DATA SCIENCE JOURNAL

Blackwell, C and Smith, N 2016 Modeling Citable Textual Analyses for the *Homer Multitext*. *Data Science Journal,* 15: 17, pp. 1–11, DOI: http://dx.doi.org/10.5334/dsj-2016-017

PRACTICE PAPER

Modeling Citable Textual Analyses for the *Homer Multitext*

Christopher Blackwell[1] and Neel Smith[2]

[1] Furman University, Greenville, SC, USA

[2] College of the Holy Cross, Worcester, MA, USA

Corresponding author: Christopher Blackwell (christopher.blackwell@furman.edu)

Keywords: rdf; greek; text; canonical; citation

The *Homer Multitext* project (HMT) is documenting the language and structure of Greek epic poetry, and the ancient tradition of commentary on it. The project's primary data consist of editions of Greek texts; automated and manually created readings analyze the texts across historical and thematic axes. This paper describes an abstract model we follow in documenting an open-ended body of diverse analyses. The analyses apply to passages of texts at different levels of granularity; they may refer to overlapping or mutually exclusive passages of text; and they may apply to non-contiguous passages of text. All are recorded in with explicit, concise, machine-actionable canonical citation of both text passage and analysis in a scheme aligning all analyses to a common notional text. We cite our texts with urns that capture a passage's position in an *Ordered Hierarchy of Citation Objects* (OHCO2). Analyses are modeled as data-objects with five properties.

We create collections of 'analytical objects', each uniquely identified by its own URN and each aligned to a particular edition of a text by a URN citation. We can view these analytical objects as an extension of the edition's citation hierarchy; since they are explicitly ordered by their alignment with the edition they analyze, each collection of analyses meets satisfies the (OHCO2) model of a citable text. We call these texts that are derived from and aligned to an edition 'analytical exemplars'.

Summary

This article describes the approach to textual analysis adopted by the Homer Multitext Project (hereafter HMT) taking advantage of the CITE/CTS architecture (Blackwell and Smith, 2012*a*). The aim is to enable declarative models of textual analysis, with the particular goals of analyzing syntactic structure, historical orthographies, text-reuse, and of creating machine-actionable (but technologically agnostic) critical apparatus. This data-model is independent of technology, but currently implmented through an archive of XML texts and plain-text tabular data, and served as RDF in `.ttl` format.

We treat every analysis as a *reading* of a text (Ramsay, 2011). We separate the concerns of *citation*, *manipulation of textual content*, *alignment of analysis with textual content*, and *identification of acts of analysis*. All of these concerns can be documented concisely and explicitly, with simple tabular structures. This approach to analysis yields what we call 'analytical exemplars', coherent readings of a specific edition which can be treated as texts in their own right, identified by canonical citation. This approach has proved useful in our work on the Homeric tradition, but also for other problems in classical philology (Berti et al., forthcoming, 2016).

Background

The *Homer Multitext* is an ongoing collaborative project in editing and analyzing the primary source documents for Greek epic poetry, particularly the Byzantine manuscripts from the 10th through the 13th Centuries CE that preserve versions of the Homeric *Iliad* and the ancient tradition of commentary from Alexandrian and Roman scholars. The project's primary data consists of digital images of manuscript folios and diplomatic editions of the poetic and commentary texts that they contain.

The editions are encoded in XML validated against the Text Encoding Initiative's P5 schema (TEI Consortium Editors, 2016), but the HMT uses only a very small subset of the TEI's tagset, limited to three semantic areas:

1. markup applying a citation scheme to the text
2. markup documenting the editorial status of portions of the text (such as text erased, added or corrected by the original scribe)
3. markup identifying textual tokens that are not valid Greek lexical forms (such as Greek letters used as numbers, or regular abbreviations expanding to full forms)

This approach to editing is well understood today, but how best to model and organize *analyses* is an open question. We want to produce and publish citable analyses of our textual data, and be able to associate additional information with passages of text at different levels of granularity.

