Misinformation and Bias in Metadata Processing: Matching in Large Databases

Gail Thornburg and
W. Michael Oskins

This article discusses structural, systems, and other types of bias that arise in matching new records to large databases. The focus is databases for bibliographic utilities, but other related database concerns will be discussed. Problems of satisfying a "match" with sufficient flexibility and rigor in an environment of imperfect data are presented, and sources of unintentional variance are discussed.

Editor's note: This article was submitted in honor of the fortieth anniversaries of LITA and *ITAL*.

Sameness is a sometime thing. Libraries and other information-intensive organizations have long faced the problem of large collections of records growing incrementally. Computerized records in a networked environment have encouraged the recognition that duplicate records pose a serious threat to efficient information retrieval.

Yet what constitutes a duplicate record may be neither exact nor completely predictable. Levels of discernment are required to permit matches on records that do not differ significantly and records that do.

Initial definitions

Matching is defined as the process by which additions to a large database are screened and compared with existing database records. Ideally, this process of matching ensures that duplicates are not added, nor erroneous replacements made of record pairs that are not really equivalent.

OCLC (Online Computer Library Center, Inc.) is a nonprofit organization serving member libraries and related institutions throughout the world. It is the chief database capital of the organization, and it is "owned" in a sense by the member libraries worldwide that use and contribute to it. At this writing, it contains over seventy-three million records. This discussion focuses chiefly on OCLC's Extended WorldCat (XWC), though many of the issues are common to other bibliographic databases. Examples of these include the Research Libraries Group's Research Libraries Information Network (RLIN) database, PICA (a European cooperative of libraries headquartered in the Netherlands), and other union catalogs. The literature will demonstrate that the problems described exist in many if not most large bibliographic databases.The database contents are representations or surrogates of the objects in shared collections. Individual records in XWC are complex bibliographic representations of physical or virtual objects—books, films, URLs, maps, slides, and much more. Each of these records consists of metadata, i.e., "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource"[1](appendix A). The records use an XML variation of the MARC communications format.[2] For example, a record for a book might typically contain such fields for author, title, publisher, and date, and many more in addition. The representation of any one object can be quite complex, containing scores of fields and subfields. Such a record may be quite brief, or several thousand characters long. The depth and richness of the records varies enormously. They may describe materials in more than 450 languages. This is a database against which millions of searches and millions of records are processed, each month.

Why is matching a challenge? Two records describing the same intellectual creation or work (e.g., Shakespeare's *Othello*) can vary by physical form and other attributes. Two records describing both the same work and exactly the same form can differ from each other if the records were created under different rules of record description (cataloging). Two records intended to describe the same object can vary unintentionally if typographical or other entry errors are present in one or both. Thus sorting out significant from insignificant differences is critical. An example of the challenges of developing matching software in the Metadata Capture Project is described elsewhere.[3]

The scope of misinformation is limited to information storage and retrieval, and specifically to comparison of incoming records to candidate matches in the database. The authors define misinformation as follows:

1. Anything that can cause two database records, i.e., representations of different items to be mistaken as representations of the same item. These can lead to inappropriate merging or updates.
2. The effect of techniques or processes of search that can obscure distinctions in differing items.
3. Any case where matching misses an appropriate match due to nonsignificant differences in two records that really represent the same item.

Note that disinformation (the intentional effort to misrepresent) is not considered in scope for this discussion. The assumption is that cooperation is in the interests of all parties contributing to a shared database. We do not assume that all institutions sharing the database have the same goals.

Gail Thornburg (thornbug@oclc.org) has taught at the University of Maryland and the University of Illinois, and served as an Adjunct Professor at Kent State University, and as a senior-level Software Engineer at OCLC. **W. Michael Oskins** (oskins@oclc.org) has worked as a Developer and Researcher at OCLC for twenty years.

What is bias? Bias can be defined as factors in the creation or processing of database records that feed on misinformation or missing information, and skew characterizations of the database records in question.

Context—Matching and bias

How are matching and bias related to each other? The growth of a database is in part a function of the matching process. If matching is not tuned correctly, the database can grow or change in nonoptimal ways.

Another way to look at the problem is to consider the goal of success in searching, and the need to know when to stop. Human beings recognize that failure to find the best information for a given problem may be costly. Finding the best information when less would suffice may also be costly. Systems need to know this. For a large shared database, hundreds of thousands of records may be processed in a day; the system must be as efficient as possible.

What are some costs? Fail to match when one should, and duplicates may proliferate in the database. Match badly, and there is risk of merging multiple records that do not represent the same item.

A system of matching can fail in more than one way. Balance is needed.

1. Searches, which are based on data in the incoming record, may be too precise to find legitimate matches. Loosen the criteria too much, and the search may return too many records to compare.
2. Once retrieved, candidate matches are evaluated. Compare candidates too narrowly, and records with insignificant differences will be rejected. Fail to take note of salient differences between incoming record and database record, and the match will be wrong, undetected, and potentially hard to detect in the future.

The goals vary in different matching projects. For some projects, setting "holdings," the indication that a member library owns a copy of something, is the main goal of the processing. This does not involve adding, replacing, or merging database records. For other projects, the goal is to update the database, either by replacing matched records, merging multiple duplicate records into one, or by adding new records if no match is found in the database. For the latter, bad matching could compromise database contents.

