Implications of MARC Tag Usage on Library Metadata Practices

Karen Smith-Yoshimura
OCLC

Catherine Argus
National Library of Australia

Timothy J. Dickey
OCLC

Chew Chiat Naun
University of Minnesota

Lisa Rowlison de Ortiz
University of California, Berkeley

Hugh Taylor
University of Cambridge

A publication of OCLC Research in support of the RLG Partnership

Implications of MARC Tag Usage on Library Metadata Practices
Smith-Yoshimura, et al., for OCLC Research and the RLG Partnership

March 2010

OCLC Research
Dublin, Ohio 43017 USA
www.oclc.org

ISBN: 1-55653-378-0 (978-1-55653-378-5)
OCLC (WorldCat): 547584991

Please direct correspondence to:
Karen Smith-Yoshimura
Program Officer
smithyok@oclc.org

Suggested citation:
Smith-Yoshimura, Karen, Catherine Argus, Timothy J. Dickey, Chew Chiat Naun, Lisa Rowlinson de Ortiz, and Hugh Taylor. 2010. *Implications of MARC Tag Usage on Library Metadata Practices*. Report produced by OCLC Research in support of the RLG Partnership. Published online at: www.oclc.org/research/publications/library/2010/2010-06.pdf.

RLG Partnership MARC Tag Usage Working Group

- Catherine Argus
  National Library of Australia
- Peter Hirsch
  New York Public Library
- Chew Chiat Naun
  University of Minnesota
- Lisa Rowlison de Ortiz
  University of California, Berkeley
- Hugh Taylor
  University of Cambridge
- Karen Smith-Yoshimura
  OCLC Research
- Timothy J. Dickey
  OCLC Research

The working group expresses its gratitude to Jenny Toves, OCLC Research, for supplying the WorldCat™ data used in its studies.

Contents

Executive Summary: Implications of MARC Tag Usage on Library Metadata Practices 8

Key Findings from our Studies 11

Implications for Library MARC Metadata Practices 13

MARC's Future? 14

1. Requirements for Enhanced Library Data Mining, *Timothy J. Dickey* 15

2. MARC Tag Usage in WorldCat, *Karen Smith-Yoshimura* 17

Caveats 18

Key to Tables 22

Observations and Implications 27

Full Data Tables Related to MARC Tag Usage in WorldCat 36

3. MARC Fields and Subfields Used in Machine Matching, *Hugh Taylor* 37

Notes Accompanying "MARC Fields/Subfields Used in Matching" Table 38

Conclusions 38

4. Comparison of Search Interfaces and Data Elements, *Catherine Argus* 48

Methodology 48

Search Interface Comparison 48

Sample Record Analysis 51

Final Comments 51

Contents (continued)

5. Encoding Level and Tag Occurrences in WorldCat, *Chew Chiat Naun* 60

Methodology 60

General Observations 61

Encoding Level Blank (Full) 63

Encoding Level 1 (Full, Material Not Examined) 63

Encoding Level 3 (Abbreviated) 63

Encoding Level 4 (Core) 64

Encoding Level 7 (Minimal) 65

Encoding Levels I, K, L, and M 65

Further Questions 67

Full Data Tables Related to Encoding Levels in WorldCat 68

6. Relator Terms and Form/Genre Designations in MARC Tagging, *Timothy J. Dickey* 69

Introduction 69

MARC Tags Studied 70

Methodology 70

Limitations 71

Results 71

Tables

Chapter 2

Table 2.1: MARC tags occurring in 20% or more of WorldCat records 19

Table 2.2: MARC tags occurring in 10% to 20% of WorldCat records 20

Table 2.3: WorldCat records and average holdings, by format..... 21

Table 2.4: Number of MARC tags representing 10% or more of records, by format 21

Table 2.5: Computer files: Tags used more heavily than in WorldCat as a whole 22

Table 2.6: Maps: Tags used more heavily than in WorldCat as a whole 23

Table 2.7: Mixed materials: Tags used more heavily than in WorldCat as a whole..... 23

Table 2.8: Scores: Tags used more heavily than in WorldCat as a whole 23

Table 2.9: Serials: Tags used more heavily than in WorldCat as a whole 24

Table 2.10: Sound recordings: Tags used more heavily than in WorldCat as a whole 24

Table 2.11: Visual materials: Tags used more heavily than in WorldCat as a whole..... 25

Table 2.12: MARC 21 fields added 2001-2008: Presence in WorldCat..... 26

Table 2.13: Fields used in WorldCat, by tag 28

Table 2.14: MARC 21 fields little or not used 32

Full Data Tables Related to MARC Tag Usage in WorldCat (Tables 2.15–2.25) 36

Tables (continued)

Chapter 3

Table 3.1: Core MARC fields and subfields used in matching across five databases 39

Table 3.2: MARC fields/subfields used in matching 40

Chapter 4

Table 4.1: Selected search options 49

Table 4.2: Limit options 50

Table 4.3: Sort options 51

Table 4.4: MARC tags used for searching and limiting 53

Table 4.5: MARC tags used for searching specific types of records 59

Chapter 5

Table 5.1: Average number of characters in summary field by encoding level 62

Table 5.2: Comparison of tag occurrences for encoding levels I, K, and L in visual materials 66

Full Data Tables Related to Encoding Levels in WorldCat (Tables 5.3–5.5) 68

Chapter 6

Table 6.1: Use of form/genre and relator terminology in NYPL and OCLC WorldCat 72

Executive Summary:
Implications of MARC Tag Usage on Library Metadata Practices

Karen Smith-Yoshimura, Lisa Rowlison de Ortiz, and Timothy J. Dickey

RLG Partners participating in discussions about metadata management hoped that evidence could be gathered that could inform more efficient and effective MARC metadata creation practices. The working assumption: analysis of existing data could identify "good enough" cataloging.

"Good enough" for what? Discovery of known or unknown items? Machine or human matching? Discovery of all manifestations of a given work? Interpreting the potential value of an item for a user's needs? Limiting or faceting search results? Delivering content? Facilitating machine processing and manipulation? Different aspects of the MARC record move into and out of focus depending upon the answer to this question.

In 2008 and 2009, the RLG Partnership MARC Tag Usage Working Group gathered and analyzed a considerable amount of evidence. We analyzed the MARC field occurrence in WorldCat™ as of September 2009, when it contained 146 million records, which parallels the research William Moen did on an earlier version of WorldCat as of May 2005, when it contained 56 million records. In his presentation for a NISO Webinar in October 2009, *Data-Driven Evidence for Core MARC Records,*[1] Moen noted that statistical analysis of field utilization based on frequency counts provide empirical evidence of catalogers' use of MARC content designation. This evidence could improve decision-making about cataloging practices and guide decisions on what constitutes a core record.

Our work also complements that reported in *Online Catalogs: What Users and Librarians Want,*[2] which looked at two categories of MARC metadata usages. Users were surveyed to discover which

1. Bibliographic Control Alphabet Soup: AACR to RDA and Evolution of MARC, October 14, 2009. Slides from the events are available at: http://www.niso.org/news/events/2009/bibcontrol09/.

2. Calhoun, Karen, and Diane Cellentani. 2009. Online catalogs: What users and librarians want: an OCLC report. Dublin, Ohio: OCLC.

data elements are most essential and what data quality enhancements would be most helpful for identifying the items they need. The results suggest trends for MARC usage. Of the top five most essential data elements identified by users, three are impacted by cataloging decisions: author, item details, links to online or full text content. Of the top five desired data quality enhancements identified by users, all impact cataloging decisions: more links to online content/full text, more subject information, added summaries/abstracts, added tables of content, and more information in the "details" tab. This corresponds to the OCLC Research synthesis of user studies on archives and special collections: "Archivists and librarians have often focused on what collections are made up of (*Ofness),* while many users prefer to learn what collections are about (*Aboutness).*"[3]

The report acknowledges that "Many of the users of library online catalogs are librarians and staff—these individuals form an important catalog user community themselves. Therefore, just as catalog end users (e.g. citizens, students, and faculty) have information needs, preferences, and expectations that need to be supported by catalog data, so do librarians who get their work done using the data underpinning the catalog" (p. 3). As such, librarians and library staff were also queried regarding desired data quality enhancements, with the top five being: merge duplicate records, add tables of content, add summaries, fix typos, upgrade brief records.

In terms of MARC, the emphasis within a catalog record would be on main entry fields (1XX), formatted contents note (505), summary, etc. (520), subject access fields (65X), and electronic location and access (856).

The RLG Partnership MARC Tag Usage Working Group's efforts focus further on a third category of MARC metadata users: machine applications. It provides a different outlook on what is "good enough cataloging" and on the nature of the standard record. Studying WorldCat MARC tag occurrences by format reveals a complex picture of MARC usage: more than half of the MARC 21 fields are used in at least 10% of a given format (or more than 1% in books, representing 1.2 million records.) Since MARC is the foundation of most library catalogs, we also looked at machine matching and indexing in several aggregate databases. Reliable bibliographic data elements are required for machine matching to present users with an overview of available works from thousands of possible manifestations as well as for a variety of internal processes.

3. Schaffner, Jennifer. 2009. The Metadata is the Interface: Better Description for Better Discovery of Archives and Special Collections : Synthesized from User Studies. Dublin, Ohio: OCLC Programs and Research. http://www.oclc.org/programs/publications/reports/2009-06.pdf. p. 6.

Each working group member did a separate analysis on one of five topics, producing the detailed reports that follow:

- Karen Smith-Yoshimura (OCLC Research) analyzed the occurrences of MARC 21 tags in the WorldCat database of 145 million bibliographic records, as of September 2009.
- Hugh Taylor (University of Cambridge) analyzed the MARC tags used for matching records while building five aggregated databases—the Research Libraries UK's union catalog for record retrieval, COPAC (the public union catalog derived from the RLUK database), WorldCat, the former RLG Union Catalog, and Libraries Australia, and compared the tags used with those mandated in the Program for Cooperative Cataloging's BIBCO and CONSER standards and OCLC Level 3 records.
- Catherine Argus (National Library of Australia) analyzed the MARC tags that are indexed in five aggregate databases: AMICUS (the national union catalog of Canada, hosted by the Library and Archives Canada), COPAC, Libraries Australia, WorldCat.org and OCLC's FirstSearch.
- Chew Chiat Naun (University of Minnesota) analyzed the MARC fields represented in WorldCat records bearing different encoding levels.
- Timothy J. Dickey (OCLC Research) collaborated with Peter Hirsch (The New York Public Library) to compare the use of form/genre designations and relator terms in the NYPL's local catalog and in WorldCat.

The occurrences of MARC tags in large aggregated databases indicate to some degree the value the creators of MARC records attach to the fields represented. However, not all systems support all the fields described in MARC 21 documentation. Even in the largest aggregate database—WorldCat—there are some tags that never occur because the system doesn't support them and they are dropped in batchloading.[4] Our study into search interfaces and machine matching highlight that for a wide range of MARC tags there is little consistency beyond a core set.

4. MARC 21 Bibliographic Data Elements not Implemented by OCLC. Available: http://oclc.org/us/en/support/documentation/worldcat/records/notimplemented/.

Key Findings From Our Studies

- Only a small subset of MARC 21 fields are used in WorldCat.

  Even when considering the MARC fields that are heavily used in non-book formats, there are only 21 to 30 tags that occur in 10% or more records.

- There is little consistency between the various routines used by machines in matching based upon MARC data elements.

  The fields shared by machine standards for matching of records sampled in this study are limited to five elements from the leader; four MARC fields—fixed length data elements (008—selected bytes only), Library of Congress Control Number (010), International Standard Book Number (020), and International Standard Serial Number (022); and a sampling of other "core" bibliographic data—main entry fields (1XX), title statement and varying form of title (245, 246), edition statement (250), and imprint statement (260).

  The "good news" is that all of these fields used for matching tend to be required on input, ensuring that the matching fields will contain metadata.

- Although machine-matching systems generally use a core of fields and subfields, there is clear evidence that some services need to go beyond that core to verify the accuracy of the match.

  The complexity of the algorithms using MARC data for matching cannot be overestimated. This includes routines operating to produce different levels of matching confidence. It also means that each system will potentially "use" a large number of discrete fields and subfields at some point in their algorithm, and the total array of fields potentially used by matching systems in the aggregate is quite large.

- Only a subset of fields is indexed by common library search systems.

  There is a mismatch between MARC tags' potential use by end-users, and their assignment of content in actual records. A study of several common search interfaces showed remarkably few MARC fields are indexed by all of them. In addition, numerous fields are strongly associated with, and thus potentially very useful for searching within, a specific format, but they are not indexed by a majority of library systems in the present study.

- Note fields are in common use, but machines are not necessarily good at interpreting the free text which generally characterizes their contents.

  The general note (500) field is among the most common "optional" MARC data elements. However, machine interfaces are not necessarily good at interpreting free text, in contrast to (for instance) name fields, where authority control may be more often followed.

  This lack of functionality is reflected throughout the present study: relatively few note (5XX) fields were used by *any* system in the present study for machine matching, and they were absent from a majority of search interface indexes.

- There is not at present much variation in indexing and display among different formats in different aggregate databases.

  Many common library search interfaces are currently undergoing redesign processes, though there appear to be no great differences among them, either in basic display interfaces, or in MARC field indexing.

- The availability of limits provides the opportunity for offering facets in search result displays.

  Limiting functions currently available in common search interfaces tend to rely on similar combinations of MARC fields. As an offshoot to the commonalities among the systems, all library search interfaces have the potential to leverage common MARC fields for enabling common faceted browsing tasks and to provide guidance on important metadata for future search systems.

- Encoding level as a criterion for selecting the "most complete" record is far from reliable.