Some examples can serve to illustrate the challenge. The text of the first two lines of the Homeric *Iliad*, as it appears in one manuscript, reads:

μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος
οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,

One obvious analysis of these lines might attach lexical information to each word-token. The first word-token is "μῆνιν". Another anlysis might attach metrical information to each poetic foot. The first poetic foot corresponds to the text "μῆνιν ἄ". A syntactic analysis might want to identify the first noun-phrase of the *Iliad*, which is "μῆνιν οὐλομένην" (the first word of the first line, and the first word of the second line, but nothing in between).

This kind of analysis cannot be embedded in an XML edition. As a practical matter, the examples given above represent overlapping hierarchies that, even if they can be encoded in XML, would quickly make the document unmanageably complex. As a scholarly matter, we assume that there are an unlimited number of possible analyses, and we would like our approach to data to accommodate them all.

Annotation Standards

The HMT's workflow and data model for annotation takes advantage of several existing standards for annotation of digital texts. Our editors capture transcriptions of the textual content of manuscripts as TEI-conformant XML files. In our current implementation, all project data (except for binary image files) is stored on the server as RDF statements, using some established namespaces, and some project-specific namespaces.[1]

As described above, TEI, and any XML-specific standard (such as XPointer[2]) for annotation was deemed insufficient and impractical for capturing multiple, concurrent or even competing, textual analyses. The Open Annotation working group's Web Annotation Standard, a W_3C Candidate Recommendation as of September, 2016[3], provides a framework for annotation built using Linked Data fundamentals. As a generic framework, Open Annotation's vocabulary could express many of the relationships we describe below. Because our textual analyses reflect and extend a specific model of "text", we have chosen to use project-specific RDF with an `hmt` namespace.

[1] The HMT's data is transformed to `.ttl`, to be served by an RDF triple-store. The project's `.ttl` data includes project-specific vocabularies from the **cts**, **cite**, and **hmt** namespaces. It also includes vocabularies from the following namespaces: **dcterms** (http://purl.org/dc/terms/), **rdf** (http://www.w3.org/1999/02/22-rdf-syntax-ns#), **xsd** (http://www.w3.org/2001/XMLSchema#), **olo** (http://purl.org/ontology/olo/core#), **lex** (http://data.perseus.org/rdfverbs/), **rdfs** (http://www.w3.org/2000/01/rdf-schema#), **owl** (http://www.w3.org/2002/07/owl#).

[2] https://www.w3.org/TR/WD-xptr.

[3] https://www.w3.org/TR/annotation-model/.

An Abstract Model of 'Text'

We treat citable texts as an **ordered hierarchy of citation objects** (OHCO2) (Smith and Weaver, 2009). (The '2' recognizes the earlier model proposed by DeRose, et al., of an 'ordered hierarchy of *content* objects,' or OHCO (DeRose et al., 1990).) OHCO2 defines a citable text as a set of citable nodes that:

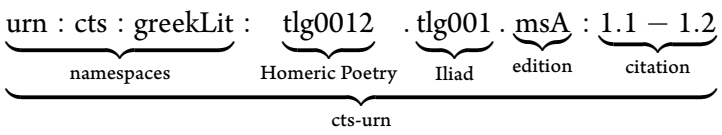
- belong to a bibliographic hierarchy
- belong to a citation hierarchy
- are ordered

This model frees us to convert representations of a citable text to any format determined to be equivalent under OHCO2. When analyzing textual content, a tabular representation is often convenient; when integrating textual content with other related material, we use directed graphs. For editing, we instantiate a tree model using TEI-XML (a widely used standard for encoding humanist texts) with associated metadata.

Canonical Text Services URNs

The CTS URN captures the bibliographic hierarchy of a text—text-group, work, version—and the hierarchy of citation, in a concise, machine-actionable canonical citation (Blackwell and Smith, 2012*b*)[4]. This CTS URN identifies the two lines of the *Iliad* quoted above:

```
urn:cts:greekLit:tlg0012.tlg001.msA:1.1-1.2
```



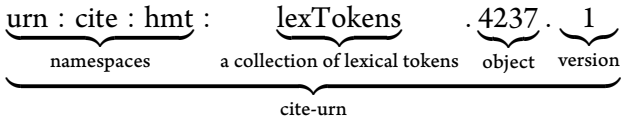
By manipulating the citation element of the URN, we can identify any passage of text: individual citable nodes (*e.g.* '. . . :1.1'), ranges ('. . . :1.1–1.2'), containing-elements ('. . . :1', identifying all citable nodes in Book 1 of the *Iliad*), or 'mixed ranges' like '. . . :1–2.2', 'all of Book 1, and lines 1–2 of Book 2'.