Background

Hickey and Rypka provide a good review of the problems of identifying duplicates and the implications for matching software.[4] Their study notes concerns from a variety of library networks including that of the University of Toronto (UTLAS), Washington Library Network (WLN), and Research Libraries Group (RLIN). They also reference studies on duplicate detection in the Illinois statewide bibliographic database and at Oak Ridge National Laboratories. Background discussion of broader misinformation issues in shared library catalogs can be found in Bade's paper.[5] A good, though dated, review of duplicate record problems can be found in the O'Neill, Rogers, and Oskins article.[6] The authors discuss their analysis of differences in records that are similar but not identical, and which elements caused failure to match two records for the same item. For example, when there was only one differing element in a pair, they found that element was most often publication date. Their study shows the difficulties for experts to determine with certainty that a bibliographic record is for the same item.

Problems of typographical errors in shared bibliographic records come under discussion by Beall and Kafadar.[7] Their study of copy cataloging errors found only 35.8 percent were corrected later by libraries, though the ordinary assumption is that copy cataloging will be updated when more information is available for an item. Pollock and Zamora report on a spelling error detection project at Chemical Abstracts Service (CAS) and characterize the types of errors they found.[8] Chemical Abstracts databases are among the most searched databases in the world. CAS is usually characterized as a set of sources with considerable depth and breadth. Of the four most common typographical errors they describe, errors of omission are most common, with insertion second, substitution third, and transposition fourth. Over 90 percent of the errors they found were single letter errors. This is in agreement with the findings of O'Neill and Aluri, though the databases were substantially different.[9] Another study on moving-image materials focuses on problems of near-equivalents in cataloging.[10] Yee suggests that cataloging practice tends to lead to making too many separate records for near equivalents. Owen Gingerich provides insight in the use of holdings information in OCLC and other bibliographic utilities such as RLIN for scholarly research in locating early editions of Copernicus' *De Revolutionibus*.[11] Among other sources, he used holdings information in multiple bibliographic utilities to help in collecting a census of copies of *De Revolutionibus*, and plotting its movements through Europe in the sixteenth century. His article highlights the importance of distinguishing very similar items for scholarly research. Shedenhelm and Burk discuss the introduction of vendor records into OCLC's WorldCat database.[12] Their results indicate that these minimal-level records increase the duplication rate within the database and can be costly to upgrade. (See further discussion in the section Change in Contributor Characteristics below.) One problem in analysis of sources of mismatch in previous studies is that there is no good way to detect and characterize typos that form real words. Jasco reviews studies characterizing types and sources of errors.[13]

Sheila Intner compares the quality issues in the databases of OCLC and the Research Libraries Group (RLG) and finds the issues similar.[14] Intner used matched samples of records from both WorldCat and RLIN to list and compare types of errors in the records. She noted that while the perception at that time was that RLIN had higher-quality cataloging, the differences found were not statistically significant.

Jeffrey Beall, while focusing in his study on the full-text online database JSTOR, notes the commonality of problems in metadata quality.[15] In addition, he discusses the special quality problems in a database of scanned images. The scanning software itself may introduce typographical errors. Like XWC, the database changes rapidly. O'Neill and Visine-Goetz present a survey of quality control issues in online databases.[16] Their sections on duplicate detection and on matching algorithms illustrate the commonalities of these problems in a variety of shared cataloging databases. They cite variation in title as the most common reason for failure to identify a duplicate record that should match. Variations in publisher, names, and pagination were noted as common. Lei Zeng presents a study of Chinese language records in the OCLC and RLIN databases.[17] Zeng discusses quality problems including (1) format errors such as field and subfield tagging and incorrect punctuation; (2) content errors such as missing fields and internal record inconsistencies; and (3) editing and inputting errors such as spacing and misspelling. Part 2 of her study presents the results of the prototype rule-based system developed to catch such errors.[18] While the author refrains from comparing the quality of OCLC and RLIN Chinese language catalog records, the discussion makes clear that the quality issues are common to a number of online databases.

More work is needed on quality and accuracy of shared records in non-Roman scripts, or in other languages transliterated to Roman script.

Types of bias to be considered

Specific factors that may tend to bias an attempt to match one record to another include:

1. Violated expectations—system software expects data it does not receive, or data received is not well formed.
2. Temporal bias—changes in rules and philosophies of record creation over time.
3. Design bias—choices in layout of the records, which favor one type of record representation at the expense of another.
4. Judgment calls—distinctions introduced in record representations due to differing but legitimate variation in expert judgment. OCLC is a multinational cooperative and there is no universal set of standards and rules for creating database records. Rules of cataloging most widely used are not absolutely prescriptive and are designed to allow local deviation to meet local needs.[19]
5. Structural bias—process and systems bias. This category reflects internal influences, inherent in the automatic processing, storage, and retrieval of large numbers of records.
6. Growth of the database environment—whether in raw numbers of records, numbers of specific formats, numbers of foreign languages, or other characteristics that may affect efficient location and comparison of records.
7. Changes in contributor characteristics—in the goals or focus of institutions that contribute to the database.

Violated Expectations

Data may not conform to expectations.

Expectations about the nature of records in the databases are frequently violated. What seem to be good rules for matching may not work well if the incoming data is not well formed, or simply not constructed as expected.

