Description in the Electronic Environment

Abstract: The significant differences that exist between the print and digital worlds are sometimes felt to diminish the need for bibliographic description in the electronic world. An analysis of these differences, especially with respect to (1) the control of production and distribution of documents and (2) the need for software intermediation, coupled with a discussion of the functions of bibliographic description in the task of document retrieval argue, however, for an increased role for bibliographic description in the electronic world.

1. Introduction

The bibliographic world has long recognized the necessary role played by description in implementing bibliographic control. For example, the entries of an analytical bibliography display a high degree of concern for descriptive detail, and even now, the only real international standard in the bibliographic control arena is one dealing with description. Description has yet to blossom, however, in the shorter history of the electronic environment; for example, search engines on the World Wide Web (e.g., Lycos, Yahoo, Excite, InfoSeek, WebCrawler, Open Text Search)—the Internet equivalents of library catalogs—give quick access to the resources of the Web, but are short on meaningful description. This development evidences a lack of understanding of the functions played by description, functions that are at least as critical to successful retrieval in the electronic environment as they are in the print environment. My intent here is to explore the need for descriptive surrogates of document-like objects in the digital world.

This exploration will proceed down three related paths. First, I will examine the role of bibliographic description in a general retrieval model for document-like objects. This will establish the default assumption that every retrieval system needs to incorporate bibliographic description. Second, I will introduce the general range of options respecting fullness of description used in retrieval systems for both print and electronic materials. This will show that the need for bibliographic description in the electronic environment is not universally accepted; indeed, the assumption evidenced by most current Internet search engines is that true bibliographic description is unnecessary. Third, I will examine some of the significant differences between print and electronic documents, exploring how they impact the need for bibliographic description, given the retrieval model previously introduced. This will demonstrate that the need for bibliographic description is heightened, not diminished, in the electronic environment.

2. The Role of Description in a General Model of Retrieval

In the most basic information retrieval model, two input streams—one representing users and their needs, the other representing knowledge and information sources—are compared so as to bring relevant texts (i.e., those that can effectively help respond to an information need, original or evolved) to the attention of the user. In order to identify these relevant texts, we form representations of the users and their needs on the one hand and of knowledge and information...
sources on the other. These representations may involve natural language terms in user statements and in documents, terms assigned from controlled vocabularies to represent the aboutness of user queries and documents, or classes of related queries or documents, etc.

We portray the ideal retrieval system as retrieving all, but only, the documents that a user would select were he or she to personally inspect all documents available. In the past we tended to envision this ideal retrieval system operating in a single-pass mode, but increasingly we have come to recognize that truly effective information retrieval is both iterative and interactive. Rather than picture a one-step retrieval process in which the retrieval system does all the work (after user needs and documents are captured by some sort of representation), we must revise our model to show retrieval to be a multi-step process. In the initial step, the retrieval system identifies a set of potentially relevant documents. In a subsequent step, the user de-selects documents that, although they match certain criteria, appear not to be useful to that particular user, at that particular time, in that particular situation, for that particular need. Alternatively, in the second step, the user identifies especially relevant documents. Various characteristics of these documents become clues used in further searching.

Where does bibliographic description fit in this picture? In answering this question, I will take the term "description" to mean something a little broader than what we might normally assume bibliographic description to refer to. I will take description to be nothing more or less than a factual representation of something, especially if given in words. A bibliographic description is then, in its broadest sense, whatever representation of document-like objects is used in a retrieval system. The term is especially relevant when the representation is verbal. Further, it is assumed that the description is a level removed from the object of interest, since representation implies a substitution of one object for another. Thus, a document-like object cannot serve as a surrogate for itself; the representation/description is something other than the thing represented/described.

Although we often make a distinction between description and access—as when we distinguish between descriptive cataloging and subject cataloging—the above presentation should make it clear that the initial retrieval stage, in which various access points (e.g., name, title, subject) are used to retrieve potentially relevant documents, makes critical use of description. This is especially true regarding the use of subject access points, which are chosen to describe salient topics, issues, or needs addressed by a document. Perhaps we hesitate to speak of access points as involving description because we associate some degree of verbosity with description, while access points are the more efficient, the more compact they are. Nevertheless, access points do represent salient features of document-like objects and therefore qualify as elements of description. All access points stand as surrogates for their respective documents, quite in keeping with their representation here as descriptive elements.