  Encoding level encapsulates multiple aspects of record quality in a single byte, and library systems have used it as a ranking criterion for selecting among records representing the same item or to parse work among cataloging staff. However, record content is determined more by method of input than by ostensible quality as defined by encoding level. Encoding level that is assigned at a batchload or project level is particularly suspect.

- Uses of specific MARC tags can be significantly different in a specific local catalog than as represented in an aggregate database like WorldCat.

  The Dickey-Hirsch study illustrates that specialized data—relator terms and form/genre designations—are far more consistently applied by the New York Public Library's Performing Arts Library than by the profession at large as reflected in WorldCat.

- Search log data currently captured by library systems usually cannot provide enough information on user behavior.

  To take one instance, although we can tell whether a value is "searchable" in any given database/system, we often lack information about how often they are searched. System logs are not sophisticated enough, and the indexing used in systems is not well enough documented; this situation makes it impossible to determine in many common systems which fields users are actually retrieving information from, and whether the results satisfy their query. The working group would have benefitted if systems provided search logs and circulation data that met Timothy J. Dickey's "Requirements for Enhanced Library Data Mining" (p. 15).

Implications for Library MARC Metadata Practices

Based on our research, the working group offers the following factors to consider when making decisions about your own MARC metadata practices:

- Strive for consistency in the choice and application of a field. Splitting content across multiple fields will negatively affect indexing, retrieval, and mapping to other encoding schemas.
- Respond to local user needs. Would your users rather that you spend time counting the number of plates in a book, or linking to the table of contents or full text?
- The number of full-text documents available on the Web will substantially increase over the next few years, and the need for surrogate "descriptive metadata" will decrease. Focus instead on the authorized names, classifications, and controlled vocabularies that key word searching of full-text will not provide.
- Use the appropriate fields to reflect the resource. Use specific MARC fields for particular types of note if they are available rather than the general 500 note.
- MARC data cannot continue to exist in its own discrete environment, separate from the rest of the information universe. It will need to be leveraged and used in other domains to reach users in their own networked environments. The 200 or so MARC 21 fields in use must be mapped to simpler schema.
- MARC data is used for far more than user retrieval and identification: machine matching, linking, machine manipulations, harvesting, collection analysis, ranking, systematic views of publications. Accuracy of fields that are used in machine matching becomes more important

in environments using linked data to leverage fuller descriptions and other related information generated from other sources.

MARC's Future?

Libraries rely on MARC data for library inventory control, but users do their discovery elsewhere.[5] Delivery of the inventory from the library will likely be mitigated by the availability of digitized works, especially for those in the public domain. The RLG PartnersHIP MARC Tag Usage Working Group's view on MARC's future:

- MARC is a niche data communication format approaching the end of its life cycle.
- Future systems, if they are to be able to meet users' needs in the ways documented in the *Functional Requirements for Bibliographic Records*[6] and to take advantage of linked data as envisioned by the new Resource Description and Access standard, will need a more relational approach to data storage. MARC is not the solution.
- Future encoding schemas will need to have a robust MARC crosswalk to ingest the millions of legacy records we now have.
- Ask ourselves: How would we create, capture, structure, store, search, retrieve, and display objects and metadata if we didn't have to use MARC and if we weren't limited by MARC-centric library systems?
- Consider how best to take advantage of linked data and avoid creating the same redundant metadata in individual records. Consider sources outside the traditional library environment.
- Rather than enhancing MARC and MARC-based systems, let's give priority to interoperability with other encoding schemas and systems. We need to meet the demands that have arisen from the rest of the information universe.

5. De Rosa, Cathy. 2005. Perceptions of libraries and information resources: a report to the OCLC membership. Dublin, Ohio: OCLC Online Computer Library Center. Eighty-nine percent of college students begin their search for information on a particular topic on a search engine; only 2% begin with a library catalog (pp. 1–17).

6. IFLA Study Group on the Functional Requirements for Bibliographic Records. 1998. Functional requirements for bibliographic records: final report. UBCIM publications, new ser., v. 19. München: K.G. Saur.

1. Requirements for Enhanced Library Data Mining

Timothy J. Dickey

Generally speaking, system logs are not particularly sophisticated, and the indexing used in systems is not well enough documented. This situation makes it impossible to determine in many common systems which fields users are actually retrieving information from, and whether the results satisfy their query.

The lack of adequate search log data from library systems could arise from at least three discrete limitations:

- Library OPAC search data and circulation transaction data may be collected by a wide variety of disparate systems, in different data formats, even among the different ILS modules used by a single library.
- Libraries traditionally focus on bibliographic data, and ignore the "business intelligence" more commonly collected by commercial providers of online search or e-commerce. In addition, libraries—as non-profit organizations—usually do not have the resources to collect, preserve, and mine large amounts of transactional data.
- Libraries value the profession's ethical standards of patron privacy, and many choose not to collect user data which might be perceived as violating a patron's rights.

However, with library catalogs often in direct competition with commercial search engines and e-commerce sites, user data—without compromising user privacy—can offer library systems invaluable assistance in meeting users' needs.

- Library search interfaces could be made more responsive to user behaviors, with transactional evidence about how users search within them, and how they evaluate search success.

- The same interfaces could provide more "value-added" content in terms of recommender services (a feature commonly desired in library systems, but infrequently realized to users' satisfaction).
- Library metadata practice could be improved by understanding which fields, and which types of data within a field, are the most helpful to users.

Such enhanced data mining could be possible with the following requirements:

- More complete transaction logs captured from library search interfaces, including
    - Anonymous session ID, for linking transactions
    - Referring IP address (even noting internal/ external)
        - Distinction between machine and human search requests
    - Search string(s) and index(s)
    - Qualifiers for use of facets and internal links
    - Indication of results retrieved
    - Path within site (view results, limit search, click see-also reference, move to request item, transition to DRM module for download, abandon search, etc.)
- Circulation data with greater commonality of data format, including
    - Transaction-level data that is anonymous but linked:
        - Distinction between internal (processing, reserves) and patron circulation
        - Coordination of data regarding items circulated together
        - Coordination of circulation data to user actions in OPAC module
    - Aggregate data on holds placed for copies of an item
    - Aggregate data on length of circulation

2. MARC Tag Usage in WorldCat

Karen Smith-Yoshimura

WorldCat offers a unique perspective on which MARC tags are used by practitioners. The following tables are derived from a snapshot of WorldCat in September 2009, when WorldCat contained 145.7 million bibliographic records and 1.5 billion holdings.

Only the MARC tags defined in MARC 21 documentation are listed—the MARC tags you would find in other library catalogs. Local fields are also excluded. Each MARC field and subfield exists because some constituency at some time advocated a need for it. The analysis looked at both the occurrences of a field tag and the occurrence of the field tag when weighted by the number of holdings for the record with that tag. The set of tables that accompanies this report can be used to compare the tag occurrences within local catalogs. The full set of statistics, along with occurrences of MARC subfields, from the September 2009 snapshot is available at: http://outgoing.typepad.com/outgoing/files/bibstats.200909.xls

We can infer how widely a MARC tag is used overall by this weighted occurrence. For example, 6% of WorldCat records had alternate graphic representations (880), indicating the presence of non-Latin scripts currently supported in WorldCat—Arabic, Bengali, Chinese, Cyrillic, Devanagari, Greek, Hebrew, Japanese, Korean, Tamil, and Thai. When weighted by holdings, however, alternate graphic representations represent less than 2% of the total, indicating that generally local catalogs have less representation of non-Latin scripts. This difference—the percentage of holdings with the field divided by the percentage of records with the field—is represented by a "gap" of 0.3 in the table. On the other hand, Library of Congress control numbers (010) appear in 11% of WorldCat records, but in 60% of the holdings. The gap of 5.7 would indicate that many of the contributing libraries to WorldCat are likely to use Library of Congress control numbers to a greater extent than could be inferred from just the tag's occurrence in WorldCat master records.

Although the number of MARC tag occurrences in WorldCat changes daily, the percentages are less likely to change significantly from one year to the next.

In his presentation for a NISO Webinar in October 2009, *Data-Driven Evidence for Core MARC Records,*[1] William Moen noted that statistical analysis of field utilization based on frequency counts provide empirical evidence of catalogers' use of MARC content designation. This evidence could improve decision-making about cataloging practices and guide decisions on what constitutes a core record. Moen did his study of a WorldCat database that contained 56 million records as of May 2005, a size just 38% of its size when this new analysis was undertaken.

Caveats

- The presence of a field does not necessarily tell us anything about the utility of the data within the field for retrieval, matching, or intellectual use.
- Occurrences of MARC tags depend also on system requirements. In WorldCat, some fields are system supplied (such as control numbers) and others may be generated only on export (such as system control numbers, control number identifiers, and date and time of transaction).
- Not all systems support all MARC tags. This is true even for WorldCat.[2] Some fields that occur in local systems are dropped during batchloading if the field is not supported. We cannot tell which fields have low occurrences because of the local system constraints in which the record was created.
- Tag occurrences depend on the standards that are followed. For example, the Program for Cooperative Cataloging's suite of CONSER and BIBCO[3] standard record guidelines mandate certain fields, taking format into account.
- The content of a field may be split among different tags. For example, the series statement/added entry—title (440) is now obsolete and descriptive series content is now entered only in the series statement (490). Both tags currently occur in WorldCat; if combined their aggregate would be among the top 10 tags occurring in WorldCat with 25% of all occurrences. Similarly, correct application of library cataloging standards means that a

1. Bibliographic Control Alphabet Soup: AACR to RDA and Evolution of MARC, October 14, 2009. Slides from the events including Moen's presentation are available at: http://www.niso.org/news/events/2009/bibcontrol09/.

2. See MARC 21 Bibliographic Data Elements not Implemented by OCLC at http://www.oclc.org/us/en/support/documentation/worldcat/records/notimplemented/. Accessed December 11, 2009.

3. The suite of BIBCO core record standards are available at http://www.loc.gov/catdir/pcc/bibco/coreintro.html; documentation of CONSER standard records are available at http://www.loc.gov/acq/conser/issues.html.

meeting name could be provided as a main entry (111) or added entry (711). Separately their occurrence falls within the "little used" MARC tags; combined, they occur in more than two million records and would be part of the list of most used tags.

- Not all the records represented by "weighted by holdings" may include the specific field. Some local records may lack fields that are in the WorldCat Master Record, and some will have fields that are absent in the WorldCat Master Record.

- Tag occurrences provide only a context. Some tags are applicable only to certain types of materials. A low occurrence may simply mean that the tag was not applicable.

In WorldCat, all records must have a control number, fixed-length data elements, cataloging source, and title statement (001, 008, 040, 245); the 001 and 040 are system-supplied. Of the 200 MARC tags analyzed, only 11 tags occur in 20% or more of WorldCat records (listed in descending order of occurrences).

Table 2.1: MARC tags occurring in 20% or more of WorldCat records

MARC Tag	Description	Occurrence
001	Control number	100%
008	Fixed-length data elements	100%
040	Cataloging source	100%
245	Title statement	100%
260	Imprint statement	96%
300	Physical description	91%
100	Main entry – personal name	61%
650	Subject added entry - topical term	46%
500	General note	44%
700	Added entry – personal name	28%
020	International Standard Book Number	23%

Another 11 tags occur in 10% or more of WorldCat records, but less than 20%.

Table 2.2: MARC tags occurring in 10% to 20% of WorldCat records

MARC Tag	Description	Occurrence
050	Library of Congress call number	20%
043	Geographic area code	19%
710	Added entry – corporate name	19%
504	Bibliography, etc. note	16%
082	Dewey Decimal classification number	14%
440	Series statement/Added entry - title	14%
250	Edition statement	13%
007	Physical description fixed field	12%
490	Series statement	12%
010	Library of Congress control number	11%
016	National bibliographic agency control number	10%

The "Overview of MARC Tag Usage in WorldCat" table[4] lists the 69 tags that occur in 1% or more of WorldCat records—1% representing over one million records. Another eleven tags occur in more than 1% of the holdings represented in WorldCat (15 million records) but in less than 1% of the WorldCat occurrences.

As the "Grid of MARC Tag Usage in WorldCat by Format" table shows,[5] some fields are much more heavily used in specific formats than in WorldCat as a whole. Separate tables show the MARC tags used in 10% or more WorldCat records in each format, except for books, which is by far the largest segment of WorldCat with 123 million records, where the tags that occur in 1% or more WorldCat records are shown. The average number of holdings per format also varies.

[4] Table 2.15, available online at http://www.oclc.org/research/publications/library/2010/2010-06a.xls (Microsoft Excel format), sheet (tab): All Formats, 1% or more.
[5] Table 2.16, available online at http://www.oclc.org/research/publications/library/2010/2010-06a.xls (Microsoft Excel format), sheet (tab): Grid All WC Fields.

Table 2.3: WorldCat records and average holdings, by format

Format	Records	Average Holdings
Books	123,032,672	10.80
Serials	5,871,037	8.06
Sound recordings	4,749,945	7.54
Visual resources	4,570,531	6.58
Scores	3,104,902	4.65
Maps	2,289,888	3.29
Computer files	1,068,793	3.37
Mixed materials	947,922	0.86
Integrating resources	22,948	9.16

The count of tags representing 10% or more of records is fairly consistent across formats.

Table 2.4: Number of MARC tags representing 10% or more of records, by format

Format	Number of Tags
Books	21
Computer files	30
Integrated resources	29
Maps	28
Mixed materials	28
Scores	21
Serials	28
Sound recordings	27
Visual resources	29

Although Moen's study showed that there were 17 fields in books, pamphlets, and printed sheets that accounted for 80% of occurrences in WorldCat in 2005, not including system-supplied fields, in our current analysis there are just four: fixed-length data elements, title, imprint statement, and physical description (008, 245, 260, 300).

Some tags are specific to characteristics of a given format.

