The Errors of Our Ways: Using Metadata Quality Research to Understand Common Error Patterns in the Application of Name Headings

Katherine M. Wisser

School of Library and Information Science, Simmons College, Boston, Massachusetts, USA
wiser@simmons.edu

Abstract. Using data culled during a metadata quality research project for the Social Network and Archival Context (SNAC) project, this article discusses common errors and problems in the use of standardized languages, specifically unambiguous names for persons and corporate bodies. Errors such as misspelling, qualifiers, format, and miss-encoding point to several areas where quality control measures can improve aggregation of data. Results from a large data set indicate that there are predictable problems that can be retrospectively corrected before aggregation. This research looked specifically at name formation and expression in metadata records, but the errors detected could be extended to other controlled vocabularies as well.

Keywords: Metadata, Quality assessment, Authority control, Data utilization, MARC, Encoded Archival Description, Encoded Archival Context – Corporate Bodies, Persons and Families.

1 Introduction

The dream of aggregating data and providing seamless access to metadata has been realized. That dream, though, illustrates very real issues of data quality that confront the library and archival professions. Incomplete or inaccurate metadata have been a topic of conversation in the bibliographic cataloging world for many years. As early as 1987, aggregating services such as OCLC's World Cat and the former aggregation, RLIN, were assessed for data quality.[1] Efforts to improve the overall quality of the metadata being added to these aggregations were made and overall quality was improved. More recently, large-scale research on the use of content designation in the MARC environment exposed the actual use of specific fields and subfields to open up discussions for extending that functionality. [5] In contrast to the focus on metadata code, though, a large-scale analysis on controlled vocabulary application provides different perspectives on metadata quality.

Research on metadata quality is not new. With the introduction of integrated library systems and the use of digital technologies to represent materials in collections, the quality of metadata has been scrutinized and analyzed and further identified as a problem space that requires attention. As Yasser notes, much recent research has examined various aspects of metadata quality, primarily in the digital

S. Closs et al. (Eds.): MTSR 2014, CCIS 478, pp. 83–94, 2014.

libraries arena. Yasser identifies five categories of metadata problems, including incorrect values, elements, missing information, information loss and inconsistent value representation. Yasser goes on to assert that the identification of problem areas provides metadata projects with the ammunition for preventive and/or corrective measures to enhance their metadata quality. [9] Aggregating data outside of local implementation has exacerbated these issues. Shreeves, Riley and Milewicz coined a new term – "shareable metadata" – to deal with the very real problems associated with aggregating metadata from multiple local implementations. [7] The concept of shareable metadata came from the significant aggregation in the IMLS-funded Illinois Digital Collection and Content project, which explored the use of OAI Metadata Harvesting Protocol to bring together disparate digital collections. [1]

The importance of accuracy in authority data cannot be overstated. Without accuracy, the purpose of authority work is undermined. As Jeng notes in discussing the purpose of authority control, "to be full, useful, and best is to be accurate." [4] Authority control does not operate in a vacuum, however. As Hearn notes, authority records are dynamic as information and perception evolves. [3]

The large data set produced through the Social Network and Archival Context (SNAC) project allows for in-depth analysis of metadata quality, particularly regarding common errors in name formation. Errors such as misspelling, qualifiers, format, and miss-encoding point to several areas where quality control measures can improve aggregation of data. These common error patterns can also be applied to other uses of controlled vocabularies.