Each citable node contains text content, which may be structured as the editor of a specific edition chooses. The text contents of one edition might follow a plain-text model; others might used a mixed-content model represented by markdown or XML. This distinction between the canonical citation object, a fundamental component of the text, and its text content has implications for how we represent analyses of a text.

The Canonical Text Services (CTS) protocol[5] defines a networked service for retrieving passages of text identified by CTS-URN. CTS separates the concern of *retrieval by canonical citation* from the related but subsequent concern of *analyzing* the text content. We will now consider how we can work with analyses that do not align with the boundaries of citable nodes retrievable in the CTS protocol

Citation of Other Data

The HMT models data other than texts as versioned records in a collection of objects with common properties. We refer to these objects with CITE URNS (Blackwell and Smith, 2012*a*). A CITE URN identifies a collection of objects unique within a namespace, a uniquely identified object in the collection, and a version-identifier for that object.



[4] In this discussion, 'canonical' does not necessarily mean 'traditional', but rather 'independent of any particular representation of the text.' *Iliad* 1.1, as the citation to the first line of Book 1 of the poem, is both traditional and canonical. Some classical texts have traditional citation schemes based on particular printed editions (*e.g.* Plato); those often fail to serve well as canonical citations.

[5] http://cite-architecture.github.io/cts/.

Declaration vs. Alignment

Expressing the analysis of a passage of text can be reduced to associating a CITE URN with a CTS URN. The CTS URN can point to any passage: a single citable node in the text, a range of citable nodes, or a larger citation unit, '*Iliad* Book 2', for example.

When we analyze the *contents* of a citable node, however, we may need to work with textual contents that do not correspond perfectly to one or more citable nodes. We can *align* our analysis with characters in a particular edition by means of an extension to a CTS URN identifying an indexed substring within a citable portion of specific text. `urn:cts:greekLit:tlg0012.tlg001.msA:1.1@μῆνιν[1]`, for example, points to the first instance of the string 'μῆνιν' in Book 1, line 1, of the 'msA' edition of the *Iliad*.

In the CITE architecture, @ extends a URN with a type-specific subreference. CTS URN subreferences index substrings from an origin of 1. The subreferenced string is based on the CDATA content of the identified version of the text (excluding any markup), in the character encoding of that version. They can be treated naively as strings, and they can serve for more sophisticated comparisons using language-aware methods.[6]

Requirements for Declarative Analysis

For our editorial and analytical work to constitute a foundation for further scholarship, we want to emphasize the *declarative* over the *procedural*. That is, while computational processes such as search, tokenization, or difference operations are tools for scholarship, we want subsequent scholars to be able to point, explicitly and unambiguously, to any results of those operations.

With URN citations, we can construct what we call an 'analysis-relation' associating some data with a span of text. Each analysis-relation is a member of a collection of analysis-relations; the collection is ordered by the document order of the text analyzed. The textual component of the analysis-relation may be defined at any scale (a single character, for example, or, 'Book 2 of the *Iliad*'). The data component may be unique to this analysis-relation ('the first scribal correction on manuscript A') or applied to many analysis-relations ('dactyl').

Every analysis 'deforms' the text, to borrow Stephen Ramsay's portmanteau term for the activity of reading: 'deformance' (Ramsay, 2011). An analysis of `urn:cts:greekLit:tlg0012.tlg001.msA:1.1@μῆνιν` as a lexical token might deform the text of that version simply as 'μῆνιν' with some conversion of Unicode characters to ensure precombined accents; it might deform it to 'mh=nin', using an ASCII representation of polytonic Greek. A lemmatized analysis of this same text might deform the text to 'μῆνις' (the nominative, singular 'lexicon form'); a metrical analysis might deform it to ¯ ˘.

