Manuscript Studies

Volume 1
Issue 1 *Spring 2017*

Article 2

2017

Citation and Alignment: Scholarship Outside and Inside the Codex

Christopher Blackwell
Furman University, cwblackwell@gmail.com

Christine Roughan
New York University, cmr639@nyu.edu

Neel Smith
College of the Holy Cross, nsmith@holycross.edu

Citation and Alignment: Scholarship Outside and Inside the Codex

Abstract

We describe a hierarchical approach to modeling text that allows machine-actionable canonical citation of text at many levels of specificity. This model address the problem of overlapping or mutually exclusive analyses. In turn, this flexibility in citation allows rich linking of textual transcriptions and other data to regions-of-interest on digital images, of particular value to codicological and paleographic study. Our examples are from work on Byzantine manuscripts containing Greek epic poetry and scholarly commentary, but our approach can apply to any image-based project in documenting books, manuscripts, inscriptions or other text-bearing surfaces.

Keywords

Homer Multitext project, Digital humanities, palaeography, Homer, Iliad, Venetus A, digital research, digital editions, transcription, citation, manuscript studies, epic, greek, codicology, rdf, xml, linked open data

Manuscript Studies

A Journal of the Schoenberg Institute for Manuscript Studies

VOLUME 1, NUMBER 1

(*Spring 2016*)

Manuscript Studies (ISSN 2381-5329) is published semiannually
by the University of Pennsylvania Press

The Schoenberg Institute
for Manuscript Studies
University of Pennsylvania Libraries

MANUSCRIPT STUDIES

VOLUME 1, NUMBER 1
(Spring 2016)

ISSN 2381-5329

Published by the University of Pennsylvania Press,
3905 Spruce Street, Philadelphia, PA 19104.

Printed in the U.S.A. on acid-free paper.

Manuscript Studies brings together scholarship from around the world and across disciplines related to the study of premodern manuscript books and documents, with a special emphasis on the role of digital technologies in advancing manuscript research. Articles for submission should be prepared according to the *Chicago Manual of Style*, 16th edition, and follow the style guidelines found at http://mss.pennpress.org.

2016 SUBSCRIPTION INFORMATION:
Single issues: $30
Print and online subscriptions: Individuals: $40; Institutions: $90; Full-time Students: $30
International subscribers, please add $18 per year for shipping.
Online-only subscriptions: Individuals: $32; Institutions: $78

Please direct all subscription orders, inquiries, requests for single issues, address changes, and other business communications to Penn Press Journals, 3905 Spruce Street, Philadelphia, PA 19104. Phone: 215-573-1295. Fax: 215-746-3636. Email: journals@pobox.upenn.edu. Prepayment is required. Orders may be charged to MasterCard, Visa, and American Express credit cards. Checks and money orders should be made payable to "University of Pennsylvania Press" and sent to the address printed directly above.

One-year subscriptions are valid January 1 through December 31. Subscriptions received after October 31 in any year become effective the following January 1. Subscribers joining midyear receive immediately copies of all issues of Manuscript Studies already in print for that year.

Postmaster: send address changes to Penn Press Journals, 3905 Spruce Street, Philadelphia, PA 19104.

Visit *Manuscript Studies* on the web at mss.pennpress.org.

Citation and Alignment

Scholarship Outside and Inside the Codex

Christopher Blackwell
Furman University

Christine Roughan
College of the Holy Cross

Neel Smith
College of the Holy Cross

Introduction

Digital information technology is extending the study of manuscripts to new audiences and widening the range of questions that scholars can ask about a codex. With a change of potential audience and a wider range of questions, digital technology also allows, and demands, new approaches to scholarship. Since 1999, the Homer Multitext has explored all of these possibilities through a collaborative project in documenting the scholarly history of Greek epic poetry. The work presented in this paper is the result of collaboration between undergraduate students and faculty working outside traditional institutional settings for research or teaching. We first show how we organize and publish digital research so that we can unify contributions from team members with different interests and areas of expertise within a single project, and so that we can integrate material from multiple independent projects studying manuscripts of widely varying dates and contents. We then introduce a new application of the

CITE architecture's[1] Uniform Resource Names (URNs) for citation[2] that simplifies identifying complex relations among the many possible perspectives on a codex—editorial, codicological, paleographic, historical, etc.

1. *Scholarly Research as Graphs*

A focus of the work of the Homer Multitext is the oldest complete manuscript of the *Iliad*, known as the Venetus A (Venice, Biblioteca Marciana, Codex Marcianus Graecus 454 = 822). The digital representation of the codex includes the following: a suggested date of the tenth century, and a possible origin in Constantinople; a quire analysis relating each of folios 12–327 to its position in its quire; and two alternate models relating the first eleven folios (which their current eighteenth-century binding confuses) to possible original positions.

Editorial teams work page by page to record every feature they observe: texts, graphic elements, or other features such as quire numbers. Each page is related to one or more photographs in the project archive, and individual features are documented with references to specific regions of a standard image. In figure 1, a selection of features appearing on folio 24 recto is highlighted with colors distinguishing the type of feature: here, for example, the principal scholarly annotations, or *scholia*, are highlighted in blue, while *scholia* distinguished both by their placement in the inner gutter and by their contents are highlighted in red. Our edition of the texts is also related to the physical folio: here we explicitly record that *Iliad* 1.602–2.10 appears on folio 24 recto.[3]

---

1 The CITE architecture home page on GitHub: http://cite-architecture.github.io/.

2 The CTS URN specification: http://www.homermultitext.org/hmt-docs/specifications/ctsurn/.

3 A guide for editors contributing to the Homer Multitext project describes this process in more detail. At the time of writing this paper, a revised edition of the guide for work in summer 2014 was in preparation for posting at http://www.homermultitext.org/hmt-docs/. This image shown in figure 1 was derived from an original ©2007 Biblioteca Nazionale Marciana, Venezie, Italia. The derivative image is ©2010 Center for Hellenic Studies. Original and derivative are licensed under the Creative Commons Attribution-

