Event Gazetteers for Navigating Humanities Resources

Ryan Shaw
School of Information
University of California, Berkeley
Berkeley, California, USA 94720
ryanshaw@ischool.berkeley.edu

ABSTRACT

In history and the other humanities, events and narrative sequences of events are often of primary interest. Yet while named events sometimes appear as subject headings, systems for knowledge organization generally do not provide facilities for identifying and disambiguating events as they do for person or place names. As a result opportunities for collocating resources that pertain to specific events have been limited, and support for navigating among related resources by way of the various relationships represented by events has been weak. The accelerating digitization of document and artifact collections and the ongoing development of digital metadata infrastructure make this an excellent time to address this oversight. This paper describes ongoing research to develop gazetteers for representing events and their relationships and best practices for using such gazetteers to enhance digital resources and information services.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*thesauruses*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*web-based services*

General Terms

Design, human factors, standardization

Keywords

Events, knowledge organization, digital humanities

1. INTRODUCTION

Well-established techniques for controlling proper names of persons and places exist, and work has been done to develop similar techniques for time periods [8, 12]. Yet events have not been accorded the same kind of systematic treatment, even though events and narrative sequences of events are often of primary interest in history and the other humanities. As a result it is often difficult to select resources that pertain to specific events, and tools for browsing digital resources cannot help users situate those resources in a historical context by relating them to contemporary events.

The solution proposed here is to develop gazetteers for enumerating and disambiguating events and mapping their relationships to people, places, times, artifacts, and other events. These event gazetteers would be made available as web services, integrated with complementary web services such as person name authorities and place name gazetteers, and used to annotate resources in order to enable event-centric finding and browsing.

2. EVENT GAZETTEERS

The objectives of an event gazetteer are improved *selection* of and *navigation* among resources. An event gazetteer identifies and disambiguates historical events, enabling collocation of resources that relate to a specific event. This list of events is augmented by semantic relations among events and related entities, enabling navigation among resources via the events to which they are related. An event gazetteer can include two main categories of semantic relations: *factual* and *colligatory.* The first category includes the "what, where, when, and who" relations that position events in a basic factual context [6]. The second category includes relations that situate events within explanatory narrative structures.

2.1 Factual relations

Factual relations with events can be characterized in terms of the "4 Ws:" *what* happened, *where* did it happen, *when* did it happen, and *who* was involved. These links among events, people, places, and times are those relations about which a stable consensus has been reached, so that it is relatively unproblematic to assert them in a general way. Their defining characteristic is that most people agree that they occurred. Whether such relations are considered to be empirically given or rhetorically produced will depend on one's epistemological stance. That should not prevent them from being used as a way to link people, things and activities to particular times and places, allowing users to navigate from resources that depict or describe any of these elements to resources that depict or describe any other of these elements.

2.2 Colligatory relations

Factual relations within and among events are intended to represent "consensus reality" and thus are not necessar-

PIKM'08, October 30, 2008, Napa Valley, California, USA.

ily associated with a particular perspective or interpretation (though they must be amenable to change over time as consensus changes). Colligatory relations, on the other hand, are intended to explain events by categorizing them and relating them to other events, placing them in the context of larger structures, processes or narratives [17]. Different storytellers may choose to link events in different ways, depending on their rhetorical purpose and on what overarching concepts they are trying to illuminate. Representing these links in a semantic tool should enable users to navigate among resources that present different interpretations of the events by situating them in different narrative contexts.

Several researchers have called for a multi-perspectival approach to information organization, but few actual implementations exist [16]. Information visualization researchers have developed tools for multi-perspectival exploration of quantitative data sets [13], and recently have added a collaborative dimension to these tools, allowing users to create and share visualizations of the patterns and correlations they find.[1] But few such tools exist for exploring qualitative information using narrative modes of explanation. News article aggregators like Google News, by bringing together news coverage of the same events from a diverse array of places and publications, hint at the potential of such systems, but without better representation of event structure are limited to mere juxtaposition. A system that allowed users to assert various kinds of colligatory relations among events could become a powerful tool for navigating through collections of resources in history and the humanities.