We fully document an analysis with five pieces of information:

1. The **analyzed text** is the citation to a version of the text, a CTS URN, perhaps with an aligning sub-reference, identifying the text we are analyzing.
2. The **analysis** is a CITE URN identifying the data resulting from this analysis.
3. The **sequence** is the index of this analysis-relation in the ordered collection of analyses.
4. The **text deformation** is a string of characters expressing the reading resulting from this analysis.
5. The **analysis record** is the canonical identifier for this cluster of objects: the five items documenting this instance of applying a specific analysis to a specific passage of text.

A Simple Example

Each of these five pieces of information is necessary, even in such an apparently simple case as tokenizing a text by word.

The *Homer Multitext*'s edition of the *Iliad* as it appears on the Venetus A manuscript—*Marcianus Graecus* Z454 [=822]; for an overview of this manuscript, see Dué (2008)—documents the text and editorial status of Book 2, line 4, with this TEI XML markup:

> τιμήσ<choice><reg>ει</reg><orig>η</orig></choice>· ὀλέσῃ δὲ πολέας ἐπὶ νηυσὶν Ἀχαιῶν·
>
> — *Iliad* 2.4 (msA)

[6] For the HMT, we have published a code library for working with ancient Greek text: http://neelsmith.github.io/greeklang/. This library can identify equivancies between Greek strings in different encodings—*e.g.* "μῆνιν" and "`mh=nin`"—and translate among them.

The scribe wrote the verb-form τιμήσῃ, but added an alternate ending -ει supralinearly. For our lexical analysis, our project's convention is to capture the original text. The first analysis, then, analyzes the textual content of our XML edition from the first instance of the character *tau* through the first instance of the combined *eta with iota-subscript* (τιμήσῃ). We can express this with the CTS URN `urn:cts:greekLit:tlg0012.tlg001.msA:2.4@τ[1]-2.4@ῃ[1]`. That URN accurately identifies the textual content under analysis, but it does not itself represent a coherent text. Naively resolved, it would result in 'τιμήσ`<choice><reg>`ει`</reg><orig>`ῃ', which is neither Greek nor well-formed XML. Resolved without markup, it would result in 'τιμήσειῃ', which is not a Greek word. In either case, a sensible reading of the resulting text would require further, potentially complex, processing.

By specifying a text-deformation for this particular analysis, we can state explicitly that `urn:cts:greekLit:tlg0012.tlg001.msA:2.4@τ[1]-2.4@ῃ[1]`, *in this analysis*, is read as 'τιμήσῃ'. Our citable analysis can be declarative rather than procedural.

We identify explicitly the position of this analysis in the sequence of all lexical tokens in this version of the *Iliad*. We treat the lexical token 'τιμήσῃ' as a data-object in a collection of lexical tokens, and identify it with a unique CITE URN; every instance of τιμήσῃ in our *Iliad* will be analyzed with this URN.

In this analysis by lexical token we choose to ignore editorial markup, but because its tokens are still be aligned to the Edition, the editorial status of any given token—unclear, supplied, *vel sim.*—can be determined, even though the reading that results from the analysis is straightforward Greek without markup.

Field	Value
Analysis Record	`urn:cite:hmt:VenA_lexical_analysis.4237.1`
Sequence	`4237`
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:2.4@τ[1]-2.4@ῃ[1]`
Analysis	`urn:cite:hmt:lexicalTokens.788.1`
Text-Deformation	τιμήσῃ

Table 1: Aligning a single lexical token to a form correct on the manuscript.

Field	Value
Analysis Record	`urn:cite:hmt:VenA_lexical_analysis.4238.1`
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:2.4@ὸ[1]-2.4@ῃ[2]`
Sequence	`4238`
Analysis	`urn:cite:hmt:lexicalTokens.1988.1`
Text-Deformation	ὸλέσῃ

Table 2: The next lexical token in sequence.

We can document the first metrical foot with the following data (**Table 3**):

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1@μ[1]-1.1@ᾰ́[1]`
Sequence	`1`
Analysis Record	`urn:cite:hmt:metricalAnalysis.1.1`
Analysis	`urn:cite:hmt:meter.dactyl.1`
Text-Deformation	μῆνιν ᾰ́

Table 3: The first metrical foot of *Iliad* 1.1.

