Editorial Control over Linked Data

Ryan Shaw
University of North Carolina at Chapel Hill
100 Manning Hall, Chapel Hill, NC 27599-3360
ryanshaw@unc.edu

Michael Buckland
University of California, Berkeley
102 South Hall, Berkeley, CA 94720-4600
buckland@ischool.berkeley.edu

ABSTRACT
Linked Data promises to facilitate the collaborative production and use of structured information about historical people, places, organizations, events, and ideas. But few processes have been established to assess and improve the quality of Linked Data. Documentary editors can potentially contribute to this effort by assessing Linked Data quality, publishing "gold standard" datasets that meet their high standards, and connecting assertions to bibliographic descriptions of evidential resources. The Editors' Notes project is exploring these possibilities by integrating tools for harvesting and editing Linked Data into the research processes of documentary editing projects.

Keywords
Linked Data, documentary editing, digital history.

LINKED DATA
Linked Data is gaining currency as a set of techniques and technologies for publishing and connecting structured data on the Web (Bizer, Heath & Berners-Lee, 2009). Recently libraries have become interested in publishing as Linked Data information about their collections and the people, places, organizations, events, and other subjects related to them. The Library of Congress has begun publishing its authorities and vocabularies as Linked Data,[1] the Virtual International Authority File (VIAF) publishes as Linked Data the results of its efforts to merge various national library authorities,[2] and the W3C has formed a Library Linked Data Incubator Group to examine how other kinds of library data can best be published on the Web.[3] Archives and museums have been slower to embrace Linked Data, but the recent International Linked Open Data in Libraries, Archives, and Museums Summit[4] demonstrated that the interest is there.

Meanwhile, outside the world of libraries, archives and museums, new open sources of structured data have emerged. The DBpedia project mines information from Wikipedia "infoboxes" and publishes it as Linked Data (Auer et al., 2007). The GeoNames geographical database integrates various geographical data sources, incorporates corrections and additions from users, and publishes the results as Linked Data.[5]

Finally, scholarly projects that produce databases as their primary products, such as prosopographies and historical gazetteers, are also beginning to experiment with publishing Linked Data. One example is the Pleiades project, which essentially does for ancient places what GeoNames does for modern ones, except with an added layer of scholarly editing and control.[6]

The exciting prospect of integrating structured information about people, places, organizations, events, and ideas from libraries, archives, museums, scholars and the general public is driving much of the interest in Linked Data. However there are also concerns about the quality of the data being produced as these various sources are merged (Dodds et al., 2011). Scholarly projects can potentially provide "gold standard" Linked Data, but many of these projects do not yet focus on linking to or assessing data from other sources.

DOCUMENTARY EDITING
Documentary editors prepare "editions" of documents such as letters, diaries, and essays that have value as evidence for political, intellectual, or social history (Kline & Perdue, 2008). Documentary editors communicate that value by contextualizing these documents, identifying and explaining the people, places, organizations, events and ideas to which they refer. In edited volumes, these various entities are linked to each other and to sources in libraries and archives via a rich web of explanation and commentary.

Because documentary editors seek to provide the best possible information about the entities relevant to the scope of their projects, they are ideally suited to act as filters and expert editors of Linked Data. By integrating tools for working with Linked Data into their normal research and fact-checking processes, editorial projects can potentially become sources of very high-quality Linked Data, where every assertion has been checked and either associated with a credible source or flagged as "dubious."

ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

[1] http://id.loc.gov/
[2] http://viaf.org/
[3] http://www.w3.org/2005/Incubator/lld/
[4] http://lod-lam.net/
[5] http://www.geonames.org/
[6] http://pleiades.stoa.org/

Providing quality control for Linked Data is useful enough, but integrating Linked Data into editorial practices has additional benefits as well. First, there are benefits for the editors. Editors are interested in using digital tools to aid their work and produce additional forms for disseminating it. For example, editorial projects maintain comprehensive chronologies listing where and when people were. This kind of data is ideal for presentation through interactive maps and timelines.[7] But entering data such as coordinates of places can be tedious. If this kind of data can be pulled from external Linked Data sources, it can save editors and their assistants from unnecessary data entry. Likewise, structured data from external sources can be used to add features like faceted browsing of topics, without editors having to add all the facet values (birth dates, death dates, locations, nationalities, etc.) themselves.