- *Computer file characteristics* (256) occurs in 11% of computer file records but represents 40% of all occurrences of that field in WorldCat.

- *Number of musical instruments or voices codes* (048) occurs in 28% of scores records (51% of scores holdings) but represents 79% of all occurrences of that field in WorldCat. Sound recordings represent 21% of the remaining occurrences of the 048 field.
- *Creation/production credits note* (508) occurs in 24% of visual materials records (59% of visual materials holdings) but represents 89% of all occurrences of that field in WorldCat.

These are the MARC tags that are much more heavily used in non-book formats compared to the tag's occurrences in WorldCat as a whole (sorted by the number of the tag's occurrences).

Key to Tables

Percent: Number of records in that format where the tag occurs at least once as a percentage of the total number of records in that format.

Weighted %: Number of records in the format where the tag occurs at least once multiplied by the number of holdings attached to the records as a percentage of the total number of holdings in that format.

WC %: Number of records in WorldCat where the tag occurs at least once as a percentage of the total number of WorldCat records.

WC Gap: Percentage of records in the format containing the tag divided by the percentage of records with the tag in all of WorldCat. Numbers below 1.0 indicate that the tag occurs less frequently in a particular format than in WorldCat as a whole.

[Format] rep: Number of records in that format where the tag occurs as a percentage of the total number of WorldCat records where the tag occurs.

Table 2.5: Computer files: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	Weighted %	WC %	WC Gap	Files Rep
538	System details note	44.11	48.10	2.27	19.4	14.3%
516	Type of computer file or data note	15.99	19.16	0.46	34.8	25.6%
256	Computer file characteristics	11.30	10.53	0.21	53.8	40.3%

Conversely, main entry—personal name (100), Library of Congress call number (050), and geographic area code (043) occur only half as frequently compared to WorldCat as a whole.

Table 2.6: Maps: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	Weighted %	WC %	WC Gap	Maps Rep
255	Cartographic mathematical data	75.85	81.49	1.22	62.2	97.6%
034	Coded cartographic mathematical data	54.40	72.89	0.86	63.4	99.4%
052	Geographic classification	35.70	62.93	0.63	56.7	88.4%

Conversely, main entry—personal name (100), International Standard Book Number (020), Dewey Decimal classification number (082), and Bibliography, etc. note (504) occur only half or less as frequently compared to WorldCat as a whole.

Table 2.7: Mixed materials: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	WC %	WC Gap	Mixed Rep
545	Biographical or historical data	33.87	0.38	89.1	57.5%
555	Cumulative index/finding aids note	28.56	0.30	95.2	61.3%
541	Immediate source of acquisition note	19.25	0.49	39.3	25.7%
351	Organization and arrangement of materials	14.82	0.14	105.9	69.1%
524	Preferred citation of described materials note	14.13	0.15	94.2	60.6%
583	Action note	13.13	0.26	50.5	33.4%
561	Ownership and custodial history	10.17	0.37	27.5	17.8%

Since mixed materials are by nature describing unique materials, they are not weighted by holdings. Not surprisingly, imprint statement (260) occurs only 10% as frequently as in WorldCat as a whole.

Table 2.8: Scores: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	Weighted %	WC %	WC Gap	Scores Rep
240	Uniform title	44.08	58.46	3.75	11.8	25.1%
028	Publisher number	39.62	54.38	3.37	11.8	25.1%
048	Number of musical instruments or voices codes	27.60	50.90	0.75	36.8	78.5%
045	Time period of content	12.54	31.62	1.07	11.7	25.0%

Not surprisingly, International Standard Book Number (020) occurs only 30% as frequently as in WorldCat as a whole.

Table 2.9: Serials: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	Weighted %	WC %	WC Gap	Serials Rep
362	Dates of publication and/or sequential designation	64.06	84.48	2.64	24.3	97.8%
310	Current publication frequency	41.59	70.47	1.71	24.3	97.9%
022	International Standard Serial Number	21.25	59.21	0.96	22.1	89.1%
042	Authentication code	20.31	64.45	3.53	5.8	23.2%
780	Preceding entry	19.72	34.35	0.82	24.0	97.4%
785	Succeeding entry	17.67	25.18	0.72	24.5	98.3%
222	Key title	12.40	40.43	0.51	24.3	98.9%
130	Main entry – uniform title	12.11	23.60	1.21	10.0	40.3%
850	Holdings information	11.58	47.00	0.56	20.7	83.7%
515	Numbering peculiarities note	9.97	18.97	0.43	23.2	93.2%
210	Abbreviated title	7.64	32.12	0.31	24.6	98.1%
580	Linking entry complexity note	7.24	15.21	0.82	8.8	35.6%
550	Issuing body note	6.43	16.72	0.37	17.4	70.8%
321	Former publication frequency	3.03	14.44	0.12	25.3	98.7%
030	CODEN designation	1.25	10.79	0.10	12.5	48.3%

Table 2.10: Sound recordings: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	Weighted %	WC %	WC Gap	Recordings Rep
028	Publisher number	62.27	76.32	3.37	18.5	60.3%
511	Participant or performer note	59.43	83.69	2.76	21.5	70.2%
518	Date/time and place of event note	19.10	27.65	0.93	20.5	66.9%
306	Playing time	8.55	23.58	0.34	25.1	82.1%
033	Date/time and place of an event	6.82	14.39	0.56	12.2	39.5%
048	Number of musical instruments or voices codes	4.84	11.00	0.75	6.5	21.0%

Conversely, International Standard Book Number (020), subject added entry—geographic name (651), and edition statement (250) occur only half or less as frequently compared to WorldCat as a whole.

Table 2.11: Visual materials: Tags used more heavily than in WorldCat as a whole

Tag	Description	Percent	Weighted %	WC %	WC Gap	Visual Rep
538	System details note	29.55	78.23	2.27	13.0	40.8%
655	Index term - genre/form	27.93	41.18	4.27	6.5	20.5%
511	Participant or performer note	25.06	63.51	2.76	9.1	28.5%
508	Creation/production credits note	23.61	59.15	0.84	28.1	88.5%
540	Terms governing use and reproduction note	13.27	3.51	0.64	20.7	65.3%
521	Target audience note	8.44	36.83	0.57	14.8	46.8%

Please note: The full lists of MARC tags that occur in 10% or more in WorldCat by non-book format (and 1% or more in books) are in the full data tables.[6]

Some fields are not necessarily specific to a particular format, but we see practitioners using that field predominantly in only one or two formats. For example, the subject added entry—personal name (600) tag occurs less than 10% in most formats, but in 40% of mixed materials records. Other examples include:

- Target audience note (521) occurs in 0.57% of all WorldCat records, but in 8% of computer file and visual material records. When weighted by holdings, the percentages increase to 13% and 37% respectively.
- System details note (538) occurs in 2% of all WorldCat records, but in 44% of computer file records and in 30% of visual material records. When weighted by holdings, the percentages increase to 48% and 78% respectively.
- Index term—genre/form (655) occurs in 4% of all WorldCat records, but in 53% of mixed material records and 28% of visual material records. When weighted by holdings, the percentage for visual material records increases to 41%.
- Added entry—uncontrolled related/analytical title (740) occurs in 4% of all WorldCat records, but in 9% of maps records and 12% of recordings records.
- Additional physical form entry (776) occurs in 2% of all WorldCat records, but in 22% of computer file records and in 7% of serials records. When weighted by holdings, the percentages increase to 11% and 32% respectively.

[6] The full data tables related to MARC tag usage in WorldCat are available online at http://www.oclc.org/research/publications/library/2010/2010-06a.xls (Microsoft Excel format).

We used a conservative benchmark that a 10% occurrence in either bibliographic records or representation in a format's holdings represents "usage"—or a 1% occurrence in records or holdings for books or WorldCat overall. That resulted in a total of 102 tags. Table 2.13 (p. 28) lists these fields, in tag order, with the number of occurrences in WorldCat, the percentages, the representations weighted by holdings, and the gaps represented by the differences in percentages of occurrences and holdings.

These tags can be considered of value by the catalogers who added them or the systems that generated them. We have general guidance from user studies reporting that users value information about what collections are about, authors, abstracts, summaries, and links to online or full text content. These would map to main entry fields (1XX), formatted contents note (505), summary, etc. (520), subject access fields (65X), and electronic location and access (856). We have no statistical evidence to what extent the contents of specific MARC fields lead to better user discovery or item identification.

That leaves 86 tags that are little used, or not used at all, as listed in the "MARC 21 fields little or not used" table (Table 2.14, p. 32). Of these infrequently occurring fields, 16 are fields that were introduced between 2001 and 2008. Three of these fields (highlighted in orange) have no occurrences in WorldCat since OCLC has no plans to implement them.

Table 2.12: MARC 21 fields added 2001-2008: Presence in WorldCat

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Year Added
026	Fingerprint identifier	25,515	0.02	26,392	0	2002
031	Musical incipits information	100,581	0.07	295,761	0.02	2004
038	Record content licensor	0	0	0	0	2002
083	Additional Dewey Decimal Classification number	7	0	380	0	2008
085	Synthesized classification number components	2	0	2	0	2008
258	Philatelic issue data	2	0	191	0	2004
363	Normalized date and sequential designation	2	0	6	0	2007
365	Trade price	0	0	0	0	2003
366	Trade availability information	0	0	0	0	2003
542	Information relating to copyright status	536	0	565	0	2008
563	Binding information	74,918	0.05	100,623	0.01	2002
648	Subject added entry - chronological term	103,559	0.07	241,362	0.02	2002
662	Subject added entry - hierarchical place name	728	0	753	0	2005
751	Added entry - geographic name	4	0	4	0	2007
882	Replacement record information	0	0	0	0	2007
887	Non-MARC information field	23	0	33	0	2001

Observations and Implications

1. Given that only just over half of MARC 21 fields have been supplied in more than 1% of WorldCat records (or 10% in a non-book format), there are few justifications to continue adding more *new* MARC fields.

   New fields were added to the MARC 21 format in 2009 to accommodate the new Resource Description and Access standard. We will likely continue to have MARC around at least as a data communication format for some years. But given the need to expose our data to where our users are—who search first using search engines rather than library catalogs—the focus should be on new data structures that can accommodate linked data from other sources.

2. Library practices are often constrained by what their own system supports. If a MARC tag cannot be used locally, then its value in an aggregated, networked system is diminished. Conversely, if a MARC tag used in local systems is not supported in an aggregated, networked system, its value beyond the local scope disappears.

3. MARC fields that focus on description of a textual document will become unnecessary as more titles become available in full-text on the Internet as a result of mass digitization efforts. Users can click to see if the document is what is wanted rather than relying on the object's metadata.

4. MARC was designed to provide a machine version of the traditional catalog card which used to be the only metadata that could lead users to the physical item within a library. A degree of redundancy was inherited, based on where the entry would have appeared at the top of a card. This redundancy imposes maintenance overhead, makes mappings to other, simpler schemas more difficult, and hampers taking advantage of metadata from other sources.

5. MARC itself is arguably too ambiguous and insufficiently structured to facilitate machine processing and manipulation.

6. MARC fields designed to facilitate local inventory management may be better served by using other metadata structures for such management.

7. With more text being indexed by search engines, focus should be on the authorized names, classifications, and controlled vocabularies that key word searching of full-text will not provide.

8. We need to consider the future of our metadata content outside of MARC. If several MARC fields will in due course be mapped to a single element in another, simpler schema, how many of the MARC tags we're using now are really needed?

Table 2.13: Fields used in WorldCat, by tag

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Gap
001	Control number	145,658,639	100.00	1,468,380,998	100.00	1.0
006	Fixed-length data elements—additional material characteristics	5,224,974	3.59	80,598,278	5.49	1.5
007	Physical description fixed field	17,661,122	12.13	180,449,919	12.29	1.0
008	Fixed-length data elements	145,658,639	100.00	1,468,380,998	100.00	1.0
010	Library of Congress control number	15,403,625	10.58	878,889,963	59.85	5.7
015	National bibliography number	7,979,794	5.48	252,612,486	17.20	3.1
016	National bibliographic agency control number	15,163,184	10.41	160,821,172	10.95	1.1
017	Copyright or legal deposit number	2,159,168	1.48	2,327,306	0.16	0.1
020	International Standard Book Number	33,383,073	22.92	870,140,517	59.26	2.6
022	International Standard Serial Number	1,400,621	0.96	28,345,554	1.93	2.0
024	Other standard identifier	7,414,990	5.09	38,039,004	2.59	0.5
028	Publisher number	4,903,433	3.37	51,431,784	3.50	1.0
030	CODEN designation	151,414	0.10	5,205,216	0.35	3.5
033	Date/time and place of an event	820,140	0.56	6,859,019	0.47	0.8
034	Coded cartographic mathematical data	1,255,642	0.86	5,532,352	0.38	0.4
037	Source of acquisition	2,745,549	1.88	64,490,585	4.39	2.3
040	Cataloging source	145,658,639	100.00	1,468,380,998	100.00	1.0
041	Language code	12,385,529	8.50	105,005,943	7.15	0.8
042	Authentication code	5,138,438	3.53	235,238,453	16.02	4.5
043	Geographic area code	28,050,302	19.26	564,326,681	38.43	2.0
044	Country of publishing/ producing entity code	1,873,423	1.29	3,783,733	0.26	0.2
045	Time period of content	1,559,210	1.07	15,777,151	1.07	1.0
048	Number of musical instruments or voices codes	1,091,977	0.75	11,301,078	0.77	1.0
050	Library of Congress call number	28,868,349	19.82	1,155,103,134	78.67	4.0
052	Geographic classification	924,529	0.63	4,996,641	0.34	0.5
055	Classification numbers assigned in Canada	2,257,083	1.55	19,771,152	1.35	0.9
060	National Library of Medicine call number	1,318,668	0.91	65,848,454	4.48	4.9
070	National Agricultural Library call number	792,357	0.54	33,468,607	2.28	4.2
072	Subject category code	1,651,701	1.13	39,798,167	2.71	2.4

Table 2.13: Fields used in WorldCat, by tag (continued)

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Gap
074	GPO item number	730,040	0.50	51,180,927	3.49	7.0
080	Universal decimal classification number	4,095,727	2.81	7,042,615	0.48	0.2
082	Dewey Decimal classification number	20,953,255	14.39	1,007,384,646	68.61	4.8
084	Other classification number	10,504,834	7.21	196,217,625	13.36	1.9
086	Government document classification number	2,546,876	1.75	67,616,131	4.60	2.6
100	Main entry—personal name	88,327,070	60.64	1,077,151,051	73.36	1.2
110	Main entry—corporate name	12,767,602	8.77	78,413,273	5.34	0.6
130	Main entry—uniform title	1,763,334	1.21	21,670,118	1.48	1.2
210	Abbreviated title	457,096	0.31	15,260,816	1.04	3.4
222	Key title	736,620	0.51	19,192,962	1.31	2.6
240	Uniform title	5,460,464	3.75	65,313,432	4.45	1.2
245	Title statement	145,658,646	100.00	1,468,380,998	100.00	1.0
246	Varying form of title	13,438,950	9.23	117,359,067	7.99	0.9
250	Edition statement	19,116,097	13.12	291,786,693	19.87	1.5
255	Cartographic mathematical data	1,779,008	1.22	6,226,959	0.42	0.3
256	Computer file characteristics	299,902	0.21	1,401,141	0.10	0.5
260	Imprint statement	139,777,516	95.96	1,459,254,154	99.38	1.0
300	Physical description	132,199,166	90.76	1,431,641,957	97.50	1.1
306	Playing time	494,429	0.34	9,594,817	0.65	1.9
310	Current publication frequency	2,493,621	1.71	33,791,554	2.30	1.3
321	Former publication frequency	180,039	0.12	6,853,778	0.47	3.9
351	Organization and arrangement of materials	203,369	0.14	185,300	0.01	0.1
362	Dates of publication and/or sequential designation	3,844,862	2.64	40,423,981	2.75	1.0
440	Series statement/ Added entry—title	20,086,063	13.79	240,540,930	16.38	1.2
490	Series statement	17,435,350	11.97	214,970,867	14.64	1.2
500	General note	63,427,638	43.55	648,963,939	44.20	1.0
502	Dissertation note	11,540,519	7.92	23,850,586	1.62	0.2
504	Bibliography, etc. note	23,896,747	16.41	617,136,191	42.03	2.6
505	Formatted contents note	7,762,565	5.33	198,321,546	13.51	2.5
506	Restrictions on access note	3,035,334	2.08	11,754,383	0.80	0.4
508	Creation/production credits note	1,219,462	0.84	20,104,189	1.37	1.6
510	Citation/references note	2,238,409	1.54	21,324,277	1.45	0.9
511	Participant or performer note	4,021,973	2.76	49,212,294	3.35	1.2
515	Numbering peculiarities note	627,906	0.43	9,131,656	0.62	1.4

Table 2.13: Fields used in WorldCat, by tag (continued)

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Gap
516	Type of computer file or data note	666,970	0.46	2,716,093	0.18	0.4
518	Date/time and place of event note	1,356,481	0.93	11,616,897	0.79	0.8
520	Summary, etc.	8,669,392	5.95	168,190,897	11.45	1.9
521	Target audience note	824,852	0.57	19,820,802	1.35	2.4
524	Preferred citation of described materials note	220,875	0.15	214,237	0.01	0.1
530	Additional physical form available note	3,866,934	2.65	121,410,341	8.27	3.1
533	Reproduction note	10,216,381	7.01	96,206,934	6.55	0.9
538	System details note	3,306,716	2.27	52,086,634	3.55	1.6
540	Terms governing use and reproduction note	929,027	0.64	1,714,808	0.12	0.2
541	Immediate source of acquisition note	711,391	0.49	953,044	0.06	0.1
545	Biographical or historical data	558,382	0.38	555,782	0.04	0.1
546	Language note	4,124,646	2.83	30,436,797	2.07	0.7
550	Issuing body note	532,838	0.37	8,535,396	0.58	1.6
555	Biographical or historical data	441,914	0.30	3,004,993	0.20	0.7
561	Ownership and custodial history	541,899	0.37	776,863	0.05	0.1
580	Linking entry complexity note	1,192,281	0.82	9,602,496	0.65	0.8
583	Action note	372,813	0.26	908,548	0.06	0.2
600	Subject added entry—personal name	10,311,495	7.08	180,274,938	12.28	1.7
610	Subject added entry—corporate name	7,707,544	5.29	79,494,396	5.41	1.0
630	Subject added entry—uniform title	1,475,625	1.01	21,984,315	1.50	1.5
650	Subject added entry—topical term	67,056,447	46.04	1,153,889,751	78.58	1.7
651	Subject added entry—geographic name	14,460,979	9.93	267,959,899	18.25	1.8
653	Index term—uncontrolled	8,795,207	6.04	93,333,774	6.36	1.1
655	Index term—genre/form	6,219,255	4.27	145,545,124	9.91	2.3
700	Added entry—personal name	41,249,927	28.32	499,048,757	33.99	1.2
710	Added entry—corporate name	27,663,715	18.99	255,452,605	17.40	0.9
730	Added entry—uniform title	1,919,693	1.32	21,427,840	1.46	1.1
740	Added entry—uncontrolled related/analytical title	6,260,167	4.30	69,473,663	4.73	1.1
752	Added entry—hierarchical place name	1,603,694	1.10	8,738,214	0.60	0.5

Table 2.13: Fields used in WorldCat, by tag (continued)

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Gap
773	Host item entry	5,452,146	3.74	8,024,916	0.55	0.1
776	Additional physical form entry	2,919,440	2.00	64,285,347	4.38	2.2
780	Preceding entry	1,188,657	0.82	16,535,025	1.13	1.4
785	Succeeding entry	1,055,276	0.72	12,069,892	0.82	1.1
800	Series added entry—personal name	588,115	0.40	21,538,189	1.47	3.7
810	Series added entry—corporate name	1,776,400	1.22	20,459,229	1.39	1.1
830	Series added entry—uniform title	8,752,426	6.01	116,390,596	7.93	1.3
850	Holdings information	811,980	0.56	22,618,149	1.54	2.8
856	Electronic location and access	8,738,346	6.00	252,747,946	17.21	2.9
880	Alternate graphic representation	8,621,428	5.92	24,354,264	1.66	0.3

Key to Colors

Yellow:	Repeatable fields
Rose:	Obsolete field
Green:	Gap (differences between occurrences and holdings percentages) is double or more
Orange:	Gap is half or less

Table 2.14: MARC 21 fields little or not used

Notes

- This table lists the 86 tags occurring less than 1% in WorldCat or Books and less than 10% in any other format.
- Tags added in 2001–2008 are displayed in **bold** font.
- 27 tags have "0" occurrences, even when weighted by holdings (shown in lavender).
- 3 tags (in orange) are the ones OCLC has no plans to implement.

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Occurs Most In	Occurrences
013	Patent control information	507	0	1,034	0	Books	486
018	Copyright article-fee code	13,151	0.01	15,267	0	Books	13,138
025	Overseas acquisition number	381,677	0.26	3,841,183	0.26	Books	358,028
026	Fingerprint identifier	25,515	0.02	26,392	0	Books	25,493
027	Standard technical report number	510,206	0.35	1,165,565	0.08	Books	509,369
031	Musical incipits information	100,581	0.07	295,761	0.02	Books	96,500
036	Original study number for computer data files	4,799	0	16,561	0	Books	1,218
038	Record content licensor	0	0	0	0		
046	Special coded dates	61,630	0.04	101,699	0.01	Books	47,015
047	Form of musical composition code	271,638	0.19	4,088,566	0.28	Recordings	209,892
051	Library of Congress copy, issue, offprint statement	76,122	0.05	5,895,575	0.40	Books	72,123
061	National Library of Medicine copy statement	9	0	311	0	Books	8
071	National Agricultural Library copy statement	90	0	135	0	Mixed	50
083	Additional Dewey Decimal Classification number	7	0	380	0	Books	7
085	Synthesized classification number components	2	0	2	0	Books	2
088	Report number	579,314	0.40	12,562,146	0.86	Books	556,528

Table 2.14: MARC tags little or not used (continued)

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Occurs Most In	Occurrences
111	Main entry—meeting name	1,180,252	0.81	13,192,425	0.90	Books	1,103,394
242	Translation of title by cataloging agency	603,056	0.41	1,054,695	0.07	Books	561,706
243	Collective uniform title	74,222	0.05	159,844	0.01	Books	66,090
247	Former title	90,119	0.06	588,961	0.04	Serials	82,797
254	Musical presentation statement	130,426	0.09	727,924	0.05	Scores	129,739
258	Philatelic issue data	2	0	191	0		
261	Imprint statement for films, pre-AACR2	72,316	0.05	159,048	0.01	Visual	72,305
262	Imprint statement for sound recordings, pre-AACR2	162,042	0.11	1,987,104	0.14	Recordings	162,042
263	Projected publication date	537,279	0.37	10,180,089	0.69	Books	518,574
270	Address	141,443	0.10	304,521	0.02	Books	59,863
307	Hours, etc.	2,731	0	2,215	0	Books	2,447
340	Physical medium	82,060	0.06	304,896	0.02	Comp Files	32,264
342	Geospatial reference data	5,862	0	7,340	0	Maps	5,733
343	Planar coordinate data	29	0	477	0	Maps	28
352	Digital graphic representation	1,072	0	2,924	0	Maps	964
355	Security classification control	0	0	0	0		
357	Originator dissemination control	0	0	0	0		
363	Normalized date and sequential designation	2	0	6	0	Serials	2
365	Trade price	0	0	0	0		
366	Trade availability information	0	0	0	0		
501	With note	510,735	0.35	1,609,559	0.11	Books	346,276
507	Scale note for graphic material	124,958	0.09	377,917	0.03	Maps	109,302
513	Type of report and period covered note	99,871	0.07	790,078	0.05	Books	95,626
514	Data quality note	561	0	1,898	0	Books	477
522	Geographic coverage note	87,621	0.06	177,478	0.01	Maps	20,150
525	Supplement note	96,132	0.07	2,636,802	0.18	Serials	71,312
526	Study program information note	6,488	0	1,810,228	0.12	Books	6,396
534	Original version note	769,334	0.53	2,446,667	0.17	Books	635,182

Table 2.14: MARC tags little or not used (continued)

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Occurs Most In	Occurrences
535	Location of originals/duplicates note	802,790	0.55	1,450,720	0.10	Books	689,604
536	Funding information note	309,705	0.21	1,450,800	0.10	Books	258,097
542	Information relating to copyright status	536	0	565	0	Books	535
544	Location of other archival materials note	38,838	0.03	36,372	0	Mixed	23,686
547	Former title complexity note	22,280	0.02	61,395	0	Serials	17,088
552	Entity and attribute information note	260	0	1,619	0	Books	113
556	Information about documentation note	5,438	0	32,613	0	Comp Files	3,608
562	Copy and version identification note	118,241	0.08	118,974	0.01	Books	85,609
563	Binding information	74,918	0.05	100,623	0.01	Books	70,826
565	Case file characteristics note	5,018	0	10,731	0	Comp Files	4,867
567	Methodology note	18,310	0.01	56,944	0	Comp Files	16,051
581	Publications about described materials note	80,032	0.05	81,846	0.01	Books	53,545
584	Accumulation and frequency of use note	7,701	0.01	6,986	0	Mixed	7,623
585	Exhibitions note	19,031	0.01	54,541	0	Books	11,990
586	Awards note	35,571	0.02	5,405,729	0.37	Books	17,605
611	Subject added entry—meeting name	210,866	0.14	2,441,612	0.17	Books	166,746
648	Subject added entry—chronological term	103,559	0.07	241,362	0.02	Books	101,867
654	Subject added entry—faceted topical terms	57,217	0.04	70,674	0	Visual	38,677
656	Index term—occupation	117,315	0.08	115,991	0.01	Mixed	72,819
657	Index term—function	30,918	0.02	29,131	0	Mixed	29,324
658	Index term—curriculum objective	131	0	2,225	0	Books	72
662	Subject added entry—hierarchical place name	728	0	753	0	Visual	466
711	Added entry—meeting name	1,057,274	0.73	6,337,744	0.43	Books	968,597
720	Added entry—uncontrolled name	904,707	0.62	1,517,283	0.10	Books	703,532
751	Added entry—geographic name	4	0	4	0	Books	2
753	System details access to computer files	79,697	0.05	382,879	0.03	Comp Files	76,076
754	Added entry—taxonomic identification	1,930	0	2,419	0	Books	1,893
760	Main series entry	286,508	0.20	873,240	0.06	Books	239,749
762	Subseries entry	1,598	0	18,390	0	Serials	1,387

Table 2.14: MARC tags little or not used (continued)

Tag	Description	Occurrences	Percent	With Holdings	Weighted %	Occurs Most In	Occurrences
765	Original language entry	140,849	0.10	337,238	0.02	Books	134,043
767	Translation entry	36,050	0.02	105,386	0.01	Books	30,148
770	Supplement/special issue entry	58,741	0.04	1,584,437	0.11	Serials	49,855
772	Supplement parent entry	261,455	0.18	940,323	0.06	Books	173,338
774	Constituent unit entry	145,150	0.10	149,713	0.01	Books	73,869
775	Other edition entry	234,826	0.16	1,739,733	0.12	Serials	131,416
777	Issued with entry	28,085	0.02	359,134	0.02	Serials	14,152
786	Data source entry	121,200	0.08	108,159	0.01	Visual	89,212
787	Other relationship entry	927,861	0.64	4,994,482	0.34	Books	503,453
811	Series added entry—meeting name	12,854	0.01	202,074	0.01	Books	10,844
882	Replacement record information	0	0	0	0		
886	Foreign MARC information field	746,194	0.51	1,409,924	0.10	Books	618,728
887	Non-MARC information field	23	0	33	0	Books	10

Full Data Tables Related to MARC Tag Usage in WorldCat

The following data tables were too extensive to be included in the body of this report, but are available online (Microsoft Excel format) at http://www.oclc.org/research/publications/library/2010/2010-06a.xls. Each table is under a separate spreadsheet (tab), which is listed below with the table name.

- Table 2.15: Overview of MARC tag usage in WorldCat (Sept 2009)
  Sheet: All Formats, 1% or more
- Table 2.16: Grid of MARC tag usage in WorldCat by format (Sept 2009)
  Sheet: Grid All WC Fields
- Table 2.17: MARC tags in WorldCat—Books (1% or more)
  Sheet: Books
- Table 2.18: MARC tags in WorldCat—Computer files (10% or more)
  Sheet: Computer Files
- Table 2.19: MARC tags in WorldCat—Integrating resources (10% or more)
  Sheet: Integrating Resources
- Table 2.20: MARC tags in WorldCat—Maps (10% or more)
  Sheet: Maps
- Table 2.21: MARC tags in WorldCat—Mixed materials (10% or more)
  Sheet: Mixed Materials
- Table 2.22: MARC tags in WorldCat—Scores (10% or more)
  Sheet: Scores
- Table 2.23: MARC tags in WorldCat—Serials (10% or more)
  Sheet: Serials
- Table 2.24: MARC tags in WorldCat—Sound recordings (10% or more)
  Sheet: Sound Recordings
- Table 2.25: MARC tags in WorldCat—Visual materials (10% or more)
  Sheet: Visual Materials

3. MARC Fields and Subfields Used in Machine Matching

Hugh Taylor

The purpose of this study was to compare the MARC fields and subfields used by machine matching profiles of selected union catalogs or other aggregation databases. A comparison was then made with the input requirements of a small number of cataloging environments, in order to determine the degree of consistency between the need (expressed in the matching profiles) for MARC fields and subfields to be provided (where appropriate to the resource being described) and the requirements of the cataloging programs that such data always be provided.

Matching profiles from the following were used in this study:

- RLUK (Research Libraries UK) Database (a staff tool used for copy cataloging)
- Copac National, Academic, & Specialist Library Catalogue (public union catalog derived from the RLUK database)
- WorldCat
- RLG Union Catalog (service now discontinued)
- Libraries Australia

Input requirements from the following standards were recorded:

- BIBCO Core Record standards
- CONSER (excluding special formats)
- OCLC Abbreviated Level (Level 3) guidelines

Notes Accompanying "MARC Fields/Subfields Used in Matching" Table

The matching profiles are tabulated in the second through sixth columns and the input requirements in the seventh through ninth columns. The last column is used to record notes, observations, etc., that supplement the tabulation of the other columns.

Generally, how something has been represented in the table reflects whatever the documentation actually says. Occasionally, where one document gives a field and another a particular element and I knew that the latter covers the whole field I've collapsed them into a single entry.

I've assumed that no service would ever use $6 and $8 in matching. These assumptions are built into the table, even if documentation doesn't make this explicit (which it rarely did).

A bold, italic, red ***Y*** indicates that the requirement for a particular piece of data is restricted to a particular format(s).

It is impossible to represent the complexity of most matching algorithms in a table such as this. A number operate on various levels ("simple" match, confirmation of "simple" match, "complex" match, etc.) which the software works through in sequence. Basically, any field or subfield used for any reason in a matching algorithm receives an entry in the table. This tends to "flatten" out the importance (or otherwise) of fields and subfields. Even in avowedly egalitarian societies some people are usually more important than others.

Some of the documentation surely can't reflect what's intended. Situations I've spotted during the course of this study have been noted, but this cannot hope to be exhaustive.

Requirements that involve local fields and/or subfields have been omitted on the grounds that these are meaningless outside the context of a particular project or program.

On the input side, no attempt has made to distinguish "mandatory" fields or subfields from "required if applicable." If the input standard says the field or subfield is required in some situation or other then that data element receives an entry in the table.

Conclusions

The following table is somewhat subjective, as reference back to the longer table will show. It attempts to identify where there's unanimity or near-unanimity amongst the services or standards sampled. If everyone's agreed that tag X is vital for matching purposes, then there's a Y in the "Matching" column—and similarly for input. Of course, there aren't huge numbers of either sort to work with.

What's striking is that most of the core elements for matching are amongst those identified as important on the input side. So most of the data that the majority of services have identified as relevant to their matching needs should be present in the majority of records (or those created since the idea of mandatory data elements was formulated). But beware: it may be that the matching services use the elements that they use simply because they can expect them to have been provided already—because the input standards require them. It could just be a circular relationship that's producing this consistency!

The only data element used by a majority of the matching routines surveyed and not required during record creation is the *Other Standard Identifier,* 024 (BIBCO requires it only for Music).

Amongst the odd things that this or that matching routine uses to help achieve a "match" in the longer Table 3.2 (p. 40):

a. there's **no** consistency between the various routines sampled, and

b. there's no requirement to provide the data in records.

Table 3.1: Core MARC fields and subfields used in matching across five databases

Field etc.	Matching	Input	Notes
Leader/06	Y	Y	
Leader/07	Y	Y	
Leader/17		Y	
Leader/18		Y	
008/06	Y	Y	
008/07-10	Y	Y	
008/11-14	Y	Y	
008/15-17		Y	
008/35-37		Y	
008/39		Y	
008/23		Y	Certain formats only
010	Y	Y	
020	Y	Y	
022	Y	Y	
024	Y		Three services utilize 024, but no consistency re the specific subfields of interest
1XX	Y	Y	
245	Y	Y	
246		Y	
250	Y	Y	Matching on 250/a only
260/a b	Y	Y	
260/c	Y	Y	
700-730		Y	

Table 3.2: MARC fields/subfields used in matching

Key: **Y** = Required or Required if applicable (no distinction in this document)
Y = The requirement for a particular piece of data is restricted to a particular format(s).

Note: WorldCat also uses material type terms generated as for mt index, but no specific details available

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
	Used in matching	Used in matching	Used in matching	Used in matching	Used in matching	Required on input	Required on input	Required on input	RLG UC details here exclude AMC; BIBCO details here exclude Collections, Multiple character sets,
Leader/06	**Y**		**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	WorldCat: uses subset of dt index (created from Leader/06-07)
Leader/07	**Y**		**Y**	**Y**	**Y**	**Y**	**Y**	**Y**	WorldCat: uses subset of dt index (created from Leader/06-07)
Leader/17						**Y**	**Y**	**Y**	
Leader/18						**Y**	**Y**	**Y**	
001			**Y**		**Y**				WorldCat: uses Local system number (assumed here to be 001) for Institution Records only
006					**Y**				
006/06								*Y*	OCLC 3: BKS, CNR, MIX, SCO, REC
006/12								*Y*	OCLC 3: MAP, VIS
006/17				*Y*					RLG: CR only
007					**Y**				
007/00						*Y*			BIBCO: For microforms, code for Books, Rare Books, Cartographic, and Music
007/00						*Y*			BIBCO: For Electronic Resources, code for Electronic Resources, Books, Rare Books, Cartographic, Music
007/00						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/01						*Y*			BIBCO: For microforms, code for Books, Rare Books, Cartographic, and Music
007/01						*Y*			BIBCO: For Electronic Resources, code for Electronic Resources, Books, Rare Books, Cartographic, Music

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
007/01						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/03						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/04						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/05						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/06						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/07						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/08						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/12						*Y*			BIBCO: For Sound Recordings, code for sound recordings
007/13						*Y*			BIBCO: For Sound Recordings, code for sound recordings
008/00-05					Y				
008/06				Y	Y	Y	Y	Y	
008/07-10	*Y*	*Y*	Y	Y	Y	Y	Y	Y	COPAC/RLUK: Pre-1800 never matches; also Not used for periodicals
008/11-14	*Y*	*Y*	Y	Y	Y	Y	Y	Y	COPAC/RLUK: Pre-1800 never matches; also Not used for periodicals
008/15-17			Y			Y	Y	Y	
008/35-37						Y	Y	Y	
008/38						Y	Y		
008/39						Y	Y	Y	
008/18-20						*Y*			BIBCO: Graphic, Moving image
008/18-19							Y		
008/20						*Y*			BIBCO: Sound recordings; Music
008/21							Y		
008/22						*Y*	Y		BIBCO: Books and Rare Books

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
008/23					Y	*Y*	Y	*Y*	BIBCO: Books and Rare Books; Electronic Resources; Sound recordings; Music; OCLC 3: BKS, CNR, MIX, SCO, REC
008/24-29						*Y*			BIBCO: Sound recordings; Music
008/25						*Y*			BIBCO: Cartographic
008/26						*Y*			BIBCO: Electronic Resources
008/28						*Y*			BIBCO: Books and Rare Books; Graphic, Moving image; Cartographic
008/29					Y	*Y*		*Y*	BIBCO: Graphic, Moving image; Cartographic; OCLC 3: MAP, VIS
008/30-31						*Y*			BIBCO: Sound recordings; Music
008/33						*Y*			BIBCO: Books and Rare Books; Graphic, Moving image
008/34				*Y*		*Y*	Y		RLG: CR only; BIBCO: Books and Rare Books; Graphic, Moving image
010		Y			Y	Y	Y	Y	
010/a	Y		Y	Y					
010/z			*Y*	Y					WorldCat: Serials only
015					Y				
016					Y				
016/a			Y						
017					Y				
018					Y				
020		Y			Y	Y		Y	COPAC: Probably matches only on 020/a (minus any parenthetical stuff)
020/a z	Y		Y	*Y*					RLG: Excl CR
022		Y	Y		Y	Y	Y	Y	COPAC: Probably matches only on 022/a (minus any parenthetical stuff)
022/a z	Y			*Y*					RLG: CR only
024					Y				

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
024/a			Y						
024/a z				Y					
024.2		Y				*Y*			BIBCO: Music
027			Y		Y				
028					Y	*Y*		*Y*	BIBCO: Sound recordings; Music; Graphic (with qualifications); OCLC 3: Music publisher's no. only
028/a			*Y*	*Y*					WorldCat & RLG: Music and Sound recordings only
028/b				*Y*					RLG: REC & SCO only
030					Y				
030/a			Y						
034/a b c h s t						*Y*			BIBCO: Cartographic
034/b		Y							
035					Y				
035/a			*Y*						WorldCat: OCLC number and (for Institution Records only) RLG number
037					Y				
037/a						*Y*			BIBCO: Graphic (with qualifications)
040						Y			
040/b			Y						
042						Y	Y		BIBCO: Required for BIBCO records only
052/a b				*Y*					RLG: MAP only
050, 082, 086, 090, etc.						*Y*	Y		BIBCO: One number from standard system required for Books, Music, Cartographic
074							Y		
086					Y				
088			Y		Y				

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
1XX		Y			Y	Y	Y	Y	COPAC: one or more author words is additional match if first threshold limit exceeded; later on, fuzzy match (removing corporate author stopwords) may be applied
100/a	Y		Y	Y					RLUK: Only used if title consists entirely of generic word(s)
110/a b			Y						
110/a b d	Y			Y					RLUK: Only used if title consists entirely of generic word(s); Certain stopwords ignored
111/a			Y						
111/a b e	Y			Y					RLUK: Only used if title consists entirely of generic word(s); Certain stopwords ignored
130/a p s			Y						
130/a	Y			Y					RLUK: Certain stopwords ignored
222/a			Y						
222/a b				*Y*					RLG: CR only
240						Y	Y		BIBCO: If known or can be readily inferred
240/h s		*Y*							COPAC: Music only
240/s			Y						
245		[Y]	Y		Y	Y	Y		COPAC: either title key (like a fingerprint) or, if single word title, whole subfield; also for Music only for score type check; WorldCat: derives title key from field; BIBCO: For Sound recordings & Music, in cases of multiple parallel titles, MINIMALLY include the first title and any English parallel title.
245/a b	Y							Y	RLUK: creates 3,2,2,1 key from this
245/a b f g k p		*Y*							COPAC: Excl periodicals; fuzzy matching applied
245/a b f k n p			Y						
245/a b n p	Y			Y					

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
245/a p		*Y*							COPAC: Periodicals only; fuzzy matching applied
245/c		Y						Y	COPAC: Used only if fuzzy match on author words not possible because no author fields to check
245/h				*Y*				Y	RLG: VIM only
246			Y			Y	Y	Y	WorldCat: derives title key from field; also uses 246/a separately at other point in matching; BIBCO: Use judgment
247			Y						WorldCat: derives title key from field
250					Y	Y	Y	Y	
250/a	Y	Y	Y	Y					RLUK: If 250/a begins with number, use that; otherwise whole subfield
254/a			Y	*Y*	Y				RLG: SCO only
255					Y				
255/a			*Y*						WorldCat: Cartographic only
255/a b						*Y*			BIBCO: Cartographic
256/a				*Y*					RLG: ELE only
260					Y	Y	Y		
260/a b	Y	Y	Y	Y				Y	COPAC/RLUK: Partial match after common stopwords removed
260/c	*Y*	*Y*	Y	Y				Y	COPAC/RLUK: Pre-1800 never matches; also Not used for periodicals
260/f			Y						
261/a b e f			Y						
261/a d f				*Y*					RLG: VIM only
262/a b c			Y						
262/c				*Y*					RLG: REC only
300		*Y*			Y	Y			COPAC: Music only
300/a	Y	*Y*	*Y*	*Y*				Y	COPAC: Excl periodicals; WorldCat: Excl serials; RLG: Excl CR