Complementary Examples

Metrical Feet

The meter of Greek epic poetry disregards word-boundaries. A compresensive documention of poetic meter, encoded in XML markup without violating the rule against overlapping hierarchies would result in an XML edition so complex as to be unusable, especially when combined with markup documenting editorial status. Here is the first line of the *Iliad*, divided into six metrical feet, with the analysis of each foot (**Table 1, 2**):

μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος
μῆνιν ἄ | ειδε θε | ὰ Πη | ληϊά | δεω Ἀχι | λῆος
dactyl | dactyl | spondee | dactyl | dactyl | spondee

Syntax

Syntactical analysis may also fail to align with word-boundaries. The Greek word 'οὔτε', for example, performs two syntactic functions. The 'οὔ' is an adverb, and the 'τε' is a coordinator. One response to this problem has been to create editions tailored to this specific type of analysis by inserting editorial word divisions into forms like οὔτε. For our work, we want to avoid introducing non-standard orthography into our edited texts merely to serve a single kind of analysis. By generating a 'syntax-token analysis' we can leave our edition intact, while precisely assigning syntactical roles to parts of words where needed.

ἵν᾽ οὔτε φωνὴν οὔτε του μορφὴν βροτῶν — Aeschylus, *Prometheus Bound* 21

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0085.tlg003:21@οὔτε[1]`
Sequence	`N`
Analysis Record	`urn:cite:fu:pvSyntax.45.1`
Analysis	`urn:cite:fu:syntaxTokens.2345.1`
Text-Deformation	οὔ

Table 4: Analyzing οὔτε as an adverb, οὐ.

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0085.tlg003:21@οὔτε[1]`
Sequence	`N+1`
Analysis Record	`urn:cite:fu:pvSyntax.46.1`
Analysis	`urn:cite:fu:syntaxTokens.2346.1`
Text-Deformation	τε

Table 5: Analyzing οὔτε as an coordinator, τε.

Field	Value
Sequence	`13`
Analysis Record	`urn:cite:histfragDipl:arist.577`
Analysis	`urn:cite:histfrag:arist.577`
Analyzed Text	`urn:cts:greekLit:tlg0007.tlg012.perseus-grc1:26.3@ὑπὸ[1]-26.3@πρότερον[1]`
Text-Deformation	ὑπὸ τοῦ Μελίσσου καὶ Περικλέα αὐτὸν ἡττηθῆναι ναυμαχοῦντα πρότερον

Table 6: Identifying and reuniting non-contiguous reported speech.

Both analyses, above, align to the same string of characters in the original edition. But in the colletion of analyses, they are two elements, one following the other in sequence.

Non-contiguous Text

Modelling analyses of 'text reuse'—quotation, paraphrase, allusion—can be challenging because it is often necessary to treat non-contiguous spans of text as a single unit. Here, **bold type** highlights 'text reuse' (**Table 4, 5, 6**):

> **ὑπὸ** δὲ **τοῦ Μελίσσου καὶ Περικλέα** φησὶν **αὐτὸν** Ἀριστοτέλης **ἡττηθῆναι ναυμαχοῦντα πρότερον** — Plut. *Per.* 26.3
>
> But Aristotle says that **Pericles, too, fighting in a previous naval battle, was defeated by Melissos.**

While the reused text (an indirect quotation from Aristotle) is contiguous in English translation, it is not in the Greek edition of Plutarch's *Life of Pericles*.

In this example, we analyze a string of text from our Edition, associating it with an Analysis URN that identifies an instance of text-reuse. For the **text-deformation** of our analytical exemplar, however, we choose to omit the *verbum dicendi* and speaker-attribution (*i.e.* 'φησὶν. . . Ἀριστοτέλης'), and the sentence-adverbial ('δὲ'), which are not actually part of the quotation. We have kept this analysis separate from our base edition, but we can present our reading of quotation as we choose, and attach commentary to the object pointed to by the Analysis URN.

Analytical Exemplars

Every analysis of a text is a *reading* tokenizing a text into a series of analyzed units. Our approach to documenting analyses results in an ordered collection of readings, aligned to the citation hierarchy of a version (through the analyzed text CTS URN). Each of these analyses has its own text content (the text deformation), which is controlled and defined by its association with the analysis URN). So we essentially have the necessary components for a new text, in the OHCO2 model.