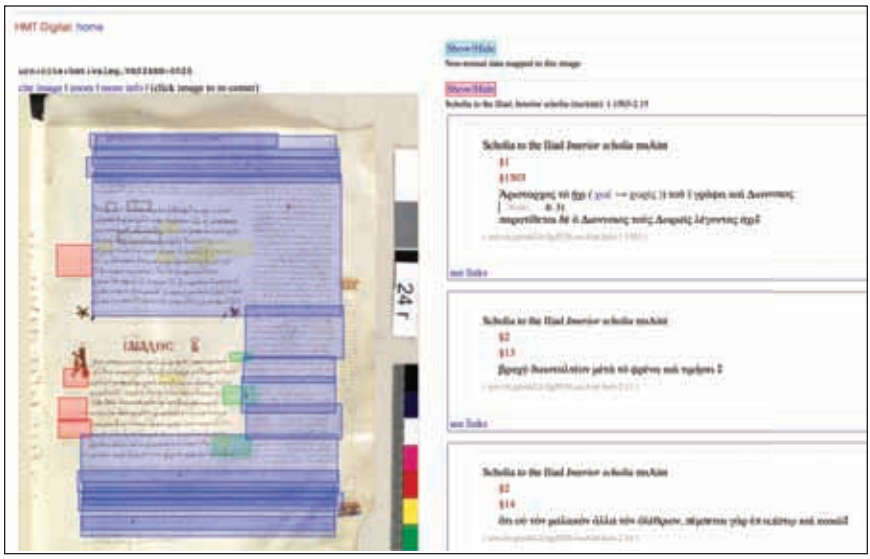

FIGURE 1. View showing record of paleographic observations on fol. 24r of the Venetus A manuscript (Venice, Biblioteca Marciana, Codex Marcianus Graecus 454, now 822).

But the Venetus A manuscript is of interest to scholars outside the Homer Multitext project. As an example of a *de luxe* product of its period and place, it also has an important place in Greek paleography, for instance. A digital record of paleographic observations about a single glyph or character could note the passage of the *Iliad* where the letter occurs, and further link the observation to the visual evidence of a photograph. Figure 1 illustrates this schematically: some of the important relations recorded in the regular work of the Homer Multitext project are shown in black, while the notes of a separate paleographic project are in blue. The two projects are linked by their study of the same codex: implicitly, everything that the Homer Multitext project records is related to the paleographic note; and conversely, all the paleographic project's information can be linked to the Homer Multitext project's editions. How do we translate this conceptual unity into a digital design?

Noncommercial-Share Alike 3.0 License. The CHS/Marciana Imaging Project was directed by David Jacobs of the British Library.

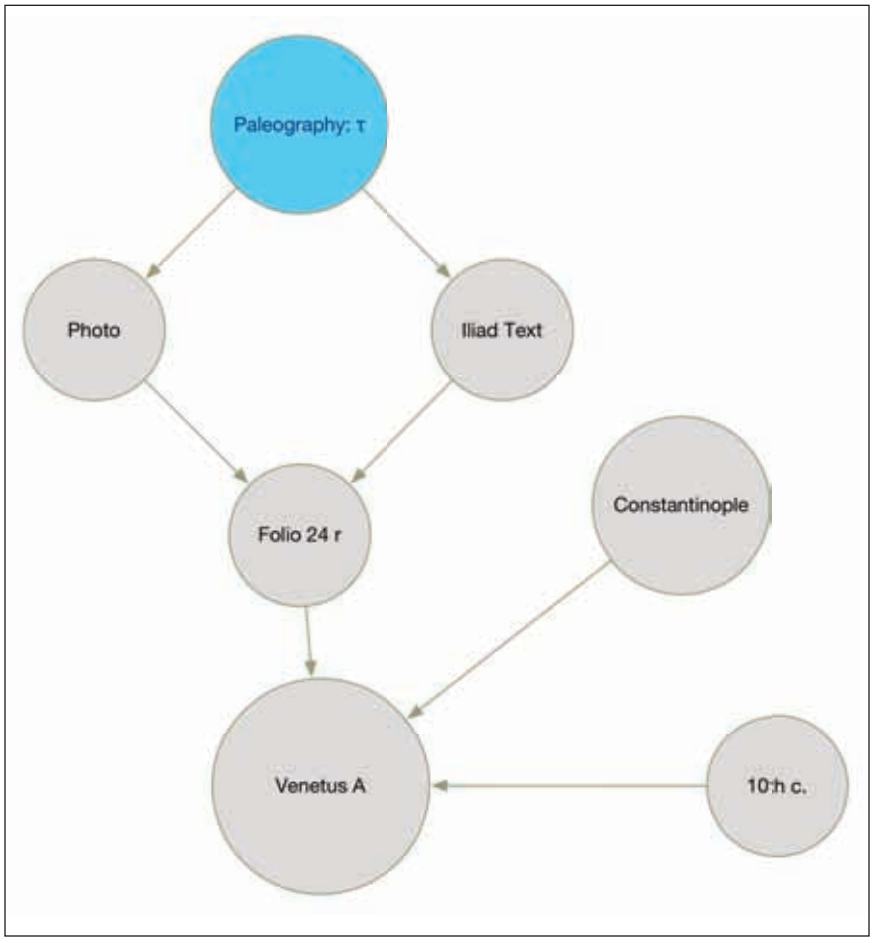

Figure 2. A paleographic observation is implicitly related to everything known about fol. 24r of Venetus A where it appears.

