

Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture*

D. Neel Smith¹ and Gabriel A. Weaver²

¹ College of the Holy Cross, Department of Classics, Worcester MA 01610, USA

² Dartmouth College, Department of Computer Science, Hanover NH 03755, USA

Dartmouth Computer Science Technical Report TR2009-649

Version of June 1, 2009

Abstract. Domain knowledge expressed in structured citation formats can be exploited in data mining. We propose four structural properties of canonically cited texts, then look at two classic problems in the study of the *scholia*, or ancient scholarly commentary, found in the manuscripts of the *Iliad*. We cluster citations of *scholia* to analyze their distribution in different manuscripts; this leads to a revised view of how the manuscripts' scribes drew on their source material. Correlated frequencies of named entities suggest that one group of manuscripts had access to material more closely based on the work of the greatest Hellenistic editor of Homer, Aristarchus of Samothrace.

1 Introduction

The World Wide Web is undoubtedly the world's largest collection of freely available digital documents: one recent estimate places the number of unique URLs above one trillion (1,000,000,000,000).[1] While this vast resource has certainly motivated much work in data mining, the comparatively free form of most documents on the web has encouraged approaches that view documents as unstructured bags of words (which may be related to each other in ways that data mining applications can recover by considering information like hypertext links in HTML documents). In this paper, we consider a different situation: mining an essentially fixed corpus of texts cited by canonical reference, such as the surviving corpus of ancient Greek.³ We begin with an explicit theory about the structure of canonically cited texts, and how data mining applications can exploit the domain knowledge embedded in the practice of canonical citation. We then apply this knowledge to two classic problems in the study of the *scholia*, or ancient

* The authors' work was supported by non-residential fellowships from the Center for Hellenic Studies, Harvard University.

³ Much of this research has grown out of our involvement with the Center for Hellenic Studies' "First Thousand Years of Greek" project. We are especially grateful to the Center's director, Gregory Nagy, for his support of our work.

scholarly commentary, found in the manuscripts of the *Iliad*. Finally, we suggest how this kind of domain knowledge could, in future research, be extrapolated to web-oriented data mining problems.

2 Canonically Cited Texts

Both theoretical work and hands-on experience with digital texts over the past twenty years lead us to propose that all canonically cited texts possess four properties offering potentially useful information to data mining applications, namely, that:

1. citable units of a text are ordered
2. citable units of a text are organized in a (possibly flat) hierarchy
3. versions of a text are related to a notional text in a conceptual hierarchy
4. citable units may include mixed content

In a seminal article, DeRose, Durand, Mylonas and Renear in 1990 boldly proposed that the true nature of text was an “ordered hierarchy of content objects” (OHCO).[2] Basing their theorizing on experience in textual markup, three of the original authors soon modified this strong statement[4] to address cases where different analytical perspectives focus on structures best expressed in markup through concurrent, overlapping hierarchies. Poetic enjambement, for example, occurs when syntactic and prosodic structures overlap each other, so that in these cases sentence structure and units of verse cannot be expressed in a single, contained hierarchy of “Chinese box”-like elements. In discussing possible approaches to problems of overlapping hierarchies in markup, Renear et al. point out that one option, “the most radical,” is “simply to pick a single hierarchy as the ‘real’ document hierarchy, and flatten all other hierarchies.” This approach is radical for the question Renear et al. propose: what is the “real” nature of text? But in actual practice, this is exactly what scholars routinely do: in scholarly discourse, we cite references in a canonical citation scheme, to which we map other analytical structures. Classicists discussing the syntax of the *Iliad* will not be troubled by the fact that sentences and clauses do not necessarily align with individual hexameters, for example; they will cite the *Iliad* by book and poetic line, as will scholars approaching the *Iliad* from any other analytical perspective. To a greater extent than many disciplines in the humanities, classicists have tended to establish logical, hierarchically organized citation schemes (like the books and lines of the *Iliad*) for the texts they study. Our common denominator for reference generally satisfies the modified OHCO hypothesis that meaningful units of text are organized in an ordered hierarchy or content objects. These citation systems express domain knowledge about the structure of a text.