We implement this by defining an 'analytical exemplar', derived from a specific version, with an additional level added to the bibliographic hierarchy of a CTS URN:[7]

> `urn:cts:greekLit.tlg0012.tlg001.msA`: (The MS. A edition of the Homeric *Iliad*)
>
> `urn:cts:greekLit.tlg0012.tlg001.msA.lexTokens`: (An **analytical exemplar** derived from the MS. A edition of the Homeric *Iliad*)

The analytical exemplar's citation scheme follows that of the version it analyzes, but its citation hiearchy adds a further level:

> `urn:cts:greekLit.tlg0012.tlg001.msA:2.4` (Book 2, line 4, of the MS. A edition of the Homeric *Iliad*)

The most finely grained unit of citation from the Edition resolves to:

> τιμήσ<choice><reg>ει</reg><orig>ῃ</orig></choice>· ὀλέσῃ δὲ πολέας ἐπὶ νηυσὶν Ἀχαιῶν·

We can resolve our analytical exemplar more finely, while being sure to get meangful text content:

[7] A CTS URN captures a bibliographic hiearchy that is similar to that defined by the FRBR (Functional Requirements for Bibliographic Records) recommendation of the International Federation of Library Associations and Institutions (IFLA) (http://www.oclc.org/research/activities/frbr.html). FRBR asserts a hierarchy of: work, expression, manifestation, item. The last, "item", is defined as "a single exemplar of a manifestation". CTS's "exemplar" aligns with FRBR's "item" for physical volumes—*e.g.* Homer, *Iliad*, editio of Villoison, **Thomas Jefferson's personal copy thereof**. For digital texts, CTS defines an "exemplar" as "a specific transformation explicitly derived from a specific version of a text."

`urn:cts:greekLit.tlg0012.tlg001.msA.lexTokens:2.4.1` (Book 2, line 4, lexical token 1 of the 'lexical tokens analytical exemplar' derived from the MS. A edition of the Homeric *Iliad*)

This resolves to:

τιμήσῃ

By creating an analytical exemplar, we can separate concerns more effectively. For an analysis of the language of the *Iliad*, an analytical exemplar of lexical tokens provides clear, editorially controlled, data. Our text-mining for lexical forms or morphology need not work around paleographical or codicological issues. At the same time, every citation in an analytical exemplar is aligned to the version from which it is derived, and so all analytical exemplars are implicitly aligned to each other.

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1-1.2@οὐλομένην[1]`
Sequence	`1`
Analysis Record	`urn:cite:hmt:clauses.1`
Analysis	`urn:cite:hmt:clauses.1`
Text-Deformation	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.1.1`
Next Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.2.2`

Table 7: Aligning the first grammatical clause of the *Iliad* to the text. Record 1 of 2.

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1-1.2@οὐλομένην[1]`
Sequence	`2`
Analysis Record	`urn:cite:hmt:clauses.1`
Analysis	`urn:cite:hmt:clauses.1`
Text-Deformation	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.2.1`
Next Exemplar urn	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.2.2`

Table 8: Aligning the first grammatical clause of the *Iliad* to the text. Record 2 of 2.

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.2@ἡ[1]-1.2@ἔθηκε[1]`
Sequence	`3`
Analysis Record	`urn:cite:hmt:clauses.2`
Analysis	`urn:cite:hmt:clauses.2`
Text-Deformation	ἡ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.2.2`
Next Exemplar urn	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.3.1`

Table 9: Aligning the second grammatical clause of the *Iliad* to the text.

Clauses

Analytical exemplars also allow us to read and cite a text according to analytical tokenizations. For example, it might be desirable to 'read' the *Iliad* in chunks defined not by poetic line, but by grammatical clauses. By analyzing the text by clauses and creating an analytical exemplar, we can make this possible.