2.3 Semantic infrastructure

Events and resources can be related in different ways. Buckland [4] lists three kinds of relationships that may obtain between events and resources. First, there may be objects such as bloodstains or footprints that are produced during an event and serve as evidence for it. Second, there may be firsthand representations of an event in the form of eyewitness accounts, memoirs, photographs, audio tapes, video footage, and so on. Third, there may be reconstructions which abstract some important aspects of an event from a particular perspective. These typically are discourses that analyze the event secondhand, but they needn't be limited to written or spoken forms: maps, diagrams, and computer simulations are also reconstructions. To these can be added a fourth kind of relation: patterns in data may reveal effects of or be correlated with an event.

Thus references to any one particular event may be spread across and embedded within many separate resources. An event gazetteer takes these widely distributed references, makes them explicit, and connects them with a web of relationships. Suspending resources within such a web enables people to use descriptions of events to find resources (inverting the typical relationship in which resources record or serve as evidence for such descriptions). Thus event descriptors and the relations among them can be considered a form of semantic infrastructure for aiding search and mapping out patterns among resources [5].

3. RELATED WORK

Subject authority files such as the one maintained by the Library of Congress include entries for both historical and recurring events, but these are not treated any differently from other subjects. In particular, there is no effort to establish relationships between event subjects and authority records for the places at which they occurred or the people who participated in them.

Recognizing a need for knowledge structures that do for time periods what gazetteers do for place names, a few researchers have done preliminary work to develop thesauri of time periods and their relationships. Petras et al. proposed a simple standard for time period directories that identify named time periods and associate them with locations and date ranges [12]. Doerr et al. created a multilingual thesaurus of time period names with the objective of helping to resolve disagreements among different communities of archaeologists regarding the definitions of time periods [8]. The primary focus of both these projects is on spans of time rather than individual events.

Allen has developed an interface for a gazetteer of Civil War events and described how it might be used to facilitate navigation of a newspaper archive [2]. Allen's event gazetteer is a web application intended for direct use by people as standalone resource or in conjunction with with a particular document collection. In contrast, the event gazetteer described here is intended to be an open, machine-usable web service that could be used to enhance any information service or application that deals with events.

Research into topic detection and tracking seeks to build systems that can automatically group documents pertaining to events of interest [1]. The focus is on classification and clustering rather than identifying, disambiguating and representing events *per se*. Moreover, much of this work has focused on news articles, which have very different characteristics from scholarly resources in the humanities. In particular, the entire content of a news article tends to be "about" the same time and place, so that a single event can be associated with a single document. This is not the case for humanities literature, which may discuss many spatially and temporally separated historical events in a single document [15]. Nevertheless, some of the techniques that have been developed for extracting event descriptions and semantic relations from texts are expected to be useful for the initial seeding and ongoing maintenance of an event gazetteer.

Finally, archivists, historians, genealogists, and journalists have developed a number of standards for representing information about events [14], and there are a number of Semantic Web ontologies that pertain to events. One goal of my research is to extract from these various standards a common core of event properties to be used in event gazetteers, so that records from an event gazetteer might be easily mapped to various domain-specific representations.

4. ONGOING RESEARCH

To demonstrate the validity of the proposed approach, I am developing an event gazetteer to provide event-oriented access to a large collection of materials being digitized as part of the *Digital library of core e-resources on Ireland* project at The Queen's University, Belfast. This project comprises the digitization of 100 key journals, 205 monographs and 2,500 manuscript pages from core Irish Studies collections. Given the cultural and geographical focus of the collection, it provides an ideal testbed for developing a tool for navigating among resources in terms of events–in this case events in Irish history and culture.

[1] E.g. `http://services.alphaworks.ibm.com/manyeyes/`

Designing and engineering an event gazetteer for organizing Irish Studies resources involves three parallel and interrelated activities: domain analysis, system building, and evaluation. The challenge of building information systems would be made much easier if these three activities could be cleanly separated. Unfortunately this is not possible: building infrastructure requires constant movement between social, technical, local, and global perspectives [9].