Since the work of DeRose, Durand, Mylonas and Renear, research in the library cataloging community has made us more aware of another hierarchy that humanists have traditionally expressed in their citation practice. In the model of the Functional Requirements for Bibliographic Records (FRBR), a notional work

may be recognized in multiple expressions (such as different editions, or translations), which may themselves be represented by distinct items (e.g., a physical copy of a book).⁴ While citation in the OHCO model expresses knowledge about the structure of a text, identification of a reference within a FRBR-like hierarchy captures the relations among versions of a text. Data mining applications can leverage this knowledge both for intertext comparison of different versions of a text, and for cross-lingual alignment of versions in different languages.

Finally, in classics, as in other areas of the humanities, a great deal of work has been dedicated to analyzing otherwise undifferentiated chunks of text. In the data mining problems discussed in the following section, for example, we draw on results of the morphological parsing system developed at the Perseus Project. Thanks to the Perseus parser, it is equally easy to view the tokens in a passage of text as literal strings, or as instances of a lemmatized lexical entity.⁵

We have been part of a technical working group at Harvard University's Center for Hellenic Studies that has defined a suite of network services codifying these ideas about the structure of texts. At the core of the group's work are conventions for canonical reference to passages of texts, and to discrete objects. Texts are cited with a Uniform Resource Name (URN) notation that expresses in a simple string the position of a cited passage of text both in the OCHO hierarchy of a canonical citation scheme, and in a FRBR-like conceptual model of a work. (Originally developed for use with the Center's Canonical Text Services, or CTS, the notation is therefore called a CTS URN.)^[5] Discrete objects are cited by namespace-qualified identifiers; an extension mechanism allows the definition of more detailed type-specific reference schemes (e.g. to a rectangular region of interest on an image identified by a namespace-qualified identifier). Simple pairings of canonical references can then associate any two citable objects. A set of network services then support the discovery and retrieval of objects by canonical reference. Acronym-mongers have coined the label CITE for this architecture.⁶

For any text known to a CTS, we can discover all valid citation values, expressed as CTS URNs, in their document order, and of course we can retrieve

⁴ The formal description of the model is available from <http://www.ifa.org/VII/s13/frbr/frbr.pdf>. For current information about FRBR, and ongoing activity in the very active FRBR community, see the FRBR blog at <http://www.frbr.org/>.

⁵ Although the source code for the parser has not been published, Peter Heslin has run the parser over the word list of nearly one million unique strings distributed with CD versions of the Thesaurus Linguae Graecae corpus, and made the results freely available as part of the "expert" package of his Diogenes software. The result is that virtually any known word token can be looked up and a list of one or more valid morphological analyses retrieved. See <http://www.dur.ac.uk/p.j.heslin/Software/Diogenes/> (with links to download).

⁶ The principal network services support discovery and retrieval of Texts, Collections of discrete objects, Extended citation of specific object types, and Indexes relating pairs of objects: or, reordered to make a pronounceable acronym, Collections, Indexes, Texts and Extensions, the CITE architecture. For a fuller introduction to the CITE architecture, see^[6]

the contents of a passage, given a CTS URN. Since the CTS URN by itself explicitly places a passage of text within both the OHCO citation hierarchy and the FRBR-like document hierarchy, we can trivially collect text nodes for data mining with three of the four properties of canonically cited text (citation order, citation hierarchy and document hierarchy) by walking through the CTS requests for a series of nodes, and recording in sequence the CTS URN along with the text content of the node. For the text mining problems discussed in the next section, we then tokenize the content of each citable node, and query a Collections service to analyze each token sequentially (again, keeping track of their order in the document). We assign each token to a type, and possibly attach further analytical information (as explained further below). The result is a simple tabular structure: a sequence number, a CTS URN, a token, and analytical data about the token. This structure is easily adapted to a variety of data mining queries, while it fully preserves the four properties of a citeable text that provided our point of departure. In fact, if a continuous XML text were desirable for other applications, we could easily generate one from this tabular structure (although to maintain the analytical information about each word, we would either have to relegate that data to standoff markup, or burden the XML with tags on every token in the text). The abstraction of these definitive properties of citable texts is attractive in part precisely because they provide criteria for assessing the expressive equivalence of different representations of a text: we initially acquired the data to mine through queries to CITE services; we constructed tabular data from the results of these queries; we could equally well generate XML texts encoding the same information; and in each transformation of the data, we can test the preservation of these four dimensions of the text's structure.

