Identities and relationships: parallels between metadata and professional relevance

To appreciate the gist of my thinking, some exploration of current metadata issues is in order. Two aspects of metadata—identities and relationships—are central to initiatives both within and outside the library profession. They parallel the familiar cross-references—“see,” identifying a preferred synonym, and “see also,” indicating a relationship—recorded in authority records and controlled vocabularies. Answers to the questions below explore how establishing identities is a fundamental prerequisite for building reliable relationships between specific instances of various kinds of metadata. Following these basics, I delve into current controversies involving relationships—both as a basis for organizing content on the web and for improving cataloging infrastructure—because these have become so entangled. Then, I discuss keywords, often raised as an alternative to metadata, before returning to my suggestion that medical libraries would fare better—together.

Why are identities important?

In any large corpus of material, clarity of reference becomes an issue. To uniquely identify a specific instance of something, assigning a widely accepted label or code serves as shorthand that allows people and computers to easily and consistently refer to the same thing. Notable examples are CrossRef’s digital object identifier (DOI) for intellectual property management and the PubMed unique identifier (PMID) for journal articles. The scope of what is covered by the assigning agency may be a limiting factor, but the need for specific identifiers is universal, extending to events, organizations, works, and especially people.

What are identities for?

Establishing clear identities for people facilitates proper credit for and citation of their work but is a tougher proposition than for articles. While MEDLINE contains more than eighteen million citations, it contains more than sixty-four million occurrences of author names. People have the same or similar names; change names due to marriage, divorce, and gender changes; write under different names for different purposes; use inconsistent forms of their names on different publications; vary the order of forenames and surnames; and use aliases and nicknames. Names also change via translation, transliteration, and typographical error. The potential for confusion increases as people change institutions, participate in interdisciplinary studies, and collaborate interinstitutionally and internationally. The challenge of disambiguating author names will continue to grow as the world’s population accelerates toward eight billion.

What efforts are underway?

Enserink highlighted the challenges to achieving a universal, trusted identifier for researchers in Science [1], resulting in special sections and even issues of journals devoted to this topic [2, 3]. These cite progress, albeit slow, since introduction of a national digital author ID (DAI) in the Netherlands. Commercial systems (e.g., ResearcherID) devised algorithms to help consolidate name variants but resort ed to authors’ voluntary input due to software limitations. The need for broader solutions prompted CrossRef’s contributor ID; the international standard name identifier (ISNI), focused on the media industry; and open researcher and contributor ID (ORCID), aimed at scientific and academic publishing. Expected to debut in 2012, the ORCID identifier registry will emphasize an open and transparent linking mechanism with other schemes. Another contender featuring author input and linking is the PubMed Author ID, on hold as of October 2011 [4].
Don’t libraries already do this?

In response to the Science article, a metadata librarian noted the overlooked Library of Congress (LC) authority file with its LC control number (LCCN) identifier [5]. He contrasted weaknesses in algorithmic and artificial intelligence solutions with the deterministic approach of humans providing precise disambiguation. Library authority records often contain meager information but do reflect an author’s predominant usage over time. Going further, the Virtual International Authority File (VIAF) consolidates authorities from more than 20 additional sources and contains over 14 million name clusters. Although not a panacea, this treasure trove of metadata must be part of a more universal solution.

What about metadata in the broader library context?

Libraries are collections of content organized for use, rather than for sale. These deliberate collections may be virtual (separate, but appearing as one), their content digital (or physical or a mix), and the user unseen, fickle, and changing, but the need for content organization remains constant. The content itself is fixed by the author at the time of creating a work, and fidelity to the original is paramount for honest academic discourse. Updates, derivatives, translations, and new editions constitute different, related works. An explicit title, proper credit for the effort involved, and inclusion of the date of creation make a good start as each work enters the “scholarly record” and serves as fodder for the distillation of knowledge. Additional metadata help reflect the author’s intention and increase the potential for the work’s retrieval in the future, whether tomorrow via a smartphone or 100 years later in a literature search.

Aren’t metadata esoteric?

Most metadata about content (as opposed to technical metadata concerning characteristics of digital files) constitute a limited number of categories of information. The XOBIS project identified ten categories: concepts (topics and categories), strings (keywords and phrases), languages, organizations, events, times, places, beings (people), works, and physical objects [6]. Each instance of any of these needs an identifier and a preferred name, and may have variant names. The resulting metadata are essentially a collection of relationships between one of these elemental categories and any of the ten where applicable. For example, a document’s group of metadata (a catalog record) might include the triplet: Paget disease of bone [work] authored by [relationship] Ladd, Amy L [person]; whereas a surgeon’s group of metadata (an authority record) might include the triplet: Ladd, Amy L [person] specializes in [relationship] orthopedics [concept], and so forth.

Aren’t relationships more complicated?

