Statistical Analysis of Geographic and Language clues in the MARC record

Technical Report for Going Places in the Catalog: Improved Geographical Access Project. Supported by the IMLS National Leadership Grant for Libraries, award LG-02-02-0035-02

Dec. 08, 2004

Vivien Petras, vivienp@sims.berkeley.edu
School of Information Management & Systems, University of California Berkeley

Summary

This report looks at roughly 5 million unique MARC records from Melvyl, the University of California library catalog. We identified 10 MARC fields containing geographic information and 3 MARC fields containing language information. Some of these fields were never or rarely used in our 5 million record sample pointing to underutilized clues in the MARC record, which could assist in search if enlivened. We've found correlations between place of publication, geographic area and language that could be used to enrich catalog records with missing fields.

Introduction

Library catalog records contain much more than the author, title, and subject keyword fields commonly available to library users for search. Information about geographic scope and about language is available but not searchable although they contain important clues about the subject area of an item. This report looks at geographic and language clues in MARC records (http://www.loc.gov/marc/) and analyzes whether the data is consistent enough to be exploited in search.

We looked at 5,065,574 MARC catalog records from the University of California MELVYL catalog. The entry dates (008$00-05) for these records ranged from 1968-2000 with roughly 160,000 records per year. Most of the records (3.2 million) list as primary cataloging agency (040$a) the Library of Congress, whereas the UC Berkeley Main Library is only listed 18,400 times, showing that most of the records are only adapted for local use.

Relevant geographic MARC fields

We identified 10 MARC fields that could potentially contain geographical information. Table 1 lists the fields and the number of records in our five million sample that contain them.

008	Control field 008		
$15-17	Place of publication, production or execution	5065574	100%
033	Date/Time and Place of an event		
$b	Geographic classification area code	0	
$c	Geographic classification subarea code	0	
043	Geographic area code		
$a	Geographic area code	2341152	46.22%
$b	Local GAC code	61	
$c	ISO code	0	
044	Country of publishing/producing entity code		
$a	MARC country code	16	
$b	Local subentity code	0	
$c	ISO country code	0	

052	Geographic classification	412	0.008%
260	Publication, distribution, etc. (Imprint)		
$a	Place of publication, distribution, etc.	5048397	99.66%
518	Date/time and place of an event note		
$a	Date/time and place of an event note	4	
522	Geographic coverage note	2	
$a	Geographic coverage note		
650	Subject added entry – topical term		
$c	Location of event	122	
$z	Geographic subdivision	1767693	34.90%
651	Subject added entry – geographic name		
$a	Geographic name	919137	18.14%
$z	Geographic subdivision	42136	0.83%

Table 1. Geographical Marc fields. Number of MARC records containing them.

Most fields are repeatable so that the total number of geographic subdivision statements (650$z), for example, is much higher than the number of records containing them.

One can see from table 1 that certain fields are unused, whereas the publication information (008$15-17, 260$) and the geographic coverage information (043$a, 650$z, 651$a) are better represented.

Table 2 shows the top 10 ranked codes occurring in the 008$15-17 Place of publication field and contrasts to it the top 10 ranked phrases in the 260$a Place of publication field. The 008$15-17 field is probably more useful for comparison to geographic coverage (where the book is published compared to the geographic area the book describes). The 260$a field contains the place of publication as it is found in the imprint of the book, this can be more specific than the place of publication codes in the 008$15-17 field.

From 008$15-17 Place of publication				From 260$a Publication		
nyu	656054	New York (State)	12.95%	New York	558238	11.02%
enk	351841	England	6.95%	London	217387	4.29%
gw#	316208	Germany	6.24%	Paris	148452	2.93%
fr#	200586	France	3.96%	Washington	146770	2.90%
ii#	198140	India	3.91%	Moskva	90013	1.78%
dcu	164333	District of Columbia	3.24%	Berlin	54527	1.08%
it#	147982	Italy	2.92%	Boston	40736	0.80%
cau	141036	California	2.78%	Madrid	38150	0.75%
sp#	114428	Spain	2.26%	München	33837	0.67%
mau	111520	Massachusetts	2.20%	Milano	32253	0.64%
Total	2402128		47.42%		1360363	26.86%

Table 2. Top 10 008$15-17 and 260$a Place of publication.

Table 2's 260 column only contains numbers where the city appears in this form, however, imprints can contain more specific location information like a particular institution in a city etc. – this means that there are a lot more books published in these cities than these numbers

state. The top 10 codes in the 008$15-17 cover about half of the records whereas the 260 field is, as could have been expected from the more flexible and specific entries, more diverse.

