Integrating Tools for Synthesis into Digital Libraries

Ryan Shaw, Michael Buckland and Ray Larson
School of Information
University of California, Berkeley
Berkeley, California, 94720
{ryanshaw, buckland, ray}@ischool.berkeley.edu

ABSTRACT

Second-generation digital libraries aim to go beyond providing access to resources, toward integrating tools and services for exploring the content of resources. Much of the work in this area focuses on tools for analyzing data sets such as numeric observations or text corpora. Our work, in contrast, focuses on tools for synthesizing an understanding of mentioned but unexplained names or terms. We argue that these *self-service reference tools*, by making connections across separate collections, complement analytic tools that uncover patterns within a collection. Finally, we demonstrate working prototypes that illustrate the basic principles of self-service reference.

1. ANALYSIS AND SYNTHESIS

Research into integrating digital library content with computational tools and services has been primarily concerned with examining, analyzing, and finding patterns within a digital data set. A digital data set is ordinarily enhanced by adding formal elements to aid analysis. These elements allow researchers to identify patterns within a data set and to navigate among these patterns and structures. Such elements include markup, such as TEI, or features designed for statistical machine learning, data-mining, and other techniques. With markup, the patterns are explicitly revealed in databases representing structures of relationships, while statistical techniques uncover latent patterns among the defined features.

Traditional library work, however, tends to focus on synthesis. Some work is concerned with analyzing individual data sets, as when cataloging a collection or compiling a bibliography, but the focus on a data set is qualified by an effort to make what is inside the data set interoperable with materials elsewhere through standardized formats and standard descriptions. Other library work, notably reference work and much bibliographical work, looks outwards to find related material elsewhere.

Like the elements added to enable analysis, the markup and metadata added by librarians not only describe documents but also form structures for finding documents, as in information retrieval, or for discerning patterns within a population of documents, as in bibliometrics. Some useful kinds of metadata (such as titles and bibliographical citations) are not added by librarians but can be usefully exploited by them, for example, expanding subject access with keywords derived from titles and creating citation indexes to identify related prior work, co-citing research fronts, and interdisciplinary connections.

Scholarship involves an interplay between analytic and synthetic modes, between representing the structure of data and building structures to connect data [5]. There is a large opportunity for linking analysis-oriented tools and services with the synthesis-oriented tendencies of library and bibliographical efforts. In both cases markup and metadata play dual roles as descriptions and also as facilitating infrastructure [2].

A major barrier to closing this gap is that researchers typically focus their structure-building efforts on individual data sets, while librarians strive to create structures that traverse a wider documentary universe. Researchers might focus on producing a data set that is richly structured internally, yet relatively disconnected from the external universe of documents. Librarians, on the other hand, tend to treat such a data set as a unit, adding some descriptive metadata to give it a place within the larger set of resources, but failing to exploit the richness of its internal structure. Changing this situation requires better practices on both sides. Researchers need to think more about how the elements they produce will participate in the world outside the scope of their current project, while librarians need to be more aware of the internal structure of the resources they organize. Both need to appreciate the potential benefits.

2. SELF-SERVICE REFERENCE

Our current research is based on the principle that the difference between seeing and understanding lies in knowing the context and relationships of whatever is of interest. This implies that identifying relationships between items in a data set is important, but understanding also depends on being able to relate any item to resources in the wider environment outside the data set capable of providing explanations and relationships.