> 1 μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος
> 2 οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,
> . . .
> — *Iliad* 1.1–1.2

The first grammatical clause of the *Iliad* is 'μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην'. This includes all of 1.1, and the first part of 1.2. The second is 'ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε', the rest of 1.2. This would present a problem of overlapping hierarchies, if we were to embed this analysis in the XML of the edition. The following tables add two additional properties to each analysis, the **analytical exemplar URN**, by which tokens in this new reading can be cited, and for each citeable node of the exemplar, the **next** citable node. This satisfies the OHCO2 requirements, documenting citable nodes in a citation hierarchy, and their sequence (**Table 7, 8, 9**).

There are two clauses, identified by the Analysis URNS: `urn:cite:hmt:clauses.1` and `urn:cite:hmt:clauses.2`. There are *three* entries in our record of these two clauses. The first two both have `urn:cite:hmt:clauses.1` as their Analysis Record and their Analysis (because in this case, the analysis is unique: the first clause of this edition of the *Iliad*).

The Analytical Exemplar URNS are the key for understanding why we have two entries for the first clause. This analytical aligment is creating an exemplar that is tokenized and citeable according to clauses. The Analytical Exemplar URNS, and the aligned analyses, make the following identifications:

- The first citable analysis of *Iliad* 1.1 is `clauses.1.`
- The first citable analysis of *Iliad* 1.2 is `clauses.1.`
- The second citable analysis of *Iliad* 1.2 is `clauses.2.`

If we were to navigate our Edition via a CTS service, the following URNS would return the following text-content (**Table 10**):

Edition-level CTS-URN	**Text-Content**
`urn:cts:...msA:1.1`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος
`urn:cts:...msA:1.2`	οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,

Table 10: Citation of an Edition of the *Iliad*, two citable nodes.

If we were to navigate our Analytical Exemplar via a CTS service, the following URNS would return the following text-content (**Table 11**):

Edition-level CTS-URN	**Text-Content**
`urn:cts:...msA.clauses:1.1.1`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,
`urn:cts:...msA.clauses:1.1`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,
`urn:cts:...msA.clauses:1.2.1`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,
`urn:cts:...msA.clauses:1.2.2`	ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,
`urn:cts:...msA.clauses:1.2`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,
`urn:cts:...msA.clauses:1.1.1-1.2.1`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,
`urn:cts:...msA.clauses:1.1.1-1.2.2`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,
`urn:cts:...msA.clauses:1.1-1.2`	μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,

Table 11: Citation of an Analytical Exemplar at different levels of the citation hierarchy, and the resulting text.

A CTS implementation recognizes that for each position in the sequence of a text, there can be only one citable node containing content. A request for '. . . :1.1.1' will yield the correct text, as will a request for '. . . :1.2.1', but a request for '. . . :1.1.1–1.2.1' will not repeat the same text-content.

Editing, Archiving, and Publication

Our reliance on abstract data models allows the HMT to use the most appropriate tools and formats for each of the tasks of editing, archiving, and publishing. Editors transcribe texts into TEI XML files, taking advantage of automated validation. Analytical work is captured in plain-text `.csv` files; our editors often work via the GitHub web interface, which offers validation of tabular data.

Data is archived as tabular text files; the project's archival repository (https://github.com/orgs/homermultitext) saves versions of texts as both XML and tabular files derived from them.

The project's online publication is through the *HMT Digital* web-application, which uses an RDF database (currently Apache Fuseki) as a datastore. The archival data is transformed with our CITE Archive Manager utility into RDF statements. *HMT Digital* accepts queries on URNs and allows browsing of texts, delivering data either as raw XML or JSON objects, or as XML transformed to HTML for human readers.[8]

Conclusion

This approach to managing analytical data affords a number of benefits:

- It allows us to separate the concern of editing a text in machine-readable form from the concern of publishing analyses of that text.
- It allows us to record an open-ended number of analyses. We are not limited to any given XML vocabulary.
- It allows us to *cite* our analyses at a very granular level.
- It explicitly aligns each analysis to the edition analyzed, and so implicitly aligns all analyses to each other.
- Its simple structure can be represented by plain-text tabular data files, clear to read, easily repurposed and shared.
- It frees us to represent Greek using different encodings or orthographies without losing the connection to the primary source evidence of our manuscripts.
- It supports more complex sets of analyses than anything that can be embedded in xml markup.