As figure 2 suggests, our scholarly work can be thought of as a kind of network, or, in mathematical terms, a directed graph. A directed graph is a very general data structure that can fully represent information managed in databases, texts, or others formats. It is not surprising, therefore, that when the World Wide Web Consortium set a standard for its Resource Description Framework (RDF) to be a "standard model for data interchange on

the Web,"[4] it chose to define this model as a directed graph. Over the past two years, Blackwell and Smith have developed an automated build system, citemgr for managing archives of data stored in simple formats and using URNs[5] to identify citable resources. The citemgr build system can construct a single graph in RDF syntax that expresses all the information contained in the archive. One metaphor for graphs imagines *nodes* (represented by icons in fig. 2) connected by *edges* (represented by arrows). Alternatively, a pair of nodes joined by an edge can be thought of as a simple sentence in subject-verb-object (SVO) order: the subject and object correspond to nodes, while the verb corresponds to an edge. In this metaphor, a directed graph is a set of SVO statements, so the pair of joined icons in figure 2 labeled "Iliad text" and "folio 24 r" could be represented by a statement like: iliad_1.509–2.10 appears_on folio_24_recto .

In fact, one syntax for an RDF graph, the Terse Triple Language (TTL), is formatted as just this sort of human-readable SVO statements (called "triples").[6]

What is distinctive about the graph built by citemgr is that all nodes in the graph (or all subjects and objects in the TTL expression of the graph) are identified by URNs. The URN naming scheme is an Internet Engineering Taskforce standard. Both the syntax and semantics of a given type of URN are defined, and are "intended to serve as persistent, location-independent, resource identifiers."[7] They are therefore ideal for referring to scholarly resources we expect to outlast ephemeral technologies. In our work on the CITE architecture, we have defined two types of URNs: the Canonical Text Service (CTS) URN for citing passages of text, and the CITE Object URN for citing discrete objects of any kind. We have also defined and implemented network services that retrieve digital representations of objects cited by URN. Applications working with the RDF graph created by citemgr can now resolve the URNs to URLs pointing to an installation

4 The RDF standard: http://www.w3.org/RDF/.

5 The URN specification in RFC 2141: http://www.ietf.org/rfc/rfc2141.txt.

6 The TTL syntax definition: http://www.w3.org/TeamSubmission/turtle/.

7 The URN specification in RFC 2141: http://www.ietf.org/rfc/rfc2141.txt.

Figure 3. A survey of five centuries of Greek paleography, created from a single textual citation.

on the web of one of the CITE architecture services, or for that matter to any other sites that recognize scholarly citation by URNs.

It is important to distinguish citemgr's graph from the kind of "linked data" an application might superficially use.[8] The simple text format of

8 Tim Berners-Lee is perhaps the most prominent advocate of "linked data" using URLs as identifiers, but some enthusiasts for his linked data model overlook the fact that Berners-Lee is explicitly arguing for linked data *on the World Wide Web*, not for a scholarly web designed to persist beyond the lifetime of ephemeral technologies. His famous "5-star" statement (T. Berners-Lee, "Linked Data," edition of 18 June 2009 (http://www.w3.org/DesignIssues/LinkedData.html) directly violates foundational principles

TTL, when all subjects and objects are expressed as URNs, gives us a single summary of an entire scholarly archive in a form that is independent of any specific technology, but easily parsed by machines. Beyond merely identifying something that might be retrieved on the World Wide Web, the URNs we use capture the semantics of traditional scholarly citation practice.[9]

This is especially clear for Canonical Text Service URNs. CTS URNs follow an abstract model of canonically citable texts as an "ordered hierarchy of citation objects."[10] A passage always has a location in the text's order (e.g., *Iliad* 1.610 is followed by *Iliad* 1.611, which in turn is followed by *Iliad* 2.1). Passages may be identified at any level of specificity (e.g., a work of prose cited by book, chapter, and section), and may refer to works as notional works (e.g., "the *Iliad*"), or specific versions (e.g., "the translation by Fitzgerald," or "the edition by Villoison"), or even specific physical copies (e.g., "the copy of Villoison's edition that belonged to Thomas Jefferson and is now in the Library of Congress"). One important consequence for editors of manuscript material is that any edition citable by CTS URN is automatically aligned with other

---

of digital scholarship, since its use of HTTP URIs expresses *identification* with a syntax designed to express *location* (as if a bibliographic citation were given with reference to a single library's catalog number, rather than to generally applicable bibliographic information), and, more seriously, because it does not adequately capture the meaning of a scholarly citation, as explained below.

9 For a fuller overview of the CITE architecture and of citation by URN in particular, see D. N. Smith and C. W. Blackwell, "Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture," in *Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum*, ed. V. Bers, D. Elmer, and L. Muellner (Washington, DC: Center for Hellenic Studies, 2012), http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=4846.

10 In the "OHCO2" model, a citable document consists of a set of nodes with four properties: (1) they are ordered within the document; (2) they are at a leaf node of a (possibly flat) citation hierarchy; (3) they are cited within a work hierarchy; (4) the digital representation of the citable node may include markup or other rich content. N. Smith and G. Weaver, "Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture," in *Text Mining Services: Building and Applying Text Mining Based Service Infrastructures in Research and Industry*, ed. Gerhard Heyer (Leipziger Beiträge zur Informatik 14; Leipzig: Leipzig University, 2009), 129–39 (Reprinted in Dartmouth College Computer Science Technical Report series, TR2009-649, June 2009).

editions and translations, an especially significant advantage when we want to work computationally with a corpus of related material. Now that both the Perseus Digital Library[11] and the Leipzig Open Greek and Latin Project[12] have standardized CTS URNs for referring to texts, editors of manuscripts with ancient Greek or Latin texts can reasonably expect that over the next five years, essentially any Greek or Latin text previously published in print will be available online with CTS URN references.