2 Sample Description

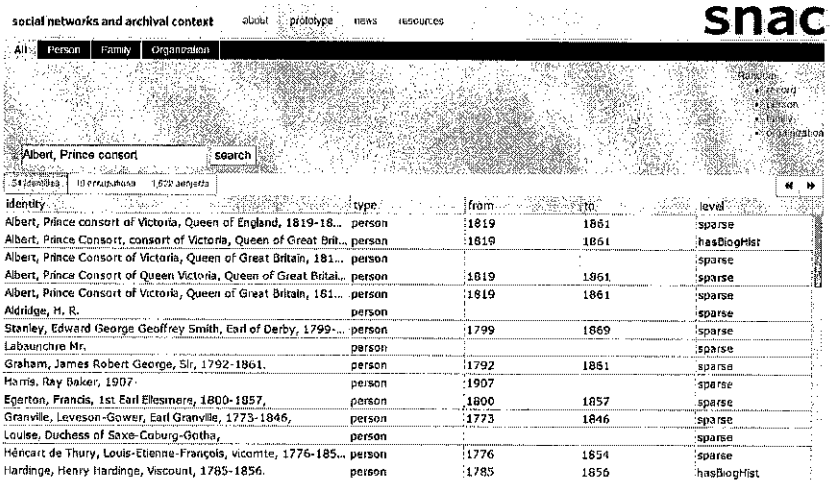
The Social Network and Archival Context (SNAC) project "aims to not only make the records more easily discovered and accessed but also, and at the same time, build an unprecedented resource that provides access to the socio-historical contexts (which includes people, families, and corporate bodies) in which the records were created." [8] SNAC uses automated extraction and merging to generate records to describe corporate bodies, persons, and families. It extracts names from encoded documents provided by large-scale repositories and aggregators. These records are provided in Encoded Archival Description and MARC formats and generally comply with the use of controlled vocabularies. The advantage of this process is that the data can be predictable in form and format. SNAC targeted specific areas of these records for extraction, including those that would most likely use a controlled form of the name (<origination> and <controlaccess> in EAD, 1xx and 7xx fields in MARC) but also other areas where name references would be more freeform (e.g., names within the <dsc>). SNAC stores the information in records using Encoded Archival Context – Corporate Bodies, Persons and Families (EAC-CPF) and creates relationship structures between entities and other entities and entities and resources.

As part of this work, SNAC relies on algorithms and n-gram matching techniques to decrease the number of duplicative records for the same entity. To test the effectiveness of these techniques, two phases of research were conducted. The first phase examined the success of extraction techniques for targeted names and the

accuracy of merging records that represent the same entity. The results showed a high success in extraction and some problem areas for merging (reported in an unpublished technical paper for the project). The second phase, conducted in late 2013, reexamined the merging protocols to see if adjustments improved the merging process. Additionally, new strategies were employed to examine the undermatching of names. One of these strategies entailed the examination of more than 26,000 names in browsing lists.

3 Methodology

A visual scan (and count for descriptive statistical purposes) of alphabetically organized lists of headings provided a specific view of the names in the SNAC data. Recording headings that appeared to represent the same entities illustrated some common patterns to name formation. Figure 1 demonstrates the view of the data as collected to create the sample.



identity	type	from	to	level
Albert, Prince consort of Victoria, Queen of England, 1819-18...	person	1819	1861	sparse
Albert, Prince Consort, consort of Victoria, Queen of Great Brit...	person	1819	1861	hasBiogHist
Albert, Prince Consort of Victoria, Queen of Great Britain, 181...	person			sparse
Albert, Prince Consort of Queen Victoria, Queen of Great Britai...	person	1819	1861	sparse
Albert, Prince Consort of Victoria, Queen of Great Britain, 181...	person	1819	1861	sparse
Aldridge, H. R.	person			sparse
Stanley, Edward George Geoffrey Smith, Earl of Derby, 1799-...	person	1799	1869	sparse
Labaunchre Mr.	person			sparse
Graham, James Robert George, Sir, 1792-1861.	person	1792	1861	sparse
Harris, Ray Baker, 1907-	person	1907		sparse
Egerton, Francis, 1st Earl Ellesmere, 1800-1857,	person	1800	1857	sparse
Granville, Leveson-Gower, Earl Granville, 1773-1846,	person	1773	1846	sparse
Louise, Duchess of Saxe-Coburg-Gotha,	person			sparse
Héricart de Thury, Louis-Etienne-François, vicomte, 1776-185...	person	1776	1854	sparse
Hardinge, Henry Hardinge, Viscount, 1785-1856.	person	1785	1856	hasBiogHist

Fig. 1. Example of undermatching browsing approach, Albert, Prince Consort...

