Towards a Full-Text Historical Digital Library

Robert B. Allen and Yoonmi Chu

Yonsei University, Seoul, Korea
rba@boballen.info, yoonmichu@gmail.com

Abstract. A new generation of digital libraries is now possible based on the large amount of open-access full-text and other rich-media materials available. Such content can be more richly modeled and cross-linked than is possible for traditional document-level digital libraries. For collections which include details of events such as collections of newspapers, structured descriptions could be developed to focus on events. For higher-level historical analysis a combination of content and discourse descriptions is needed. Prior work on composite hypertexts has focused almost exclusively on the relationship of the discourse terms without considering the semantics of the content. Here, we describe a framework and interface widgets that support interaction with a historical text which incorporates both discourse and content descriptions. Further, we consider broader issues of interaction based on rich description of content.

Keywords: Adversarial Argumentation, Community Model, Claims Browser, Digital History, Discourse, Footnotes, Historiography, Human-Information Interaction, Long-form Historical Analysis, Model-based Argumentation, Semantic Microworlds, Tables of Contents, Widgets.

1 Introduction

A great many full-text repositories of digitized historical texts from archives, libraries, historical societies, and publishers are now available online. [4] has called for rich linking of full-text scientific research reports. Here we consider the potential for full-text and rich-media collections of historical materials.

Historical collections are composed of many types of content. There are primary sources such as letters and oral histories but also historical analyses which evaluate the credibility of these primary sources as well as interpreting them. The analyses vary greatly in complexity including book-length long-form arguments. We examine how the layers of content in a historical digital library can be coordinated in a structured way. Ultimately, such structures should be useful in developing services for casual readers, students, and historians.

2 Information Organization for Historical Materials

As with traditional document-level libraries, structured descriptions are a central concern for full-text libraries. Indeed, determining the organizational structure which is

K. Tuamsuk et al. (Eds.): ICADL 2014, LNCS 8839, pp. 218–226, 2014.

needed should precede indexing, text extraction, development of interactive services, and personalization for complex material.

The value of semantic description is increasingly recognized with projects such as schema.org and dbpedia.org along with broader work on ontologies and linked open data. Our approach is related to these projects in seeking rich description but we view them as occupying different points in a space with dimensions of scope, completeness, and formality. In this paper, we consider the coordination of semantics with discourse while other forthcoming work explores semantic microworlds for relatively detailed representations for historical Community Models. In terms of semantics, our approach follows "realist" ontologies such as the Basic Formal Ontology (BFO) which is an upper ontology that is widely used in biomedicine. Beyond ontologies, we also argue that modeling should include an explicit representation of states [3, 4, 6].

Discourse is interpretation about entities and events. For historical analysis, we attempt to separate descriptions of entities and events from discourse about them [5]. However, there are many ambiguous cases; for instance, there is a vigorous debate about whether causation is fact or interpretation. We do not resolve that here.

Rhetorical Structure Theory (RST) [14] provides a framework for describing rhetorical relationships in texts. Rhetoric is related to discourse. It is an approach to the systematic and persuasive presentation of a position. In addition to relationship labels, RST proposes that there is an overall pattern for connecting discourse elements. Specifically, it proposes that some elements are part of a "nucleus" while secondary elements (those which amplify the nuclei) form "satellites". For carefully authored text the nuclei structures are claimed to be hierarchical [14].

Issue-based information systems (IBIS) interconnect concepts related to policy. Composite hypertext systems[1] have applied that approach in a variety of other domains by implementing sets of discourse relationships as labels for the links between concepts. Some of those composite hypertexts focus on supporting argumentation and, thus, are termed argumentation systems. However, these argumentation systems have consistently emphasized discourse relationships without also structuring the semantics of the content. By comparison, we consider ways that discourse and semantics can be combined. For instance, we propose a model-based argumentation approach in which models of the content are integral to argumentation.

Shum (e.g., [17]) made an early proposal for applying discourse tags in the context of a scholarly digital library. Notable as that work is, it did not provide end-user widgets. Moreover, like the argumentation systems described above it did not link the discourse labels to the semantic content in a general way.

3 Long-Form Historical Analysis

Historical analyses differs from simple recounting of historical entities and events and such analyses are often extended. Indeed, they are often long-form book-length texts. Such complex integrated works have only rarely been considered by hypertext or digital object researchers. Yet, such texts are very important in the humanities.

[1] Halasz, F. "Seven issues": Revisited (closing keynote address). ACM Hypertext' 91 Conference, San Antonio, TX.