The semantics of CITE Object URNs are simpler. Every object is a unique member of a collection (e.g., a particular digital image in an archive of images). Some collections may be ordered (e.g., a collection of manuscript pages might be ordered as they appear in the manuscript's current binding). References to unique images include an extension that optionally allows us to identify a region of interest on the image.

Because the nodes of our graph are URNs, we can take a canonical identifier from any source—an interactive user, an automated query, or other computational output—and discover related information by following links in the graph. The paleographic observation in figure 3 is linked to an image that in turn leads us to a physical page. From the image reference alone, we can infer for any given paleographic observation all the information connected to the physical artifact: that the observation appears in the Venetus A manuscript, which is dated to the tenth century, and attributed to Constantinople; that it appears on a digital photograph of the manuscript, taken in 2007 under natural light conditions; that it appears on folio 24 recto, on the ninth folio of the second quire, and even that it appears on the lower half of the page. Linking the observation to a passage of text (in this case, a single letter τ) tells us that the observation appears in the manuscript's main *Iliad* text, in book 2, line 4; that it is the first instance of a tau in that line, and the fourth instance of a tau in that book; and that it appears in the word τιμήσῃ, which is analyzed as a form of the verb τιμάω.

Conversely, if we explore the graph from a reference to a passage of text

11 The Perseus Digital Library: http://www.perseus.tufts.edu/.

12 The Open Greek and Latin Project of Leipzig University's Open Philology Project: http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/.

or a physical artifact, we can automatically find related paleographic observations in ways that would be impractical or impossible with traditional scholarly instruments. The semantics of CTS URNs allow us, for example, to cite a passage of text as part of a notional work, such as "book 2, line 4, in the *Iliad* (any version)," and use that URN to find book 2, line 4, in *all* versions of the *Iliad* known to the graph. Once other objects (such as our example paleographic observation) are related to our texts, we can exploit the power of textual citation to identify those objects, too. In the next section, we consider a shortcut for aligning textual and non-textual data that helps us do exactly that.

2. *Aligning Citations of Texts and Objects*

In 1820, Thomas Jefferson created a small book he entitled the *Life and Morals of Jesus of Nazareth*, which can now be seen online (and even ordered in full-color facsimile) thanks to recent conservation and digitization efforts by the Smithsonian Institution.[13] The pastiche physically aligns passages of texts in Greek, Latin, French, and English to create a multi-column reader. It is interesting to reflect on how we could create a digital edition with versions in parallel columns using CTS URNs. Jefferson's opening section on page 1 is taken from Luke 2: this alignment could be completely constructed from a single URN identifying Luke 2:1. Jefferson understood this: in the right hand margin of each page, he provides the canonical reference that applies to each passage. Page 27, seen in figure 4, is constructed from Luke 5:27–29, Mark 2:15–17, Luke 5:36–38, and Matthew 13:53–56.

The source texts have drawn less attention, but Jefferson's personal copy of the English translation from which he drew his excerpts is thought-provoking (see fig. 5). It is a sort of negative image of the English column of *Life and Morals*, preserving only the text passages that were *not* cut out by

13 The Smithsonian Institution's digitization of "Thomas Jefferson's Bible" (that is, his eighty-four-page *Life and Morals of Jesus of Nazareth*): http://americanhistory.si.edu/jeffersonbible/.

77.

5 And they were the more fierce, saying, He stirreth up the people, teaching throughout all Jewry, beginning from Galilee to this place.

6 When Pilate heard of Galilee, he asked whether the man were a Galilean.

7 And as soon as he knew that he belonged unto Herod's jurisdiction, he sent him to Herod, who himself also was at Jerusalem at that time.

8 And when Herod saw Jesus, he was exceeding glad: for he was desirous to see him of a long season, because he had heard many things of him; and he hoped to have seen some miracle done by him.

9 Then he questioned with him in many words; but he answered him nothing.

10 And the chief priests and scribes stood, and vehemently accused him.

11 And Herod, with his men of war, set him at nought, and mocked him, and arrayed him in a gorgeous robe, and sent him again to Pilate.

12 And the same day Pilate and Herod were made friends together: for before they were at enmity between themselves.

13 And Pilate, when he had called together the chief priests, and the rulers, and the people,

14 Said unto them, Ye have brought this man unto me, as one that perverteth the people: and, behold, I, having examined him before you, have found no fault in this man, touching those things wherof ye accuse him:

15 No, nor yet Herod: for I sent you to him; and, lo, nothing worthy of death is done unto him:

16 I will, therefore, chastise him, and release him.

Figure 4. Page 27 of Jefferson's *Life and Morals of Jesus of Nazareth*, with Jefferson's canonical citations highlighted. Division of Political History, National Museum of American History, Smithsonian Institution.

Jefferson (or were *not* printed on the opposite side of the page to a passage Jefferson excised).

In the terminology of CTS, this is an *exemplar*: a specific instance of a specific version. We could of course create a diplomatic edition of the exemplar. While it ought to agree with other exemplars of this particular translation in every preserved reading, it is unique and interesting to us for its lacunae. The lacunae, the holes where Jefferson removed his chosen verses, become an *analysis* of the source text, and it might be very interesting to follow Jefferson's selections *either* in the order he chose for *Life and Morals*, or in their position in the source text.

If we consider each snippet that Jefferson cut out to be an analysis of that

Figure 5. Jefferson's English New Testament. Division of Political History, National Museum of American History, Smithsonian Institution.

passage, we could identify the analysis with a URN (just as we do with our Greek paleographic observations), and then could align the analysis with the source edition. In this way, we could work with the collection of analyses in terms either of their own CITE Object URN, or of the CTS URN identifying the passage.

