DHQ: Digital Humanities Quarterly
Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure
2009
Volume 3 Number 1

Exploring Historical RDF with Heml

Bruce Robertson <broberts_at_mta_dot_ca>, Mount Allison University

Abstract

The Web, though full of historical information, lacks a means of organizing that information, searching on it or visualizing it. The Historical Event Markup and Linking Project (Heml) was begun six years ago to explore how disparate historical materials on the Internet can be navigated and visualized, and for the past four years has used an XML data format defined in W3C Schemas. This format aims for conforming data that can be quickly parsed but provide a variety of facets on which to search for historical materials. While the project's graphical visualizations are in some respects successful, they have revealed some deficiencies in the underlying data format: it ought to provide for nested events, it ought to represent relations of causality between events and it ought to express the varieties of scholarly opinion about the attributes of events. By encoding the Heml data in the Resource Description Framework (RDF) it is possible to undertake these improvements. Moreover, an RDF-encoded Heml process provides easier access to CIDOC-CRM data into Heml events. Finally, a historical RDF language would simplify the discovery of references to historical events in digitized texts, thereby automating a growing network of historical information on the Web.

Introduction

Beginning his work which was to be the first called a “history”, the Greek author Herodotus describes his task in simple terms: “to preserve the memory of the past by putting on record the astonishing achievements both of our own and of other peoples” (line 1.0). Within the first few chapters, though, it becomes clear that, to Herodotus, this entails something more multivalent than mere story-telling. His very next sentence buttresses the narrative with references to its various, even conflicting, sources 1.1; not much later, he corroborates his stories by referring to physical artifacts and providing careful genealogies 1.7, 1.14; and throughout he attempts to place the notable events of the past within the available chronological outlines (e.g, 16.1). From its origins, then, history has been conducted across a framework of argumentation, evidence, chronology and prosopography. 1

This intellectual approach, begun by Herodotus and his rough contemporaries around the world, has proved a durable and useful way of thinking. Moreover, as the global Internet expands, and the common software technologies of the World Wide Web provide us with unprecedented means of communication, humanity has greater access to more historical statements and more of the raw materials of history than ever before. Indeed, it might be said that today's Web has the makings of an historian's fantasy: it provides worthy encyclopedia entries on a vast array of topics; its textual editions are rapidly improving in quality and amassing in quantity; and it offers historical source material ranging from argumentation in on-line editions of the best journals to the first-hand accounts appearing in blog entries. Scholarly historical projects published uniquely on the Web are common, supported by academic centres whose purpose is to communicate history effectively in new media, and who are also addressing the general problems of adopting best standards of scholarship to the Web in projects such as Zotero. Behind the scene, data formats such as the CIDOC-CRM aim to make it possible to interchange and reconcile vast cultural archives. 2

However what today's Web offers in historical content, it lacks in organization. The multivalent nature of historical thought that we noted in one of its earliest practitioners eludes the keyword-indexed approach to the Web today on offer 3

through Google and other search engines. Though we can summon up an exhaustive list of Web resources that contain the *words* “Gallipoli” and “sources”, today's Web cannot effectively respond to a basic historical question such as, “which sources attest the Gallipoli Campaign of World War I?” much less a more advanced one such as, “what evidence is there for major architectural projects undertaken in the U.K. during the period of the Boer War, and does anyone think that these projects are influenced by the conflict?” If it were possible to conduct such a query, of course the result would not itself qualify as “history” any more than the results of a Google search constitute knowledge. However, the results of such a hypothetical search would offer a considerable aid to historical research and thinking, just as the indices, maps and, occasionally, tables that appear at the end of most monographs allow the reader rapidly to access matters of interest. Applying this analogy, what is needed in the Web is “universal back-matter”, a Web-wide equivalent to the traditional print monograph's helpful conventions.[1]

Herodotus and his successors have given us a template for such a project. Minimally, it will need to express associations between words describing the event through chronology, geography and prosopography. It will need to provide references to diverse sources and evidence, and it will need to traverse, as much as possible, the boundaries of local language and even local time-keeping techniques.[2] Indeed, over fifteen years ago, the prescient Historian's Workstation Project employed the same basic types of data [Thaller 1991]; however, the Internet has added the challenge of working with a more heterogeneous set of data sources and today's possibilities for the visualization of historical data go far beyond the Workstation's.[3] 4

The Historical Event Markup and Linking Project (Heml) has pursued the vision of a system such as this since 2001, when it began to explore the markup and transformation of historical materials on the Internet using XML tools. All schemas and code developed by this project are available at its website, http://www.heml.org/ and complete revision history is provided in an SVN repository. This paper explains Heml's next step toward the goal of an historical Web. It outlines the data format used by the project, shows the shortcomings of our previous XML-based approach to historical markup and describes the potential of an historical markup scheme based around the W3C's Resource Description Framework, or RDF, to interchange historical concepts widely and to express them in a more nuanced manner. 5