4.1 Domain analysis

Events are concepts conceived and constructed by the people narrating them [11]. For this reason, it is important to ground the development of an event gazetteer in an understanding of how events are characterized within a specific domain. To develop this understanding I am working with Irish Studies scholars and analyzing the literature and existing scholarly apparatuses (reference works, subject guides, indexes, etc.) that they use. Specific questions that I am considering include: what are the events of interest, and on what temporal and geographical scales are these events studied? Can events at different scales be nested within one another? What level of detail in event descriptions do different types of users need? To what extent is there consensus about the "facts" of historical events? Is there an identifiable set of major narratives within which events are typically presented? While ready answers may not exist, the process of trying to answer questions like these is critical for developing appropriate tools.

4.2 System building

Building the event gazetteer involves constructing a list of events and their relations using both automatic and manual techniques, making this information available as a web service, and integrating with other web services and tools.

4.2.1 Harvesting events automatically

To achieve scale it is necessary to use automated techniques to seed a gazetteer with an initial set of events. I am harvesting event names and descriptions from three sources: Library of Congress subject headings, Wikipedia, and the digitized corpus itself.

The Library of Congress subject authority file includes many subject headings corresponding to historical events, as specified in section H 1592 of the Subject Cataloging Manual [10]. The OCLC FAST project [7] has developed a faceted version of the subject authority file, including facets for time periods and events. By searching for FAST authority files within these facets it has been possible to compile a small list of important events and time periods from Ireland-related subject headings.

This list can be augmented by treating Wikipedia as an authority file for named events. Wikipedia has hundreds of articles related to Ireland and Irish History. Since Wikipedia articles are extensively hyperlinked, it is a potential source not only of events but relations to people and places as well. The DBpedia project is taking this loosely structured information from Wikipedia and making it available as linked data [3]. I am currently growing my gazetteer by harvesting event concepts and relationships from DBpedia and mapping them to FAST concepts where possible.

Finally, following [15], I plan to use natural language processing techniques to extract candidate events from the the digitized texts, looking for co-occurrences of place names and dates across documents, and then clustering on similar phrases found near these co-occurrences. This is expected to be useful for finding alternative names for the same event and separate events with similar names, as well as for further increasing the coverage of the event gazetteer.

4.2.2 Collaborative manual construction

In addition to harvesting events automatically, I am also creating tools to enable collaborative editing of the gazetteer, allowing users to add events and relations of interest. The Freebase project has demonstrated that, given suitable tools, people will voluntarily build structured representations of historical events.[2] I hope to take advantage of this either by building a custom interface geared toward Irish Studies on top of Freebase or by implementing similar capabilities in a separate system.

Developing a taxonomy of colligatory relations that can exist between events requires close cooperation with Irish Studies scholars. I plan to initially allow free creation of typed links among event concepts, to represent whatever kind of relations (causality, similarity, etc.) users find useful. The next step will be to examine how this free linking functionality is being used, to determine if it is feasible to standardize on some basic link types that could enable navigation among various interpretations and explanations of the same events.

4.2.3 Integration with services and tools

In parallel with the development of the event gazetteer, I am developing prototype tools for using the gazetteer to annotate resources (both automatically and manually), to query services for pertinent information about entities (people, places, things) related to specific events, and to navigate among resources related to the same events.

The event gazetteer needs to inter-operate with various other services, including gazetteers capable of converting between place names and geospatial coordinates, person directories capable of resolving between person names and unique identifiers, and subject indexes to explanatory resources such as catalogs, bibliographies and encyclopedias. Some of these services already exist, with GeoNames, DBpedia and Freebase being notable examples. In other cases, especially for more obscure Irish people and places, I will demonstrate that scans of non-digital reference works can be converted into queryable information services.

4.3 Evaluation

While domain analysis may provide theoretical grounding for the design of event gazetteers, and system building is expected to show that it is feasible to build them, there is still a question of whether such tools can perform effectively to meet the objectives of selection and navigation.

Evaluation of the selection objective is typically done in terms of relevance. Relevance is a vague and problematic concept [18]. One goal of my domain analysis is to determine what it means for a resource to be relevant to an event in the Irish Studies domain. If a suitable interpretation of relevance can be established, then the effectiveness of an event gazetteer for selection can be evaluated by comparing retrieval using the event gazetteer to retrieval using alternative techniques. If the event gazetteer can be shown to increase precision without severely impacting recall or *vice*

[2] See `http://www.freebase.com/view/time/event`

versa, then we can conclude that it is effective for selection of resources relevant to events.