Biasing sources in the incoming data include the following:

1. Typographical errors occur in titles and other parts of the record. Anywhere the software has to parse text, an entry error—or even correction of an entry error by a later update—could confound matching. This could confound both (a) query execution and (b) candidate comparisons. Basically the system expects textual data such as the name of a title or publisher to be correct, and machine-based efforts to detect errors in data are expensive to run. Spelling detection techniques can compensate in some ways for data problems, but will not identify cases of real-word errors. See Kukich for a survey of spelling error, real-word, and context-dependent techniques.[20]
2. There is also the issue of real word differences in similar text strings. An automated system with programmed fault tolerance may wrongly equate the publisher name "Mila" with "Mela" when they are distinct publishers. Equivalence tables can cross-reference known variations on well-known publisher names, but cannot predict merges and other organizational changes. Or consider author names: are "John Smith" and "Jon Smith" the

same? This is a major problem with automated authority control where context clues may not be trustworthy.

3. Errors of formatting of variable fields in the metadata contribute to false mismatch. The rules for data entry in the MARC record are complex and have changed over time. Erroneous placement or coding of subfields poses challenges for identification of relevant data. The software must be fault tolerant wherever possible. Changes in the format of the data itself in these fields/subfields may further complicate record comparisons. ISBNs (International Standard Book Numbers) and LCCNs (Library of Congress Control Numbers) have both changed format in the recent past.
4. Errors occur in the fields that indicate format of the information. In bibliographic records, format information is used to derive the overall type of material being described: book, URL, DVD, and so on. Errors in the data in combination can generate an incorrect material type for the record.
5. Language of cataloging: this comparison has in the past caused inappropriate mismatches. The requirements in the new matching aimed to address this.
6. Language in formation of queries: MARC records frequently are a mixture of languages. As has been seen in other projects with intensive comparison of text, overlap in languages has the potential to confuse comparisons of short strings of text.[21] The assumption made here is that the use of all possible syllables contained in the title should tend to mitigate language problems. Nothing short of semantic analysis by the software is likely to solve such a problem, and contextual approaches to detection have had most success (in the production environment) in carefully controlled cases. Matching overall must be generic in its problem solving techniques.

Temporal bias

Large databases developed over time have their contents influenced by changes in standards for record creation, changes in contributor perception of the role of the database, and changes in technology to be described. Changes may include the following:

1. Description level: e.g. changes such as book or electronic book. These have evolved from format- to content-based descriptions that transcend format. Over time, the cataloging rules for describing formats have changed. Thus a format description created earlier might inadvertently "mismatch" the newer description of exactly the same item. For example, the rules for describing a book on a CD originally emphasized the CD format, whereas now, the emphasis might be shifted to focus on the intellectual content, the fact that it is a book.
2. The role of the database once perceived as chiefly repository or even backup source for a given library has become a shared resource with responsibilities to a community larger than any one library.
3. Over time, the use of the database may change. (This is further discussed in the section on Growth of the Environment later.) Searching has to satisfy the reference function of the database, but matching as a process also relies on searching, and its goals are different.
4. Varied standards worldwide challenge cooperation. While U.S. libraries usually follow AACR2 and use the MARC21 communications format, other parts of the world may use UNIMARC and country-specific cataloging rules. For instance, the PICA Bibliotekssystem, which hosts the Dutch Union Catalog, used the Prussian cataloging rules, which tended to focus on title entries.[22] The switch to the RAK was made by the early nineties.[23]
5. Some libraries may not use any form of MARC but submit a spreadsheet that is then converted to MARC. There is some potential for ambiguities in those conversions due to lack of 1:1 correspondence of parts.
6. Even within a country, standards change over time, so that "correct" cataloging in one decade may not match that in a later period. Neither is wrong, in its own temporal context, but each results in different metadata being created to describe the same item. Intner points out that OCLC's database was initiated a full decade before RLG implemented RLIN, and RLIN started almost the same time as the AACR2 publication.[24] Thus RLIN had many fewer pre-AACR2 records in its database, while Worldcat had many more preexisting records to try to match with the newer AACR2 forms.
7. Objects referenced in the database may change over time. For instance, a record describing an electronic resource may point to a location no longer valid for that resource.
8. Vendor records are created as advance advertising, but there is no guarantee the records will be updated later. Estimating the time before updates occur is impossible.
9. Records themselves change over time as they are copied, derived, and migrated into other systems. They may be enhanced or corrected in any system where they reside. So when they return to the originating database, they may have been transformed so far as to be unrecognizable as representations of the same item. This problem is not unique to XWC;

it is a challenge for any shared database where export of records and reentry is likely.

Design bias

The title, author, publisher, place of publication, and other elements of a record, designed in a time when most of the contents of a library were books, may not appear as clear or usable for other forms of information, such as Web sites or software. There is a risk to any design of a representation for an object, that it may favor distinctions in one format over another. Or representations imported from other schemes may lose distinctions in the crosswalk from one scheme to another. A crosswalk is a mechanism for the mapping of data elements/content from one metadata scheme to another. Dublin Core and MARC are just two examples of schemes used by library professionals. Software exists to convert Dublin Core metadata to MARC format, but the process of converting less complex data to a scheme of more structured data has inevitable limitations. For instance, Dublin Core has "SUBJECT" while MARC has dozens of ways to indicate subject, each with a different kind of designation for subject aspects of an item.[25] See discussion in Beall.[26] Libraries commonly exchange or purchase records from external sources to reduce the volume or costs of in-house cataloging. If an institution harvests metadata from multiple sources, there can be varying structures, content standards, and overall quality, all of which can make record comparisons error prone. While library and information science professionals have been creating metadata in the form of catalog records for a long time, the wider community of digital repositories may be outside the LIS community, and have varied understanding of the need for consistent representations of data. Robertson discusses the challenges of metadata creation outside the library community.[27] Museums and archives may take a different view of what quality standards in metadata are. For example, for a museum, extensive detail about the provenance of an object is necessary. Archives often record information at the collection level rather than the object level; for example, a box of miscellaneous papers, as opposed to a record for each of the papers within the box. Educators need to describe resources such as learning objects. A learning object is any entity, digital or nondigital, which can be used, reused, or referenced during technology-supported learning [28] For these objects a metadata record using the IEEE LOM standard may be used.[29] While this is as complex as a MARC record, it has less bibliographic description and more focus on description of the nature and use of the learning object. In short, for one type of institution the notion of appropriate granularity of description may be too detailed or too vague for the needs of another type of institution.

Judgment calls

Two persons creating independent records for the same item exercise judgment in describing what is most important about the object. One may say it is a book with an accompanying CD, another may say it is software on a CD, accompanied by a book of documentation.

Another example of legitimate variation is the choice of use of ellipses […] to leave out parts of long titles in a metadata description. One record creator may list the whole title, another may list only the first part followed by the mark of ellipsis to indicate abbreviation of the lengthy title. Either is correct, but may not match each other without special techniques. See appendix B for the perils of ellipsis handling.

The form of name of a publisher, given other occurrences of a publisher name in a record, may be abbreviated. For instance, in one place the corporate author who is also the publisher might be listed in the author field as "Department of Health and Human Services" and then abbreviated—or not—in the publisher area as "The Department."

Note that there are limitations inherent to the validation of any system of matching, in that human reviewers may not be able to determine whether two representations in fact describe the same item.

Structural bias

1. Process bias refers to any features of the software which at run-time may change the way matching is carried out, whether by shortening or lengthening the analysis, or otherwise branching the logical flow. This can arise from many sources, including but not limited to the following factors.
   a. There is need for efficient processing of large numbers of incoming records. This can force an emphasis on speedy matching. That is, matching not required to replace records tends to be optimized to stop searching/matching as early as is reasonable. In the case where unique key searching finds a single match to an incoming record, it is fairly easy for the software to "justify" stopping. If there are multiple matches found, more analysis may be needed before the decision to stop matching can be made. Over time the numbers of records processed has increased enormously.
   b. Matching needs to exploit "unique" keys to speed searching, yet these may not prove to be unique. Though agreements are in place for use of numeric keys such as ISBNs, creation of these keys is not under the control of any one organization.

c. Problems arise when brief records are compared with fuller records. Comparisons may be biased inadvertently towards false matches. Such sparseness of data has been identified as a problem in RLIN matching as well as in XWC.

d. At the same time there is bias toward less generic titles in matching. Requirements of system throughput mandate an upper limit on the size of result set that the matching software will even attempt to analyze. This upper limit could tend to discriminate against effective retrieval of generic titles. Matching will reject very large results sets of searches. So the query that has fewer title terms may tend to retrieve too much. Titles such as "Proceedings" or "Bulletin" may be difficult to match if insufficient other information is present in the record for the query to use. Ironically this can mean addition of more generic titles to the database, since what is there is in effect less findable.

e. Transparency can contribute to bias in that, for each layer of transparency a layer of opacity may be added, when information is filtered out from a user's view. That user may be a human or an application. OpenURL access to "appropriate copy" is an example from the standards world. The complexity of choosing among multiple online copies has become known as the "appropriate copy" problem. There are a number of instances where more than one legitimate copy of an electronic article may exist, such as mirroring or aggregator databases. It is essentially a problem of where and how to introduce localization into the linking process.[30] Appropriateness reflects the user's context, e.g., location, license agreements in place, cost, and other factors.

2. Systems bias. What is this, really? The database can be seen as "agent." The weight of its own mass may affect efforts to use its contents.

   a. For maintainers of large database systems, the goals of database repository and search engine may be somewhat at odds. Yet librarians do make use of the database as reference source.

   b. Search strategies for the software that acts as a user of the database is necessarily developed and optimized at a certain point in time. Yet a river of new information flows into this database.

      1. If the numbers of types of entries in various database indexes grows nonproportionally, search strategies that worked well in the past could potentially fall "out of tune" with the database contents. See Growth of the Environment section below.

      2. Change in proportions of languages in the database may render an application's use of stopword lists less effective.

      3. If changes in technology or practice result in new forms of material being described in the database, the software searches using material type as a limiter may not work properly. The software is using abstractions provided by the database, and they need to be kept synchronized.

   c. Automated query construction presents its own problems. The use of Boolean searching [term A and term B and term C] is quite restrictive in the sense that there is no "halfway" or flex for a record being included in a set of candidates. Matching starts with the most specific search to avoid too-high numbers of records retrieved, and all it can do is drop or rearrange terms from a query in the effort to broaden the results.

   d. Disconnects in metadata object creation/revision are another problem. Links can point to broken URIs (uniform resource identifiers). Controlled vocabularies can drift or expand. Even more confusing, a URI that is not broken may point to content which has changed to the point where the metadata no longer describes the item it once did. At one extreme, Bruce and Hillmann describe the curious case of citation of judicial opinions, for which a record of the opinion may be created as much as eighteen months before the volume with the official citation is printed, and thus the official citation cannot be created.[31]

   e. Expectations for creation of metadata play a role as well. Traditional cataloging has generally had an expectation that most metadata is being created once and reused. Yet current practice may be more iterative, and must be, if such problems as records with broken Internet URIs are to be avoided.

   f. Loss of synchronization can subvert processing. Note that other elements of metadata may become divorced or out of synch with the original target /purpose. The prefix to an ISBN was originally intended to describe the publisher, but is now an unreliable discriminator. Numeric keys intended to identify items uniquely can retrieve multiple items, if the scheme for assigning them is not applied consistently. In the worst case, meaningful data elements may become so corrupted as to be useless for record retrieval or even comparison of two records.

   g. Ownership issues can detract from optimal database management. Member institutions' perceptions of ownership of individual records can conflict with the goals of efficient search and retrieval. Members may resist the idea of a "better" record being merged with a "lesser" one. So systems have ways of ranking records by source or contents with the general goal of trying to avoid losing information, but with the specific effect of refraining from actions that might be enriching in a given case.

Growth of the database environment

A shared database can grow in unpredictable ways. A change in the relative proportions of different types of materials or topical coverage can render once-effective searches ineffective due to large result sets. An example of this is the number of Internet-related entries in XWC. A search such as "dog" restricted to "Internet-related" entries in 1995 retrieved thirty-four hits. This might be a manageable number. But in 2005, 225 entries were in the result set. Similarly with subject headings, one search on "computer animation" retrieved fourteen hits in 1980, and 342 in 2005. In both cases the result sets grew from manageable to "too large" over time. The increase in the number of foreign language entries in a database can cause problems. Just determining what language an entry is in can be difficult, and records may contain multiple languages. Also, such languages as Chinese, Japanese, and Korean can overlap. Chinese syllables such as: "a, an, to, no, Jan, Ka, Jun, lung, sung, I, lo, la, le, so, sun, Juan," seen out of context might be Chinese or any one of several other languages. Determining appropriate handling of stopwords and other rules for effective title matching becomes more complex as more languages populate the database.

Changes in contributor characteristics

Copy cataloging practices in an institution can affect XWC indirectly. An institution previously oriented to fixing downloaded records may adopt a policy of refraining from changing downloaded records. Historical independence of libraries is one illustration. Prior to the 1970s, most libraries did not share their cataloging with other libraries. Many institutions, especially smaller ones, were outside the loop and did things their own way. They used what rules they felt were useful, if they used any rules at all. Later they converted sparse and poorly formed data into MARC records and sent them to OCLC for matching, perhaps in an effort to get back a more complete and useful record. Yet the matching process is not always able to distinguish or interpret these local dialects. Changes in specialization of cataloging staff at an institution, or cutbacks in staff can lead to reduced facility in providing original cataloging. Outsourcing of cataloging work can affect handling of specialized materials as well. The introduction of Vendor Records and their characteristics has been noted by Shedenhelm and Burk.[32] As they note, these records are very brief bibliographic records originally designed to advertise an item for sale by the vendor. These minimal level records have a relatively high degree of duplication with existing records (37.5 percent in their study) and because of their sparseness can increase the cost of cataloging. Changes in the proportion of contributors who create records in non-MARC formats such as Dublin Core can affect the completeness of bibliographic entries. The use of such formats, meant to facilitate the entry of bibliographic materials, does come with a cost. Group cataloging is a process whereby smaller libraries can join a larger set of institutions in order to reduce costs and facilitate cataloging. This larger group then contributes to OCLC's database as an entity. The growth of group cataloging has resulted in the addition of more records from smaller libraries, which may in the future have an effect on searching/matching in XWC WorldCat overall. Internationalization may be a factor as well. The MARC format is an Anglo-based format with English-language-based documentation. Rapid international growth thrusts a broader range of traditions into a MARC/OCLC world. The role of character sets is heightened as the database grows. A Cyrillic record may not be confidently matched to a transliterated record for the same item. Although WorldCat has a long history with CJK records, MARC and WorldCat are not yet accustomed to a wide repertoire of character sets. Now, however, XWC is an environment in which expanding character coverage is possible, and likely.

Future research

- We need more systematic study of the types of errors/omissions encountered in MARC record creation.
- How can the process of matching accomodate objects that change over time?
- How does the conversion from new metadata schemes affect matching to MARC records? Does it help to know in what format a record arrived, or under what rules it was created?
- How can we address sparseness in vendor records or legal citations? How can we deal with other advance publication issues?
- How do changes in philosophy of the database affect the integrity of the matching process?

Conclusions

In this review we have seen that characterizing metadata at a high level is difficult. Challenges for adding to a large, complex database include some of the following:

- Rules for expert creation of metadata inevitably change over time.
- The object of the metadata itself may change, more often than may be convenient.
- Comparisons of briefer records to records that are more elaborate descriptions can have pitfalls. Search and comparison strategies for such record pairs are challenged by the need to have matching algorithms that work for every scenario.
- Changes within the database may themselves contribute to exacerbation of matching problems if duplicates are added too often, or records are merged that actually represent different contents. Because of the risk, policies for merging and replacing records tend to be conservative, but this does not always favor the greatest efficiency in database processing.
- Changes in the membership sharing a database are likely to affect its shape and searchability.
- Newer schemes of metadata representation are likely to challenge existing algorithms for determining matches.