It is a pleasure to be able to offer this paper in the collection of works celebrating the achievements of Ross Scaife. Like so many efforts in computing and the Classics, this one has been nurtured by Ross' kind support and encouragement. Furthermore, I have always hoped that Heml would attain the level of openness and helpfulness that are the hallmarks of Ross' work and of his character. 6

The Heml Data Model

The goal of a searchable network of historical information can only be reached through a clear separation between the model of the data being searched, the parameters of the search and the resulting visualization of the data. Since the project intends that these visualizations be generated dynamically in response to users' queries, the data model must be specific enough that the visualization can generated quickly; likewise, since the results of users' queries could result in no events, one event, or thousands of events, and the range of values represented in these encoded events could be quite wide, the visualization processes must be written so as to generate useful views without regard for the number of events visualized or how close or far they are in time and space. For this reason, the Heml project has attempted to steer a middle path between pragmatism and markup idealism, especially regarding chronology. Thus, its data schema, on one hand, is not a complete abstraction of any way in which a person might think of the past, since this would make the task of visualization too daunting. In particular, its schema requires machine-readable data for all defined geographical and chronological data. (As shown below, this does not preclude the schema representing uncertainty.) On the other hand, the schema is not merely an API for the various visualizations offered by Heml; it is abstract enough that many other visualizations and uses could be discovered. 7

At its inception, the Heml server software transformed XML documents one-by-one [Robertson 2002]; in versions 0.4 to 0.7.2 the project adopted a distributed XML model, backed up by the eXist XML database [Robertson 2004], an approach that worked well in a collaboration with Tom Costa's project, Geography of Slavery in Virginia, where a Heml Web application backed by eXist transforms events pertaining to slaves into hundreds of dynamically generated maps 8

and timelines. In either case, the project's data model has remained the same for the past four years, and the following describes the model labelled 2003-09-17.

The Heml data format contains a collection of modelled events, each tagged with `<heml:Event>`. At their most simple, events bind an event label with a machine-readable span of time and a reference to evidence, which could be as simple as a single URN. Optionally, a `<heml:Event>` may also comprise any number of keywords, and a single location, which in turn is defined as a labelled pair of latitude and longitude coordinates. `<heml:Person>` elements may be added to events in order to represent those people who were in some way involved in the event. If the nature of this participation is known, the `<heml:Person>` element may be bound to a `<heml:Role>` element within the context of the event. Persons, roles, locations, and keywords are assigned mandatory URIs so that they may be referred to in multiple events. Finally, one or more `<heml:Evidence>` elements must be attributed to each event, and within these there is a means by which different editions and linguistic representations of the same text may be grouped together for the researcher's benefit. 9

Chronology

The most complex part of Heml's data model is the `<heml:Chronology>` element. Its model is intended to express uncertainty and ranges of time without requiring the visualization engine to have access to all data.[4] In Heml markup, chronological concepts are always built using one of the following four elements: `<heml:DateTime>`, `<heml:Date>`, `<heml:Year>` or `<IntCalDate>`. The first three of these encapsulate data encoded in the corresponding XML Schemas format; the last permits the user to encode with a non-Gregorian calendar. When used alone, these elements are meant to indicating a corresponding span of time: for instance the `<heml:Year>-31<heml:Year>` indicates an event that began on the first second of the first day of 31 BC and ended on the last second of its last day. To express a more expanded range of time, the `<heml:DateRange>` element is used, and within it mandatory `<heml:StartingDate>` and `<heml:EndingDate>` elements. This construct is parsed as indicating a span of time beginning at the beginning of the first element and ending at the ending of the last one. To express uncertainty, the `<heml:BoundedDate>` element is used. It comprises a `<heml:TerminusPostQuem>` and a `<heml:TerminusAnteQuem>` element; these express the earliest possible and latest possible time respectively of a span of uncertain time. Use alone, `<heml:BoundedDate>` makes no claim about the duration of the encoded event. However, it is permitted to use a `<heml:BoundedDate>` within the `<heml:StartingDate>` or `<heml:EndingDate>` of a `<heml:DateRange>`. 10

```
<heml:Chronology>
   <heml:DateRange>
      <heml:StartingDate>
         <heml:DateTime>1995-05-21T21:03Z</hemlDateTime>
      </heml:StartingDate>
      <heml:EndingDate>
         <heml:BoundedDate>
            <heml:TerminusPostQuem>
               <heml:Date>2005-03-21</heml:Date>
            </heml:TerminusPostQuem>
            <heml:TerminusAnteQuem>
               <heml:Date>2005-03-21</heml:Date>
            </heml:TerminusAnteQuem>
         </heml:BoundedDate>
      </heml:EndingDate>
   </heml:DateRange>
</heml:Chronology>
```

Example 1. Example of Chronological Information Expressed in Heml's 2003-09-17 Schema

Thus the XML in Example 1 encodes as an event beginning at exactly 21:03 UTC on May the twenty-first 1995 and ending at some time on March the twenty-first 2005.