This is exactly what we wanted to accomplish with our paleographic analyses, too, and it suggests a different way of thinking about a collection of analytical artifacts. Whether they are manually crafted (like Jefferson's physically excerpted passages, or the manually identified glyphs in our paleographic analysis) or automatically generated (say, a tokenization of a text into word units for morphological or syntactic analysis), when a collection of analyses surveys a specific version of a text, the selections of text it extracts create a kind of individual exemplar of that version. Jefferson's excerpts leave behind most of his source edition. A tokenization into words

to analyze morphologically might omit punctuation. A comprehensive paleographic analysis of a passage might analyze every glyph in a passage, but might treat a ligature as a single feature to analyze even if an edition of the text represents it with multiple characters. Our ability to define different analyses by assigning citations to arbitrary strings of text or regions-of-interest on images offers flexibility to address these issues.

It is particularly obvious that the analysis in Jefferson's *physical* exemplar creates a distinct version of the text, but digital analyses also identify a selection of text in a specific version, and could equally well be considered a kind of *digital* exemplar. Just as a CTS URN referring to chapter 2 of Luke as a notional work allows us to identify different versions of Luke in the graph, while a CTS URN referring to Luke 2 in the specific translation used by Jefferson leads us to its unique English text, so a CTS URN referring to Luke 2 in the specific exemplar of that version created by Jefferson would lead us only to the selection extracted by Jefferson. Since one property of CTS URNs is that they identify a passage in a document's order, almost unexpectedly, we have a way of finding the texts Jefferson analyzed by their original order in the source version. A CTS URN referring to the entire Gospel of Luke in Jefferson's exemplar would give us, in canonical New Testament order, all the passages he extracted from Luke.

If we apply this approach to paleography, we have very nearly achieved our goal of integrating data from different documents while aligning them according to various criteria such as date, alphabetic character, paleographic category, or canonical citation to text. The table in figure 6 shows five paleographic observations aligned with references to *Iliad* 2.4 in the Geneva manuscript of the *Iliad*, Bibliothèque de Genève MS Gr. 44. The table includes a citation of the visual evidence for the observation, and various paleographic notes (not all illustrated here). The CTS URN for the Geneva manuscript is urn:cts:greekLit:tlg0012.tlg001.ge; the first observation analyzes *Iliad* 2.4 in that manuscript, and more narrowly, the first occurrence of a tau in *Iliad* 2.4. From this table, we will create a new exemplar composed of paleographic analyses of the Geneva manuscript, and will identify this exemplar as urn:cts:greekLit:tlg0012.tlg001.ge.pal. A reference to *Iliad* 2.4 in this exemplar will retrieve the text of each analyzed glyph.

But it would be very convenient if, when referring to the analytical ex-

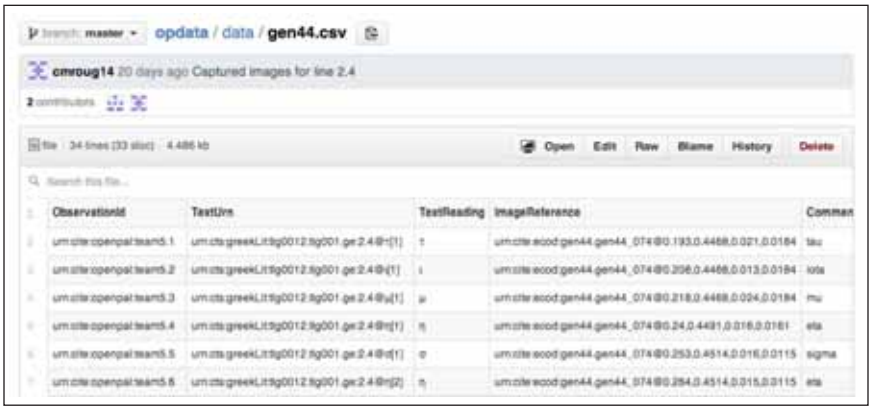

FIGURE 6. Table aligning observation with textual source for a digital exemplar.

emplar, we could follow the sequence of analyzed readings more directly, rather than having to follow a complex computation of whether the first tau in *Iliad* 2.4 precedes or follows the first occurrence of iota. The solution that almost spontaneously suggests itself is to impose an additional level of citation hierarchy that applies only to the exemplar.

That is, we want to equate a canonical reference meaning "in the Geneva Homer, the first occurrence of tau in *Iliad* 2.4" with a reference meaning "in the paleographic exemplar of the Geneva Homer, the first analyzed unit in *Iliad* 2.4". The canonical reference is expressed as a CTS URN like this

urn:cts:greekLit:tlg0012.tlg001.ge:2.4@τ[1]

and would be equivalent to a new CTS URN like this

urn:cts:greekLit:tlg0012.tlg001.ge.pal:2.4.1.

"*Iliad* 2.4" is meaningful reference to a passage in any version of the work, and in any of their exemplars including our digital paleographic exemplar; this is because all versions of the *Iliad* are organized by poetic book and poetic line. But since we have extended that citation scheme with an additional level for this paleographic exemplar, we can cite the text with greater specificity. A reference to a range, like this one

urn:cts:greekLit:tlg0012.tlg001.ge.pal:2.4.1–2.4.3

will now retrieve the text of the first three units analyzed, namely

τιμ.

The same hack could be applied to any analysis. The URN

urn:cts:greekLit:tlg0012.tlg001.msA:1.1

refers to *Iliad* 1.1 in the Venetus A manuscript of the *Iliad*, and for the text of that reference would retrieve from our graph the fully marked up diplomatic edition in TEI XML, namely

~ xml <tei:l n="1"> Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος </tei:l> ~.

