Decentralized Coordination of Controlled Vocabularies

Patrick Golden
University of North Carolina at Chapel Hill
School of Information and Library Science
ptgolden@email.unc.edu

Ryan Shaw
University of North Carolina at Chapel Hill
School of Information and Library Science
ryanshaw@unc.edu

Michael Buckland
University of California, Berkeley
School of Information
buckland@ischool.berkeley.edu

ABSTRACT

Controlled vocabularies may aid coordination within systems while simultaneously hindering coordination across them (Lancaster 1986, 181). Many solutions to this quandary have been proposed, but most of these assume a context in which there is a strong central organization able to impose a coordinating superstructure. In decentralized contexts, exhaustive full-text indexing has largely supplanted attempts to achieve cross-system compatibility of controlled vocabularies. We report on an approach to achieving some degree of cross-system compatibility of controlled vocabularies in a context where 1) there is no centralized control, and 2) full-text indexing alone is inadequate for bringing together related documents.

Keywords

Indexing, controlled vocabulary, decentralization

INTRODUCTION

Achieving compatibility among the controlled vocabularies of independent yet cooperating groups is a classic challenge of information organization (Lancaster 1986, 179–216). Highly centralized, top-down control of indexing terms produces systems that can be searched and browsed consistently, but at the expense of local indexer autonomy and with the attendant loss of specificity and precision. Decentralized, bottom-up aggregation of local indexing terms—in the extreme case, simple full-text indexing across different groups' documents—involves no loss of specificity, but searching and browsing can become more difficult, as related documents may not consistently be brought together. This will especially be the case when groups with differing interests and perspectives use different vocabularies to describe related documents.

77th ASIS&T Annual Meeting, October 31- November 4, 2014, Seattle, WA, USA.

This paper reports on our attempts to address the challenge of coordinating independent controlled vocabularies in a system for organizing and sharing scholarly projects' working research notes. We initially designed the system to have a single, shared, collaboratively authored indexing vocabulary. However, we found that forcing independent projects to share the same set of topics resulted in a number of problems, including clashes over naming policies, synecdochic use of topics, and confusion about the provenance of topics. To address these issues, we developed a new approach in which each project maintains its own independent vocabulary—thus maintaining local autonomy—but can selectively link local terms to shared "hub" topics, encouraging the discovery and shared use of research notes and reducing duplicated work.

SHARED WORKING RESEARCH NOTES

Editors' Notes is a web-based tool allowing humanities scholars to record their working notes. It is the product of collaboration between the authors and historians preparing scholarly, annotated editions of historically important documents ("documentary editions"). The editors and their assistants undertake extensive research in order to clarify the origins, context, and significance of the documents and their contents. They record their progress in working notes.

The Editors' Notes tool is organized around three kinds of records: notes, documents, and topics (Buckland et al. 2014). Notes contain text written by users of the tool and are divided into sections, each of which may cite a document. Documents are records of source material collected by users. Both note sections and documents can be assigned topics. Topics are typically proper names—of people, organizations, places, publications, or events—but they may also be terms for broader themes or phenomena, such as "Women's suffrage." Topics can be related to one another, and they may have scope notes describing them.

Although Editors' Notes provides full-text search, indexing by topic is still essential, because the topic terms often do not appear in the text of the notes or documents (and most documents do not have source text available anyway). This is because the assignment of topic terms reflects not only the content of the notes and documents, but also the scholars' emerging analytic framework for making sense of them. The network of topics is not merely a means for finding relevant content, but is itself useful and meaning content.

Editors' Notes was designed not only to give individual scholars or projects tools to manage their working notes, but also to encourage them to share those notes with others working in related areas. The original editorial projects involved in the development of Editors' Notes were chosen because they had overlapping research interests in figures belonging to related historical milieux. We saw the shared network of topics as a key factor for cross-project pollination.

PROBLEMS OF TOPIC COMPATIBILITY

In the initial implementation of Editors' Notes, four separate projects shared the same set of topics, to which any member of any project could make changes. We quickly ran into problems, however, as projects began creating topics to index their working notes. One problem was that different projects had different naming policies for topics. Naming was a point of contention because naming policies reflect not simply aesthetic preferences, but the ethos of a project. For example, both the Sanger Papers and the Cady Stanton and Susan B. Anthony Papers prefer to use the maiden names of female birth control and women's suffrage activists, where library authority files often only list them under their married names. We addressed these naming issues by allowing each topic to have multiple aliases. This did not entirely solve the problem, however, as the question then arose of which alias to display in contexts showing notes and documents from multiple projects, such as the results of a search.

Another problem was topic synecdoche, or partial overlap of topics. This occurred when different projects used the same topic but implicitly scoped it more narrowly. For example, the Sanger Papers might be interested in Havelock Ellis primarily due to his birth control activism and when creating a topic for him would write a scope note focusing on that aspect of his life. The Goldman Papers may also be researching Ellis, but with a focus on his anti-war activism. Users looking at notes indexed under "Havelock Ellis" may see a confusing mix of related material. The problem is that both projects use "Havelock Ellis" synecdochically, to refer to different parts or aspects of Ellis. One might argue that these should be two separate topics: "Havelock Ellis' birth control activism" and "Havelock Ellis' anti-war activism." However, a focus on birth control is implicit in the editorial mission of the Sanger Papers, and it is unreasonable to expect them to make this explicit in every topic they create. Nevertheless, we needed some way of indicating the different perspectives on a topic, and this was not possible in the original implementation.

RELATED WORK

One solution to the problem of cross-system compatibility of controlled vocabularies is to develop an intermediate lexicon or switching language (Coates 1970, Lancaster 1986). The basic idea is that separate vocabularies can be made compatible by mapping their terms to terms in the switching language. The assumed context is one in which indexing terms are assigned by professionals (librarians) to material that is not their own, for the benefit of a third party (library patrons). Consistency is paramount, finding

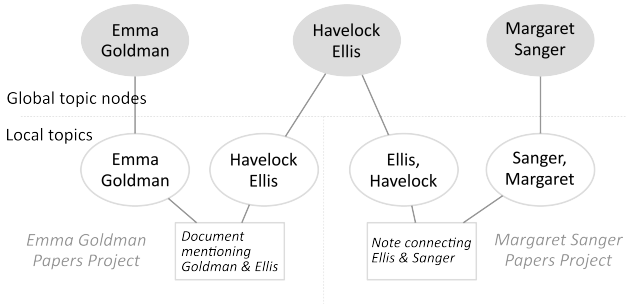

Figure 1. Topic nodes linking project-specific topics.

things is the primary goal, and alternative vocabularies are primarily viewed as routes into the single overarching organizational structure. In the case of Editors' Notes, however, professionals (historians and scholarly editors) select and assign terms to their own material for their own benefit. Though these terms still serve the purpose of helping find material, they are also important for analytic purposes, helping researchers make sense of their material and record the current state of that sense-making.

LINKING TOPICS VIA TOPIC NODES

To address these problems, we changed our data model to better reflect the independence of each project. We developed a hub-and-spoke model of topics that allows projects to maintain their own information while still linking with other projects and eliminating duplicated effort. The new model consists of *topics* and *topic nodes*.

A topic is an indexing term controlled by an individual project. As in the original data model, Topics have a preferred name, zero or more alternate names, and an optional scope note (which may include citations). This data is "owned" by a single project and can only be edited by members of that project.

Each topic has a corresponding topic node. Topics owned by different projects may "point" to the same topic node (see Figure 1). Topic nodes contain linked data assertions about topics (e.g. birth and death dates if the topic is a person) as well as identifiers from external systems such as VIAF (Loesch 2011). Topic nodes also provide read-only aggregations of the data from their associated topics (preferred and alternate names, scope notes, and assignments to notes and documents).Through topic nodes one can trace relationships among different projects' topics.

Topic management process

When adding new topics, users are prompted to connect to existing topic nodes if nodes linked to similar topics already exist. (The algorithm for determining topic similarity is described in the following section.) If the user makes a connection, the new topic will point to the existing node already in use by other projects. If no such node exists or the user declines to make a connection, then a new topic node is created along with the topic.

When a project deletes a topic, the corresponding topic node will be deleted if there are no remaining connections to any other projects' topics. Assertions relating to that node are also deleted since they are no longer being used by any project. (As Editors' Notes keeps all data under version control, deletions are reversible.)

In the course of editing data, projects may make changes to both their local topics and the corresponding topic nodes. Changes to assertions or external identifiers apply to the topic node and are propagated to all connected projects. Changes to topic names, scope notes, and topic assignments are isolated to a single project.

Projects can choose to merge their topics into existing topic nodes in order to connect to the topics of other projects. For example, a researcher might have accidentally added a new topic for Alexander Berkman even though a separate project had already created one. Merging into the existing topic node restores the benefits of connections between projects. Conversely, projects can split topics from their existing nodes. This might be the case, for example, if two topics representing two different people with the same name were erroneously pointing to the same node, or if two topics were accidentally merged.

Browsing and Autocompletion

The split between topics and topic nodes reduces confusion when browsing topics. There are two different browsing views: site-wide and project-specific. When browsing the whole site, users see a listing of topic nodes. Pages for topic nodes

show the topics connected to that node as well as the content and provenance of those topics. For example, the page for the "Havelock Ellis" topic node would separately display the birth control-related scope note and related material from the Sanger Papers and the anti-war-related scope note and material from the Goldman Papers. The name displayed in the heading of a topic node page is derived from the individual projects' naming preferences. If two projects prefer the form "Havelock Ellis" and one prefers the form "Ellis, Havelock (1859–1939)", the former will be used.

Much of the data entry in Editors' Notes relies on autocomplete functionality. To add indexing terms, users are presented with a text input that displays suggested matching topics as a user types. The algorithm used to find matches favors exact matches, but also uses the "fuzzy" text matching capabilities of the Elasticsearch search engine. We have adjusted the algorithm to present a high-recall list of potential matches to prevent missing results. We weigh results to favor a project's own topics, but return potential topic node matches in the case that no project topic matches a query. Giving users this kind of feedback as they are indexing items promotes connections between projects (Voss 2007).

ADVANTAGES OF THE NEW SYSTEM

By separating topics and topic nodes, projects can maintain lists of only those topics that they actually use to index their work, and changes to topics used by multiple projects no longer risk the possibility of accidental data loss or unwanted content. The topic nodes still allow projects to benefit from one another's curatorial work. Only one project needs to add information like name authority identifiers, dates of birth or death, or alternate name forms to a topic node. This information is then shared with all projects connected to that node via one of their topics.

FUTURE WORK

While there was significant overlap in the research domains of the initial four projects, that will not be the case going forward, as we open the site to more users. Currently when a user adds a new topic, all existing topic nodes are checked for similar topics, in order to suggest possible connections. However, this approach will not scale in an open system. To avoid "topic spam," we are exploring a different model in which projects could "follow" one another's topic nodes. When starting a new project, or at any time afterwards, users could indicate which projects they were interested in following—presumably projects closely related to their own. The set of candidate topic nodes available for linking would then consist of all the topic nodes associated with those followed projects. In this model links between topics and nodes could be asymmetric, so that even if Project B followed Project A and linked its topics to Project A's nodes, Project A could decline to follow Project B and display Project B's topics alongside its own.

ACKNOWLEDGMENTS

We are grateful for the support of the Andrew W. Mellon Foundation for supporting the Editorial Practices and the Web project.

REFERENCES

Buckland, M., Golden, P., Pateman, B., & Shaw, R. (2014). Editors' Notes: An example of changed mediation. In J. Boustany, E. Broudoux, & G. Chartron (Eds.), *La médiation numérique: renouvellement et diversification des pratiques* (pp. 143–154). Bruxelles: De Boeck.

Coates, E. J. (1970). Switching languages for indexing. *Journal of Documentation*, 26(2), 102–110. doi:10.1108/eb026489

Lancaster, F. W. (1986). Vocabulary Control for Information Retrieval (2nd ed.). Arlington, Virginia: Information Resources Press.

Loesch, M. F. (2011). VIAF (The Virtual International Authority File) – http://viaf.org. Technical Services Quarterly, 28(2), 255–256. doi:10.1080/07317131.2011.546304

Voss, J. (2007). Tagging, folksonomy & co - Renaissance of manual indexing?. *arXiv.org*, arXiv:cs/0701072v2 [cs.IR].