This approach lends itself to automated analyses, such as morphological parsing or tokenizations by word+punctuation (as required for certain approaches to documenting syntax), and to automated generation of exemplars based on the editorial status of the text. It also lends itself to analyses hand-crafted by human editors, such as analyses of speeches, extended similes, or more complex instances of text reuse.

One goal of the *Homer Multitext* is the fullest possible account of the traditional language of the Greek epic poetic tradition. Such an account does not exist, and will depend on systematic analysis of syntax, morphology, and meter across full editions of Homeric texts and quotations of Homeric texts in ancient commentaries. While our work aligning citable analyses to citable texts remains experimental, we are confident that, as the HMT's collaborators continue to expand the project's collection of digital diplomatic editions, our approach adequately serves our present scholarly requirements, and is creating durable data. We are optimistic that our simple data formats can be readily transformed as future needs dictate, and will continue to offer new opportunities for engagement with and insights from the HMT corpus.

Competing Interests

Smith's and Blackwell's research on the *Homermultitext* has received funding and other support (lodging, meals) from the Center for Hellenic Studies of Harvard University, Gregory Nagy, Director. Work on the *Homermultitext* has been funded by the National Endowment for the Humanities and the National Science Foundation. Work on CITE/CTS has been funded by the Andrew W. Mellon Foundation. M. Daniels, S. Stricklan, and K. Vincent-Dobbins, co-authors of an article cited in the bibliography, are former students of Blackwell. C. Dué, another author cited in the bibliography, is a collaborator on the *Homermultitext* with

[8] *HMT Digital*: `github.com/homermultitext/hmt-digital`. CITE Archive Manager: `github.com/cite-architecture/cite-archive-manager`.

Smith and Blackwell; Dué and Blackwell are sister- and brother-in-law. G. Weaver, co-author with Smith on a cited article, is a former student of Smith.

References

Berti, M, Blackwell, C, Daniels, M, Strickland, S and **Vincent-Dobbins, K** 2016 (forthcoming) Documenting Homeric Text-reuse in the Deipnosophistae of Athenaeus of Naucratis. *Bulletin of the Insitute of Classical Studies* .

Blackwell, C and **Smith, D** 2012a The CITE architecture. *Homer Multitext Project Documentation*. http://cite-architecture.github.io

Blackwell, C and **Smith, D** 2012b Four URLs, Limitless Apps: Separation of concerns in the Homer Multitext Architecture. In: Muellner, L (Ed.) *Donum natalicium digitaliter confectum Gregorio Nagy septuagenario a discipulis collegis familiaribus oblatum*. The Center for Hellenic Studies of Harvard University, Washington, DC. Retrieved from: http://chs.harvard.edu/wa/pageR?tn=ArticleWrapper&bdc=12&mn=4846.

DeRose, S, Durand, D, Mylonas, E and **Renear, A** 1990 What is Text, Really? *Journal of Computing in Higher Education,* 1(2): 3–26. Retrieved from: http://www.hki.uni-koeln.de/sites/all/files/courses/3226/Renear_ea-1997.pdf. DOI: https://doi.org/10.1007/BF02941632

Dué, C 2008 *Recapturing a homeric legacy: images and insights from the Venetus, a manuscript of the Iliad*, Harvard University Press.

Ramsay, S 2011 *Reading Machines: Toward an Algorithmic Criticism*, 1st edition, University of Illinois Press.

Smith, D N and **Weaver, G** 2009 Applying domain knowledge from structured citation formats to text and data mining: Examples using the CITE architecture *Text Mining Services* p. 129. Retrieved from: http://katahdin.cs.dartmouth.edu/reports/TR2009-649.pdf

TEI Consortium Editors (ed.) 2016 *TEI P5: Guidelines for Electronic Text Encoding and Inter-change.* TEI Consortium. v.3.0.0. Retrieved from: http://www.tei-c.org/Guidelines/P5/

How to cite this article: Blackwell, C and Smith, N 2016 Modeling Citable Textual Analyses for the *Homer Multitext*. *Data Science Journal,* 15: 17, pp. 1–11, DOI: http://dx.doi.org/10.5334/dsj-2016-017

Submitted: 15 June 2016 **Accepted:** 25 October 2016 **Published:** 16 December 2016

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS