing the world of the Internet and the Web under bibliographic control. If we are to ensure precision and recall

in searches, we must have controlled vocabularies, but we cannot afford to extend that control to the vast mass of marginal, temporarily useful, and useless Web documents. What shall we do?

Solutions

First, I believe that we should either abandon the whole idea of metadata as something that will ever be of utility in large databases used by librarians and library users or we should invest metadata schemes with the attributes of traditional bibliographic records. The idea of giving up on metadata is attractive, if only because it is patently obvious that such schemes as currently practiced cannot possibly succeed except in small niche databases of specialized materials. However, the idea of enriching metadata to bring it up to the standards of cataloguing may be more psychologically and politically palatable. After all, a number of influential people and organizations are involved in, or supporting, metadata schemes and programs, and it is hard to imagine them facing the reality and declaring metadata dead as far as libraries are concerned.
The dilemma with which such organizations is faced is neatly encapsulated in the report of the library at Cornell University on their participation in the Cooperative Research Cataloging project (CORC)—one of the largest metadata projects:

Many staff members are dissatisfied with the paper-based selection form we are using now to pass along selection information to acquisitions and cataloging, and having technical services staff start from scratch with each Internet resource description is wasteful. But if selectors and reference staff begin creating preliminary records, how much would be expected of them, in terms of record content? Should students and acquisitions staff be taught to use CORC and DC to obtain preliminary records?
*Related to the issue above—we are not sure how to implement the Dublin Core element set. Should there be guidelines? Would it make sense to agree on some basic guidelines for the content of a CUL DC [*Cornell Dublin Core*] record from CORC (like the University of Minnesota library has done)? We are sure we don't want using DC to be tedious, time-consuming or complicated, so if we have guidelines, they must be simple and straightforward to teach and use.*[vi]

The fact is that the use of people without the skills and experience of cataloguers to complete metadata templates will lead, inevitably, to incoherent, unusable databases. Another indisputable reality is that real cataloguing is, equally inevitably, "time-consuming" and "complicated." The world of recorded knowledge and information is complicated, and the number of complications tends to the infinite. It is impossible to conceive of a system that allows for consistent retrieval of relevant information while lacking "guidelines" (i.e., rules dictating the nature and form of the content of the records). Eventually the fact that you cannot have high quality cataloguing on the cheap will dawn on those involved in metadata schemes but not, I fear before we have gone through a long and costly process of education and re-education and experienced the failure of databases containing records without standards and authority control.

The second solution lies in a rigorous and thorough examination of the nature of electronic documents and resources. We cannot and should not catalogue the majority of electronic documents, any more than we catalogued the millions of sub-documents and ephemera found in the print world. The problems are how to identify those electronic documents of lasting worth and, once they are catalogued, how to preserve them. These are profound and complex questions, not easily answered, but they are vital to our progress in making electronic documents and resources available to all library users.[vii]

Conclusion

Authority control is central and vital to the activities we call cataloguing. Cataloguing—the logical assembling of bibliographic data into retrievable and usable records—is the one activity that enables the library to pursue its central missions of service and free and open access to all recorded knowledge and information. We cannot have real library service without a bibliographic architecture and we cannot have that bibliographic architecture without authority control. It is as simple and as profound as that.

---

[i]. In the case of many systems giving access to electronic documents and resources, the "location" is a URL or something similar that, upon being clicked, takes the user to the document or resource itself.

[ii] . Schmierer, Helen F. "The relationship of authority control to the library catalog." *Illinois libraries* 62:599-603 (September 1980)

[iii]. Burger, Robert H. Authority work. Littleton, Colo.: Libraries Unlimited, 1985. p. 5.

[iv] . I have been advocating this developed system for more than a quarter of a century. See, for instance: Gorman, Michael. Authority files in a developed machine system. *In* What's in a name / ed. and comp. by Natsuko Y. Furuya. Toronto: University of Toronto Press, 1978. pp. 179-202.

[v] See, for example among many such: Weibel, Stuart. CORC and the Dublin Core. *OCLC newsletter*, no.239 (May/June 1999)

[vi]. CORC at Cornell; final report. http://campusgw.library.cornell.edu/corc/

[vii] . For an extended exploration of this topic see: Gorman, Michael. The enduring library. Chicago: ALA, 2003. Chapter 7.