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
300/c			*Y*	*Y*					WorldCat: Excl serials; RLG: Excl CR
300/e			*Y*						WorldCat: Music only
300/a c				*Y*					RLG: REC only
305/c			*Y*						WorldCat: Excl serials
352						*Y*			BIBCO: Cartographic
362			Y				Y		
362/a				*Y*					RLG: CR only
4XX						Y	Y		
440/v		Y							
490					Y				
490/a								Y	
490/v		Y							COPAC: Use only if no 440/v
500						Y	Y		BIBCO: Source of title proper; plus various others for Electronic, Sound recordings, Music, Cartographic; CONSER: Source of title, DBO note
501						Y			
502					Y	*Y*			BIBCO: Unpublished theses only
505					Y	Y			
507/a			*Y*						WorldCat: Cartographic only
510						*Y*			BIBCO: Rare Books
511						*Y*			BIBCO: Moving image (if needed for identification), Sound recordings
520						*Y*			BIBCO: Electronic, Graphic, Moving image, Non-music sound recordings - if not obvious from remainder of record
533						*Y*			BIBCO: Books, Sound recordings, Music
533/b c d			Y						
534						*Y*			BIBCO: Cartographic
538						*Y*			BIBCO: Electronic Resources, Moving image

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

Table 3.2: MARC fields/subfields used in matching (continued)

Field, etc.	RLUK	COPAC	WorldCat	RLG UC	Libraries Australia	BIBCO	CONSER (excl special formats)	OCLC Level 3	Notes
546						*Y*			BIBCO: Graphic, Moving image; plus Sound recordings & Music (if not implied in 240/245)
552/o						*Y*			BIBCO: Cartographic
6XX						Y	Y		
6XX/v x			Y						
655						Y			BIBCO: Genre terms encouraged for Rare Books
7XX						Y			BIBCO: Should this really apply only to 700-730?
700-730		Y					Y	Y	COPAC: one or more author words is additional match if first threshold limit exceeded; later on, fuzzy match (removing corporate author stopwords) may be applied; OCLC 3: One 7XX if applicable and no 1XX
700/710/711					Y				
700/a			Y						
710/a b			Y						
711/a			Y						
720		Y						Y	COPAC: Used only if fuzzy match on author words not possible because no author fields to check; OCLC 3: One 7XX if applicable and no 1XX; last resort (prefer 700/710/711)
720/a			Y						
730/a p			Y						
752/d		*Y*							COPAC: Periodicals only
774					Y				
780							Y		
785							Y		
8XX						Y	Y		BIBCO: Should this really apply only to 800-830?; CONSER: Applies to series added entries only
800/810/811					Y				
850					Y		Y		
856						*Y*			BIBCO: Electronic Resources
856/u			Y					Y	

Key: **Y** = Required or Required if applicable (no distinction in this document); ***Y*** = The requirement for a particular piece of data is restricted to a particular format(s).

4. Comparison of Search Interfaces and Data Elements

Catherine Argus

Methodology

The object was to examine how the data recorded in a variety of bibliographic records are utilized by the search interfaces of aggregated databases.

Amicus, COPAC, Libraries Australia, Worldcat.org and WorldCat FirstSearch were selected as examples of aggregated databases. The MARC fields utilized for search and limit options in each database were mapped. However, all the databases were not originally developed for use by the public. Therefore, some search interfaces such as Expert search and Command search, were excluded to enable fairer comparison between the databases.

A sample of 52 bibliographic records for a variety of material formats was extracted from WorldCat. The fields and subfields used in each sample record were logged. However, no attempt was made to analyze if all the data recorded in those subfields were being utilized. The value recorded in most character positions in the leader and the 006, 007, and 008 fields were also noted.

The MARC fields used in the sample records were cross-matched with the MARC fields used for searching and limiting in the example databases.

Most local MARC fields were excluded.

Search Interface Comparison

Search Options

Although there was variation between the search options offered by individual search interfaces, there was a fair degree of similarity between the selected databases when their search interfaces were combined. "Any keyword", Name, Subject, Title and ISBN/ISSN search options were offered by at least one search interface to all the databases. A Publisher name search option was offered by

four databases. Three databases offered an option to search Notes fields. Libraries Australia and WorldCat FirstSearch offered options to search on Series, Language, and Place of publication.

All the databases offered at least one search option not duplicated in any of the other databases, such as the option to search by either exact or fuzzy scale in the Map search interface to COPAC.

Table 4.1: Selected search options

Search options	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	Worldcat.org (WorldCat)
Any Keyword	A	Q M Map	B A	B A	B A
Title	B A	Q M	A	B A	A
Name	B A	Q M Map	A	B A	A
Subject	B A	M	A	A	A
Publisher Name	A	M	A	A	
ISBN, ISSN etc		M Map	A	A	
ISBN	B A			B A	A
ISSN	B A			A	A
Notes	A		A	A	
Language			A	A	
Place of publication			A	A	
Series			A	A	
Place (Name. Title, Series, Subject)		Map			
Scale - Exact		Map			
Scale - Fuzzy		Map			

B = Basic search A = Advanced search Q = Quick search M = Main search Map = Map search

The MARC fields indexed for the search options were also quite similar, with most of the apparent variations due to the nature of particular interfaces. (See Appendix A) There were a number of fields not indexed for searching by any of the databases, such as 026 (Fingerprint identifier) and 366 (Trade availability information). However, many of the un-indexed fields are used to record administrative data, are MARC fields that have been defined relatively recently, or are fields that contain data that is more useful for display rather than searching, such as 307 (Hours, etc.).

No attempt was made to analyze the usage of particular subfields within indexed fields or to analyze whether the search options offered correlate with the needs of user groups.

Limit Options

All the databases generally offered limiting options through drop-down menus, “radio” buttons and check boxes. Libraries Australia and WorldCat FirstSearch appear to offer more limiting options than the other databases. However, this is because Libraries Australia and WorldCat FirstSearch mostly use radio buttons and check boxes, whereas the other databases utilize more drop-down menus.

Like the searching options, there was a fair degree of similarity between the limiting options offered by the selected databases. The most common limiting options related to content, format, date of publication, language and target audience. Most limiting options utilized coded data, frequently in combinations. Only Worldcat.org and WorldCat FirstSearch offered post-search limiting.

Table 4.2: Limit options

AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	worldcat.org (WorldCat)
Date published (A)	Date published (M Map)	Australian (B A)	Date of publication (A)	Year (A Post)
Format (A)	Place published (M Map)	Online (B A)	Language (A)	Audience (A Post)
Language (A)	Material type (M)	Government (B A)	Number of libraries (A)	Content (A Post)
Target audience (A)	Language (M)	Conference (B A)	Books (A)	Format (A Post)
Publication type (A)	Library (M Map)	In Library (held in any library) (B A)	Visual materials (A)	Language (A Post)
		Books (B A)	Computer files (A)	Author (Post)
		Journals (B A)	Internet resources (A)	Topic (Post)
		Newspapers (B A)	Serial publications (A)	
		Manuscripts (B A)	Sound recordings (A)	
		Theses (B A)	Archival materials (A)	
		Maps (B A)	Continually updated resources (A)	
		Computer files (B A)	Musical scores (A)	
		Oral histories (B A)	Maps (A)	
		Printed music (B A)	Audience (A Post)	
		Musical sound (B A)	Content (A Post)	
		Pictures (B A)	Format (A Post)	
		Film/Video (B A)	Items in my library (A)	
		Braille (B A)	Library code (A)	
		Large Print (B A)		
		Talking books (B A)		
		When published (A)		
		Where held (A)		

B = Basic search A = Advanced search Q = Quick search M = Main search Map = Map search

The MARC fields utilized for limiting were also quite similar, with the principal variations relating to limiting options offered for content and format.

Sorting

The sorting options were virtually identical. However, some of the search interfaces offered the option to sort by Title and Author in either ascending or descending order. WorldCat FirstSearch offered the option to sort by the number of holding libraries.

Table 4.3: Sort options

Sort options	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
Relevance		Default	Default	Default	Default
Title A-Z	B A	Q M Map	B A	B A	B A
Title Z-A		Q M Map	B A	B A	
Author A-Z	B A	Q M Map	B A	B A	B A
Author Z-A		Q M Map	B A	B A	
Date ascending	B A	Q M Map	B A	B A	B A
Date descending	B A	Q M Map	B A	B A	B A
No. of libraries				B A	

B = Basic search A = Advanced search Q = Quick search M = Main search Map = Map search

Display

The variations in record display between the search interfaces were not analyzed. However, anecdotal evidence suggests that in most cases only textual fields were utilized for record display. An exception was Genre and Language in COPAC, which displayed coded data in textual form.

Sample Record Analysis

Fifty-two bibliographic records were extracted from WorldCat: 21 book records, 7 visual material records, 6 continuing resource records, 4 music score and 4 sound recording records, 4 electronic resource records, 3 map records, and 3 mixed material records. Several non-English records were included.

For most material types, the only fields that were being searched by all the sample databases were ISBN and ISSN, main and added entry fields, title and variant titles, series, and subjects. Many other fields were used by three or fewer databases. There were very few fields that were not indexed for searching by any of the databases. However, some of the data recorded in 006, 007 and 008 are not used for either searching or limiting. (See analysis of records in Appendix B)

Final Comments

All the databases offer similar searching and limiting options. They also index similar MARC fields. However, these similarities are not necessarily an affirmation that current search interfaces meet user needs. Just because options are offered or fields are indexed does not mean that users find them useful or utilize them.

It is noteworthy that there is some correlation between the un-indexed fields and the MARC fields that Karen Smith-Yoshimura's snapshot of MARC tag usage in WorldCat has identified as little or not used in WorldCat (see Table 2.14, p. 32).

The limited usage of some fields and data does raise a question about the capacity of current search interfaces to make effective use of the available granular data. However, it is noted that many search interfaces are being redesigned, with a trend towards simple, visually appealing interfaces and to allowing personalization. A significant trend is to leverage data within catalog records, such as ISBNs, to deliver cover art, reviews, etc and to link to booksellers.

Table 4.4: MARC tags used for searching and limiting

Note: The data relating to the search and limit options may be out of date as this project was undertaken over an extended period and the search interfaces may have changed.

SEARCH & LIMIT USAGE	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
Leader/06	L	L	L	A L	B L
Leader/07	L	L	L	A L	B L
001	B A		B A		
005	L				
006/00 BK				L	B L
006/00 CF				L	L
006/00 MP				L	L
006/00 MX			L	L	L
006/00 MU				A L	L
006/00 CR				L	L
006/00 VM				L	L
006/01-02 MU				A	
006/02 CR			L		
006/04 CR				L	L
006/05 BK			L	L	L
006/05 CF				L	L
006/05 MU			L	L	L
006/05 CR			L		
006/05 VM				L	L
006/06 BK			L	L	L
006/06 MX				L	L
006/06 MU				L	L
006/06 CR				L	L
006/07-10 BK	L		L		
006/08 MP			L		
006/08-10 CR			L		
006/09 CF			L		
006/11 BK	L		L		
006/11 MP	L		L		
006/11 CR	L		L		
006/11 VM	L		L		
006/12 MP				L	L
006/12 VM			L	L	L
006/13-14 MU				L	L
006/16 BK				L	L
006/16 MU			L		
006/17 BK				L	L
006/17 CR			L		
007/00 MP				A	
007/00 Elec		L		A L	L

Key: B = Basic search, A = Advanced search, S = Search, Map = Map search, L = Limiting

Table 4.4: MARC tags used for searching and limiting (continued)

SEARCH & LIMIT USAGE	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
007/00 Globe				A	
007/00 Tactile				A	
007/00 PG		L		A	
007/00 MFM				A	
007/00 NPG		L		A	
007/00 Motion		L		A L	L
007/00 Kit		L		A	
007/00 MU		L			
007/00 RSI				A	
007/00 Sound		L		A L	B L
007/00 Video		L		A L	B L
007/01 MP				A	
007/01 Globe				A	
007/01 Tactile				A	
007/01 PG				A	
007/01 MFM				A	
007/01 NPG				A	
007/01 Motion				L	L
007/01 Sound	L			A L	B L
007/01 Video	L			A L	B L
007/03 Sound				A L	B L
007/03 Video				L	L
007/04 Elec				A	
007/04 Sound					B
007/04 Video				A L	L
007/06 Sound				A L	B L
007/11 MFM				A	
008/00-05	L			A	
008/06			L		
008/07-10	L	L	L	B A L	B A L
008/11-14				B A L	B A L
008/15-17	L	L	A		
008/18-19 MU	L			A	
008/19 CR			L		
008/20 MU	L				
008/21 CR	L			A	
008/22 BK	L	L	L	A L	L
008/22 CF	L	L		A L	L
008/22 MU	L	L	L	A L	L
008/22 CR			L		
008/22 VM	L	L		A L	L
008/23 BK	L	L	L	A L	L
008/23 MU	L	L		A L	L
008/23 CR	L	L		A L	L

Key: B = Basic search, A = Advanced search, S = Search, Map = Map search, L = Limiting

Table 4.4: MARC tags used for searching and limiting (continued)