Similar Schemas

11 There are other schemas from other fields that encode representations of labelled spans of time and place. Dublin Core Metadata standard includes date and location tags that, in certain uses, encode certain kinds of historical events, but these naturally revolve around the documents themselves, such as publication information; they cannot, for instance, encode the battle described therein. Similarly, the P5 edition of the TEI includes a welcomed set of biographical and prosopographical tags, including one for events. However, at present, all such P5 `<event>` elements must appear within a `<person>` or `<place>` element, expressing that the event information serves to further our information about a given person or place: events that pertain to neither at present cannot be encoded.

12 Finally, the CIDOC-CRM encodes descriptions of cultural artifacts in terms of the events they undergo — their creation, change of custody, and so forth. Though not the primary focus of this schema, historical events are quite broadly modelled in this schema. Indeed, as is shown below, the unmodified CIDOC-CRM model is too liberal for the Heml Project's purposes of generating historical visualizations, especially with reference to what entities and datatypes it accepts as chronological predicates. Nevertheless, it is likely that most of the goals of the Heml project can be fulfilled through RDFS subclassing of the the CIDOC-CRM model, ensuring more carefully data-typed chronological classes. In any case, the discussion and examples below would apply as well to such a constrained CIDOC-CRM model or to any future schema with similar properties. Indeed, the purpose of the Heml project is decidedly not to promote a particular schema; instead, its schema is intended as a tool which enables us to explore the process of combining all material encoded in all these well-entrenched and widely-adopted schemas so as to increase the pool of inter-related historical data on the Internet.

13 Technologically, there is little challenge in transferring data from one of these representations to Heml. XML representations can be exchanged through XSLT. In RDF, the SPARQL CONSTRUCT query form produces transformations from one graph to another. For example, [Robertson 2006] shows how useful a simple Heml and Dublin Core mash-up conducted in XSLT could be.

Visualizations

Heml data is encoded in order that resources pertaining to a certain historical event may be found in response to a user's query. The query might be expressed with regard to any one of the facets of historical expression encoded by the schema above, or on some combination of these. The software discovers the corresponding materials, then generates a visualization of these data according to the user's wishes. Visualizations range from simple chronological text lists in HTML to complex animated maps. Of course, since the data has been prepared without reference to any given visualization, not all visualizations are informative representations of all query results. For instance, an animated map would not be helpful if the set of events in the query response all had the same location or did not have locations encoded at all. 14

In the most recently-published version of Heml's server software (v. 0.7.2), visualizations are generated through a Java Web application based on the Cocoon Web service engine. Instructions for building and running this software locally are available at http://www.heml.org. Using this approach on locally-stored data takes full advantage of Cocoon's ability to cache server-side processes. However, it is also possible to use the server at http://www.heml.org to transform any conforming document published on the Web without building or installing software. Instructions are provided online. 15

As described above, all the visualization tools developed by the project operate fully automatically: they provide sensible chronological and geographical boundaries based only on the conforming data on their input and layout the appropriate text without clipping or abbreviation, regardless of the number of events being presented. 16

The graphical timeline view, of which Figure 2 and Figure 3 are examples, lays out event labels at a vertical distance from each other corresponding to their separation in time. It indicates spans of time through vertical coloured lines. This view was designed to be legible and to require as little user interaction as possible, emulating a graphical timeline in print. In fact, the rendering software iterates through layout scenarios in order to find one that is appropriately compact but whose text is still possible to read. It ensures that the entire text labelling an event is always visible to the user, and scrolling is not necessary to read the event label. The software also selects an appropriate range of dates or times for the events; these can range from a seconds to millennia.[5] Rendered in SVG, this view's event labels are linked to popups that lead the reader to the resources pertaining to the event. 17

This approach to timeline-generation differs from other online graphical timeline-drawing, perhaps best represented by the the Simile timeline widget. At present, the Simile timeline is not auto-ranging, and does not attempt to optimize layout. Rather, its documentation encourages users to identify so-called “hot zones”, temporal ranges comprising unusually large numbers of events so that the drawing routine can alter the scale in these ranges, thereby producing a more legible layout. The goals of the Heml project, requiring a fully automated rendering, do not make such hand-tweaking possible. However, Simile timeline's excellent use of client-side rendering and DHTML are certainly preferable to the Heml timeline approach. We hope that by modify the open-source code for Simile timelines we can adopt that project more closely to our needs, perhaps by identifying “hot zones” computationally. 18