Figure 1 illustrates the way in which the data was collected. In this browse screen, there are several entries for "Albert, Prince Consort of ..." and within them several variations of Queen Victoria. This indicates that there is the potential for multiple records for the same entity (although this example is perhaps more obvious than others encountered). For the purposes of the quality metrics for the SNAC project, this result constituted "undermatching." Over 26,000 headings for records in different initial character strings were examined. The different sets examined were determined based on initial character strings (symbol, A-Adams, Col-Cole, T., Gle, University, and US). Some of those were based on random selection while others were based on the researcher's curiosity. Table 1 indicates the breakdown of those samplings and its percentage as compared to the sample of the whole letter.

Table 1. Sample details

Grouping	Sample size	Whole letter size	Percentage: sample to whole letter
Initial character: symbol	922	922	100%
A-Adams	10,143	84,488	12.0%
Col-Cole, T.	1,525	160,830	0.9%
Gle	1,122	90,319	1.2%
University	12,020	44,756	26.9%
US	455	44,756	1.0%
Total sample	26,187	381,315	6.9%*

* This number, 381,315 of the total set (19.8%), represents the percentage of the total sample of headings examined against the total number of records in the specific letters examined. If this were broadened to the entire data set, which includes 1,922,345 data records, the percentage examined constitutes 1.4% of the data set.

Each of the character string samples demonstrated a series of errors. While SNAC handles records for corporate bodies, persons and families, the analysis of the headings focuses on corporate bodies and persons only. Families were not merged in the data set so they were not analyzed although they are included in the overall statistics reported in Table 2.

Table 2. Overall results from undermatching research

Sample range	Corporate bodies	Persons	Families	Total	
				Groupings	Headings
Symbol	19 (100.0%)	0	0	19	40
A-Adams	90 (34.0%)	174 (65.7%)	1 (0.4%)	265	746
Col – Cole, T.	11 (16.2%)	57 (83.8%)	0 (0.0%)	68	149
Gle	17 (43.6%)	22 (56.4%)	1 (2.6%)	39	106
University	107 (100.0%)	0 (0.0%)	0 (0.0%)	107	222
Us	2 (22.2%)	7 (77.8%)	0	9	20
Total	245 (48.3%)	260 (51.3%)	2 (0.4%)	507	1,283**

** Note: This sample accounts for an average of 2.5 headings in each grouping. This average has little meaning, though, given that the two family names constitute 161 headings. If the families are removed from the total number of groupings and their corresponding headings (505 and 1,122 respectively), the average drops to 2.2.

Personal name headings "groupings" outnumber corporate body headings by fifteen records. When examined more closely, though, two ranges represent only corporate body headings (Symbol and University) and when those two ranges are removed, rather than constituting nearly 48.3% of the potential errors, the number of corporate bodies drops to 31.4%. Correspondingly, personal name errors move from just over half (51.1%) to over two-thirds (67.8%). These results would indicate that while on the surface it appears that the issues were evenly spread across entity types, the issues of consistent name formation are more centered on personal names than on corporate bodies. These results are surprising given that corporate body name formation constitutes very complex rules.

Within the 507 groupings discovered in the analysis of the heading for matching, 35 pairs were exact matches. Exact matches are identical character strings. These pairs were removed from the sample before the analysis on error types was conducted. The results outlined below are based on a sample of 472 groupings and 1,213 headings.

Once the sample was established, the headings were examined for differentiations. Thirty difference types were detected. These types ranged from miss-encoding, spelling and punctuation, the presence or absence of qualifiers, abbreviations, and so forth. Groupings were examined for all instances of difference; therefore, a grouping could exhibit more than one type of difference. Multiple errors occurred in 136 groupings where between two and five differences were identified. In contrast, 336 groupings exhibited only one type of difference.

4 Results

The difference types were first examined as categories and percentages calculated (see Table 3). Encoding errors constituted the smallest percentage of differences at almost 5%; typographical problems and format problem appeared at rates over 10% and 15% respectively. Content differences constituted the largest number of errors at just over 68%. This ratio could indicate that either the actual content is the center of the problem in name formation consistency or that the categorization of problems encountered was overly oriented toward content differences.

Table 3. Differences by categories

Category	Number of occurrences	Percentage of whole sample (n=639)
Encoding	31	4.9%
Possible typographical errors	71	11.1%
Content differences	439	68.7%
Format differences	98	15.3%
Total	638	100.0%

When each category is examined more closely, some surprising issues comes to light. For instance, in encoding errors (see Table 4), MARC encoding problems within the content of the heading are prevalent, such as the presence of subfield letters. These errors indicate that the subfield and delimiter syntax was problematic in the data. The MARC encoding issues are significantly less of a problem, however, than the miss-assignment of the heading type. In the sample, nearly 90% of the encoding errors are attributed to a personal name being coded as a corporate body or vice versa. In addition, the single group that consisted of headings that were neither personal name nor corporate body name entities were encoded as corporate bodies (i.e., Account book, Account books, Account journal, and Accounts).

Table 4. Encoding errors

Specific Error	Number of occurrences	Percentage within category	Percentage of whole sample (n=639)
Erroneous encoding persname/corpname or 100/110	27	87.1%	4.2%
MARC subfield as part of heading	3	9.7%	0.5%
Not a personal name or corporate body	1	3.2%	0.2%
Total	31	100.0%	4.9%

The next category is a set of possible typographical errors, such as punctuation and spelling differences (see Table 5). In this category, there is a differentiation made between misspelled words and spelling differences. With misspelled words, it is clear that a typographical error has taken place. This is particularly true with the corporate body names, where such words as dentistry, information and veterinary are all examples of misspellings, appearing as "dentsitry," "informtion," and "veternary," respectively.

Other spelling issues were less clear. For example, these two headings are part of a group:

"Abbott, John Stephens Cabaot, 1805-1877"
"Abbott, John Stevens Cabet, 1805-1877"

In this example, two of the four names are spelled differently but it cannot be automatically assumed that the names are misspelled, although Cabaot in comparison to Cabet is a little more clear than Stephens and Stevens. Despite these differences, when examined in the light of other evidence it is suspected that these represent the same entity.

Table 5. Possible typographical errors

Specific Error	Number of occurrences	Percentage within class	Percentage of whole sample (n=639)
Punctuation differences	17	23.9%	2.7%
Spelling differences	3	4.2%	0.5%
Spacing	13	18.3%	2.0%
Misspelling	38	53.5%	5.9%
Total	71	100.0%	11.1%

Content differences constitute the largest percentage of issues within this sample (see Table 6). The content differences identified included additional parts to the name, including the inclusion of specific information such as Ltd., LLC, and Inc., and the use of different words that have similar meaning (such as University of Alabama in Birmingham and University of Alabama at Birmingham). Many of the content differences are focused on various kinds of qualifiers and the syntax of those qualifiers as they are included in the heading. In Table 6, the use of the term "qualifier" refers to those expressions in parentheses, such as (Firm) or (Ship). Other additions include the fuller form of name (as expressed in the MARC subfield $q) and dates (as expressed in the MARC subfield $d). Finally, a large number of issues involve geographic qualifiers.

Table 6. Content differences

Specific Error	Number of occurrences	Percentage within category	Percentage of whole sample (n=639)
Additional parts to the name (MARC $a)	33	7.5%	5.2%
Different dates for same entity (data discrepancies)	41	9.3%	6.4%
Presence of dates (MARC $d)	73	16.6%	11.4%
Inclusion of Inc., LLC, Ltd., etc.	21	4.8%	3.3%
Addition of geographic qualifier	37	8.4%	5.8%
Addition of other qualifier	20	4.6%	3.1%
Additional words in the name	28	6.4%	4.4%

Table 6. (*continued*)

Specific Error	Number of occurrences	Percentage within category	Percentage of whole sample (n=639)
Geographic term as part of heading rather than qualifier	6	1.4%	0.9%
Different words, similar meaning	7	1.6%	1.1%
Completeness of geographic qualifier	13	3.0%	2.0%
Co. versus & Co.	1	0.2%	0.2%
Different words	7	1.6%	1.1%
Different non-geographic qualifiers	8	1.8%	1.3%
Fuller form of name (MARC $q)	51	11.6%	8.0%
Initials/abbreviations versus spelled out name	29	6.6%	4.5%
Amount of completion of date different	43	9.8%	6.7%
Non-geographic qualifier term as part of the name (e.g., inclusion of title)	15	3.4%	2.3%
Abbreviations in geographic qualifier	4	0.9%	0.6%
Subordinate corporate body	2	0.5%	0.3%
Total	439	100.0%	68.7%

There is a broad distribution of difference types within the content differences category. The presence or absence of dates is the most common issue in this category, but it still only accounts for just under 17%. Other more common issues include the presence or absence of the fuller form of name and date completion discrepancies. These issues indicate that there is significant misunderstanding or lack of agreement on the rules for the formation of headings rather than careless application. Another explanation for these differences could result from the sources of information from which the headings are formed. Tables 8 and 9 explore in more depth the issues with geographic terms and dates respectively. They are discussed below.