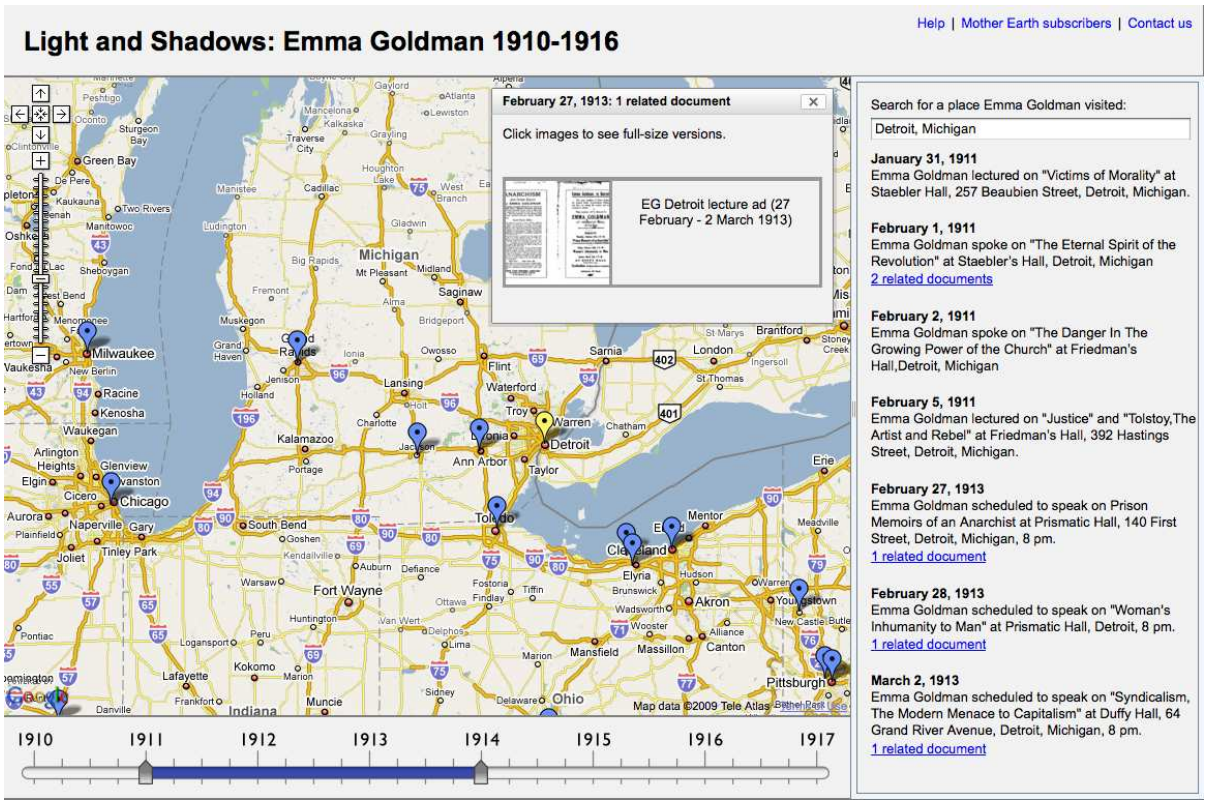

Figure 1: Interface to an RDF data set of events in Emma Goldman's life.

Reference libraries can be very good resources for acquiring some knowledge of the context and relationships of things. Through its selection and organization of resources, the reference library enables patrons to conduct initial investigations and orient themselves to unfamiliar topics. Yet research into reference service in a digital environment has mainly focused on librarians answering questions for patrons, rather than on the design of services for efficient and effective self-help [4].

Ideal self-service reference tools do not merely list resources, but empower users to select the best specific resources or fragments of resources to achieve their ends. Such tools need to perform a number of related functions. First, the tools need to indicate to users which reference resources are relatively trustworthy. Then, for any specific query, they need to indicate which resources are likely to yield useful results. In order for these results to materialize, the tools must address the differences in descriptive vocabulary between different resources. They also need to facilitate the comparison of results from several resources, so that users can verify and cross-check the explanations they receive. Finally, they must be very easy to use, or they won't be used.

3. PROTOTYPES FOR DEMONSTRATION

We are building tools for enhancing digital documents in the humanities domain by identifying references to people, organizations, places, topics, or events, disambiguating these references by linking them to identifiers from naming authorities, and using the disambiguated references to provide links to related explanatory resources available on the web. Distinguishing referenced entities by facet greatly facilitates search. Standard natural language processing techniques can analyze the text to identify entity references, and the reader can also use a web browser-based annotation tool.

Our designs are being developed in two parallel projects: one "Bringing Lives to Light: Biography in Context"[1] uses the condensed texts of biographical dictionaries; the other, "Context and Relationships: Ireland and Irish Studies"[2] uses digitized back runs of journals on Irish culture and history.

Initially we developed three "proof of concept" prototypes each focused on a different requirement. The first, based on a chronology of Emma Goldman's lecture tours, looks into a data set, adding markup to support geo-temporal navigation to identify for individual speeches illustrative materials within the Emma Goldman Papers. The Emma Goldman Papers editors, like many editors of historical papers, maintain a day-by-day chronology detailing where Emma Goldman and her associates were and what they were doing. This chronology serves as an internal reference tool for the project, allowing the editors to make inferences about when or where documents may have been produced and to check for inconsistencies in historical accounts. But as it runs to thousands of entries, it is not available in the printed volumes, despite the fact that it could be a valuable reference for Emma Goldman scholars outside the project as well.

[1]http://ecai.org/imls2006/
[2]http://ecai.org/neh2007/

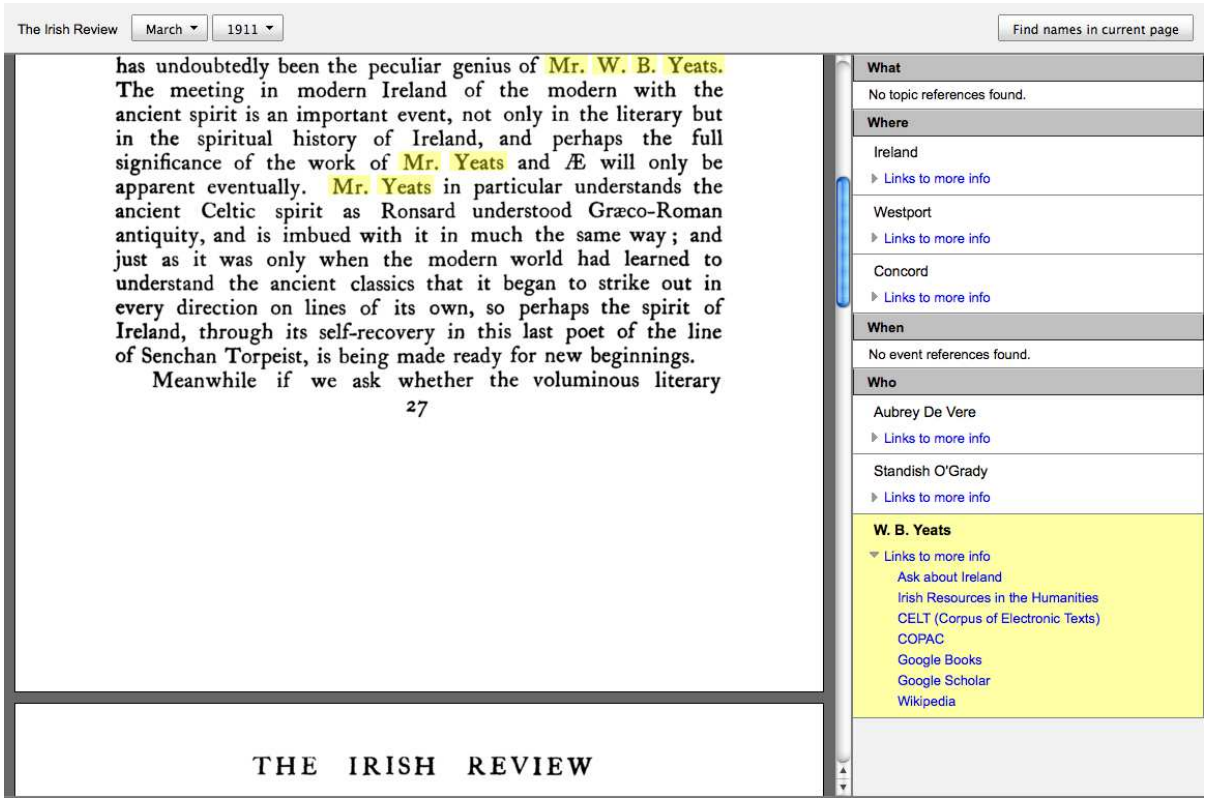

Figure 2: Integrating external resources into a digital library of full-text documents.

Starting with a text document containing the chronology for the years 1910 through 1916, we produced an RDF data set by parsing dates, geocoding place names, and disambiguating personal names by linking them to DBpedia. We then scanned a selection of documents from the Emma Goldman Papers' archives, put them online, and manually linked these scans with relevant events from the chronology. Finally, we developed a browser-based interface for exploring the RDF data set (Figure 1). All the data extracted from the chronology is encoded in the interface's HTML using RDFa [1], and all the querying is done locally. Essentially it is a self-contained, human- and machine-usable data set of information about events in Emma Goldman's life. Yet because the information is embedded using the RDFa standard, we can easily extract the RDF triples from the HTML page and do something else with it, such as merge it with another data set or use SPARQL to query it in ways that aren't possible through this interface. From the Emma Goldman Papers' point of view this approach is very low-maintenance: since this is a static HTML file it just needs to be placed on a web server; there is no need to maintain a database or a dynamic web application framework.

The second prototype (Figure 2), using pages of Irish periodicals and modeled on how reference works are used, is outward-looking, supporting real-time queries to appropriate specialized external resources whenever the reader desires an explanation of a personal name, place name, topic, or event mentioned in the text. The goal is to demonstrate how external resources might be made more readily available in a digital library of full-text documents. From our partners at Queen's University Belfast we have a collection of approximately 50,000 scanned pages of journal articles, from the 1780s to the present, focusing on Ireland. To these we have added the full text of 100 books on Irish history and culture from the Internet Archive's book-scanning project.