Figure 1. Image of a SVG Animation Generated by the v0.7.2 Heml Web application

Another event-visualization system pioneered by the Heml Project is the animated map illustrated in Figure 1. **(1)** is a set of user-operated controls familiar from audio-visual apparatus. With these, the user can stop, play pause, rewind and speed up the animation. Item **(2)** is a slider control that sets the length of the animation in seconds. Item **(3)** is a moving marker that runs along the line above the animation from left to right, thereby indicating the moment in time currently represented on the map. Above the small triangle a constantly updating text appears recording the current animation time. The event labels at **(4)** appear in red text as long as the animation is representing a date during which the labelled event is encoded as having taken place. In order that the user be able to read them, event labels do not necessarily disappear as soon as the corresponding event ends in the animated time; rather they remain on the screen in black text until the text has appeared for at least two seconds. The system designates an appropriate chronological scale to the process, as well as the geographical range of the map. 19

Areas For Improvement

Although these visualizations have fulfilled some of our goals, they have also pointed out some deficiencies in the underlying data model. Figure 1, for example, presents an apparent mishmash of events. Perhaps the user undertook a too-broad query, but when dealing with any large dataset this sort of result is common because the XML schema puts all events, no matter how trivial, on the same footing. It has therefore been suggested that events should be ranked, or should nest, with large scale events such as “The Persian Wars” expressed as parents of their composite events, like “The Battle of Thermopylae.” In response to publications and presentations of the Heml/XML model, others have expressed a desire for more complex historical relations between events to be encoded. Should not Heml express something as simple as causality, such as the fact that Ephialtes' treachery led to the defeat of the Three Hundred? Finally, it has been noted that with Heml/XML markup it is not possible to encode the variations in opinions regarding historical events. When scholars debate the date of the arrival of the Greek-speaking ancestors of the Mycenaeans, what date should be encoded for this event? 20

Some of these issues have been tackled recently by [Nakahira 2007], but in the Heml XML Schema they were intentionally left unaddressed for both theoretical and practical reasons. First, it has been this project's experience that resolving such relations with a language such as XQuery or XSLT was difficult and computationally expensive. Secondly, since the XML Schema expresses matters pertaining to an event by nested elements, it was unclear where a statement linking two elements should appear. Pertaining no more to one element than the other, ideally it should stand outside the `<heml:Event>` elements, but this raised the prospect of always new schemas required to provide 21

metadata for previous schemas. Finally, a XML schema through which scholarly debate could be encoded for any facet of an event seemed likely to be unmanageably complex.

RDF and Heml

22 Our work of the past year shows that these problems can be addressed by using the W3C's Resource Description Framework (RDF) as our means of encoding Heml events and related information.[6] As a collection of statements expressed through URIs, RDF is well suited to define metadata that expresses associations between events. Because these statements are unordered, these associations can be asserted in any source of data, anywhere. Moreover, RDF defines a method known as “reification” whereby a statement comprising Subject Verb and Object can itself play the role of subject or object in another, encompassing statement. Example 2 illustrates the process with an example from Greek prehistory. The first line makes a simple statement, without attribution, that the ancestors of the Mycenaeans arrived in the Greek mainland in 1600 BC. The second line indicates the reification of that statement with parentheses, and attributes this chronology to Drewes with the `hemlRDF:asserts` predicate. The third line refers to the same `<#Arrival_of_the_Greeks>` subject, but it shows that Renfrew assigns that subject a much earlier date.

```
Statement : <#Arrival_of_the_Greeks> <hemlRDF:simpleDate> -1600
Reified Statement A: <#Drewes> <hemlRDF:asserts>
   (<#Arrival_of_the_Greeks> <hemlRDF:simpleDate> -1600)
Reified Statement B: <#Renfrew> <hemlRDF:asserts>
   (<#Arrival_of_the_Greeks> <hemlRDF:simpleDate> -6000)
```

Example 2. RDF Reification as A Means of Encoding Scholarly Debate Regarding an Event

The URI corresponding to a scholar then becomes a part of the query, and it is possible to encode, and later summon up, an entire body of knowledge with reference to the scholar asserting it. For example, the chronology of fifth-century Athens could be encoded with reference to the arguments of Badian or those of Pritchett. This technique becomes all the more important when even more tendentious predicates are applied, such as causation. [7]

Data-Entry for HemlRDF

23 Despite these possibilities, we were concerned that the increased representational complexity of RDF would make it difficult for students and other users to encode their data. In fact, in the summer of 2007 we had excellent experience using the Semantic MediaWiki extension to MediaWiki as a HemlRDF editor. With this environment, motivated students as young as ten years old produced dozens of fully-compliant RDF-encoded events with an hour's work. The resulting RDF is not identical to the HemlRDF parsed by the Web application because certain simplifications were considered helpful to the user.[8] However, a simple SPARQL query, available in the `/sparql/wiki/` directory of the project SVN repository, transforms the data to HemlRDF. With some work, or as Semantic MediaWiki improves, it might be possible for it to generate HemlRDF directly, thereby allowing students to see automatically updated visualizations as soon as they save encoded events and related articles. A `tar.gz` SQL dump of our project's wiki, which includes the semantic metadata, is regularly generated and available at http://heml.mta.ca/releases/Wiki.

RDF-Based Nested Events

24 While building a dataset of ancient history in the Wiki, we experimented with one of the improvements suggested above, and encoded certain events as “comprising” others with a `hemlRDF:comprisesEvent` predicate. In Example 3 the nested representation indicates that the `<heml:Event>` with the URI `wiki:Samnite_Wars` is described as comprising `wiki:Battle_of_Mount_Gaurus`, etc.

```
<smw:Thing rdf:about="wiki:Samnite_Wars">
   <rdf:type rdf:resource="http://www.heml.org/rdf/2003-09-17/heml/Event"/>
   <rdfs:label>Samnite Wars</rdfs:label>
   <smw:hasArticle rdf:resource="&amp;wikiurl;Samnite_Wars"/>
   <rdfs:isDefinedBy rdf:resource="&amp;wikiurl;Special:ExportRDF/Samnite_Wars"/>
   <hemlRDF:comprisesEvent rdf:resource="wiki:Samnian_Alliance"/>
   <hemlRDF:comprisesEvent rdf:resource="wiki:Battle_of_Mount_Gaurus"/>
   <hemlRDF:comprisesEvent rdf:resource="wiki:Great_Samnite_War"/>
   <hemlRDF:comprisesEvent rdf:resource="wiki:Third_Samnite_War"/>
</smw:Thing>
```

Example 3. RDF Expressing A “Parent” Event By Means of a Series of “Child” Events

These “parent” events are ascribed chronologies, participants, and all other properties wholly in terms of their children elements. For instance, the earliest date on which any child event begins is treated as the earliest date of the parent; and the latest date on which any child ends is treated as the latest date of the parent. This process is recursive: there can be parent events that are themselves children of others. Although we performed this reasoning by parsing the RDF model in memory, it is hoped that the same effect could be made with Web Ontology (OWL) statements.

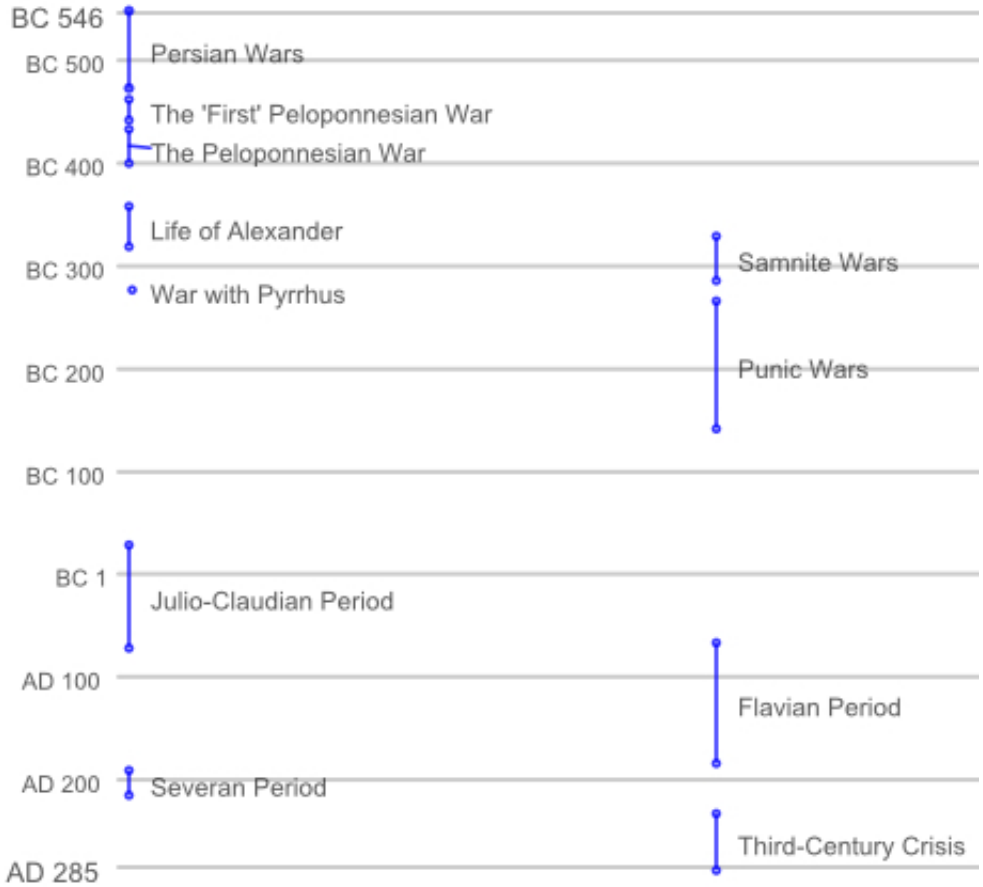

Figure 2. Timeline View of “Top-Level” Events: Top-level events in an RDF dataset accumulated with Semantic Mediawiki for a first-year Ancient History course.