References

1. National Information Standards Organization, *Understanding Metadata* (Bethesda, Md.: NISO Pr., 2004), 1. http://www.niso.org/standards/resources/Understanding Metadata.pdf (accessed Feb. 26, 2006).

2. Library of Congress, "MARC 21 Concise Format for Bibliographic Data (2002)." http://www.loc.gov/marc/bibliographic/ecbdhome.html (accessed Nov. 20, 2004).

3. Gail Thornburg, "Matching: Discrimination, Misinformation, and Sudden Death," Informing Science Conference, Flagstaff, Ariz., June 2005.

4. Thomas B. Hickey and David J. Rypka, "Automatic Detection of Duplicate Monographic Records," *Journal of Library Automation* 12, no. 2 (June 1979): 125–42.

5. David Bade, "The Creation and Persistence of Misinformation in Shared Library Catalogs," Occasional Paper No. 211, (Graduate School of Library and Information Science, University of Illinois at Urbana–Champaign, Apr. 2002).

6. Edward T. O'Neill, Sally A. Rogers, and W. Michael Oskins, "Characteristics of Duplicate Records in OCLC's Online Union Catalog," *Library Resources and Technical Services* 37, no.1 (1993): 59–71.

7. Jeffrey Beal and Karen Kafadar, "The Effectiveness of Copy Cataloging at Eliminating Typographical Errors in Shared Bibliographic Records," *Library Resources & Technical Services* 48, no. 2 (Apr. 2004): 92–101.

8. J. J. Pollock and A. Zamora, "Collection and Characterization of Spelling Errors in Scientific and Scholarly Text," *Journal of the American Society for Information Science* 34, no. 1 (1983): 51–58.

9. Edward T. O'Neill and Rao Aluri, "A Method for Correcting Typographical Errors in Subject Headings in OCLC Records," Research Report # OCLC/OPR/RR-80/3 (1980).

10. Martha M. Yee, "Manifestations and Year-Equivalents: Theory, with Special Attention to Moving-Image Materials," *Library Resources and Technical Services* 38, no. 3 (1995): 227–55.

11. Owen Gingerich, "Researching the Book Nobody Read: The *De Revolutionibus* of Nicolaus Copernicus," *The Papers of the Bibliographical Society of America* 99, no. 4 (2005): 484–504.

12. Laura D. Shedenhelm and Bartley A. Burk, "Book Vendor Records in the OCLC Database: Boon or Bane?" *Library Resources and Technical Services* 45, no. 1 (2001): 10–19.

13. Peter Jasco, "Content Evaluation of Databases," in *Annual Review of Information Science and Technology*, vol. 32 (Medford, N.J.: Information Today, Inc., for the American Society for Information Science, 1997), 231–67.

14. Sheila Intner, "Quality in Bibliographic Databases: An Analysis of Member-Controlled Cataloging of OCLC and RLIN," *Advances in Library Administration and Organization* 8 (1989): 1–24.

15. Jeffrey Beall, "Metadata and Data Quality Problems in the Digital Library," *Journal of Digital Information* 6, no. 3 (2005): 10–11.

16. Edward T. O'Neill and Diane Vizine-Goetz, "Quality Control in Online Databases," *Annual Review of Information Science and Technology* 23 (Washington, D.C.: American Society for Information Science, 1988).

17. Lei Zeng, "Quality Control of Chinese-Language Records Using a Rule-Based Data Validation System. Part 1: An Evaluation of the Quality of Chinese-Language Records in the OCLC OLUC Database," *Cataloging and Classification Quarterly* 16, no. 4 (1993): 25–66

18. Lei Zeng, "Quality Control of Chinese-Language Records Using a Rule-Based Data Validation System. Part 2: A Study of a Rule-Based Data Validation System for Online Chinese Cataloging," *Cataloging and Classification Quarterly* 18, no. 1 (1993): 3–26.

19. *Anglo-American Cataloguing Rules*, 2nd ed., 2002 rev. (Chicago: ALA, 2002).

20. Karen Kukich, "Techniques for Automatically Correcting Words in Text," *ACM Computing Surveys* 24, no. 4 (1992): 377–439.

21. Gail Thornburg, "The Syllables in the Haystack: Technical Challenges of Non-Chinese in a Wade-Giles to Pinyin Conversion," *Information Technology and Libraries* 21, no. 3 (2002): 120–26.

22. Hartmut Walravens, "Serials Cataloguing in Germany: The Historical Development," *Cataloging and Classification Quarterly* 35, no. 3/4 (2003): 541–51; *Instruktionen für die alphabetischen kataloge der preuszischen bibliotheken vom 10. mai 1899. 2 ausg. in der fassung vom 10. august 1908* (Berlin: Behrend & Co., 1909).

23. Richard Greene, e-mail message to author, Nov. 13, 2006; *Regeln für die alphabetische Katalogisierung: RAK* / Irmgard Bouvier (Wiesbaden, Germany: L. Reichert, 1980, c1977).

24. Intner, "Quality in Bibliographic Databases."

25. Richard Greene, e-mail message to author, Feb. 27, 2006.

26. Beall, "Metadata and Data Quality Problems in the Digital Library."

27. R. John Robertson, "Metadata Quality: Implications for Library and Information Science Professionals," *Library Review* 54, no. 5 (2005): 295–300.