A wider array of descriptive elements usually comes into play in the filtering stages, in an effort to enable the user to distinguish relevant documents from non-relevant documents within the set of retrieved documents. The contrast between the number of descriptive elements used for access purposes and for filtering purposes appears to be a function of the disparate roles played by different elements of description. Some, especially author, title, and subject, help define the potential relevance of a document. Many others, for example, date of publication, language, length of document, intended audience, help define why a particular document will not be useful (i.e., why the user may not find it relevant). Such descriptive elements are typically used to qualify searches, to screen out that which is unlikely to be relevant. In other words, when elements such as language or date of publication are used alongside name, title, and/or subject by the retrieval system, they almost always are used negatively, to screen out documents, not positively to retrieve
Within the contexts of retrieval and filtering, three particular functions of description are important: identification, selection, and the expression of relationships. Some retrieval operations are for known items. Those parts of the description involving identification of an item should enable the user to ascertain whether the item is or is not the known item being sought (which may in reality be only partially known). Other retrieval operations are undertaken without foreknowledge of the existence of relevant documents. In such a case, the user will need to be able to discern if a specific item retrieved is indeed a relevant item; where there are multiple relevant items, the user will also want to be able to judge which of them are most relevant. Both the identification and selection functions may be supplemented by the third function of expressing relationships between the item being described and other items. From the perspective of identification, the issue may be whether the item is the appropriate edition or version of the sought-for document; from the perspective of selection, the issue may be whether the item can be used in conjunction with other material (e.g., is the concordance to an available edition?, is the file in a format readable using available tools?).

3. Levels of Description

In implementing this general model of retrieval in the print world, different levels of bibliographic description appear to be warranted for different contexts. For example, in the world of AACR2/ISBD, a school library might adopt an abbreviated first level description as its default, while an academic library might adopt a second level description as the default for its general collection, but a full third level description for its special collections.

Bibliographic description in the electronic world similarly admits of varying levels of description. OCLC's InterCat project, which is building a database of bibliographic records for Internet resources, has adopted the traditional level of bibliographic description based on AACR2/ISBD and USMARC. The TEI community has gone a step further by encouraging the inclusion of a "header" file to accompany a digital text file, which includes the descriptive information from which an AACR2/ISBD-based description can be derived. These two phenomena represent the fullest level of description of electronic documents now in vogue. A middle level of description is represented by the "Dublin Core," a fairly small set of descriptive elements—author, title, subject, publisher, date, other agent, object type, form, relation, source, language, coverage—proposed for use in describing document-like objects (DLOs) on the Internet. A third position is advocated by those who regard electronic texts as self-documenting, such that no separate description is needed; the immediate availability of the text in the digital environment is considered as obviating the need for a document surrogate. Such a position underlies the numerous Internet search engines currently in operation and represents the de facto standard position on bibliographic description in the electronic world. In this view, access to documents can be effected entirely through surface characteristics of the document (and perhaps those of other documents). Thus, the time and effort that would otherwise go into generating bibliographic descriptions can be saved.

The seeming parallel between AACR2's definition of three levels of description and the presentation here of three approaches to level of description is deceiving. In the AACR2 world, the three levels of description exist along a single continuum, varying in terms of fullness of description. In the digital world, only the first two positions vary along the fullness continuum. The third position drinks from another trough, in that it views documents (or, as is currently the case with the typical Internet search engine, document incipits) as being able to achieve the same ends as bibliographic descriptions. The only sense of representation, and hence description, that
takes place here occurs when part of the document (the beginning, or incipit, of the document) stands for the whole document. The most universal attempts to catalog the Internet are currently based on the unexamined assumption that bibliographic description (surrogation) is unnecessary for effective retrieval when full access to document-like objects is readily available.

4. Significant Differences between the Print and Electronic Environments

The unspoken assumption behind Internet search engines (and other retrieval systems for the electronic world that eschew the use of document surrogation) is that the differences inherent in the change from print to electronic environments make the use of bibliographic description unnecessary. We can only see if this assumption makes sense by examining these differences.