Table 7 illustrates the differences in the format of the heading. The largest group of differences is the application of established abbreviations for relatively common words such as company, department, or street. Date formats constitute over 10% of the problems in this category. This refers to the use of, for example, "b. 1876" versus "1876-". Descriptive standards sanction both formats to express date information, meaning that either heading is not technically an error. Enhanced guidelines would help headings creators understand when one format is appropriate over another and help with the consistency of application.

Table 7. Format differences

Specific Error	Number of occurrences	Percentage within category	Percentage of whole sample (n=639)
Format of dates	12	12.2%	1.9%
Same word, plural/singular	13	13.3%	2.0%
Co. vs. Company, St. vs. Street, Dept. vs. Department	50	51.0%	7.8%
And vs. &	23	23.5%	3.6%
Total	98	100.0%	15.3%

Nearly a quarter of the format differences were identified as the difference between an ampersand and the word "and." Logically, these concepts are exactly equivalent and should be automatically recognized as equivalent. The number of times this issue occurs, therefore, is surprising. Particularly surprising is that 13 of the 23 instances (56.5%) of this issue constitute the only difference detected between the two headings.

Table 8. Differences within geographic terms as part of a heading

Specific Error	Number of occurrences	Percentage within errors with geographic terms	Percentage within class (content differences)	Percentage of whole sample (n=639)
Addition of geographic qualifier	37	61.7%	8.4%	5.8%
Geographic term as part of the heading rather than as a qualifier	6	10.0%	1.4%	0.9%
Completeness of geographic qualifiers	13	22.7%	3.0%	2.0%
Abbreviations in geographic qualifiers	4	6.7%	0.9%	0.6%
Total	60	100.0%	13.7%	9.4%

Geographic term issues account for nearly 10% of all the issues found in the sample. The breakdown, while skewed to the presence or absence of a geographical qualifier, demonstrates a relatively even breakdown of problems. The issue least present, the use of abbreviations in geographic qualifiers (e.g., Tenn. or Tennessee), only appears four times in the sample but is indicative of the other content problems that center on the application of rules to form headings.

Table 9. Differences relating to dates

Specific Error	Number of occurrences	Percentage within errors relating to dates	Percentage within class	Percentage of whole sample (n=639)
Different dates for the same entity, data discrepancies	41	24.3%	9.3% (n=439, content differences)	6.4%
Presence of dates	73	43.2%	16.6% (n=439, content differences)	11.4%
Amount of completion of date different	43	25.4%	9.8% (n=439, content differences)	6.7%
Total for content differences	157	92.9%	35.8% (n=439, content differences)	24.6%
Date formats different	12	7.1%	12.4% (n=97, format differences)	1.9%
Total	169	100.0%	NA	26.4%

Date differences focus more on the actual content rather than the ways in which that content is expressed. Data discrepancies and completion constitute nearly one half of the data issues found. The presence or absence of dates constitutes over 40% and the rest of the date issues are in the ways in which they are formatted for expression. It is clear that aside from accuracy issues (e.g., clear typographical errors such as 100-1993, 1900-1993) are bound to occur and are accounted for as a data discrepancy, but some discrepancies are surprising. For example, John Quincy Adams appears with four different headings, and the dates that appear are listed as: 1767-1848, 1767-1848, 1787-1848, 1797-1848. Given how much is known about the sixth President of the United States, it is hard to reconcile these discrepancies.

5 Discussion

Assessing metadata quality is a challenge for researchers. As Hearn suggests, much metadata quality research is done through the analysis of individual records. [2] Large-scale aggregated data provides an alternative view of data. That view allows for the assessment of common issues. Once those issues are brought to light, data providers can employ local preventive measures before sharing their data in an aggregation. Human error will always be a factor in metadata creation, whether it is through carelessness or lack of standards application. Nonetheless, understanding the nature of errors does provide insight that can help improve the overall quality of the data.

Highlighting error patterns can point to corrective measures that can be taken at a local level to benefit the quality of the data sent to the aggregator, such as better standards adherence, better education of standards, and quality control measures. There are, though, issues that can handled by the aggregator. The latter category includes recognizing equivalences such as "&" and "and." These issues could be resolved automatically in the aggregation. If visual integrity to the original source is a desire in the aggregation, behind-the-scenes equivalence can take place. The same could be said for equivalences between abbreviations and fully spelled-out words (such as Dept. and Department). This approach can be overused, however. Recent justifications in descriptive standard rules, for instance, demonstrate the danger in making assumptions about abbreviations: in the English language, "St." can and does stand for multiple words, such as street and saint. Careful consideration of any automated corrective measures should take place to mitigate the possibilities of erroneous equivalences.

More problematic are the issues that cannot be easily corrected through automated means post-aggregation. There are a myriad of content rules that provide conflicting guidelines on the addition and format of dates, geographic qualifiers, or other types of information. In order to ensure that aggregators are cognizant of the guidelines followed by a particular data provider, it would be useful to know which rules were being followed to establish a particular heading. While many current metadata standard implementations provide this specificity at a record level, the implementation of data components such as the second indicator (and possible corresponding subfield 2) in a MARC 6xx field or the source attribute in Encoded Archival Description can be leveraged to lessen the impact of data that accurately follows disparate guidelines. The use of these data components, while not new to data structure standards, would enhance the possibilities for recognizing that differently structured headings according to different rules belong to the same entity.

6 Conclusion

Aggregating data is one way to address the dispersion of information resources through technological means. But the success of aggregation is dependent on the quality of the data being aggregated. Complicating this process is the very human

element of data creation, standards adherence and the myriad of standards currently in use. These notions are not new. As noted in Shreeves, Riley, and Mileczek, conformance to standards, including descriptive content standards, enhances the shareability of that information. [7] Often, though, this advice is only part of a larger critique of metadata quality rather than the center of it. As data sets get larger, more and more work needs to go into quality control mechanisms and more information needs to be attached to smaller data units. Tools have been developed to assist repositories in the use of standardized vocabularies, but tools alone cannot mitigate against the errors that can occur when data is considered outside of its initial context. A better understanding of the sources of data and the decisions that go into the creation of that data will empower the reuse of that information in multiple contexts.

References

1. Cole, T.W., Shreeves, S.L.: Search and Discovery Across Collections: the IMLS Digital Collections and Content Project. Library Hi Tech 22(3), 307–322 (2004)
2. Hearn, S.: Comparing Catalogs: Currency and Consistency of Controlled Headings. LRTS 53(1), 25–40 (2009)
3. Intner, S.S.: Much ado about nothing: OCLC and RLIN cataloging quality. Library Journal 114(2), 38–40 (1989)
4. Jeng, L.H.: Why authority? Why control? Cataloging & Classification Quarterly 34(4), 91–97 (2002)
5. Moen, W.E., Benardino, P.: Assessing Metadata Utilization: An Analysis of MARC Content Designation Use. In: 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and Application, Seattle, Wash. (2003), http://www.unt.edu/wmoen/publications/MARCPaper_Final2003.pdf
6. Moen, W.E.: Examining MARC records as Artifacts that Reflect Metadata Utilization Decisions. First Monday 11(8) (2006), http://www.firstmonday.org/issues/issue11_8/moen/index.html
7. Shreeves, S.L., Riley, J., Milewicz, L.: Moving towards Shareable Metadata. First Monday 11(8) (2006), http://www.firstmonday.org/issues/issue11_8/shreeves/index.html
8. Social Network and Archival Context, http://socialarchive.iath.virginia.edu/index.html
9. Yasser, C.M.: An Analysis of Problems in Metadata Records. Journal of Library Metadata 11, 51–62 (2011)