28. IEEE. Learning Technology Standards Committee, "WG12: Learning Objects Metadata." http://ltsc.ieee.org/wg12 (accessed Feb. 26, 2006).

29. Ibid.

30. Orien Beit-Arie et al., "Linking to the Appropriate Copy: Report of a DOI-Based Prototype," *D-Lib* 7, no. 9 (Sept. 2001).

31. Thomas R. Bruce and Diane I. Hillmann,"The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice* (Chicago: ALA, 2004), 238–56.

32. Shedenhelm and Burk, "Book Vendor Records in the OCLC Database."

Appendix A. Sample CDFRecord Record from the XWC Database

```
<CDFRec db=”fs-xwc”><a>cgm 7a </a>
<c001>27681290</c001>
<c007>vf bcahru</c007>
<c007>mr baaafu</c007>
<c008>920714r19551952fr 092  mleng </c008>
<v010 i1=” “ i2=” “><sa><d> 92513007 </d></sa></v010>
<v040 i1=” “ i2=” “><sa><d>DLC</d></sa><se><d>amim</d></se>
<sc><d>DLC</d></sc></v040>
<v017 i1=” “ i2=” “><sa><d>LP5921</d></sa><sb><d>U.S. Copyright Office</d></sb></v017>
<v044 i1=” “ i2=” “><sa><d>xxu</d></sa>
<sa><d>mr</d></sa></v044>
<v050 i1=”0” i2=”0”><sa><d>VBE 6360-6361 (viewing copy)</d></sa></v050>
<v050 i1=”0” i2=”0”><sa><d>FGB 5643-5647 (ref print)</d></sa></v050>
<v050 i1=”0” i2=”0”><sa><d>FPA 0621-0625 (masterpos)</d></sa></v050>
<v130 i1=”0” i2=” “><sa><d>Othello (Motion picture : Welles)</d></sa></v130>
<v245 i1=”1” i2=”4”><sa><d>The Tragedy of Othello--the Moor of Venice /</d></sa>
<sc><d>a Mercury Production, [Films Marceau?] ; directed, produced, and written by Orson Welles.</d></sc></v245>
<v257 i1=” “ i2=” “><sa><d>U.S. ; [Morocco?]</d></sa></v257>
<v260 i1=” “ i2=” “><sa><d>France :</d></sa><sb><d>Films Marceau,</d></sb><sc><d>1952 ;</d>
</sc><sa><d>[Morocco?: :</d></sa><sb><d>s.n.,</d></sb>
<sc><d>1952?] ;</d></sc><sa><d>United States :</d></sa>
<sb><d>United Artists,</d></sb><sc><d>1955.</d></sc></v260>
<v300 i1=” “ i2=” “><sa><d>2 videocassettes of 2 (ca. 92 min.) :</d></sa><sb><d>sd., b&amp;w ;</d></sb>
<sc><d>3/4 in. viewing copy.</d></sc></v300>
<v300 i1=” “ i2=” “><sa><d>10 reels of 10 on 5 (ca. 8280 ft.) :</d></sa><sb><d>sd., b&amp;w ;</d></sb>
<sc><d>35 mm. ref print.</d></sc></v300>
<v300 i1=” “ i2=” “><sa><d>10 reels of 10 on 5 (ca. 8280 ft.) :</d></sa><sb><d>sd., b&amp;w ;</d></sb>
<sc><d>35 mm. masterpos.</d></sc></v300>
<v500 i1=” “ i2=” “><sa><d>Copyright: Orson Welles; 19Sep52; LP5921.</d></sa></v500>
<v500 i1=” “ i2=” “><sa><d>Reference sources cited below and M/B/RS preliminary cataloging card list title as
Othello.</d></sa></v500>
<v508 i1=” “ i2=” “><sa><d>Photography, Anchisi Brizzi, G.R. Aldo, George Fanto ; film editors, John Shepridge, Jean Sacha, Renzo Lucidi, William Morton ; music, Francesco Lavagnino, Alberto Barberis.</d></sa></v508>
<v511 i1=”1” i2=” “><sa><d>Orson Welles, Suzanne Cloutier, MicheaÌ l MacLiamoÌ ir, Robert Coote.</d></sa></v511>
<v500 i1=” “ i2=” “><sa><d>Director, producer, and writer credits taken from Focus on Orson Welles, p. 205.</d></sa></v500>
<v500 i1=” “ i2=” “><sa><d>LC has U.S. reissue copy.</d></sa><s5><d>DLC</d></s5></v500>
<v510 i1=”4” i2=” “><sa><d>New York times,</d></sa><sc><d>9/15/55.</d></sc></v510>
<v500 i1=” “ i2=” “><sa><d>An adaptation of the play by William Shakespeare.</d></sa></v500>
<v500 i1=” “ i2=” “><sa><d>Reference sources used: New York times, 9/15/55; International motion picture almanac, 1956,
p. 329; Focus on Orson Welles, p. 205-206; Monthly film bulletin, v. 23, no. 267, p. 44; Index de la cineÌ matographie francÌ§aise, 1952, p. 496.</d></sa></v500>
<v541 i1=” “ i2=” “><sd><d>Received: 5/26/87 from LC video lab;</d></sd><s3><d>viewing copy;</d></s3>
<sc><d>preservation, made from ref print, paperwork in ACQ: Copyright--Material Movement Form file, LWO 21635;</d></sc>
<sa><d>Copyright Collection.</d></sa></v541>
<v541 i1=” “ i2=” “><sd><d>Received: 12/2/64;</d></sd>
<s3><d>ref print;</d></s3><sc><d>copyright deposit;</d></sc>
<sa><d>Copyright Collection.</d></sa></v541>
<v541 i1=” “ i2=” “><sd><d>Received: 5/70;</d></sd>
<s3><d>masterpos;</d></s3><sc><d>gift;</d></sc>
<sa><d>AFI Theatre Collection.</d></sa></v541>
<v650 i1=” “ i2=”0”><sa><d>Othello (Fictitious character)</d></sa><sv><d>Drama.</d></sv></v650>
```

```
<v655 i1=” “ i2=”7”><sa><d>Plays.</d></sa>
<s2><d>mim</d></s2></v655>
<v655 i1=” “ i2=”7”><sa><d>Features.</d></sa>
<s2><d>mim</d></s2></v655>
<v700 i1=”1” i2=” “><sa><d>Welles, Orson,</d></sa>
<sd><d>1915-</d></sd><se><d>direction,</d></se>
<se><d>production,</d></se><se><d>writing,</
    d></se>
<se><d>cast.</d></se></v700>
<v700 i1=”1” i2=” “><sa><d>Cloutier, Suzanne,</d></
    sa><sd><d>1927-</d></sd><se><d>cast.</d></
    se></v700>
<v700 i1=”1” i2=” “><sa><d>Mac LiammoÌ ir, MicheaÌ
    l,</d></sa>
<sd><d>1899-1978,</d></sd><se><d>cast.</d></
    se></v700>
<v700 i1=”1” i2=” “><sa><d>Coote, Robert,</d></
    sa><sd><d>1909-1982,</d></sd><se><d>cast.</
    d></se></v700>
<v710 i1=”2” i2=” “><sa><d>Copyright Collection
    (Library of Congress)</d></sa><s5><d>DLC</
    d></s5></v710>
<v710 i1=”2” i2=” “><sa><d>AFI Theatre Collection
    (Library of Congress)</d></sa><s5><d>DLC</
    d></s5></v710>
<v740 i1=”0” i2=” “><sa><d>Othello.</d></sa></
    v740>
</CDFRec>
```