3 Two Examples

With this theoretical framework in place, let us turn now to concrete problems concerning a very difficult set of texts: the scholarly notes, or *scholia*, on the *Iliad* that are found in the margins or between the lines of a number of medieval manuscripts of the *Iliad*. These texts are of unique importance for several reasons. In the first place, they are primary documents for the history of literary scholarship in the middle ages, but classicists are often more interested in the *scholia* as evidence for earlier scholarship, since they certainly drew on and often even cite by name famous scholars of antiquity whose works no longer exist. These ancient scholars, in turn, read and debated what they found in a variety of versions of the *Iliad* that do not survive today, so that at one further remove the *scholia* become evidence for the earlier transmission of the *Iliad*.

Scholars' hope of peeling back the historical layers of evidence preserved in the *scholia* has led to perennial debate over two questions we can approach through data mining. First: what is the relationship among the principal manuscripts with extensive *scholia* on the *Iliad*? And second: can we attribute the sources of specific notes to specific ancient scholars?

One obstacle to scholarly consensus on these questions is the format of the standard printed edition of the major *scholia*.^[3] The editor, H. Erbse, chooses to group together scholia from different manuscripts that comment on the same passage of the *Iliad*. While this makes it very easy to look up a given passage of the *Iliad* to see what *scholia* exist for that passage, it makes it essentially impossible to determine systematically from the print edition what contents are included in a given manuscript. For the analyses described here, we worked with an in-progress digital edition of the texts, based on readings in Erbse's printed text, but breaking out a total of ten distinct sets of scholia in the six major manuscripts.⁷ The texts are available from a CTS, cited by CTS URNs. When an identical *scholion* appears in two or more of the ten documents, it shares the same citation value, just as lines of the *Iliad* are numbered the same way in different editions. When an edition of the *Iliad* omits or transposes lines from their canonical position, the text of the *Iliad* reflects this with omissions and transpositions: the lines that are present carry their familiar identifiers, whether or not they run in numerical order. Similarly, the *scholia* that are present in a given manuscript carry their normal identifiers, even if other scholia are absent, or appear in a different sequence.

For this study we used complete texts of the *scholia* to five of the twenty-four books of the *Iliad*: books 2, 6, 7, 8 and 10. This group includes approximately four thousand distinct *scholia* (out of a total we estimate to be around 20-22,000 for all of the *Iliad*), so it is large enough to give us meaningful results. The content is also varied enough to give us confidence that results for these five books are likely to be valid for the *scholia* on the *Iliad* overall. Our results should be considered as preliminary, however, and should be retested once the digital edition is complete.

3.1 Relations of Manuscripts of the Scholia

If we want to understand the evidence of the *scholia* for earlier scholarship and earlier versions of the *Iliad*, we need to understand how the ten groups of comments in our six manuscripts relate to each other. The approach to textual criticism that evolved in the nineteenth century reconstructs the history of a text by placing versions in a sort of family tree. When different texts are compared, the assumption is that, since each manuscript attempts to reproduce its source as faithfully as possible, differences between two manuscripts must reflect an error in copying by one (or both!) of the manuscripts' traditions.

This traditional stemmatic recension has been unthinkingly applied to the Iliadic *scholia*, but the briefest overview of our digital text is enough to show

⁷ Manuscripts A and T each have a principal set of scholia in the main margins of the codex. T has a further, distinct set written between lines of the text. In manuscript A, we treat as distinct documents the main scholia in the exterior margin, a set of interlinear scholia, a set of scholia written secondarily between the main scholia and the text (the intermarginal scholia), and a set of scholia in the interior margin. The other manuscripts with a single set of marginal scholia are the closely related group composed of B, C, E3 and E4.