It is unfortunate that the publishing place codes from 008$15-17 and the geographic area codes from 043$a are different standards and consequently cannot be directly compared. However, we made some interesting observations: we compared the country of publishing code with the geographic area code for books published in Brazil, China and Germany. For all three samples, the overwhelming number of geographic area codes from 043$a was from the continent in which the book was published (95.64% of the 043$a fields of the books published in Brazil contained the code for South America, 96.78% of the fields of the books published in China contained the code for Asia, and 88.9% of the fields of the books published in Germany contained the code for Europe). Not surprisingly for a US-American library, the number for books published in the United States is lower but still quite high: 77% of all books published in the United States with a 043 code given have North America for their geographic area code (74% have the United States for their area code). This gives reason to believe that the place of publishing might be a first approximation of the geographic area the book might cover. Because all records contain publishing information but only half of the records contain geographic coverage information, this indicator might be especially important for finding and enhancing more geographic clues in the record. Nevertheless, we couldn't determine how many of the records not containing a geographic coverage note were describing library items containing geographically relevant descriptions, i.e. it is not clear how many records need to be enhanced.

Tables 3 and 4 show the top 10 ranked terms for the geographic area code (043$a) and the geographic subject headings respectively. Geographic area codes (043$a) are provided in a three-level hierarchy, of which table 3 shows the top level only. It would be interesting to know whether the geographic area code (if given) and the geographic subject heading (if given) contain the same content (one as code and one as term). This is more complicated to prove because the 650 and 651 fields can be repeated whereas the 043 field cannot. Therefore, a catalog record could contain several geographic subject headings but only one geographic area code making the comparison difficult. A quick sample of the top 20 codes in the 043 field showed however that at least one of the 650$z or 651$z fields (if available) will contain a phrase describing the geographic area that is indicated in the geographic area code.

e	944228	40.33%	Europe
n	761018	32.51%	North America
a	321555	13.73%	Asia
s	114697	4.90%	South America
f	106082	4.53%	Africa
u	37883	1.62%	Australasia
c	17380	0.74%	Intercontinental areas (Western Hemisphere)
d	12169	0.52%	Developing countries
p	7331	0.31%	Pacific Ocean
m	6693	0.29%	Intercontinental areas (Eastern Hemisphere)

Table 3. Top 10 top level geographic area codes. Percentages over all records containing 043 fields.

	650$z			651$a			651$z		
Records		1767693	34.90%		919137	18.14%		42136	0.83%
Total		3666252			1345697			87130	
Unique		66922			87754			2774	
Top 10	United States	566970	15.46%	United States	99537	7.40%	United States	13599	15.61%
	Germany	125468	3.42%	Great Britain	35979	2.67%	Soviet Union	4989	5.73%
	France	105696	2.88%	France	29260	2.17%	Great Britain	3071	3.52%
	Great Britain	104807	2.86%	Soviet Union	26239	1.95%	Germany	2959	3.40%
	Italy	102180	2.79%	India	22351	1.66%	China	2476	2.84%
	India	90537	2.47%	Germany	22219	1.65%	Europe	2406	2.76%
	Spain	67627	1.84%	Europe	19645	1.46%	France	2370	2.72%
	Germany (West)	59970	1.64%	Spain	16263	1.21%	Japan	1912	2.19%
	England	55037	1.50%	Italy	15766	1.17%	Poland	1637	1.88%
	Brazil	53694	1.46%	China	14965	1.11%	India	1445	1.66%

Table 4. Top 10 Geographic subject headings for different fields. Percentages over all records for first line, otherwise over the total number that this field occurs.

Almost half of the cataloging records contain a geographic coverage note and a third contain a geographic subdivision. Consistent cataloging practice would be to have both fields filled for records containing geographic information. However, this is not always the case. Table 5 shows how the data overlaps in our sample.

	AND 043$a		NOT 043$a		043$a NOT	
650$z	1650743	32.59%	116950	2.31%	690409	13.63%
651$a	843661	16.65%	75476	1.49%	1497491	29.56%
651$z	40712	0.80%	1424	0.03%	2300440	45.41%
650$z OR 651$a OR 651$z	2089971	41.26%	175177	3.46%	251181	4.96%

Table 5. Overlap between geographic area code and subject added entry – geographic name or subdivision. 1st column: 65* field and 043$a overlap. 2nd column: contains 65* field but not 043$a. 3rd column: contains 043$a but not 65* field. Percentages over all records.