The first difference is the great speed advantage maintained by the electronic world over the print world in affording access to document-like objects. Ironically, in the content of our retrieval model, the use of bibliographic descriptions in the filtering process, i.e., for the making of relevance judgments, currently has a practical speed advantage over the use of direct access to documents. If, for example, one performs a search through an Internet search engine such as Lycos and compares the amount of time it takes to access a next "description" (here a document incipit) in the Lycos output to the time it takes to follow a URL to the actual site, the average difference in access time would overwhelmingly favor the use of descriptions for the making of relevance judgments. Depending on network load, there can be a significant time difference in favor of seeing the next document surrogate over accessing the next document. We should note that in time the delay inherent in making connections across the Internet is likely to become shorter; the significant difference in access times that holds currently may eventually taper off into insignificance, but then again it may not. This discussion does not, of course, eliminate the real advantage the electronic world has in retrieving documents for use. Where the overall speed advantage lies—whether with the quicker retrieval times associated with making relevance judgments based on document surrogates or with the quicker retrieval times associated with actual access to documents—will depend not only on the difference between the speeds of those two operations in the two environments, but also on the number of documents retrieved (or at least the number of documents that the user deals with) and the precision of the retrieval (since the filtering out of irrelevant documents based on bibliographic descriptions eliminates the need to access the document at all).

Another area in which the electronic environment differs from the print environment concerns the production and distribution of documents. Production and distribution of documents operate in the print environment in such a way that document variants are fairly strictly controlled. Only a limited number of production points exist, at least in comparison to the number of production points in the electronic environment. Moreover, they distribute a limited array of types of documents. This puts a cap of sorts on the number of items to be controlled in the print environment. Document-like objects in the electronic environment are, however, currently proliferating like proverbial rabbits: over the course of the first three months of 1996, the number of URLs on Lycos nearly doubled; in the process of that growth, the Lycos collection of URLs surpassed the number of bibliographic records in the OCLC Online Union Catalog. Admittedly, the prototypical bibliographic record in OCLC represents a larger document than is represented by the prototypical URL in Lycos, but in one regard it is only the absolute number of document-like units being controlled that is at issue here.

This proliferation of document-like objects reflects the fact that not only will many more distinct resources be made available electronically than in print, but also that many more versions
of at least some resources will be made available. The differences between these versions, or variants, may be harder to discern than in the print world. If, for example, one has two print documents whose title pages agree with each other (including title, statements of responsibility, edition, and publication details), one can assume that their two texts will be identical. One cannot, however, make the same naive assumption with electronic documents, since they are so easily reproduced and altered.

The differences between the print and electronic worlds along the issue of reproduction and alteration are many. To start with, documents are generally not made widely available in the print world until their content is relatively stable. Drafts of paper documents, for instance, enjoy only limited circulation; not until the document has completed some number of rounds of revision will it be made widely available. Drafts of electronic documents, in contrast, may be made universally available much earlier in the revision process. Moreover, once a document has been made available in print, the issuance of a new version or edition generally awaits the accumulation of many changes. In the electronic world, however, new versions are routinely produced for even the most trivial of changes; since it is often easier to make a change to a digital file than to keep track that a change needs to be made to it, a new document version may frequently result from the need to make but a single change. Finally, when new print versions of documents are published, they usually come into existence alongside copies of older versions. The differences between the two can be discerned through direct comparison. While this pattern can also easily apply to the electronic world, as is the case when changes are made to copies of originals, it is just as likely that document revisions will result in newer versions altogether replacing older versions, not just in supplementing them. In summary, the revision of print documents is usually associated with stability, cumulativity of changes, and supplementation; the revision of electronic documents is much more likely to take place in the context of continuous change, in small increments, replacing previous versions. The process of revision thus typically impacts the print and electronic worlds in very different ways.

Another factor that differentiates between the print and electronic environments is the need to use special tools (including both hardware and software) to access and display documents in the digital world. This situation contrasts markedly with the print world, where eye-readable documents are instead the norm for both textual and visual materials. In other words, the document-as-stored in the print world prototypically exists in the same eye-readable sphere as the document-as-displayed. In the digital world, however, the document exists in dissimilar spheres: the base sphere is one of digital encoding, where the document-as-stored is a file of some number of bytes on a digital storage medium, while the modified sphere of the document-as-displayed may exist in time and/or space. The presence of software as an intermediary between the document-as-stored and the document-as-displayed in the electronic environment has as one possible result that the document-as-displayed may not be strictly equivalent to the content of the document-as-stored, as is the case when software is used, not merely for the purposes of accessing and displaying the digital document, but also for the purposes of manipulating it. For example, the document-as-stored may simply be a data set, but the document-as-displayed may be a map or a graph or a chart that communicates the data in graphic rather than textual form. The resultant visual image would be in some sense nothing other than another version of the data set, although the value added through the visual interpretation also suggests that it is more than the data set.