But relevance and the recall/precision paradigm are less useful for evaluating the success of an event gazetteer as a navigational device. To evaluate this aspect of the system's performance, I plan to take a qualitative approach, observing and measuring how the system is used while engaging in dialog with its users. This will require a multipronged approach to design research, including local qualitative methods (observation, semi-structured interviews, questionnaires) centered on specific communities of users and global quantitative methods (log analysis) that measure the usage of the system. The goal is to iteratively and simultaneously develop both the system and methods for understanding how it is or is not functioning to aid navigation.

5. FUTURE WORK

The current focus of this research is on improving selection and navigation of digitized resources related to historical events. The goal is to provide users with intuitive access to resources by providing them with a conceptual map of events and of what has been said about them. At present, these resources take traditional, albeit digital, forms such as journal articles, monographs, manuscripts, photographs, or images and descriptions of museum artifacts. The association of these resources with event representations is something that happens after the creation of the resources as they are cataloged and made publicly available.

Humanities scholars are increasingly interested in using formal models of historical events as a methodological tools. This suggests that event modeling is useful not only for representing knowledge in order to improve indexing of scholarly products, but also as an analytic technique for scholars to use during the production process. While this kind of event modeling bears some resemblance to the construction of an event gazetteer, the objectives are quite different: rich and and complex models of events created for analytic purposes may not be intuitively accessible to others without detailed explanation.

Despite these different objectives, as development of computational models and simulations of events becomes more common among humanities scholars, it may be fruitful to investigate how this work could feed into the ongoing development of tools like event gazetteers. Concepts such as historical events are products of a constantly developing discourse about the historical past. Currently most of this discourse takes the form of natural language, yet as humanists adopt tools for digital model-making it may become feasible for their labor to directly contribute to the enrichment of scholarly infrastructure. In the future I hope to investigate how to close this gap between scholarly model-making and the construction of tools for finding and understanding scholarly resources.

6. REFERENCES

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Springer, 2002.

[2] R. B. Allen. A query interface for an event gazetteer. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, pages 72–73, 2004.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference*, pages 722–735, 2007.

[4] M. Buckland. *Information and Information Systems*. Greenwood Press, New York, 1991.

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

[6] M. Buckland and L. Lancaster. Combining place, time, and topic. *D-Lib Magazine*, 10, 2004.

[7] R. J. Dean. FAST: Development of simplified headings for metadata. In *Authority Control: Definition and International Experiences*, 2003.

[8] M. Doerr, A. Kritsotaki, and S. Stead. Which period is it? A methodology to create thesauri of historical periods. In *Computer Applications and Quantitative Methods in Archaeology*, 2004.

[9] P. N. Edwards, S. J. Jackson, G. C. Bowker, and C. P. Knobel. Understanding infrastructure: Dynamics, tensions, and design. `http://hdl.handle.net/2027.42/49353`, 2007.

[10] Library of Congress Cataloging Policy and Support Office (CPSO). Subject cataloging manual: Subject headings. `http://desktop.loc.gov/`, 2008.

[11] L. O. Mink. Narrative form as a cognitive instrument. In R. Canary and H. Kozicki, editors, *The Writing of history: literary form and historical understanding*, pages 129–149. University of Wisconsin Press, Madison, Wisconsin, 1978.

[12] V. Petras, R. R. Larson, and M. Buckland. Time period directories: a metadata infrastructure for placing events in temporal and geographic context. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 151–160, 2006.

[13] J. C. Roberts. On encouraging multiple views for visualization. In *Proceedings of the 1998 IEEE Conference on Information Visualization*, pages 8–14, 1998.

[14] R. Shaw and R. Larson. Event representation in temporal and geographic context. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, forthcoming 2008.

[15] D. A. Smith. Detecting and browsing events in unstructured text. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 73–80, 2002.

[16] K. Tuominen, S. Talja, and R. Savolainen. Multiperspective digital libraries: The implications of constructionism for the development of digital libraries. *Journal of the American Society for Information Science and Technology*, 54:561–569, 2003.

[17] W. H. Walsh. *An Introduction to Philosophy of History*, pages 59–64. Hutchinson, London, 1951.

[18] P. Wilson. *Two Kinds of Power: An Essay on Bibliographical Control*, chapter 3. University of California Press, Berkeley, 1968.