As an example, consider Gibbon's *History of the Decline and Fall of the Roman Empire* [9]. This classic study argues that a decrease in civic virtue led to the decline and fall of the Empire. The book makes its case by describing a broad range of trends including, somewhat controversially, the growth of early Christianity. Temporal structure is often used as an organizing strategy in historical analysis. In some cases, there is also organization of the discussion by sectors of the society (e.g., military, governance).

For historical analysis, claims of causation replace logical inference which is usually considered in an argumentation system. However, the nature of causation is controversial. For broad sweeps of history, causation is often claimed for trends and generalizations rather than for specific events (see [3]). Those trends and generalizations are themselves the subject of argumentation.

4 Widgets

Techniques to support interaction and navigation in traditional paper books have evolved through hundreds of years. Given new interface technologies and the wide availability of rich digital content, [2] has proposed a new focus on "widgets" that can support interaction with digital content. Such widgets are also related to the psychological notion of "cognitive organizers" and should benefit students, lay readers, and scholars. There are many types of widgets. The Preface to a volume often helps to put the creation of the content in a historical context. A Colophon helps to embed the production of a volume within a historical context. Following work in the hypertext community, Timpany [20] discusses linking within parts of a book though she did not appear to consider tables of contents. Widgets can also be applied across works in a collection. An example is the "meta-dex" project [11] which developed a unified index for items in a collection.

Widgets require structure but no general framework has been developed. The Open Annotation framework has recently been proposed as a framework for single annotations [22]. However, annotations almost never appear in isolation so examining sets of annotations is necessary. Thus, annotation macros will have to be developed. Alternatively, perhaps "named graphs" could be structured more systematically to cover scholarly widgets.

5 Enhanced Table of Contents Interaction Widget

5.1 TOC Interaction Widget

While the features of document abstracts have been studied extensively [12], remarkably there is no theory or general model for TOCs. There can be many different types of mapping between a set of short descriptors and sections of the body of the text.

Traditional tables of contents are ad hoc but they could be more explicitly structured. They vary in resolution (e.g., chapter, section, paragraph, sentence), in the content type of the label, and in the type of user interaction they support. One common approach is a "fish-eye" or "focus+context" TOC which allow viewing of multiple levels of resolution. Further, fish-eye TOCs may allow single or multiple

foci and selected sections of the text may be highlighted. Taken together, these attributes can define a TOC structure which could also include TOC metadata attributes such as author and date of creation.

Timelines which map to text segments might also be considered a type of TOC, or perhaps a type of index, for organizing events described in a text (cf. [1]). Claim browsers, which can also be considered a kind of TOC, provide an overview of specific issues or claims and may not follow the linear structure of a document. Because that evidence may be spread across sections of the work, the claim browser could include bridging material to explain how the parts fit together.

5.2 Discourse and Semantic Descriptions in a TOC

As noted earlier many composite hypertext frameworks have focused on discourse relationships. We suggest that the semantics should also be included. As an illustration, we developed a prototype TOC with two-part labels. For our low-level TOC, we applied RST-style discourse labels and concepts which could plausibly be derived from a semantic ontology for the passage.

We developed and applied the prototype browser to a selection from Volume 1 of Gibbon's History of the Decline and Fall of the Roman Empire obtained from Project Gutenberg.[2] This volume of the "Decline and Fall" describes the Roman Empire in the 2nd and 3th centuries AD. The selection started with:

> Till the privileges of Romans had been progressively extended to all the inhabitants of the empire, an important distinction was preserved between Italy and the provinces.

Gibbon is simultaneously framing his approach and also defending it. Gibbon contrasts the status quo between Italy and the provinces. He then lists specific examples. In the following paragraphs, he examines the ways in which that convergence occurred. The first of these was by expansion of settlements; the second by extending language; and the third, by extending language and arts. Finally, there is a paragraph about the status of slaves which does not directly support the claim but completes the description of the structure of the society.

Although Gibbon does not provide a TOC at this resolution, we can create one in an electronic edition to help orient readers. As described above, indicating both the semantics and discourse should help a reader in several ways. First, knowing that this is a comparison of the two types of political units within the Roman Empire (i.e., Italy and the Provinces) suggests some of the attributes on which that comparison will be made. In addition, knowledge of the semantics helps orient the reader to the parallelism in some of the points that are made about the two different types of political units. Figure 1 shows the implementation of this TOC widget with a prototype Java applet. There are two independently scrollable panels. The left panel has the text while the right panel has the interactive discourse-semantics outline.

[2] http://www.gutenberg.org/files/731/731-h/731-h.htm#link22HCH0002

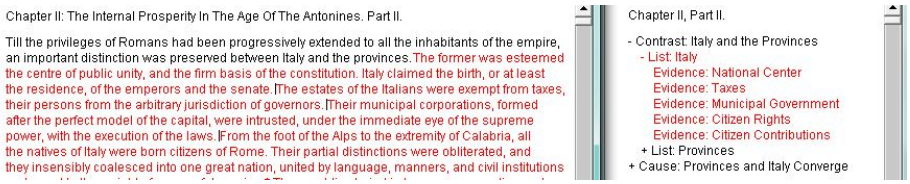

Fig. 1. Prototype semantics-discourse TOC applet for the first paragraph

In addition to the two-part TOC labels, we have added several features to enhance the usability of the TOC: (a) The TOC does not map to defined section boundaries but to conceptual units (b) The selected conceptual unit is highlighted in red (gray in the reproduction) (c) low-level subsections within the conceptual unit are indicated with vertical black tick marks rather than with still lower level TOC labels which would probably be distracting for the reader.

This interface should be especially useful for people who are not familiar with the topic. Just as interaction with a search engine teaches users about the relevant dimensions of what they are searching for even if they do not enter any of the documents, this browser can help people better understand the text.

There are several limitations to the prototype interface. First, the enhanced labels are highly abbreviated and may be cryptic for readers who are not familiar with the topic. Second, it may be difficult for readers to understand the role of this passage in the context of the entire work. Both of these issues could be addressed with additional widgets. For instance, tooltips or even audio descriptions could be used to provide richer, more complete descriptions. In addition, a graphical argumentation template could lead the reader through the narrative [7] and support more complex reasoning (cf. [15]). Finally, additional commentary could be provided about the relationship between the discourse and semantics as well as about the strategy of the author in relating the two.

6 Footnotes

There has been an evolution of humanities footnotes through time [10]. Gibbon is renown for his extensive use of footnotes and is often considered one of the first modern historians in this regard. Footnotes which include citations and the discussion of the work of other authors go beyond basic rhetorical structures in which one author is trying to make a case for a position. Indeed, citations require that a work is embedded in broader literature. While citation linking in scientific research articles has been extensively studied (e.g., [19]), footnotes in humanities have not.[3] Footnotes in humanities are often different from those in science including sometimes lengthy analysis that is not directly relevant to the main thesis of the text.

6.1 Gibbon's Footnotes

We examined the footnotes in Chapter 2, Part II of the "Decline and Fall". Gibbon's footnotes ranged from 1 to 15 lines. More than 90% included reference links to

[3] Though, see [13] for a description of the use of footnotes in fiction.

other works; there were frequently several links in a given footnote. However, one of the editors who contributed annotations mentioned the work of others but did not always provide specific citations. The majority of the footnotes was for clarification or extension of a point in the text; only about 20% of them were simple citations with no other text.

Importantly, for our ultimate goal of developing a full-text digital library the full-texts for almost all of the numerous sources mentioned in Gibbon's footnotes which we checked are available online. In many cases these were open-access clean-text English translations. However, in other cases, there was no translation, they were poor quality OCR from digitized copies, or they required a password to access.

6.2 Footnote Widget

Consider footnote #26 from Chapter 2, Part 2 of "The Decline and Fall":

> 26 The senators were obliged to have one third of their own landed property in Italy. See Plin. l. vi. ep. 19. The qualification was reduced by Marcus to one fourth. Since the reign of Trajan, Italy had sunk nearer to the level of the provinces.]

Although it is not part of the main document, this footnote provides support not just for a specific claim but also for the broader shift which may be viewed as Gibbon's main claim in this section about the spread of the rights of Citizens across the empire.

Gibbon rarely expresses direct disagreement with others; however, the Project Gutenberg edition also included annotations by H. Milman, the editor of the 1845 print edition, as well as by editors of the Project Gutenberg editions. These editors made wide ranging comments including mentioning Gibbon's support of abolitionism when commenting on his discussion of Roman slaves. In Footnote #261, Milman (identified as "M") cites the work of the German jurist Savigny who published his research several years after Gibbon wrote "The Decline and Fall".

> 261 It may be doubted whether the municipal government of the cities was not the old Italian constitution rather than a transcript from that of Rome. The free government of the cities, observes Savigny, was the leading characteristic of Italy. Geschichte des Romischen Rechts, i. p. G.—M.]

Footnotes can be considered as widgets. When encountering a footnote a reader with an enhanced interface could be given a description of the contents of a footnote with a tooltip. If the reader clicked through and found a source to follow, a browser, similar to the one in Figure 1, could be launched to organize both the discourse and the issues of the target article (cf., [4]). In other words, with a concept-based full-text digital library citations can be de-emphasized in favor of linking concepts. Footnote 261 cited above might apply extended RST terms to show adversarial argumentation. Some of the ontologies which have been proposed for characterizing citations might be a place to find such terms. Then, an interaction widget could summarize the two positions and highlight the differences.

7 Standards for a Full-Text Digital Library of Historical Materials

There are some relatively small full-text collections such as the Perseus project which collected ancient Greek writings [8]. Here, we consider developing standards which could be applied to the much larger set of Roman-era writings and, ultimately, across all of history. Because resources are spread across the web, we need a registry of definitive versions of these works to identify permissions on their use for different purposes. In addition, those repositories, as well as archives and historical collections, should adopt standard formats and a standard way of identifying anchor points[4].

Information organization is key to coordinating the contents. The upper ontology for semantics needs to be complemented by lower-level domain ontologies. Our analysis of the "Decline and Fall" could incorporate ontologies of terms relating to the Roman government, regional boundaries[5], religion, and military organization. For the work we have done with 1900's Norfolk, Nebraska, we need ontologies of Protestant traditions, small-town Midwest US government organization, and 19th century railroads. In addition, it will be helpful to have structured collections of mundane, but ubiquitous, cultural knowledge such as organizational by-laws, rosters of popular entertainers, job descriptions, lists of cities hosting sports teams, and national holidays.

We also need a comprehensive discourse ontology suitable for history and standards for widgets such as those described above. Finally, we need standards to incorporate non-textual materials such as data sets and videos.

8 Discussion

The development of the full-text and rich-media digital libraries we propose here and in [5] is a grand challenge. While some of that challenge is coordinating publishers and collections, the bigger challenge is developing a broad range of interoperable information organization structures to support user services. Consider an interface to support interaction with a collection of digitized historical newspapers. Such an interface should be based on a rich model of the community which evolves through time. Communities as described in newspapers are very complex but they are also highly structured and we should be able to capture that structure. In addition, the interface could be personalized for the background and interests of the reader and and description could be presented as a narrative.

We have focused on discourse in this paper but we should note the relationship between discourse and logic. While discourse can be simply a style of presentation, generalization and inference are related to logic. Generalizations can be seen as induction or as a weak form of predicate-logic. The role of logic in human argumentation is widely debated especially with respect to inference in science. In sciences such as biology, the goal is to identify universals through research and theory. In humanities and social science finding universals is much more difficult, if it is possible at all. In history,

[4] See `http://en.wikipedia.org/wiki/Fragment_identifier`

[5] See `http://orbis.stanford.edu`

the notion of "covering laws" is no longer widely accepted (see [16]); rather, history is much more likely to be based on generalizations.

There is also a vigorous debate about the upper ontologies for social activities (e.g., [18]). Nonetheless, it is worthwhile to explore the development of Community Models such as might be applied to an interactive digitized newspaper interface described above. Cultural frames and a variety of sociological structures may help define aspects of the community within a given perspective.

In the approach to historical information organization we propose, all of history could be part of a single "fabric". Events, evidence, authors, editors, and collection management policies can all be represented within that fabric. This broad overview of history with links from all types of events can be viewed as an implementation of the archival Continuum Model [21]. In a different research tradition, there have been attempts to link the BFO with information resources[6].

While semantic support tools which could automate the task of creating the widgets we discuss here are likely to be developed, the implementation of these proposals will require considerable human effort. We believe such effort would be forthcoming for works such as "The Decline and Fall". After all, the editors of Project Gutenberg have already proven willing to do substantial work and it seems likely that others would be willing to further enrich such classics.

Although we have focused on the specification of information structures for historical materials and end-user services which could be applied to them, tools could be developed to support working with the structured content. Once a substantial body of structured descriptions is available, tools could check for consistency in an author's arguments.

In addition, while this paper has focused on re-mapping existing texts, these structures should also be useful for developing models of and widgets for entities and events which are not anchored in text. Rather, the structure itself could organize all of the evidence and argumentation and services could be developed to support such model-based authoring.