Figure 2 is a timeline generated entirely from top-level parent events encoded in the wiki. This provides the user with a more comprehensible view of the data.

Furthermore, in some future system that provides an appropriate control interface, the user could “open” these

composite events and explore more deeply the nested layers therein. Figure 3 is an expanded view of the “Samnite Wars” event drawn in the top right corner of Figure 2 and whose encoding appeared in Example 3.

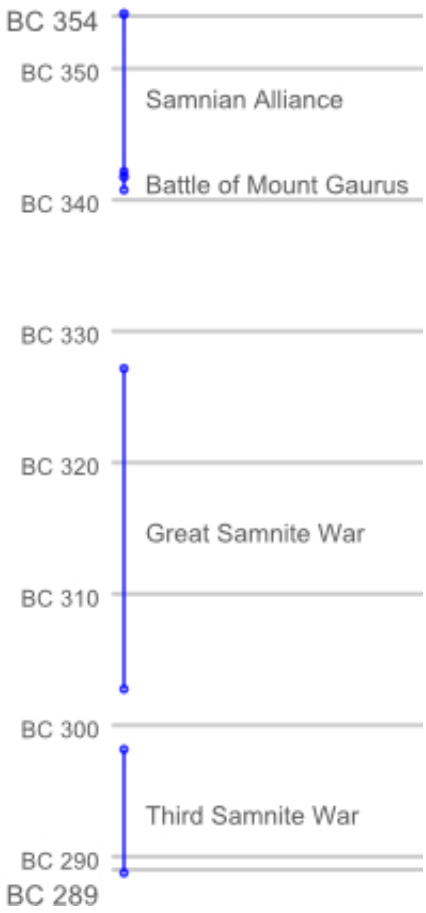

Figure 3. Graphical Timeline Representing the Children Events of the “Samnite Wars” Top-Level Event: The events which composed the “Samnite Wars” top-level event in in Figure 2 are here shown individually.

HemlRDF and the CIDOC-CRM

One of the most exciting possibilities we anticipated in the historical RDF realm was visualizing CIDOC-CRM data through Heml tools. Because the CIDOC-CRM defines objects in terms of the events they undergo, CIDOC-CRM potentially is a treasure trove of well-encoded events related to material culture. Robert Kummer of the Perseus Project kindly provided us with the Perseus art and archaeology database encoded in RDF/XML. Example 4 is a small portion of this, in which the item known as 'BCMA 1919.58.81' is declared to have been produced no earlier than 480 BC.

```
<crm:E12.Production rdf:about="http://perseus.tufts.edu/Production_of_BCMA_1919.58.81">
<dc:title>Production of BCMA 1919.58.81</dc:title>
<crm:P4F.has_time-span>
        <crm:E52.Time-Span>
        <crm:P82F.at_some_time_within>
                <crm:E61.Time_Primitive rdf:about="http://perseus.tufts.edu/date/-480"/>
<hemlRDF:SimpleTime
rdf:dataype="http://www.w3.org/2001/XMLSchema#gYear">-480</hemlRDF:SimpleTime>
        </crm:P82F.at_some_time_within>
        <crm:P2F.has_type rdf:resource="http://perseus.tufts.edu/starts_with"/>
        <crm:P3F.has_note>ca.</crm:P3F.has_note>
    </crm:E52.Time-Span>
</crm:P4F.has_time-span>
```

Example 4. Perseus CIDOC-CRM Data With Added Elements

Note that the non-heml date elements encode time using a URI. Since the purpose of CIDOC-CRM is to identify and coordinate the resources, this is a wholly appropriate use. However, any RDF application that is acting on this as a date, Heml included, will need to access it as a literal, preferably with appropriate datatyping. We used Perl regex to insert a sister element to the `Time_Primitive`. Since this might not always be easy to do, developers of CIDOC-CRM data who wish to explore the data in systems like Heml would be well advised also to generate literals for dates and locations that employ standard datatypes such as the `xsd:gYear` illustrated above. 27

A SPARQL query — available in our SVN server — transforms the rest of the CIDOC-CRM into HemlRDF, which then can be processed by the visualization tools and viewed alongside events encoded in other schemas. 28

Heml's Future

From its inception, the XML-based Heml has used the Apache Cocoon Web service framework to store data, to set controller values and to transform data into the views shown above. In the experiments regarding the potential of RDF illustrated above, we grafted the Jena RDF parser into this system, using it as a means of down-converting Heml RDF data into the XML format around which the Heml Web service framework grew.[9] The resulting system and its critical components are illustrated in Figure 4. This is a heavy-weight, server-side toolkit out of keeping with modern trends in Web software. Although the project strives to make its software freely available through GPL licensing, the general effectiveness of Heml has been reduced because of the project's server-side orientation. 29

Figure 4. Block Diagram of the Components of the Current (v 0.7.2) Heml Server Software. The Current Heml Software Relies on the Client-Side XML Pipeline Technologies of Apache Cocoon.