Such triplets are insufficient alone as relationships often need attributes for precision and clarity. When relationships change, duration matters (e.g., “Dr. Ladd joined Stanford in 1991”), as can their type and strength (e.g., primary versus secondary Medical Subject Headings [MeSH]). Relationships also need their own identifiers and metadata, for example, to indicate a decision to prefer one value over another. MeSH in particular illustrates how metadata with extensive hierarchical (broader/narrower) relationships underpin powerful retrieval mechanisms. Temporal relationships (earlier/later) are useful in tracking changes in names of organizations and places and help avoid anachronism when referring to them. While populating simple structures can become daunting, a good start exists—in library authority files, classification schemes, controlled vocabularies, and field-specific ontologies. Lastly, a schema franca is needed to articulate these identities and relationships to permit distributed metadata creation and to avoid incompatible information silos. The result should be a flexible substrate, allowing creativity within bounds, rather than imposing a rigid, monolithic solution.

What about the semantic web?

To harness the power of unbridled hyperlinking, semantic web initiatives seek to promote more consistent practices on the web to underpin computerized processes. Hyperlinked text references to other web pages are handy for humans but do not convey the “meaning” of the relationship, nor the kind of data in the linked text. To remedy this, a suite of technologies has been developed, starting around 2004. The RDF models information using subject-predicate-object expressions called triples and largely uses identifiers called uniform resource identifiers (URIs) to represent the subject and predicate. To make these simple triples useful, overlaid ontologies using the Web Ontology Language (OWL) are necessary for grouping, ordered sequences, durations, and so on. To avoid duplication, ontologies often reuse other ontologies but can inherit their deficiencies. For example, the Friend of a Friend (FOAF) ontology is a very limited representation of people and relationships and requires extensions to accommodate research interests or graduate students supervised. Overall, the decentralized application of RDF technologies can easily result in vocabulary soup—exposing inconsistencies, unevenness of application, and the resulting uncertainties in data retrieval inherent in such large, cobbled-together structures.

One project to coordinate and aggregate current information about individuals’ identities and research interests is VIVO, which uses RDF. Funded by a large federal stimulus grant, VIVO strives to enable collaboration and discovery among scientists across all disciplines. It builds on a group of academic institutions’ previous efforts to feature their faculty, physicians, and researchers. Bridging differences in practice is inherently challenging. Because the same researcher may have different identifiers at different institutions, the name identifier problem discussed above must be solved. VIVO’s aggregation of ontologies...
allows the same information to be encoded in different ways by various participants and makes it difficult to map data into the structure. The resulting inconsistencies and varying data granularity severely complicate searching of the merged data. How well RDF helps or hinders solution of these problems remains in question.

Is there an alternative to resource description framework (RDF)?

Bing, Google, and Yahoo launched Schema.org in June 2011. This pragmatic approach provides a set of schemas for web editors to embed semantic microdata in web pages for near-term improvement in search engine capabilities (e.g., display). The schemas include events, organizations, places, real and fictional persons, and specific types of works. Although prescriptive, the schemas provide focus, simplicity, and clarity, making them more appealing than RDF-based ones—if fears of centralization can be averted. Of course, there was a subsequent uproar in the RDF community. The schemas’ significance relates less to the adequacy of the schemas, which are compatible with RDF, than to the choice by the world’s three largest search engines not to adopt RDF. This decision is significant for libraries, especially in light of the following.

How do the new cataloging developments fit in?

In October 2011, LC announced its plan for a replacement for MARC that would use RDF as a basic data model, part of a new bibliographic framework for the digital age [7]. This overshadowed the June recommendation to implement the controversial new cataloging rules, resource description and access (RDA) [8]. The equivocal endorsement of RDA included nine changes that “must be completed or underway” for implementation no earlier than January 2013. To accommodate RDA, the already immense MARC formats grew by 15 new fields; the bibliographic format alone comprises around 200 fields and 3,000 subfields. The promise of web-friendly, open, linked data is alluring [9]. However, it is unclear how effectively LC’s initiative can reconcile the complexities of both MARC and RDF. Now would be an opportune time to explore simpler alternatives for mapping data from MARC in the spirit of the new framework, but without the intricacy of RDF. It will be a while before the dust settles.

Why not just rely on keywords?

Keyword searching may be our best friend and our worst enemy. Everyone knows the perils of misunderstandings that occur in conversations, the loss of context in email and texting, and the frustrations of many keyword searches. The ultimate ambiguity may be that works cited and cited works refer to the same title. The difficulties encountered in computer-disambiguation of personal names, and the promotion of the semantic web by its advocates, indicate that “natural” language alone remains insufficient for serious purposes. Much apparent success in keyword searches relies on underlying and not always obvious metadata.

Entering “cancer” in PubMed yields more than 2.5 million citations, although only about 1 million contain the word “cancer.” “Search details” shows precisely how the search was automatically transformed and executed. In addition to improving retrieval, the rich metadata constituting the database allow the latest citations to display first with options to sort by recently added, first/last author, journal title, or article title.

Contrast PubMed’s transparency with a Google “Everything” search, where “cancer” occurs 678 million times. Google’s mysterious proprietary relevance ranking relies on links to a page and more than 200 other “signals.” The Google Books and Google Scholar silos work better because metadata make it possible to distinguish citations from the work cited and to limit by date, but they have their own challenges [10, 11]. Other than the most popular resources, results remain unpredictable.

How does all this relate to our professional identities and relationships?

I hope that the answers to the questions above illustrate the pivotal role that metadata have in the design of flexible and effective interfaces to all kinds of resources. Web interfaces have become surrogates for many library services, and our visibility and reputations will rise or fall with how well these interfaces meet our users’ expectations—heightened by what they see elsewhere. We are at risk of becoming marginalized if we cannot figure out the intersection between the different types of users’ needs and the interfaces necessary to match these precisely to the most appropriate resources—via metadata. By continuing to act separately, I believe we are diluting our ability to accomplish this critical task and thereby flirting with professional irrelevance.

To achieve critical mass and provide a uniform infrastructure supporting such user-centered interfaces, I believe that medical libraries need to virtually consolidate metadata management. Currently, our metadata are either diluted in an ocean or fragmented by institutional vanity. Imagine how significant the results of our combined efforts could be and how that could improve our reputations. To showcase our collaboration, one such resource focusing on medical practice might integrate:

- recent relevance-filtered cataloging
- enhanced authority data, including MeSH
- the relevant subset of recent PubMed citations in English
- faculty profiles from multiple institutions merged with name authorities
- metadata linking to a clinical information repository contributed to and updated by faculty—addressing spiraling costs, while recognizing and sharing faculty expertise...
a mechanism to update or transfer records to a companion retrospective store of lesser-used publications, inactive faculty profiles, and so on.

Thinking bigger and building cooperative relationships could provide direction for our independent efforts to produce intentionally interlocking suites of resources—simultaneously more comprehensive, less redundant, more coherent to users, more economically sustainable, and more rewarding for librarians. Fewer, better solutions are more likely to enhance our current and future prospects.

Quality necessitates coherence. The reproducibility of research is a hallmark of the scientific method. Carefully recorded procedures permit other researchers to evaluate a given study to confirm its veracity and replicate it when there is doubt. Likewise, formal searches of the resulting content and associated metadata, especially those used in evidence-based inquiries, indicate specific sources and provide detailed search strategies to allow others to assess their thoroughness and completeness. Quality metadata are of lasting value. However, metadata formats change over time, much as content carriers change. Considering that our value as librarians relates to how well we can wield our metadata and not to metadata format, perhaps we should ask ourselves:

Are we compromising our professional relevance by ignoring the real challenges?

Dick R. Miller, MLS, dick@stanford.edu, Director for Resource Management, Lane Medical Library, Stanford University, 300 Pasteur Drive #L109, Stanford, CA 94305-5123

Received November 2011; accepted November 2011

References

1. Enserink M. Are you ready to become a number? Science [Internet]. 2009 Mar 27;323(5922):1662–4 [cited 8 Nov 2011]. DOI: http://dx.doi.org/10.1126/science.323.5922.1662.
2. Arendt J, ed., Special section: author identifiers. Cat Classif Q. 2011;49(6):500–37.
3. Topic: organization and people identifiers. Inf Stand Q [Internet]. 2011 Summer;23(3):cited 8 Nov 2011]. <http://www.niso.org/apps/group_public/download.php/7250/isq_v23no3>.pdf.
4. National Library of Medicine. PubMed® Author ID project. NLM Tech Bull [Internet]. 2010 Nov–Dec;(377):e2 [cited 8 Nov 2011]. <http://www.nlm.nih.gov/pubs/techbull/nd10/nd10_pm_author_id.html>.
5. Beall J. Cataloguing names the old-fashioned way. Science [Internet]. 2009 Jun;19:1514–5 [cited 8 Nov 2011]. DOI: http://dx.doi.org/10.1126/science.324.5934.1514-b.
6. Miller DR. XOBIS—an experimental schema for unifying bibliographic and authority records. Cat Classif Q. 2005;39(3(4)):285–303. DOI: http://dx.doi.org/10.1300/J104v39n03_18.
7. Library of Congress. A bibliographic framework for the digital age [Internet]. The Library; 31 Oct 2011 [cited 8 Nov 2011]. <http://www.loc.gov/transition/news/framework-103111.html>.
8. National Library of Medicine. NLM among three national libraries endorsing recommendations for implementation of new RDA cataloging rules [Internet]. The Library; 14 Jun 2011 [cited 8 Nov 2011]. <http://www.nlm.nih.gov/news/rda_cataloging_rules.html>.
9. Council on Library and Information Resources. Linked data for libraries, museums, and archives: survey and workshop report [Internet]. (CLIR pub 152). Washington, DC: The Council; Oct 2011 [cited 8 Nov 2011]. <http://www.clir.org/pubs/abstract/pub152abst.html>.
10. Nunberg G. Google’s book search: a disaster for scholars. Chron High Educ [Internet]. 2009 Aug 31 [cited 8 Nov 2011]. <http://chronicle.com/article/Googles-Book-Search-A/48245/>.
11. Jacso P. Metadata mega mess in Google Scholar. Online Info Rev [Internet]. 2010;34(1):175–91 [cited 8 Nov 2011]. DOI: http://dx.doi.org/10.1108/14684521011024191.