Appendix B. The Perils of Judging Near Matches

A. Challenges of Handling Ellipses in Titles Thought to be Similar

Incoming title: General explanation of tax legislation enacted in ... / prepared by the staff of the Joint Committee on Taxation
Match: General explanation of tax legislation enacted in the 104th Congress prepared by the staff of the Joint Committee on Taxation

Incoming title: General explanation of tax legislation enacted in ... / prepared by the staff of the Joint Committee on Taxation
Match: General explanation of tax legislation enacted in the 106th Congress prepared by the staff of the Joint Committee on Taxation

Incoming title: General explanation of tax legislation enacted in ... / prepared by the staff of the Joint Committee on Taxation
Match: General explanation of tax legislation enacted in the 107th Congress prepared by the staff of the Joint Committee on Taxation

Incoming title: General explanation of tax legislation enacted in ... / prepared by the staff of the Joint Committee on Taxation
Match: General explanation of tax legislation enacted in the 108th Congress prepared by the staff of the Joint Committee on Taxation

B. Partial Matches in Names Which Might Represent the Same Publisher

Publisher comparison is challenging in an environment where organziations are regularly merged or acquired by other organziations. There is no real authority control for publishers that would help cataloguers decide on a preferred form. When governmental organizations are added to the mix, the challenges increase. Below are some examples of non-matching text of publisher names in records, which might or might not considered the same by a human expert. (The publisher names have been normalized.)

1. Publisher name may be partially or differently recorded in two records

   Incoming publisher: konzeptstudien kantonale planungsgruppe
   Match: kantonale planungsgruppe konzeptstudien (word order different)

   Incoming publisher: institut francais proche orient
   Match: institut francais darcheologie proche orient

   Incoming publisher: u s dept of commerce national oceanic and atmospheric administration national environmental satellite data and information service
   Match: national oceanic and atmospheric administration

2. Publisher name may have changed due to acquisition by another organization

   Incoming publisher: pearson prentice hall
   Match: prentice hall

   Incoming publisher: uxl
   Match: uxl thomson gale

   Incoming publisher: thomson arco
   Match: arco thomson learning

3. One record may show "publisher" which is actually government distributing agency or clearinghouse such as the U.S. Government Printing Office or National Technical Information Service (NTIS), while the candidate match shows the actual government agency. These can be almost impossible to evaluate.

   Incoming publisher: u s congressional service
   Match: supt g p o
   (Here the distributor is the Government Printing Office, listed as the publisher)

   Incoming publisher: u s dept of commerce national oceanic and atmospheric administration national environmental satellite data and information service
   Match: national oceanic and atmospheric administration

   Incoming publisher: u s gpo
   Match: u s fish and wildlife service

4. The publisher in a record may start with or end with the publisher in the second record.
   Should it be called a match?

   Good: Incoming publisher trotta
   Match: editorial trotta
   Incoming publisher wiley
   Match: john wiley

   Questionable? Incoming publisher prentice hall
   Match: prentice hall regents canada

   Incoming publisher geuthner
   Match: orientaliste geuthner

   Incoming publisher oxford
   Match: distributed royal affairs oxford

   Incoming publisher: pan union general secretariat organization states
   Match: social science section cultural affairs pan union