Since the advanced features of Heml described above will require RDF technologies, our project proposes next to remove the XML format entirely from its processing pipeline. Ideally, we wish to follow the general trends of Web computing and pass all processing over to the client; it will occasionally call its server for additional RDF data. The proposed architecture is illustrated in Figure 5. 30

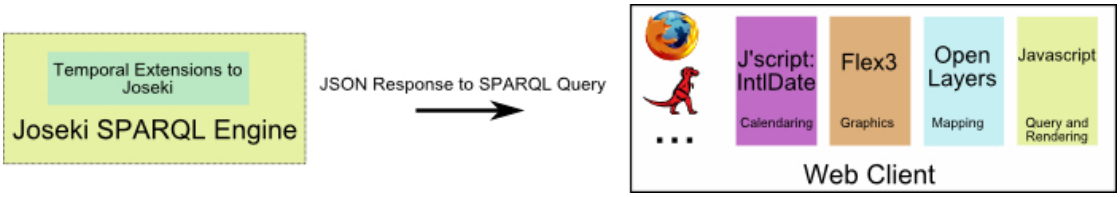

Figure 5. Block Diagram of the Components of the Projected Heml Software. Future Heml Software Development Will Aim to Do All Processing on the Client.

Although its data store and communication are modelled in RDF, this system still can consume any (non-RDF) XML that is of historical interest, such as the biographical P5 TEI tags. An appropriate XSLT stylesheet converts such materials to XML-encoded RDF. Finally, since it appears that declarative animation in the SVG format is being abandoned, Heml will explore other vector graphics formats and likely use the popular cross-platform Flex3 platform from Adobe. We are also adopting of third-party visualization tools that have arisen in recently, such as Simile Timeline and the OpenLayers mapping suite.

The combination of the SPARQL query language and RDF is reasonably powerful, but it lacks some temporal capabilities which we have had to add through extension functions written for the Joseki SPARQL server engine. For instance, while the SPARQL specification permits the query to request a response that is ordered by one or more datatyped values, the only temporal datatype that it will order in this way is `xsd:DateTime`, that is, dates specified to the second; the specification does not indicate how to order dates specified to the year or day, much less to compare between these datatypes. Our extension function uses SPARQL Update to apply to the in-memory RDF graph a starting and ending `xsd:DateTime` corresponding to the beginning and ending of the chronological ranges indicated by these other datatypes. Further improvements in chronological analysis could be made through the use of the CIDOC-CRM's predicates based on Allen temporal logic [Allen 1983] and further use of SPARQL/Update.[10] 31

Projected Work

Finally, we plan to undertake the automated discovery of references to historical events in unexplored digitized texts. We can expect these texts to be created in ever-increasing numbers, either by large-scale endeavours like the Open Archive Initiative, by the Google Books project, or through local efforts such as the one described in [Carrera 2005], which become more practical as improved Optical Character Recognition software is developed in large-scale endeavours. The method described below was inspired by Greg Crane, whose published thoughts on the topic began in [Crane 2006]. 32

[Smith 2002] and the work of the TimeML group show how effectively extracted named entities and natural language processing techniques can be used to identify events. While these natural language processing approaches discover hitherto unidentified events — typically from newspaper articles — we intend to apply some of the same machine learning techniques to discover references to *canonical* events, already encoded and stored in HemlRDF. (Such canonical events would be extracted from standard reference works.) This system thus could train linear classifiers with the persons and places associated with an event, as well as other additional features encoded in HemlRDF. For instance, the appearance of a reference to a primary source already associated with an event would be a strong indication that that event is being discussed. While in some disciplines it would be a considerable task to accurately identify such references with the one on record in the RDF database, Smith's work on Canonical Text Services promises to simplify the task. The event's label will also serve as a useful feature, especially if it were parsed with the Wordnet database, which itself is now available in a RDF/OWL format. 33

Conclusion

RDF technologies are providing Heml with several new avenues for research and a plan for reinventing its means of delivering historical visualizations. If it is true that “...people have achieved their humanity in part by attaining a fuller comprehension of their own place in time and space” [Clark 1992, xi], we might hope that the historical potential of the 34

Web will some day give us an unprecedented understanding of our place in time and a greater sense of our own humanity.

Acknowledgements

The RDF data in this paper was produced through a Curricular Innovation Grant at Mount Allison University. The National Research Council of Canada funded the development of CIDOC-CRM and Heml XML/RDF bridging code through its Industrial Research Assistance Program.