In our work on the Homer Multitext project, we create a tokenization of our edition of the Iliadic text of the Venetus A manuscript into individual words for morphological analysis. If we create a digital exemplar with those analyses,

urn:cts:greekLit:tlg0012.tlg001.msA.tokens:1.1.1

will retrieve the first word token of *Iliad* 1.1, namely

Μῆνιν

(with no markup). If for a syntactic analysis, we wanted to select all the word tokens in the first clause, we could cite

urn:cts:greekLit:tlg0012.tlg001.msA.tokens:1.1–1.2.1,

which means "in the tokenization of MS A, all the tokens from *Iliad* 1.1 through the first token of *Iliad* 1.2." This would yield the following text (with all punctuation omitted):

Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην

Our experience has been that because some analytical exemplars make such valuable complements to a full diplomatic edition, we "promote" them in order to work with them as distinct editions. Tokenizations of the text into words are one instance. For example, by identifying the Venetus A manuscript as

urn:cts:greekLit:tlg0012.tlg001.msA

we "promote" the word tokenization of the Venetus A to the status of an edition, identified with the URN

urn:cts:greekLit:tlg0012.tlg001.msAtokens.

Now, a reference to *Iliad* 1.1 in the diplomatic edition returns the full XML source text (as illustrated above), while *Iliad* 1.1 in our tokenized edition returns simply

Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος.

Our "promoted" analytical edition supports the additional citation level, so we can use

urn:cts:greekLit:tlg0012.tlg001.msAtokens:1.1.1

to find the single word token,

Μῆνιν.

Applying the Analytical Exemplar to Paleography

If we return to the situation initially represented in figure 3, where paleographic observations cite passages of text from richly documented editions, we can now consider how treating the paleographic observations as an exemplar of the text can simplify exploration of the scholarly graph. In each of the following examples, we illustrate the selected paleographic observation with the visual reference cited as evidence.

Paleography Identified by Text Reference

A query for *Iliad* 2.4 can lead us to paleographic observations on that passage of text; if we want to cite "promoted" editions with word tokens, we can even compare the paleography of a given "word." In figure 3, we look at the first word token in *Iliad* 2.4 in seven manuscripts of the *Iliad*.

The paleographic exemplar maintains document order, so we can align tokens across manuscripts. (Three of the manuscripts write the first word with seven letters, four manuscripts with only six.) If we add a secondary criterion to our graph query, we can sort the results by the chronology (here, just to the century) of the codex. A single graph query in effect surveys over a span of five centuries how you could write the first word of *Iliad* 2.4.

Paleography Identified by Text Content across Physical Location

It is also helpful to see how similar content is written within a single codex (see fig. 7). If we work with our tokenized edition, it is straightforward to find the paleography of a word like "πατήρ" in different locations within the

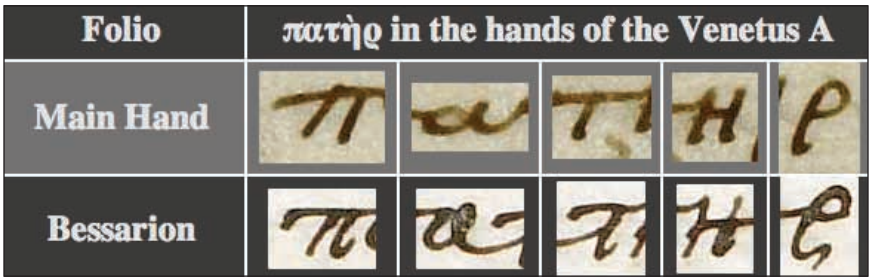

Figure 7. The word πατήρ, from a 10th-century and 15th-century hand on the same manuscript.

same manuscript. One striking feature of the Venetus A manuscript of the *Iliad* is that several pages are later replacements, lacking *scholia* and written in an obviously different and later hand. The later hand has been attributed to Cardinal Bessarion, who owned the manuscript in the fifteenth century and included it in the great library he donated to the people of Venice.[14] Here we again find the paleography of a word, and sort by a property of the digital codex model, as in figure 7.

Paleography Identified by Text Content across Documents

Similar content is written differently not only in instances across physical folios, but also across distinct types of texts. Here, a secondary criterion restricts our results to a single folio, 12 recto, but finds the paleography of the same word (ΑΛΦΑ) in two different documents: the heading at the beginning of book 1, and a separate metrical summary.[15] The word is the same, but the context is different. The paleographic comparison reveals that in the Venetus A manuscript, different forms of script appear in different textual settings (see fig. 8).

14 C. Blackwell and C. Dué, "Homer and History in the Venetus A," in *Recapturing a Homeric Legacy*, ed. C. Dué (Washington, DC: Center for Hellenic Studies, 2009), 1–19.

15 On the metrical summaries in manuscripts of the *Iliad*: http://homermultitext.blogspot.com/2011/09/metrical-book-summaries-on-two.html.

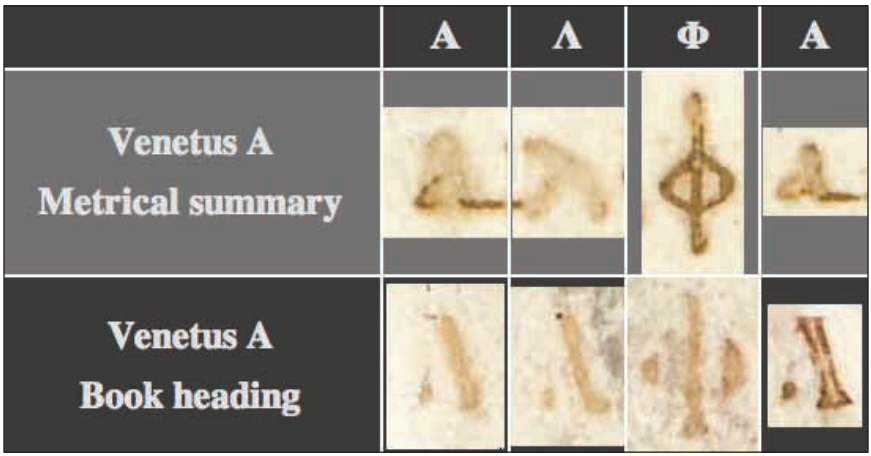

FIGURE 8. The word ΑΛΦΑ on the same manuscript, from the same century, but in different contexts.