table above summarizes the results as we progress from an initial value of $N = 2$, until at $N = 8$, the clustering process stabilizes and is unable to subdivide the clusters further. At $N = 2$, the T interlinear *scholia* (Til) are proposed as a single cluster (11%), with all others forming a second group. At $N = 3$, the four manuscripts of Erbse's **b** group, namely B, C, E3 and E4, cluster together with T. (Some *scholia* that also appear in Til fall in this category: the isolated Til cluster drops to 9%.) At $N = 4$, A emerges as a large, distinct group, that remains an independent cluster with about 24% of our instances, while at $N = 5-7$, subgroups of the larger B, C, E3, E4, T group appear. In the final iteration, manuscript A's intermarginal *scholia* (Aim) and interior *scholia* (Aint) are distinguished.

These results suggest a historical interpretation that resembles traditional text critical claims about the relation of these manuscripts in certain points, but offers a clearer picture of how the scribes drew on their source material.

First, we observe that the secondary interlinear and intermarginal *scholia* Til, Aim and Aint are clearly distinguished from the main marginal *scholia* (A, B, C, E3, E4 and T). The clustering algorithm finds the secondary of *scholia* readily in part because of the well understood fact that within a single manuscript, the main marginal *scholia* and the secondary *scholia* are almost perfectly complementary sets: if a *scholion* is in T it will not be in Til, and vice versa. But Til is the first cluster to emerge (at $N = 2$) in part because the Til *scholia* also have a complementary relation with *all* of the A *scholia*: if a *scholion* is in A, Aim or Aint, it is not in Til, and vice versa. There is some overlap between Til, and Erbse's **b** family of B, C, E3 and E4. This strongly suggests that the Til *scholia* derive from a source not used by or known to the scribe of any of the A *scholia*, used occasionally in composing the main *scholia* of B, C, E3, and E4, and reserved in manuscript T for the interlinear comments of Til (in contrast to the main marginal *scholia*, T). The secondary *scholia* of Aim and Aint, on the other hand, appear nowhere else: the scribe of A seems to have drawn them from a source not familiar to the scribes of the other manuscripts. We can see quite clearly that when scribes create a secondary set of *scholia* distinguished by position on the folio (and often by size or even hand of the script), they are drawing on a different set of sources than those used for the main *scholia* of the same manuscript.

Among the main *scholia*, it is striking how clearly A stands apart from the other manuscripts: its separation is sharper than traditional text critical summaries might suggest. B and E3, long recognized to be very close, appear to be more nearly identical than any other pair. They are also the largest sets of *scholia* within Erbse's related **b** group. B and E3 might well represent a very nearly exact copy of a common source, while the smaller sets of *scholia* in the related manuscripts C and E4 illustrate how a scribe working primarily from one source could select from and add to material relying heavily on a single source. T appears to be aware of the **b** family's source, but in the main *scholia* drew from it much more selectively, alongside material not represented in the **b** group.

We see then that A stands apart from the other manuscripts in both its main and secondary *scholia*; the **b** family draws heavily on a single source but while

B and E3 follow it closely, C and E4 adhere to it less rigidly; and T, while aware of the **b** family’s main source, uses it much more selectively in the main scholia, and includes a secondary set of *scholia* from a source that is unknown to A, and barely appears in the **b** group.

While our results are still preliminary, cluster analysis on different subsets of our data produces consistent results, that are also consistent with a historical reconstruction of the process of composition that is much more plausible than the assumptions of traditional stemmatic criticism.

3.2 Identifying Ancient Sources for the *Scholia*

Cluster analysis using only the structural information encoded in a CTS URN can help us understand the ways that medieval scribes selected material from traditional sources in composing a commentary on the *Iliad*. Can we extract further information about the sources for particular *scholia*?

With the citation information of a CTS URN, it is equally easy to extract information at any level of the citation hierarchy. An obvious approach that we experimented with extensively would be to treat each *scholion* as a virtual “document,” and to apply well known document classification techniques. For training and testing sets, we might hope to use *scholia* where an explicit reference in the text justifies our attribution of the *scholion* to a particular ancient source, and then attempt to construct a classifier that we could apply to unattributed *scholia*.