SEARCH & LIMIT USAGE	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
008/23 MX	L	L		A L	
008/24-27 BK	L	L			
008/25 MP				A	
008/26 CF				A	
008/28 BK	L	L	L	A	
008/28 CF		L		A	
008/28 MP	L	L	L	A	
008/28 CR	L	L	L	A	
008/28 VM	L	L	L	A	
008/29 BK				A	
008/29 MP	L	L		A L	L
008/29 VM		L	L	A L	L
008/30-31 MU				A L	L
008/33 BK				L	L
008/33 VM			L	A	L
008/34 BK				A L	L
008/34 VM				A	
008/35-37	L	L	A	A L	L
010		S	B A	A	
013			B A		
015		S	B A		
016			B A	A	
017		S	B A		
018	A	S	B A		
020	B A	S Map	B A	B A	B A
022	B A	S Map	B A	B A	B A
024		S	B A	A	
025		S			
027		S	B A	A	
028		S	B A	A	
030		S	B A	A	
031				B A	B A
033	A		B A		
034	A	Map		B A	B A
035		S			
037	A	S	B A	A	
040		L	L		
041		L	L	A L	L
043	L				
046		L		B A L	B A L
047				A	
050			A		
052				B A	B A
060			A		

Key: B = Basic search, A = Advanced search, S = Search, Map = Map search, L = Limiting

Table 4.4: MARC tags used for searching and limiting (continued)

SEARCH & LIMIT USAGE	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
070			A		
072			A		
080			A		
082			A		
084			A		
086			B A		
088			B A	A	
100	B A	S	B A	B A	B A
110	B A	S	B A	B A	B A
111	B A	S	B A	B A	B A
130	B A	S Map	B A	B A	B A
210	B A	S Map	B A	B A	B A
222	B A	S Map	B A	B A	B A
240	B A	S Map	B A	B A	B A
242	B A	S Map	B A	B A	B A
243	B A	S Map	B A	B A	B A
245	B A	S Map	B A L	B A	B A
246	B A	S Map	B A	B A	B A
247	B A	S Map	B A	B A	B A
250	A		B A		
254	A		B A		
255	A		B A	B A	B A
256	A		B A		
257	A		B A		
260	A	L	B A	A	
263			B A		
270			B A		
300	A		B A		B
306	A		B A		
310	A		B A		
321	A		B A		
340	A		B A		
342			B A		
343			B A		
351	A		B A		
355	A				
357	A				
362	A		B A		
490	A	S Map	B A	B A	B A
500	A	S Map	B A	B A	B A
501	A	S Map	B A	B A	B A
502	A	S Map L	B A	B A L	B A L
504	A	S Map	B A	B A	B A
505	A	S Map	B A	B A	B A

Key: B = Basic search, A = Advanced search, S = Search, Map = Map search, L = Limiting

Table 4.4: MARC tags used for searching and limiting (continued)

SEARCH & LIMIT USAGE	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
506	A	S Map	B A	B A	B A
507	A	S Map	B A		
508	A	S Map	B A	B A	B A
510	A	S Map	B A		
511	A	S Map	B A	B A	B A
513	A	S Map	B A		
514	A	S Map	B A	A	
515	A	S Map	B A		
516	A	S Map	B A		
518	A	S Map	B A	B A	B A
520	A	S Map	B A	B A	B A
521	A	S Map	B A	B A	B A
522	A	S Map	B A		
524	A	S Map	B A		
525	A	S Map	B A		
526	A	S Map	B A		
530	A	S Map	B A	A	
533	A	S Map L	B A	B A	B A
534	A	S Map	B A	B A	B A
535	A	S Map	B A		
536	A	S Map	B A	B A	B A
538	A	S Map	B A	B A	B A
540	A	S Map	B A	A	
544	A	S Map	B A		
545	A	S Map	B A	B A	B A
546	A	S Map	B A		
547	A	S Map	B A		
550	A	S Map	B A	B A	B A
552	A	S Map	B A		
555	A	S Map	B A	A	
556	A	S Map	B A		
563	A	S Map		A	
580	A	S Map	B A		
581	A	S Map	B A		
583	A	S Map		A	
600	B A	S Map	B A L	B A	B A
610	B A	S Map	B A L	B A	B A
611	B A	S Map	B A L	B A	B A
630	B A	S Map	B A	B A	B A
648			L	B A	B A
650	B A	S Map	B A L	B A	B A
651	B A	S Map	B A L	B A	B A
653	B A	S Map	B A L	B A	B A
654	B A	S Map	B A L	B A	B A

Key: B = Basic search, A = Advanced search, S = Search, Map = Map search, L = Limiting

Table 4.4: MARC tags used for searching and limiting (continued)

SEARCH & LIMIT USAGE	AMICUS	COPAC	Libraries Australia	FirstSearch (WorldCat)	WorldCat.org (WorldCat)
655	B A	S Map	B A L	B A	B A
656	B A	S Map	B A L		
657	B A	S Map	B A L		
658	B A		B A		
700	B A	S Map	B A	B A	B A
710	B A	S Map	B A	B A	B A
711	B A	S Map	B A	B A	B A
720				B A	B A
730	B A	S Map	B A	B A	B A
740	B A	S Map	B A	B A	B A
751			L		
752	B A		B A	A	
753	A		B A L	B A	B A
754	B A		B A		
770				B A	B A
773				B A	B A
780			A	B A	B A
785			A	B A	B A
787				B A	B A
800	B A	S Map		B A	B A
810	B A	S Map		B A	B A
811	B A	S Map		B A	B A
830	B A	S Map	B A L	B A	B A
852$a		L	A L	L	
856	A L		B A	A L	L
880			A		

Key: B = Basic search, A = Advanced search, S = Search, Map = Map search, L = Limiting

Table 4.5: MARC tags used for searching specific types of records

Note: The WorldCat records may have changed since they were extracted for analysis.

Key: B = Basic search, A = Advanced search, Q = Quick search, M = Main search, Map = Map search

SEARCH OPTIONS	AMICUS	COPAC	Libraries Australia	WorldCat (FirstSearch)	WorldCat (WorldCat.org)
Any Keyword	A	Q M Map	B A	B A	B A
Title	B A	Q M	A	B A	A
Name	B A	Q M Map	A	B A	A
Subject	B A	M	A	A	A
Publisher Name	A	M	A	A	
Notes	A		A	A	
ISBN/ISSN etc		M Map	A	A	
ISBN	B A			B A	A
ISSN	B A			A	A
Language			A	A	
Place of publication			A	A	
Access method ($u subfield)				A	
Accession number				A	
Amicus number	B A				
Books					B
CDs					B
Classification			A		
Corporate and conference name (Author)				A	
Corporate and conference name (Subject)				A	
Date of publication				B	
Descriptor (Subject)				A	
DVDs					B
Genre/Form (Subject)				A	
Geographic (Subject)				A	
Library (holdings)			A		
Material type				A	
Musical composition				A	
Personal name (Author)				A	
Personal name (Subject)				A	
Place (Name. Title, Series, Subject)		Map			
Related periodical			A		
Scale - Exact		Map			
Scale - Fuzzy		Map			
Series			A	A	

5. Encoding Level and Tag Occurrences in WorldCat

Chew Chiat Naun

The encoding level (character position 17 in the leader line of a MARC record) is often used as a guide to the quality of MARC records—for example, as a tool for managing records in library systems. A typical strategy implemented in library systems is to allow a ranking to be applied to encoding levels and to use that ranking as a criterion for selecting among records representing the same item. In other words, the encoding level encapsulates the multiple aspects of record quality in a single dimension. Those aspects of record quality are reflected in the definitions of the encoding levels given on the OCLC web site.[1] These definitions make reference not only to input standards but also to such criteria as the cataloging agency, the degree of verification carried out (e.g., against the actual item or a surrogate), and how the records enter the database (single manual entry or in batch). These criteria are not in themselves a direct measure of quality, but provide grounds for confidence in the quality of the records, and have the advantage of being easier to ascertain at a broad level than quality in absolute terms.

Methodology

In what ways do records at different encoding levels resemble or differ from each other, and what does that tell us about record quality? One way to get a broad view of the characteristics of the records is to look at tag occurrences.

For this project OCLC provided reports on WorldCat records in book, visual material, and music score formats giving, among other data, the number of occurrences and percentage of records with each tag at each encoding level. The data in the tables shown here were extracted from those reports. To make the tables easier to scan visually, the percentage figure for each encoding level was divided by

1. See http://www.oclc.org/us/en/bibformats/en/fixedfield/elvl.shtm and http://www.oclc.org/us/en/bibformats/en/onlinecataloging/default.shtm#BCGGBAFC Note that some of the encoding levels are OCLC-specific. The application of encoding levels in the samples examined for this report reflects the practices of OCLC and its contributors.

the overall percentage figure for the tag. A number higher than 1 means that the tag occurred with above average frequency for the encoding level in question, and these results were marked in blue. A number lower than 1 means a below average number of occurrences, and these were marked in yellow.

The OCLC reports included other data including subfield breakdowns and length of fields, and these data were consulted in a few instances. In one case OCLC also provided us with a random sample of records for a given format and encoding level.

Only encoding levels blank, 1, 3, 4, 7, I, K, L, and M were considered. Encoding levels 5 and 8 were excluded because the records may be considered to be provisional, while level 2 is used only by the Library of Congress, and levels E and J indicate processing status rather than record content.

Reading down each column of a table gives a profile of the records in each encoding level. Reading across the rows of each table gives a profile of the relative strengths of the encoding levels for a given field. Finally, comparisons may be made between the profiles for different formats.

Within each format particular attention was given to three kinds of tags: those that are generally good indicators of record content, such as 6XX subject added entry fields; those, such as the 520 summary field, which are not restricted to any particular format but may nevertheless have special significance for certain types of material; and specialized, format-specific tags such as the 048 number of musical instruments or voices code.

The conclusions drawn in this study should be read with some caution. Differences in tag distributions among encoding levels may be due not to differing degrees of thoroughness in cataloging, but rather to differences in the type of material being cataloged. For example, a difference in the number or type of subject headings found in records at two encoding levels may be due to differing proportions of fiction and nonfiction cataloged at each encoding level. In addition, the unit of measurement used for comparison here indicates only relative, not absolute, frequency of occurrence. It should also be borne in mind that whether a field is mandatory or repeatable will also affect its relative frequency. Finally, it is obvious that tag occurrences alone cannot tell us anything about the accuracy of the metadata—they do not tell us, for example, if controlled-vocabulary headings have been verified.

General Observations

Broadly speaking, the data from OCLC supports the belief that encoding level provides a good guide to record quality as measured by tag occurrences. However, the degree to which it does so varies appreciably. For example, encoding level 4 scores higher than any of the other encoding levels for the 650 topical subject heading field. Factors that influence tag distributions include the type of

material being cataloged, method of input (batchloaded or created online), original record encoding schema (native MARC or crosswalked), and whether a controlled or uncontrolled vocabulary is used.

As expected, subject access correlates strongly with level of cataloging. For books, encoding levels blank, 4, I, and L had the highest occurrence of 6XX subject added entry fields. Visual materials showed a similar pattern, with one important caveat concerning batchloaded records noted below. For scores, it was harder to discern a clear pattern, but Encoding levels blank, 4 and I were again strongest in the significant 600 and 650 fields.

In some instances a given encoding level appears to be strongly associated with certain kinds of material or certain approaches to cataloging. For example, with each of the formats under consideration encoding level 3 has a very high incidence of 720 uncontrolled name headings and 540 Terms governing use and reproduction notes. Sometimes the association is specific to a format. For example, with books and music scores encoding level 7 does appear to represent minimal-level cataloging, but this encoding level has a quite different profile with visual materials (see below).

Certain fields are strongly associated with a given format, or have a distinctive profile within a given format. With visual materials the 520 summary field was present in 63.7% of all records, much higher than the corresponding percentages for books (3.83%) or music scores (1.23%). The breakdowns by encoding level are in the relatively narrow range of 46% (level L) to 91.6% (level 1). 520 is an instance where the length of a field may be a good indicator of record quality. The table below shows the average number of characters in a 520 field at each encoding level. Encoding levels 3 (abbreviated) and 7 (minimal) have figures well below the average, while the other encoding levels are close to the average and, in the case of encoding levels blank, 4, and L, comfortably exceed it.

Table 5.1: Average number of characters in summary field by encoding level

All	Blank	1	3	4	7	I	K	L	M
222.9	472.7	193.6	127.2	303.4	158.5	217.7	218.8	298.7	224.9

For music scores 240 had a high incidence in encoding levels blank, 4 and I, as expected, but also in all batchloaded encoding levels. The relatively even distribution of 240 across encoding levels for music score records is partly a reflection of the fact that 240 is a non-repeatable field, but it tends also to suggest that the high value traditionally placed by music librarians on access by uniform title is widely reflected in music cataloging practices.

The OCLC data provide fertile ground for analysis. The observations recorded here merely scratch the surface.

Encoding Level Blank (Full)

Records with this encoding level showed the expected high occurrences in key fields.

- *Books.* The table shows high values for 6XX, suggesting, as expected, a high overall quality of cataloging. Other fields with relatively high values are 240, 505, and 856.
- *Visual materials.* The key controlled fields (440/830, 650, 655, 700/710/711) show the expected high numbers, suggesting a good general standard of cataloging.
- *Scores.* These records were notable for high values for music-specific fields: 028, 047, 048, and 306.

Encoding Level 1 (Full, Material Not Examined)

For this encoding level the results were mixed, and it is difficult to make generalizations. This is perhaps not unexpected given the origin of these records (primarily from retrospective conversion) and the vagaries of the conversion process.

- *Books.* At this encoding level records for this format showed somewhat higher than average values for some 6XX fields (600, 650, 651) but not others (630, 610). The records were of high quality overall but it is hard to draw firm conclusions.
- *Visual materials.* Fields with very high numbers include 050, 082, and 830; the numbers are also high for 520, 650 and 710 (but not 700 or 711). Again it is difficult to draw conclusions.
- *Scores.* There was a relatively small number of records with this encoding level, and their profile is somewhat unusual. There were very high values for 533, 776, 007, and unexpectedly low ones for 300, 246, and 700. However, the music-specific fields 047 and 306 had the expected high occurrences, and the solid numbers for 100, 240, 600, 650, and 830 are as expected.