Paleography Distinguished by Context within a Document

In the same way that the URN reference alone distinguished whether paleographic observations on a word ΑΛΦΑ occurred in the heading of *Iliad* 1 or in the metrical summary of *Iliad* 1, the CTS URN is enough to tell us what section of a work a cited passage of text appears in. This example draws on two manuscripts of Archimedes: Cologny, Fondation Martin Bodmer, MS 85, and the famous Archimedes Palimpsest (private collection). These mathematical texts include labeled figures; the labels are referred to in the text. In the same way that an analytical exemplar can reduce a complex diplomatic edition to a series of word tokens, an analysis of the labeling text in these manuscripts can reduce a complex edition to a series of labels, and the CTS URN will clearly identify what section of a document a text belongs to.

The table in figure 9 breaks out a sample of occurrences of alpha, beta, and gamma in the two manuscripts' labels into two groups: labels attached to the mathematical figure, and the corresponding labels cited in the text. We can immediately see that while the figures in the Archimedes Palimpsest appear to have been done by the same scribe who copied the text, the figures in Bodmer 8 were separately drawn and labeled with a very different hand.

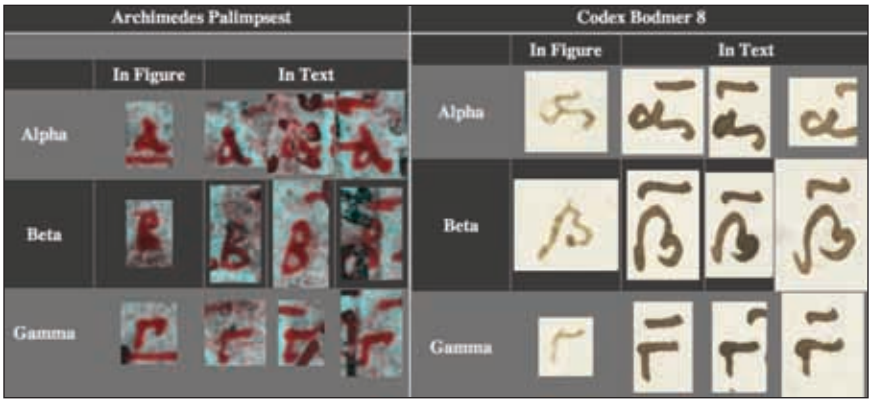

Figure 9. A view organized by alphabetic character, manuscript, and context within manuscripts.

Paleographic Features Grouped by Document

The preceding examples differ from traditional tools for paleography in identifying paleographic observations by textual context or content. Even when we want to use purely paleographic criteria for finding observations, however, it can be useful to organize the results by textual context. The following examples illustrate graphemes characterized visually as composed with a diagonal slash that rises from left to right, and are restricted to surveyed manuscripts dated to the tenth century. The results in figure 10 come from only two manuscripts, the Archimedes Palimpsest and the Venetus A, but the material from the Venice manuscript has been further broken out to distinguish the main Iliadic text from *scholia*. We see from this grouping that while shared abbreviations tagged as "composed with vertical slash" are very similar in form in both manuscripts, the repertory of those abbreviations is different not only from the Archimedes Palimpsest to the Venetus A, but within the Venetus A, from the Iliadic text to the *scholia*.[16]

16 The collection of paleographic observations we used for this example was neither comprehensive, nor a random sample, so we should be careful not to generalize too quickly about the complete repertory of graphemes or the frequency of individual graphemes in the complete repertory. Rather, this points to the interesting possibility that in

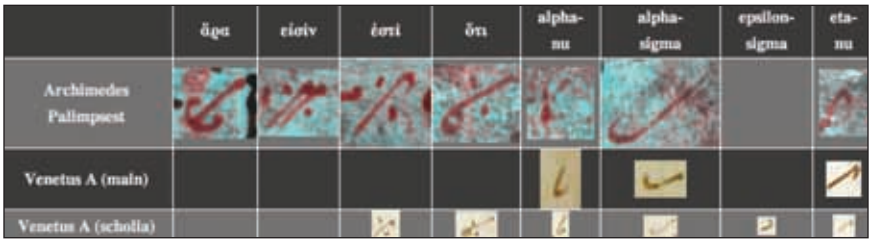

FIGURE 10. Paleography aligned by abbreviations and ligatures, grouped by MS.

Complex Graph Relation Illustrated by Paleography

It is equally possible to discover interesting patterns in paleography when the graph is being explored for other purposes. The Venetus A manuscript includes multiple sets of *scholia*, and sometimes includes more than one *scholion* on a single passage. The relations of the *scholia* to the text they comment on are represented in our graph, so that we can easily discover that the first word token of *Iliad* 2.4 in the Venetus A is commented on by three *scholia*. Editors and Homeric scholars might normally look only at the text content of the *scholia*, but now we can also find out how they relate paleographically. In figure 11, we see once more the first word token of *Iliad* 2.4, but now we align it with sections of three *scholia*: a superscript *scholion* offering an alternate reading for the final two letters of the word, and two comments in the margins that quote the word in linking introductory *lemmata*. It is quite apparent that the superscript variants are done in the same hand as the main text, but the *lemmata* of the marginal *scholia* are treated differently. Although these *lemmata* may have been written by the same scribe, they are written in a semi-uncial script that stands apart not only from the manuscript's poetic text, but also from the very cursive script of the commentary.

---

the future, we could organize collaboratively collected digital paleographic observations designed to support statistical inferences about the frequency and composition of a repertory of graphemes.

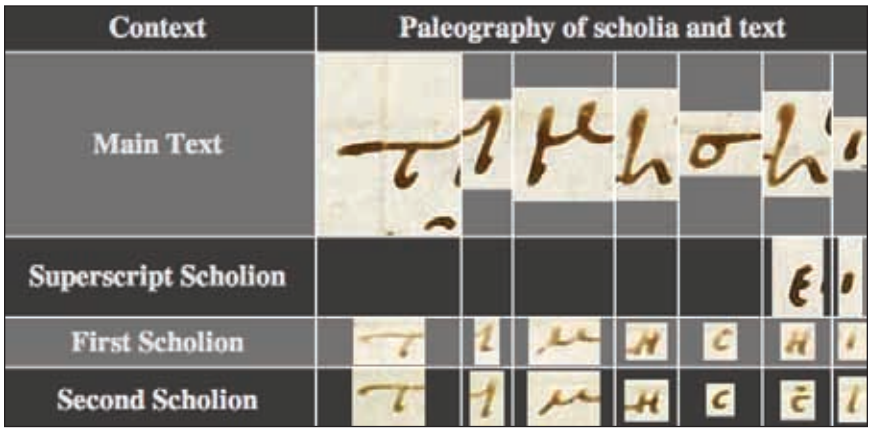

Figure 11. A complex issue of transmission, and tradition, expressed in a poetic text and two commentary texts, viewed paleographically.

Summary

With CTS and CITE URNs, we can concisely identify our objects of study, at any level of specificity, from "The *Iliad*" down to a single glyph, from a physical "codex" to a tiny rectangle on a particular digital image. These conventions also allow us to create, cite, and work with new information: commentaries, virtual "rebindings" of folios, and so forth. With these canonical citations, we can identify and record relationships among objects. With citemgr, we can generate a graph relating every scholarly statement that can be made about a digital archive at any point in time. This graph, expressed in standard RDF format, is also concise and explicit, and can serve as the basis for end-user applications and further analysis.

In this paper, we have described a solution using the CITE architecture to apply the semantics of textual citation to analyses of a text. Given a text cited with CTS URNs, we can align any analysis (citable by CITE URNs) to create *digital exemplars*. The textual content of the exemplar is drawn directly from the version it analyzes, but the exemplar organizes the analyzed selections in strings of text that can be cited more precisely than the text of the version it derives from.

This simple convention requires no changes to the protocols or tooling of the CITE architecture, and can capture any analysis of a text. Analyses are not exclusive, and by virtue of their citable relationship with the original edition, are automatically aligned with the analyzed text and with each other.

Furthermore, we can "promote" a digital exemplar to the status of an "edition"—e.g., a morphologically tokenized edition of the *Iliad*. This can, in turn, serve as the basis for further digital exemplars.

A few obvious applications for digital exemplars are:

- analysis of a text into morphologically meaningful tokens
- analysis of a word into paleographically meaningful glyphs
- analysis of a range of morphological tokens into syntactically meaningful spans
- analysis of a poetic line into syllables, and into metrical structures
- analysis of a prose text into possibly discontinuous spans of reused text, or an analysis of literary "fragments"

All of these can exist in parallel, or indeed in diverse competing analyses—analyses that in effect argue with each other—uniquely identified, cleanly separated from their source texts but fully aligned with them.

Conclusions

The implementation of the architecture we have described rests on widespread, generic technologies. RDF is the language of shared data: our CTS and CITE utilities generate RDF, while our implementations of CTS and CITE services consume RDF data.

Our use of the CITE architecture cleanly separates scholarly concerns. The task of editing a codex is separate from paleographic analysis, poetic analysis, syntactic analysis, or commentary. The data resulting from each of these scholarly acts can be separately maintained (e.g., in independent version-control systems) and published. For an individual project, a source archive and its CITE graph together provide a complete record of the proj-

ect at any moment of time, and can be published and cited as such.[17] A major design goal of RDF was to make it easy to aggregate and work with graphs from many sources. As a result, we can automatically create a unified and comprehensive representation of any published RDF graphs referring to the same material.

This approach has already shown its value in revealing patterns and generating insights that we could not hope to discover were we working exclusively in print media. But the reality is that even for the most thoroughly studied codices we have worked with, much of the information we draw on has never been published in print media. By returning to the evidence of the codex and attempting to create an explicit digital representation of every scholarly observation we make, we are beginning to reverse the limitations on the scholarly record that print publication has imposed for the past five centuries. We have had no alternative in the past but to accept what amounts to a kind of *de facto* censorship: varieties of readings reduced to selections in a critical apparatus, evidence of visual observations asserted without illustration, and publication of "significant" groups of *scholia* with no accompanying record of what is unpublished, to name a few common examples.

For the future of "scholarship outside the codex," perhaps the most important aspect of this work is the community of scholars it is helping to foster. The Homer Multitext project has profited from the labor, passion, insight, and creativity of a diverse group of scholars whose participation would not so much have been undervalued but impossible to achieve in an earlier generation. All paleographic observations in the examples we have presented in this paper were contributed by undergraduate members of the Manuscripts, Inscriptions, and Documents Club at the College of the Holy Cross;[18] at the time of their editorial work, the highest earned degree among these students was a high school diploma.[19]

---

17 See, for example, http://homermultitext.blogspot.com/2014/02/publishing-hmt-archive.html, which describes the January 2014 release of all data from the Homer Multitext.

18 Home page of the Holy Cross Manuscripts, Inscriptions, and Documents Club: http://shot.holycross.edu/hcmid.

19 Undergraduate contributors to the Homer Multitext project have included students at Holy Cross, Brandeis, the University of Washington, the University of Houston, Fur-

These young intellectuals and their older professional colleagues alike can be confident in the quality of their work because in a scholarly environment founded on canonical citation, no assertion need be an appeal to authority. As we look deep inside and widely outside the codex, we follow a trail of canonical citations that makes humanist scholarship reproducible, falsifiable, and thus durable.

---

man University, Gustavus Adolphus College, Trinity University, and the University of Leiden.