We tried several variations on this approach. We used as our basic data set the tabular representation of a text described in the opening section of this paper, and derived from the “First Thousand Years of Greek” project. For each white-space delimited token, the attributes we record are: a sequence number, a CTS URN, and assignment to a token type of “parseable word,” or “named entity.” Parseable words are further mapped to an identifier for a lexical entity; named entities are mapped to identifiers uniquely identifying the named entity. (There is only one Aristarchus in the corpus we studied, but several ancient scholars named Ptolemy, for example.) Based on this source, we can construct a bag of words at any level of the citation hierarchy, composed of any combination we choose of surface string forms, identifiers for lexical entities, and identifiers for named entities.

We tried using different thresholds for the minimum number of occurrences of a token throughout the corpus to be considered, and for a maximum to define stop words, but ultimately had to abandon document classification at level of individual *scholia* because the average comment is simply too brief. The longest *scholion* in our five-book sample is 1351 words, but the average length is only 45 words (and the minimum, one word). Only 482 *scholia* had more than 100 words; only 113 had more than 200 words.

If individual *scholia* are too brief to treat as virtual documents, can we detect patterns in groups of *scholia*? At the other extreme of the citation hierarchy, if we consider an entire set of *scholia*, the problem becomes how to find diagnostic material in the contents of its entries. We extracted all the named entities in each

document group, and were immediately struck by the fact that the distribution of particular entities was very uneven.

Among the named Hellenistic scholars cited or referred to in the *scholia*, by the most frequently named was the most famous and influential ancient scholar of Homer, Aristarchus of Samothrace. In the table below, we summarize how many times his name appears in each of our ten document groups, how many tokens are found in each document, and in the final column, a normalized value showing how often Aristarchus is mentioned in 100 words of text.

MS	Aristarchus	Size of text	Normalized
Aim	95	2204	4.31
Aint	31	1440	2.15
Til	32	2278	1.4
A	298	25589	1.16
T	103	32464	0.32
C	38	27530	0.14
E3	42	30894	0.14
B	42	30914	0.14
E4	23	24335	0.09

Occurrences of Aristarchus in the *Scholia*

The first point to emphasize is the range of the normalized frequency. It is quite extraordinary that Aristarchus' name should appear more than four times per hundred words in the intermarginal notes of manuscript A when we can see that Aristarchus is mentioned less often than once in per thousand words in E4. There are three quite groups: A, together with all of the secondary *scholia* in A (Aim and Aint) and in T (Til) refer to Aristarchus often; the **b** family of B, C, E3 and E4 refer to him infrequently; and the main *scholia* of T occupy something of an intermediate position. Particularly when we bear in mind the complementarity of main *scholia* and secondary *scholia* we mentioned before, the absence of Aristarchus from the **b** group is striking.

The second point to emphasize is that this clear triage is the more notable because it corresponds precisely to the pattern we independently established in our cluster analysis. In the previous section, we suggested that the secondary *scholia* of A derived from a unique source, and that the main *scholia* of A were more independent than usually imagined from the other manuscripts with Iliadic *scholia*; we further noted the distinctive place of Til. The manuscripts that were singled out by our cluster analysis, and that we interpreted historically as drawing on distinct sources, refer to the most Homeric important scholar of the ancient world ten times more often than the other texts available to us.

This also hints at possibilities for future research. To improve our chances of attributing groups of *scholia* to their ancient sources, we might be able to identify specific lexical or named entities that, like the named entity "Aristarchus,"

are not evenly distributed across the manuscripts and that correlate with the results of our independently derived cluster analysis. Perhaps we could use a well chosen set of such entities in the text as specific attributes to focus on in a document classification of individual *scholia*, where a more open-ended bag-of-words classification cannot separate the signal of significant terms from the surrounding noise in very small documents.

Our preliminary work here suggests that in the distribution of scholia on the *Iliad* across our manuscripts, we can see how scribes used their source material; in the content of manuscripts, such as references to named entities, we can see an indirect reflection of the content, interests and concerns of these sources. This illustrates how such classic problems in the humanities as analyzing the relation of manuscript sources for a text, or attributing secondary use of material to earlier sources, can be systematically addressed using data or text mining techniques.