Another benefit is that by mapping their topics and entities to identifiers for those topics and entities elsewhere, editorial projects can make their research products more widely accessible. Documentary editors produce a wealth of material in the form of research notes, reports, and chronologies. Most of this material never leaves the disk drives or manila folders of editorial projects. Limitations on space in paper-based editions force editors to pare their contextual contributions down to a highly polished minimum. Left on the cutting room floor are "dubiosa" (statements that have not been verified to the editors' satisfaction) and material deemed not sufficiently relevant to that project's focus.

In a networked, digital environment, it often makes sense to publish dubiosa in the hopes that someone else may be able to find the missing evidence needed to verify or falsify them. Even when no such evidence can be found, dubiosa may prove useful to others despite their uncertain status. Likewise, what are ephemera from the perspective of a given editing project may prove critical to others. The Web and Linked Data have the potential to dissolve the silos that isolate research in separate editing projects and to enable the serendipitous discovery of useful material by other scholars and the wider public. By linking their research notes to external identifiers for people, places, organizations, events and ideas, editorial projects make it far easier for their work to be incorporated into other scholarly projects, library catalogs, archival finding aids, and open knowledge projects such as Wikipedia.[8]

THE EDITORS' NOTES PROJECT

Editors' Notes is an experimental hosted service, funded by the Mellon Foundation, that documentary editors can use to manage the process of collaboratively discovering, assessing, and disseminating this contextual material. Just as in paper-based editions, the site allows editors to publish the transcribed texts of original documents, with lengthy footnotes explaining anything that needs explaining according to the editors' judgment (see Figure 1). In addition to the footnotes, editors can publish dedicated articles on specific people, places, organizations, events, and ideas. Finally, the site enables high-resolution scans of the documents to be published as well.

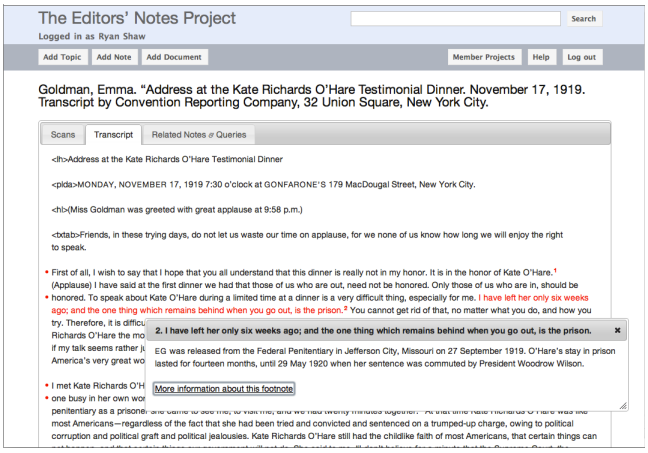

Figure 1. A footnoted transcript in *Editors' Notes*.

However, the focus and purpose of the site is not to simply reproduce in digital form the features of paper-based documentary editions. It is intended to be a platform for publishing, sharing, and editing notes throughout the research process, from the initial posing of specific research questions all the way through to the authoring of polished footnotes and articles. The site is organized around *Notes*, *Documents,* and *Topics* (see Figure 2).