Notes

[1] It might be observed that since Wikipedia has made great strides in organizing historical information on the Web, the (seemingly inevitable) continued growth of this resource will fulfil this goal. However, as an encyclopedia, Wikipedia does not aim to be exhaustive. Its notability guidelines rightly forbid a great deal of material to which historians might wish to have ready access. Moreover even those Wikipedia article which do explain historical events cannot comprise exhaustive lists of references, secondary sources and other evidence. As described later in this paper, the Heml Project has has good experiences using the Semantic MediaWiki extension to its Mediawiki software as a means of RDF data entry.

[2] [Carrera 2002] defines the set of “cultural semantics” as those concept which requires the localization of language and terminology. [Drucker and Noviskie 2004] make it clear that the problem of time-keeping and calendars, personal or collective, should be added to the list of “cultural semantics.”

[3] The Electronic Cultural Atlas Initiative has developed a coordinated network of databases, whose limited access helps to ensure the authority and quality of the aggregated data. A windows application is available for entering conforming data, and the excellent TimeMap software produces rich historical GIS visualizations in the client's browser.

[4] This precludes the use of the exhaustive representation of temporal expressions in natural language that the TimeML markup language provides.

[5] See the timeline views of the `columbia_accident` and `greek_prehistory` at http://www.heml.org for examples of very brief and very long events rendered in this view.

[6] Tim Berners-Lee's Primer on RDF is a good introduction for those new to the subject.

[7] Two widely recommended means of RDF reification that appear to be sufficient for Heml are the appropriate use of `rdf:ID` and predicate subclassing. For our purposes, as for most humanists', the choice will depend on support for these in query languages and their implementation.

[8] In particular, since Semantic Media Wiki treats each node is an article, nodes which are usually treated as “blank nodes” in RDF are difficult to deal with in this Wiki extension.

[9] While appropriate XML can be converted to RDF with an XSLT transformation (and the Heml Project has therefore provided such a transformation for many years), the opposite is not true: there is no XML-based tool that can reliably convert RDF (in its standard XML format) into a sister, non-RDF, XML format.

[10] It is also likely that the temporal built-in fuctions for SWRL provided with the Protégé ontology editor could be modified for use as extension functions in SPARQL queries.

Works Cited

Allen 1983 Allen, James F. “Maintaining Knowledge about Temporal Intervals”. *Communications of the ACM* 26 (1983), pp. 832-843.

Carrera 2002 Carrera, F. “Challenges for a Semantic Web”. Presented at *Semantic Web Workshop*. *Proceedings of the International Workshop on the Semantic Web 2002* (2002), pp. 16-22.

Carrera 2005 Carrera, F. “Making History: an emergent system for the systematic accrual of transcriptions of historic manuscripts”. *Proceedings of the Eighth International Conference for Document Analysis and Recognition* (2005), pp. 543-547.

Clark 1992 Clark, Grahame. *Space, Time and Man: A Prehistorian's View*. Cambridge and New York: Cambridge University Press, 1992.

Crane 2006 Crane, Gregory. “What Do You Do with A Million Books?”. *D-Lib Magazine* 12: 3 (2006). http://www.dlib.org/dlib/march06/crane/03crane.html.

Drucker and Noviskie 2004 Drucker, Johanna, and Bethany Noviskie. “Speculative Computing: Aesthetic Provocations in Humanities Computing”. In Susan Schreibman Ray Siemens and John Unsworth, eds., *A Companion to Digital Humanities*. New York: Blackwell Publishing, 2004.

Nakahira 2007 Nakahira, Katsuko T., Masashi Matsui, Kazutoshi Aboko and Yoshiki Mikami. “The Use of XML to Express a Historical Knowledge Base”. Presented at *WWW 2007*. *Proceedings of the Sixteenth International World Wide Web Conference* (2007).

Robertson 2006 Robertson, Bruce. “Visualizing an Historical Semantic Web with Heml”. Presented at *WWW 2006* (May 23-26 2006). *Proceedings of the 2006 W3C Conference* (2006).

Robertson 2002 RobertsonB. “An Overview of the Historical Event Markup and Linking Project”. In Niccolucci Franco, ed., *Dalla Fonte alla Rete: Il linguaggio XML e la codifica dei documenti storici, archeologici e archivistici*. *Quaderni. Centro di Ricerche Informatiche per i Beni Culturali*. Pisa: 2002.

Robertson 2004 Robertson, Bruce. “Improving Ancient History Online with Heml”. *Classics@* 2 (2004).

Smith 2002 Smith, David A. “Detecting Events with Date and Place Information in Unstructured Text”. Presented at *JCDL 2002*. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* (2002), pp. 191-196.

Thaller 1991 Thaller, Manfred. “The Historical Workstation Project”. *Computers and the Humanities* 25 (1991), pp. 149-162.