Overall, there are opportunities for a new generation of digital libraries of historical materials which integrate and support access to full text and other types of rich media. Indeed, perhaps texts and other materials from all fields could be unified with rich linking.

Acknowledgment. We thank Yonghwan Kim for his assistance.

References

1. Allen, R.B.: Timelines as Information System Interfaces. In: Proceedings International Symposium on Digital Libraries, pp. 175–180 (1995), http://boballen.info/RBA/PAPERS/TL/isdl.pdf
2. Allen, R.B.: Weaving Content with Coordination Widgets. D-Lib Magazine (2011), doi: 10.1045/november2011-allen
3. Allen, R.B.: Visualization, Causation, and History. In: iConference (2011), doi:10.1145/1940761.1940835

[6] That work has not yet considered collections and standards for organizing information resources such as OAIS and FRBR could be also be applied.

4. Allen, R.B.: Model-Oriented Information Organization: Part 1, The Entity-Event Fabric. D-Lib Magazine (July 2013), doi: 10.1045/july2013-allen-pt1
5. Allen, R.B.: Rich Linking in a Digital Library of Full-Text Scientific Research Reports. In: Columbia Research Data Symposium (2013), http://hdl.handle.net/10022/AC:P:19171
6. Allen, R.B.: Frame-based Models of Communities and their History. In: Nadamoto, A., Jatowt, A., Wierzbicki, A., Leidner, J.L. (eds.) SocInfo 2013. LNCS, vol. 8359, pp. 110–119. Springer, Heidelberg (2014)
7. Allen, R.B., Acheson, J.A.: Browsing the Structure of Multimedia Stories. ACM Digital Libraries, 11–18 (2000), doi:10.1145/336597.336615
8. Crane, G.: The Perseus Project and Beyond. How Building a Digital Library Challenges the Humanities and Technology. D-Lib Magazine, doi:10.1045/january98/01crane.html
9. Gibbon, E: The Decline and Fall of the Roman Empire. Harper, New York (1782/1845), http://www.gutenberg.org/files/731/731-h/731-h.htm with production notes for the electronic edition at http://www.gutenberg.org/files/25717/25717-h/25717-h.htm
10. Grafton, A.: The Footnote: A Curious History. Harvard University Press, Cambridge (1999)
11. Hugget, M., Rasmussen, E.: The Meta-Dex Suite: Generating and Analyzing Indexes and Meta-Indexes. ACM SIGIR, 1285–1286 (2011), doi:10.1145/2009916.2010162
12. Lancaster, F.W.: Indexing and Abstracting in Theory and Practice, 3rd edn. University of Illinois Press, Champaign (2003)
13. Maloney, E.: Footnotes in Fiction: A Rhetorical Approach. Dissertation, Ohio State University (2005), https://etd.ohiolink.edu/!etd.send_file?accession=osu1125378621
14. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a Function Theory of Text Organization. Text 8(3), 243–281 (1987)
15. Reed, C.: Wigmore, Toulmin, Walton: The Diagramming Trinity and their Application in Legal Practice. In: Cardozo Conference on Graphic and Visual Representations of Evidence and Inference in Legal Settings (2007), http://tillers.net/reed%20diagramming%20trinity.pdf
16. Roberts, C.: The Logic of Historical Explanation. Pennsylvania State University Press, State College (1995)
17. Shum, S.B., Motta, D., Dominguez, J.: ScholOnto: An Ontology-based Digital Library Server for Research Documents and Discourse. International Journal of Digital Libraries (2000), http://oro.open.ac.uk/23353/
18. Smith, B., Searle, J.: The Construction of Social Reality: An Exchange. American Journal of Economics and Sociology 62, 285–309 (2003)
19. Teufel, S., Siddharthan, A., Tidhar, D.: An annotation Scheme for Citation Function. In: ACL SIGdial Workshop on Discourse and Dialogue, pp. 80–87 (2006), doi:1654595.1654612
20. Timpany, C.: Designing the Printed Book as an Interactive Environment. The International Journal of the Book 7(1), 11–28 (2012), http://hdl.handle.net/10289/6592
21. Upward, F.: Structuring the Records Continuum, Part One: Postcustodial Principles and Properties. Archives and Manuscripts 25(1) (1997), http://www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-fupp1.html
22. W3C: Open Annotation Data Model (2013), http://www.openannotation.org/spec/core/20130208/oa.owl