A *Note* is any kind of research note written by an editor. The text is stored as HTML, so it may have hyperlinks and all the other features that HTML enables. Editors can use a WYSIWYG interface to easily edit Notes, and all past

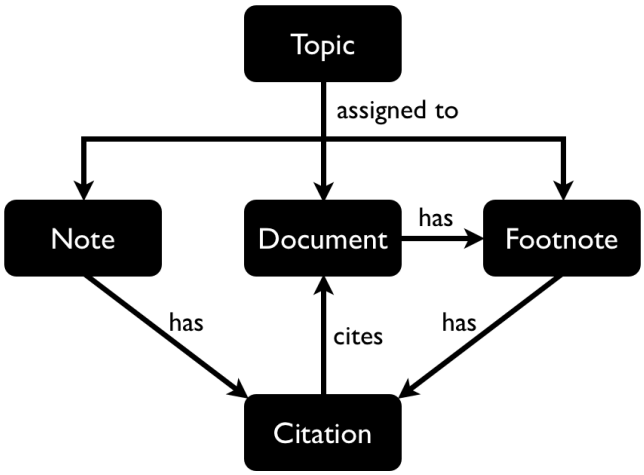

Figure 2. The Editors' Notes data model.

[7] See http://metadata.berkeley.edu/emma/ for an example from the Emma Goldman Papers.

[8] In a previous project, we observed the need for better use of shared identifiers to enable linking across diverse genres of scholarly resources (Buckland, 2011).

versions of edited notes are saved. A *Document* is anything that an editor is editing (e.g. the letters, diary entries, speeches, etc. that are the focus of the project) or is citing (any supporting evidence found in the course of the editor's research). As described above, Documents may have transcripts or scans associated with them. Transcripts can be footnoted, and footnotes can cite other Documents. A *Topic* is a controlled vocabulary term such as a person name, an organization name, a place name, an event name, a publication name, or the name of a topic or theme. Notes and Documents can have multiple Topics assigned to them, and relations among Topics are created via Notes.

HARVESTING AND EDITING LINKED DATA

The tools for working with Linked Data in *Editors' Notes* consist of a harvester and an editor. The harvester runs periodically, looking for new Linked Data related to Topics. For each Topic name, the harvester queries a *reconciliation service* to obtain sets of candidate identifiers (URIs) for that name. A reconciliation service is a web service that, given a name or label, returns identifiers for entities that potentially match that name or label. In this case, we use the SameAs reconciliation service[9] to obtain sets of candidate identifiers (URIs) for each Topic name.

The SameAs service provides zero or more sets of URIs in response to each query. Each set contains URIs that have been asserted to refer to the same entity. For example, when queried with the name "Emma Goldman," the first set of URIs returned by the SameAs service includes the identifiers for Emma Goldman from VIAF[10] and the Freebase structured data repository.[11] For each set in the response, each URI in the set is dereferenced and the resulting data is examined. If, for any of the dereferenced URIs, this data includes a valid label, and the value of this label matches the Topic name, then the whole set of candidate URIs is accepted. Otherwise, the set is rejected and the next set is examined.

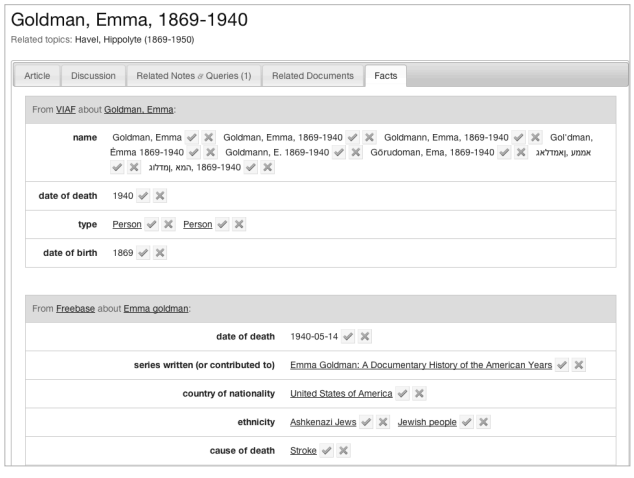

Figure 3. Assertions harvested for the Topic *Goldman, Emma, 1889-1940*.

In this way, the harvester obtains a set of zero or more URIs for each Topic. The harvester then stores all assertions obtained by dereferencing the URIs. Each assertion is stored in two separate graph databases: one database contains all the candidate assertions about a given Topic, while the other contains all the assertions from a given source (e.g. DBpedia, VIAF, Deutsche Nationalbibliothek, etc). The Topic-specific databases make it simple to display all the assertions found for a given Topic (see Figure 3), while the source-specific databases make it easy to request fresh data from a given source.

Once a set of candidate assertions about a Topic has been obtained, editors can use the Linked Data editor to accept or reject them. Accepting and rejecting assertions can happen at different levels of specificity. An editor might reject a single assertion that she judges to be inaccurate. Or she may choose to reject all assertions that share a given predicate that is judged irrelevant to the editing project. For example, many DBpedia resources contain assertions about what templates are used on their corresponding Wikipedia pages, and this information is not likely to interest editors. Finally, she may accept all the assertions about a given Topic, or all the assertions from a given source.

When an editor accepts assertions from a given source, this is treated as evidence that the identifier from that source refers to the same entity, and an `owl:sameAs` assertion is created linking the Topic to that identifier. Thus, the process of accepting assertions has the effect of linking *Editors' Notes* Topics to standard identifiers in external systems.

Accepted assertions are inserted into a graph database associated with the editor who accepted them. This way the provenance of published assertions can be made clear, and editors can choose whether they need to further assess assertions accepted by less expert contributors (i.e. student assistants).

FUTURE WORK

The integration of Linked Data tools into Editors' Notes is relatively new. At this point, we are still assessing the usefulness of this approach, and whether the benefit of additional structured data is worth spending editors' (or their interns') time accepting or rejecting it.

If the approach is deemed useful, there are a number of areas for future research and improvements. First, we are eager to analyze what kinds of assertions are accepted and what kinds are deemed irrelevant or trivial. Grand claims are often made about Linked Data but there have been few studies looking at just what kinds of data are worth linking.

Second, we would like to develop easy-to-use tools for adding additional assertions, beyond simply accepting or rejecting existing assertions. Many of the Topics in the current system are esoteric and do not have any

[9] http://sameas.org/
[10] http://viaf.org/viaf/39377930
[11] http://rdf.freebase.com/ns/en.emma_goldman

counterparts in existing Linked Data sets. We hope to use the set of harvested predicates to create a simple auto-completion interface that allows editors to quickly add structured data about Topics without having to learn complex vocabularies.

Finally, we would like to experiment with ways of linking accepted and added assertions to bibliographic citations. Editorial projects are scrupulous about always citing their sources, and this practice should apply to their Linked Data as well. *Editors' Notes* integrates with the Zotero bibliographic data management services, so it is possible to imagine linking individual assertions about topics to Zotero records for individual sources. This could give students and researchers a powerful way of assessing the credibility of given source as judged by expert editors.

ACKNOWLEDGMENTS

We are grateful to the Andrew W. Mellon Foundation for funding "Editorial Practices and the Web" (http://ecai.org/mellon2010). The project has benefited greatly from the contributions of Patrick Golden. And of course, none of this work would be possible without the cooperation and feedback of our colleagues at the Emma Goldman Papers, the Margaret Sanger Papers, the Elizabeth Cady Stanton and Susan B. Anthony Papers, and the Labadie Collection.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A nucleus for a Web of open data. In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, et al. (Eds.), *The Semantic Web* (Vol. 4825, pp. 722–735). Berlin: Springer. doi: 10.1007/978-3-540-76298-0_52

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. doi: 10.4018/jswis.2009081901

Buckland, M. (2011). Final Performance Report on *Context and Relationships: Ireland and Irish Studies*, NEH award PK-50027-07. University of California, Berkeley. http://metadata.berkeley.edu/neh2007finalreport.pdf

Dodds, L. et al. (2011). Quality indicators for linked data datasets. *SemanticWeb.com.* Accessed July 21, 2011. http://answers.semanticweb.com/questions/1072/quality-indicators-for-linked-data-datasets

Kline, M.-J., & Perdue, S. H. (2008). *A Guide to Documentary Editing*. Charlottesville: University of Virginia Press.