We aim to integrate three kinds of functionality into this library, which we call context finding, context building, and context providing. Context finding means a reader should be able to move from a name in a text to resources that describe, explain, or otherwise provide resources for understanding that name. One might think of it as moving from names in texts to entries in reference works that explain those names. Context building means the ability to add links to such resources to the texts. In our systems context building is done both manually (via annotation) and algorithmically (using named-entity detection). Finally, context providing means aggregating across a corpus of texts all the references to a particular name. This is an inversion of context finding, and one might think of it as a union index that supports moving from an name in a reference work to a list of locations in texts where that name is used.

Searches depend on unambiguous queries, so the third prototype, applied to biographical articles in Wikipedia and similar sources, focuses on invoking established naming authorities to disambiguate personal names (e.g. *Which King Charles?*). To resolve ambiguities a simple interface links to the appropriate authoritative identifiers. For names of persons and organizations, these include Library of Congress and Deutsche Nationalbibliothek authority files and World-

Cat Identities URIs. For place names we use identifiers from the GeoNames geographical database. We are using Wikipedia and Freebase URIs to identify historical events for which no established naming authority currently exists. The system is extensible: new sources of identifiers can be added.

After entities have been unambiguously linked to authoritative identifiers, these identifiers are used to find additional information about them. For example, a linked map and timeline displayed alongside a document may indicate the locations of places and the locations and times of events referenced in the document. Access to unstructured information and related documents can be provided by using identifiers or aliases to construct dynamic links representing queries on appropriate reference sources or special collections. We are currently working on representing all this auxiliary information—disambiguated names and relationships, factoids, and links to relevant resources—using the same Linked Data[3] approach taken in the Emma Goldman prototype. This will enable us to combine the functionality of all three prototypes in future versions.

4. FACILITATING SYNTHESIS

Identifying and disambiguating references to entities within data sets, so as to represent the interrelationships among these various entities and thus provide richer context for interpretation, is common practice in computational research projects. Typically such projects mint their own identifiers for these entities, which is sufficient for identifying entities within the scope of that project and within that data set, but fails to support their use to create bridging links to other, external resources concerning those entities. Use of shared naming authority services would go a long way towards ameliorating this situation. Librarians have deep experience building such services, yet matching queries to unfamiliar vocabularies on external network-accessible resources has received relatively little attention [6, 3].

In the humanities, place name gazetteers, encyclopedias, biographical directories, dictionaries of concepts, and other long-established reference resources are services for finding and providing contextual background. The elements needed for such services, including identifiers for entities and concepts of interest and various kinds of semantic relationships for linking them to one another, are being produced by digital humanities scholars, but they aren't yet being aggregated into coherent or interoperable wholes. If they were, the fruits of humanities scholarship would be accessible to a wider audience and opened to wider participation in such scholarship. Generating such services must yield tangible benefits for scholars, without straitjacketing them.

The affordances of the classic reference tools of the humanities have been dominated by the structure of the codex and the high cost of printing. Neither constraint applies in a digital environment. As a result a full agenda emerges for systematically redesigning traditional tools for a new and different environment and reconstructing the amenity of a reference collection in the digital library environment [4].

[3]http://en.wikipedia.org/wiki/Linked_Data

5. CONCLUSIONS

Understanding any datum depends on knowing the context and background. For this reason alone, the development of better tools for linking any datum to explanatory resources outside as well as inside the data set of which it forms a part is important for both research and education. Further, doing this has the potential for repositioning how we understand the data set itself in relation to a wider context and, thereby, also suggesting relationships between data sets that are unlikely to be evident so long as the focus remains on what is inside the data set.

6. ACKNOWLEDGMENTS

The work discussed in this paper has been generously funded by the Institute of Museum and Library Services and the National Endowment for the Humanities.

7. REFERENCES

[1] B. Adida and M. Birbeck. RDFa primer. W3C working draft, W3C, June 2008. http://www.w3.org/TR/2008/WD-xhtml-rdfa-primer-20080620/.

[2] M. Buckland. Description and search: Metadata as infrastructure. *Brazilian Journal of Information Science*, 0, 2006.

[3] M. Buckland, A. Chen, H. Chen, Y. Kim, B. Lam, R. Larson, B. Norgard, and J. Purat. Mapping entry vocabulary to unfamiliar metadata vocabularies. Technical report, Corporation for National Research Initiatives, 1999.

[4] M. K. Buckland. Reference library service in the digital environment. *Library & Information Science Research*, 30(2):81–85, June 2008.

[5] W. McCarty. What's going on? *Literary & Linguistic Computing*, 23(3):253–261, Sept. 2008.

[6] C. Plaunt and B. A. Norgard. An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science*, 49(10):888–902, 1998.