Encoding Level 3 (Abbreviated)

Records at this encoding level showed a distinctive profile, low in most access fields but high in those for uncontrolled terms. This encoding level had the highest value for 856 in the music score and visual materials formats, and the second highest in books. Evidently encoding level 3 is heavily used for cataloging projects involving electronic resources.

- *Books.* Records for books showed very high values for the uncontrolled fields 653 and 720. There were high occurrences also of 520 and 856, suggesting projects with specialized focus.

The values for the access fields 6XX, 246, 1XX/7XX, 490/830 were notably low, as they were for 505.

- *Visual materials.* The results were broadly consistent with those for books. There were very high numbers for the uncontrolled fields 653 (19.8%) and 720 (20.6%), and also for 042 (18.0%) and 856 (9.3%). These figures suggest special projects rather than general cataloging. An examination of a sample of encoding level 3 records supplied by OCLC showed some idiosyncratic use of MARC fields, e.g. corporate body names in 260 $a. One may conjecture about the reasons. A possible explanation is that many encoding level 3 records in this sample were not "born MARC" but have been crosswalked from other encoding schema.

- *Scores*. Again there was a relatively small number of records, and they had an unusual profile. There was a high number of 653 and 720 uncontrolled headings, and also of 856 fields. The values for all other 1XX, 6XX, and 7XX fields were low, with the exception of 710. Compared to the other encoding levels for this format, with the sole exception of level 7, level 3 shows a low value for 240. Surprisingly, it has a high number for 246.

Encoding Level 4 (Core)

Encoding level 4 raises interesting questions because although characterized as "core" or less-than-full, many of these records are created under the PCC/BIBCO program, which has a clearly defined set of input standards and requires participants to undergo training and review in their application. The data suggest that encoding level 4 records are in some significant respects of higher quality than most categories of "full" cataloging records. In all three formats studied, encoding level 4 had higher values for 650 than encoding levels blank, 1, I, or L.

- *Books.* Among book records this encoding level had decidedly the highest numbers for key access points in 6XX and 440/830. The values were also high for 880, suggesting a large proportion of foreign-language content, and for 246, 505, and 050. Since 246 often reflects cataloger judgment, the high value for this field may suggest a high degree of expert intervention on the part of catalogers. Ideally we would have studied separate sample of records where the 042 field contained the "pcc" code indicating Program for Cooperative Cataloging (PCC) origins, but we did not acquire such a sample. It seems likely that the overall high standard is attributable to presence of large numbers of PCC Core records.

- *Visual materials.* Results were generally consistent with those for books. The high number for 042 fields (14.6%) again suggests a preponderance of PCC/BIBCO records. Encoding level 4 had the highest combined number for 440/830 and for 650 for the visual materials format. 246 again scored highly.

- *Scores.* Again the high occurrence of 042 fields may indicate a high proportion of BIBCO records. The records appear to reflect the provisions of the BIBCO core standard in the high numbers for the 028 and 505 fields (mandatory if applicable for this format in the BIBCO core standard), but also in the low value for the non-mandatory 047 and 048 fields. 6XX headings other than 630, 653, 655 have high occurrences, as do 8XX series and 246 variant title fields, again suggesting a high overall standard of cataloging. As with most encoding levels for this format, the 100 and 240 fields are relatively well-populated.

Encoding Level 7 (Minimal)

OCLC's definitions allow cataloging agencies considerable latitude in deciding the content of minimal-level records. Nevertheless, the data examined in this study confirm that records with this encoding level include lower numbers of key access tags than full- or core-level records.

- *Books.* The flexibility of OCLC's definition of "minimal" notwithstanding, this term aptly describes the records in this category. They showed very low numbers for 6XX subject fields, whether controlled or uncontrolled, and also low values for 246 and 505. The relatively high values for 007 and 520 suggest materials with particular characteristics, but no attempt was made to sample individual records.

- *Visual materials.* Again there were high numbers for certain fields, such as 655 (3.4%) and 856 (4.9%), suggesting special projects or materials. This conjecture appears to be supported by a random sampling of 1000 records, which showed that 313 records had supplied (devised) titles, as indicated by the presence of surrounding brackets. The records for this format at this encoding level are not particularly deficient in their content. They have more 650 fields than encoding level L or M records do, and they have by far the highest occurrence of 655 of any of the encoding levels for this format.

- *Scores.* Again, the term "minimal" is apt for these records. They have low values for 6XX and 7XX fields, although values for 440/490/830 series fields are relatively high. As with encoding level 3, there is a low occurrence of 240 fields. 246 occurrences are low also. However, 505 (0.90%) has a respectable score. It is difficult to draw overall conclusions about the use of this encoding level for this type of material.

Encoding Levels I, K, L, and M

Encoding levels I, K, L, and M make an interesting comparison because they are distinguished by method of input. I and K represent full- and less-than-full cataloging from direct input into OCLC; L and M are respectively full- and less-than-full cataloging input from a batch process. (Although the

parallels are not exact: for example, I and K are reserved for OCLC participants, while L and M are not.)

In general, irrespective of format, full-level records entered directly into WorldCat by OCLC participants (level I) have more content in fields for significant access points than full-level records added by a batch process. Each of the formats shows high values in most 6XX fields as well as in specialized fields such as 028, 048 and 306 (for music scores), and 520 (for visual materials). Given that catalogers have more control over which encoding level they assign when cataloging online than when loading in batch, this result is unsurprising. But all else being equal, one would expect the strongest similarities to be between I and L on the one hand, and K and M on the other.

This proves indeed to be true for books, and it is partly true for music scores. Encoding levels I and L are strong in the music-specific 048 and 306 fields, for example, and in some of the 6XX fields, notably 600 and 655. Elsewhere, on the other hand, for example in 650 and 700, the differences among encoding levels are not especially pronounced. As we saw with uniform titles, it may be that music cataloging is simply more consistent across encoding levels than is the case with other formats.

Visual materials, however, exhibit quite a different pattern from books or music scores. By far the strongest similarity is between encoding levels I and K. Compare values for the following fields:

Table 5.2: Comparison of tag occurrences for encoding levels I, K, and L in Visual Materials

Field	I (online)	K (online)	L (batch)
043	1.23	1.1	0.87
050	0.82	0.83	0.32
130	0.64	0.64	5.46
246	1.3	1.33	0.89
505	1.15	1.27	0.95
650	1.27	1.1	0.6
655	0.88	1.0	2.73
700	1.23	1.17	0.99
730	1.04	1.2	2.51
856	0.45	0.47	0.05
880	1.66	1.54	0.74

Numbers higher than 1 indicate higher than average frequency;
numbers lower than 1 indicate lower than average frequency.

These results suggest the following conjectures for records representing this format:

- Record content is determined more by method of input than by ostensible quality as defined by encoding level.
- Encoding level may not be an exact indicator of the content of individual records because it may be assigned at a batch or project level.
- Encoding levels L and M may not accurately reflect the full/less-than-full split because of the vagaries of the uploading process.

Even if these conjectures prove to be correct, however, it would still need to be explained why they are true of visual materials and not of the other formats under consideration.

One interesting data point is that encoding level L records have by far the highest occurrence of uniform titles (130: 5.46, 240: 4.69) of any encoding level, and also the highest occurrence of field 306 (playing time). An examination of a sample of level L records might suggest an explanation for these numbers, but such an analysis was not attempted.

Further Questions

Although the results of the present study suggest that the encoding level serves adequately in most situations as an indicator of record quality, there appears to be scope for improvement. Questions for further consideration and research might include the following.

- What rules do libraries apply—and how are they implemented in library systems—when assigning or changing encoding levels during various kinds of cataloging activity, such as original cataloging at different levels, record upgrades, authority control, and automated enrichment?
- Could there be better—simpler, or more explicit—ways of recording information about record quality? If so, would they take the form of a revised set of best practices for assigning encoding levels, or would they be better implemented using a different data element or set of data elements altogether?

Full Data Tables Related to Encoding Levels in WorldCat

Three data tables underpinning this research were too extensive to be included in the body of this report, but are available online (Micorsoft Excel format) at http://www.oclc.org/research/publications/library/2010/2010-06b.xls. Each table is under a separate spreadsheet (tab), which is listed below with the table name.

- Table 5.3: Encoding levels in WorldCat—Books
  Sheet: Books
- Table 5.4: Encoding levels in WorldCat—Visual resources
  Sheet: Visual
- Table 5.5: Encoding levels in WorldCat—Scores
  Sheet: Scores

6. Relator Terms and Form/Genre Designations in MARC Tagging

Timothy J. Dickey

Introduction

The project was initiated by Peter Hirsch of The New York Public Library (NYPL). His experience and interest has been in archival manuscript collections and audiovisual materials, both commercial and archival. The use of relator and form/genre terminology (in fields 655$a; $e for 100/110/700/710) had been a topic at NYPL at various cataloging practices discussions, without much consensus on how they are best used and whether their value equals the effort expended in adding them to a record.

The issues, however, are not confined to the NYPL; both anecdotal evidence and the work of OCLC Program Officers Jackie Dooley and Jennifer Schaffner show a lack of unified approach to such specialized access points in the profession. (See, for instance, Schaffner 2009[1]).

Relatively current developments in cataloging and catalogs may enhance the value of form/genre information, while relator terms will be of more use as cataloging moves in a work-centered direction, where it will likely be increasingly important and useful to give a more granular picture of individuals appearing in 1XX and 7XX fields. Thus, the content of these MARC fields should be an informative case study in actual cataloging usage, within the cataloging community at large and the relatively specialized group of catalogers for whom they potentially have the most value.

1. Schaffner, Jennifer. 2009. The Metadata is the Interface: Better Description for Better Discovery of Archives and Special Collections : Synthesized from User Studies. Dublin, Ohio: OCLC Programs and Research. http://www.oclc.org/programs/publications/reports/2009-06.pdf.

MARC Tags Studied

- 655$a—"A term indicating the form, genre and/or physical characteristics of the materials being described."
- 100$e; 110$e; 700$e; 710$e (as well as the $4 for each field above)—"A designation of function that describes the relationship between a name and a work (e.g. collector, com., defendant, ed., ill., joint author or tr.)."

Methodology

The study used data mining from two different sources: the catalog of the NYPL Performing Arts Library (approximately 750,000 records), and the WorldCat database (130 million records from libraries worldwide at the time). The datasets were then limited to records for works in the following Material Types: music recordings (type j), projected graphics (type g), mixed archival materials (type p), and manuscripts (types t and d).

- OCLC records were only surveyed from 2000-2008, to reflect the state of current cataloging.
- NYPL records were only surveyed from the Performing Arts Library, to capture a data subset in which staff have invested greater effort (by policy) in these tags.

Staff at both NYPL and OCLC mined the content of the MARC fields listed above, to study the presence and character of the metadata:

- Percentage of records in each database which used the fields at all.
- List of all the distinct terms entered by catalogers in them.

The project planned to proceed to compare a list of strings against all subject (and perhaps keyword) searches over a designated period of time, in both databases. This was not to check searches on a complete thesaurus of controlled vocabulary terms (which would evaluate the usefulness of that list). Rather, we hoped to see how many of the actual form/genre and/or relator terms *present in each catalog* (CATNYP and WorldCat) have been searched, and what percentage of total searches they represent. This could demonstrate the usefulness of the specialized metadata that NYPL catalogers have been using to the users, or not.

Limitations

Time pressures on staff, including a complete redesign of the NYPL library system, unfortunately made it impossible to collect data for the second part of the methodology: the search string analysis. Even had time been available, current policies on what transaction log data are collected by both NYPL and OCLC make any conclusions on user preferences difficult. (See "Requirements for Enhanced Library Data Mining," p. 15.)

Results

- In almost every instance, the catalogers at NYPL were using the MARC fields in question with much greater frequency than the profession at large (as reflected in WorldCat).
- The single exception is for the material type "manuscript," for which the aggregate WorldCat data include relatively high use of the fields in question. This may be a reflection of greater specialization across the field for anyone cataloging this material type.
- In addition, the NYPL data were much more streamlined in their choice of terms, using far fewer distinct tags than observed in WorldCat.
- This unfortunately suggests a widespread lack of consistency within the profession at large in thesaurus selection.
- Future research within this dataset could elucidate incidence patterns, potentially to standardize adoption levels.
- These results, positively speaking, reflect the value attached to the presence and consistency of specialized metadata on the part of the NYPL, *by policy in the Performing Arts Library*.
- The researchers believe that this metadata investment is justified by specialized user needs, but cannot at present prove this is so.

Table 6.1: Use of form/genre and relator terminology in NYPL and OCLC WorldCat

	Music Recording	Projected Graphic	Mixed Materials	Manuscript
NYPL Records	137,730	29,837	40,239	20,086
Records w/ 655 field	380	17,808	28,287	14,828
% with 655 field	0.28%	59.68%	70.30%	73.82%
Distinct 655 contents	12	63	54	20
Records w/ relators	111,196	27,078	1216	422
% with relator term	80.73%	90.75%	3.02%	2.10%
OCLC Records (2000-2008)	3,754,098	3,565,310	860,953	100,000
Records w/ 655 field	39,070	390,368	120,410	70,508
% with 655 field	1.04%	10.95%	13.99%	70.51%
Distinct 655 contents	2010	5727	7730	3867
Records w/ relators	276,052	99,506	16,861	58,528
% with relator term	7.35%	2.79%	1.96%	58.53%
Distinct relators	357/788	1744/1722	272/62	348/106