Table 5 shows that half of the records (2,516,329) contain either the geographic area code or a geographic name or subdivision. The overlap between 043$a and the LCSH fields is 83%. Of the records containing geographical information of this type, 10% only contain the geographic area code and 7% only contain a geographic subject heading. However, if one compares single geographic subdivision fields like 650$a to 043$a, then the overlap is only 67%. Here, an automatic enhancement of cataloging records by filling in the empty fields could be useful.

Relevant language MARC fields

We identified 3 MARC fields that could potentially contain language information. Table 6 shows how many records in our 5 million sample contain them.

008 $35-37	Control field 008 Language	5065574	100%
041 $a	Language code Language code of text/sound track or separate title	375589	7.41%
546 $a	Language note Language note	109409	2.16%

Table 6. Language MARC fields. Number of records containing them.

Below are the top ten numbers for language codes in MARC field 008$35-37 ranked by number of records.

eng	English	2872628	56.71%
ger	German	395689	7.81%
spa	Spanish	305580	6.03%
fre	French	272852	5.39%
rus	Russian	189920	3.75%
ita	Italian	139671	2.76%
por	Portuguese	91610	1.81%
pol	Polish	56555	1.12%
dut	Dutch	54436	1.07%
ara	Arabic	46319	0.91%

Table 7. Top 10 languages in the 008$35-37 field.

There are 377 unique language codes in the sample (also counting discontinued codes). The MARC code list for languages contains 486 codes (also counting discontinued codes). More than half of the items described in our sample are in the English language.

041$a is the language code for multiple language entries. The first language code in subfield a is also contained in field 008$35-37. In our sample, 375,589 out of 5,065,574 records contained field 041. However, 128,737 records contained only 1 language code (possibly erroneous entries). Below is a ranked list showing how many records contained how many language codes.

2 languages	213679	4.22%	6 languages	903	0.02%
3 languages	22417	0.44%	7 languages	19	0.00%
4 languages	7106	0.14%	8 languages	2	0.00%
5 languages	2425	0.05%			

Table 8. Number of languages represented in a single language code (041$a) field.

Looking at a sample of the top 10 publishing locations from control field 008$15-17, place of publication and language are clearly correlated. If the book is published in a non-English speaking country, there is a high probability for that book to be published in the official language of that country. Table 9 shows this for 4 non-English speaking countries.

008$15-17				008$35-37		
nyu	New York (State)	656054	12.95%	eng	649867	99.06%
enk	England	351841	6.95%	eng	347878	98.87%
gw#	Germany	316208	6.24%	ger	287112	90.80%
fr#	France	200586	3.96%	fre	186036	92.75%
dcu	District of Columbia	164333	3.24%	eng	160330	97.56%
cau	California	141036	2.78%	eng	139532	98.93%
it#	Italy	147982	2.92%	ita	134992	91.22%
mau	Massachusetts	111520	2.20%	eng	110586	99.16%
ilu	Illinois	110689	2.19%	eng	109521	98.94%
sp#	Spain	114428	2.26%	spa	93566	81.77%

Table 9. Contrasting place of publication and language. Column 008$15-17 shows the top 10 listed place of publication codes in the sample and the percentage over all records. Column 008$35-37 shows the language most often co-occurring in records with this publication place code and the percentage of the records in this language over the all records with this publication place code.

One can also compare language and geographic area code, which we did for Brazil, China and Germany. Here, the situation is not as clear-cut as it is when we compare geographic area to place of publication. Whereas 90% of the books with a geographic area code for Brazil are in Portuguese and 83% of the books with a geographic area code for Germany are in German, only 17% of the books with a geographic area code for China are in Chinese, but 59% of those books are in English - pointing towards an English language bias in this area in the library. However, 82% of the books written in Chinese, for which a geographic area code is given, have a geographic area code for China, 62% of the books written in German have a geographic area code for Germany (and 7% for Austria) and 74% of the books written in Portuguese have a geographic area code for Brazil (and 17% for Portugal) providing an interesting insight into the relationship between language and geographic coverage. For the books published in English with a 043 field filled, 50.06% have a geographic area code for the United States and 10.64% one for the United Kingdom. Besides these 2 countries, over 900 other ones appear in the 043 field for books published in English – the prominent language for a US-American library.

Conclusion

Although there is a surprising high overlap between geographic area code and geographic subject headings, space for improvement is still left, especially if one looks at individual subject fields like the geographic subdivision field 650$z. Both place of publication and language seem to be a good indicator for the geographic area, if this is applicable for an item.

The subject entry fields, place of publication and language could be used to enliven the 043 field and vice versa. However, it is not determined what percentage of records is missing geographical clues (043, 650$z, 651 fields) that should have one of these fields filled.