5. The Need for Description in the Electronic Environment

How do these differences between the print and electronic environments impact retrieval?
The proliferation of sources, many of which are interrelated by being versions of the same works, requires finer distinctions to be made if individual sources are to be uniquely identifiable. To borrow database terminology, this can either be done by assigning an arbitrary key attribute to all document-like objects, so that each object has a unique identifier for this attribute, or by recording further descriptive data for each object, the combination of which will be unique. While the first alternative is the more efficient for the purposes of identification, it overlooks the opportunity to supply meaningful identifying information that will show relationships between objects; moreover, it is unlikely to be a feasible solution, seeing we have been only partially successful in implementing such a system (i.e., ISBN) in the more controlled print environment. The identification issue is made even less manageable by the fact that we can regularly expect some percentage of documents to which descriptions refer to disappear completely, for example, through replacement. Unless descriptions are full, it may be impossible, given a description of an item that no longer exists and faced with an existing version of a variant of the item, to determine that the description does not match the item in hand.

The selection function of description is also more difficult in the electronic environment. Given the proliferation of electronic sources, it will be more and more common for searches to yield numerically overwhelming results. Whether the retrieval system is used to improve this condition—for example, through ranked output or through using additional search clues to narrow the results—or whether the user must filter out irrelevant material, a finer level of descriptive detail is needed than if retrieval results are not numerically overwhelming. Not only is this fuller description needed for sifting through the output to determine which objects are relevant, i.e., satisfactory to the task, but it is also needed—indeed all the more so—for identifying which document-like objects are best for the task.

Finally, the relationship function of description is also adversely affected by the transition to the digital world. The proliferation of variants complicates the task of expressing relationships across works. For example, does a document at one location exactly mirror a document that purports to be the same at another location? If not, how are they related? Further, is the one document identical to the document with the same name present at that location a day/week/month ago? The kind of proliferation that causes these kinds of complications arises from the lesser control on production and distribution in the electronic world. The presence of software intermediation in the digital world brings about another kind of proliferation that also puts strain on the expression of relationships. This is the kind of proliferation that results from taking a single text and representing it variously as, for example, an ASCII text, a PostScript file, a WordPerfect document, a troff document, etc. Incompatibilities between software products and across operating system platforms may require the generation of document variants at the document-as-stored level that are equivalent at the document-as-displayed level. Again, the expression of these types of relationships will require greater descriptive detail.

6. Conclusion

If it is as important in the electronic environment as in the print environment to be able to uniquely identify document-like objects, to be able to select from among numerous possibly relevant documents those most useful for the task at hand, and to know what, if any, relationship exists between two document-like objects, then bibliographic control should play as important a role in the electronic environment as in the traditional print world. An examination of salient differences between the two worlds suggests numerous ways in which the bibliographic control of digital objects is more complex than that of print objects. Hence, the need for bibliographic
description, which is at the heart of bibliographic control, is not diminished in the electronic world; rather the need for full description of electronic objects is elevated.

Notes
1. The use of the term "document-like object" is meant to suggest that the discussion throughout is more generally applicable than in the traditional text world. Although the term "document" alone will also be used in the paper, it is not meant to be distinct in meaning from "document-like object," but is used simply for purposes of diversity and efficiency. (See note 6 for the origin of the term.)
2. In giving this description of "description," I rely heavily on the dictionary's expertise, where the first meaning given for "description" is "an act of describing" and where the first meaning given for "describe" is "to represent or give an account of in words" (Merriam-Webster's Collegiate Dictionary, 10th ed., s.v. "describe" and s.v. "description").
3. The reader may counter that such elements are indeed often used positively, for example, to retrieve documents published after a certain date or to retrieve documents in a certain language, and I will counter in turn that the user rarely wants documents after a certain date or in a certain language per se, but wants to exclude documents from before that date or written in other languages; the fact that language or date qualifiers may be added to the search with AND rather than NOT is not necessarily indicative of whether they are being used positively or negatively, where positivity refers to including something for what it is and negativity refers to excluding for what it is not.
5. The TEI guidelines can be browsed at http://etext.virginia.edu/TEI.html. Material on the header is located in part 2, chapter 5.