4 Directions for Future Research: Extrapolating from a Structured Corpus

In mining the data in the *scholia* to the *Iliad*, we took advantage of the structural information expressed in CTS URNs in two different ways. We used the information in the FRBR-like document hierarchy to compare sets of identically cited *scholia* in different versions of a document. We used the information in the OHCO citation scheme to create virtual documents, by collecting content at different levels of a citation hierarchy (individual *scholia*, or complete sets of scholia in a given document of a manuscript). Mining this structural information led us to new conclusions about the relations among manuscripts of the scholia on the *Iliad*, and helped us identify promising directions to follow to investigate further the ancient sources for the medieval *scholia*.⁸ This was only possible because the texts we examined were associated with CTS URNs. To return to the contrast with which we began this paper: does data mining of highly structured texts have any relevance for scholars who want to quarry the rapidly exploding quantities of material appearing on the World Wide Web?

We foresee two ways in which increased experience with a highly organized corpus of texts could lead to improvements in identifying citation information in a less structured mass like the Web. First, a structured corpus provides a solid basis for aligning less structured versions of the same text, whatever the source. The “First Thousand Years of Greek” project is currently building on the work of many predecessors that have created more or less structured versions of ancient Greek texts in original or in translation. The project has recently

⁸ The examples we discussed did not make use of the information in our tabular structure about the sequence of tokens, since we looked only at the occurrence or frequency of tokens, without considering their order. We note only that many text analysis applications depend on sequential structure. E.g., for context-aware disambiguation of multiple morphological analyses for a single token in our data, hidden Markov models might be applied.

completed its initial inventory of material, and is beginning to prepare the same kind of tabular data used in our data mining examples, above, for several million words of ancient Greek, explicitly keyed to CTS URNs. This corpus will provide a useful baseline for citation alignment of different versions of ancient Greek texts, whether in some other digital collection, or harvested directly from the Web. It is not farfetched to imagine a spider crawling a targeted section of the web looking for contents that closely match text in the base collection organized by CTS URN.

Unlike the open-ended production of the Web, the corpus of ancient Greek is closed. While it is too large a corpus for any individual to read, no matter how many new discoveries we make, it remains a finite amount: no new ancient Greek is being created. Depending on one's precise definition of "ancient Greek," that amount probably numbers only a few tens of millions of words—five orders of magnitude smaller than the estimated number of unique URLs on the Web today. That scale seems to represent a sweet spot for classical scholars: we need computational approaches to make sense of the whole corpus, but as a discipline we are progressing rapidly towards digital texts including explicit representations of the structured citation information used in the data mining examples discussed in this paper. That in turn may mean that in the near future classicists will be well positioned to project the structure of their their corpus on to amorphous texts, and thereby increase their ability to identify and harvest information from a continuing exponential explosion of digital texts.

References

1. Alpert, Jesse and Nissan Hajaj, "We knew the Web was big...", <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>
2. DeRose, Steven J., David G. Durand, Elli Mylonas and Allen H. Renear, "What is text, really?" *Journal of Computing in Higher Education* 1:2 (1990) 3-26.: Full text available from <http://doi.acm.org/10.1145/264842.264843>
3. Erbse, Hartmut, *Scholia Graeca in Homeria Iliadem (scholia vetera)* (Berlin, De Gruyter: 7 vols., 1969-1988).
4. Renear, Allen, Elli Mylonas, and David Durand, "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies," final version of 1993 available from <http://www.stg.brown.edu/resources/stg/monographs/ohco.html>.
5. Smith, Neel, "Citation In Classical Studies". In G. Crane and M. Terras (eds) 2009, *DHQ, Special Issue: Transforming classical Studies Through Cyberinfrastructure. Digital Humanities Quarterly* Vol 2 No 2. <http://www.digitallhumanities.org/dhq/>
6. Smith, Neel, "Digital Infrastructure and the Homer Multitext Project," submitted for publication in *Digital Research in the Study of Classical Antiquity*, edited by Gabriel Bodard and Simon Mahony (Ashgate Press, forthcoming).