1 Use Cases & Infrastructure: Analytical Editions

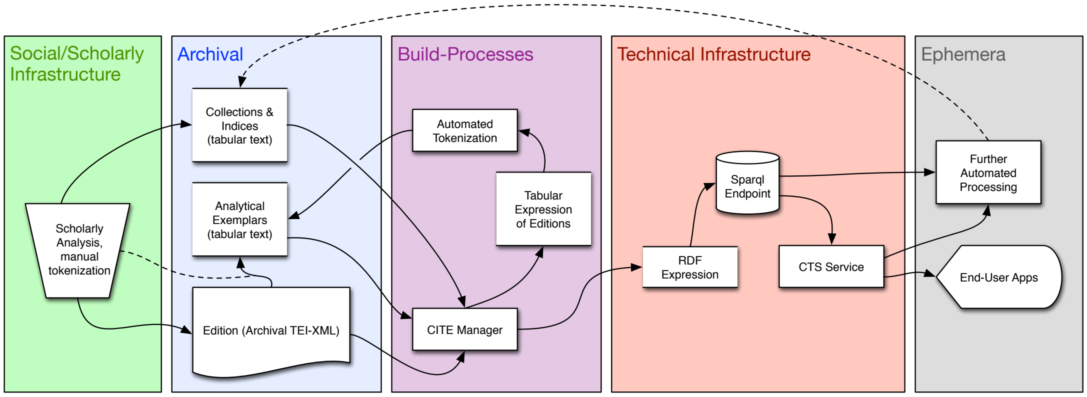
Christopher Blackwell, Neel Smith. May, 2015.

1.1 Summary

This document describes ongoing work on textual analysis for the Homer Multitext, and on text-reuse, textual history, and syntax that is a collaboration between Furman University and the Leipzig Open Greek and Latin project.

Our experience has shown that the model of "text" as and **ordered hierarchy of citation objects** (OHCO2) allows us to express the semantics of a text in many different data formats.[1] We use TEI-XML mainly as an archival format and for working with a text as it is being edited, using a very constrained subset of its elements—only those necessary for documenting the citation scheme, the editorial status of specific spans of text (unclear, added, corrected, &c.), and disamguating non-lexical content in the text (*e.g.* Greek letters used as numbers, fragments of words, personal names).

For subsequent processing, we express the texts' semantics as tabular data in plain-text files; our implementation of the CTS service uses an RDF triplestore as its back-end.



Social/Scholarly Infrastructure
Archival
Build-Processes
Technical Infrastructure
Ephemera
Scholarly Analysis, manual tokenization
Collections & Indices (tabular text)
Analytical Exemplars (tabular text)
Edition (Archival TEI-XML)
Automated Tokenization
Tabular Expression of Editions
CITE Manager
RDF Expression
Sparql Endpoint
CTS Service
Further Automated Processing
End-User Apps

Figure 1: The CITE/CTS Architecture as currently implemented for the Homer Multitext, &c.

The technological infrastructure that would most benefit this work would be an **extremely robust triplestore with a public-facing SPARQL endpoint.**

What follows is a description of the approach to **analysis** that we have been able to develop to meet our need for multiple, mutually incompatible analyses of complex texts, and our desire that those analyses align to one another. The digital editions and exemplars derived from them can be entirely expressed as RDF statments, but these will inevitably number in the hundreds of millions.

Some of these analyses will be the products of human editors. The Furman students working in Leipzig with Monica Berti are generating analyses of text-reuse in Athenaeus by hand, entering data in `.csv` files in GitHub. Others will be programmatically generated, such as lexical or metrical analyses across our corpus of Homeric epic.

1.2 Background: Analysis

In our work on the tradition of Greek Epic poetry for the Homer Multitext, and on text-reuse for the Leipzig Open Greek and Latin project, we confront the need for many kinds of *analysis* of texts and images.

By *analysis* we mean: the systematic association of metadata (commentary, cross-references, categories or labels in a controlled vocabulary) to objects of study *or parts of those objects*.

Some examples of analysis:

- Associating textual citations with regions-of-interest on an image.
- Attaching morphological identifications to lexical tokens in a text.
- Identifying syllables in a a poetic text and assingment them a metrical value.
- Documenting the syntax of a sentence.
- Identifying instances of text-reuse and assinging them citations.

There are many ways to perform these analyses. The challenge is to move these acts of analyses from the **procedural** to the **declarative**, in some manner independent of technology.

1.2.1 The Easy Part

In many ways, analysis of images is the least difficult:

- There is an image with a unique identifier.
- It is accepted that the image may be scaled, turned from a `.tif` to a `.jpg`, without losing its identity.
- We can define regions-of-interest on the image, through various schemes of citation, and link those citations to other data.
- The ROIs can overlap.
- So, a single image of a manuscript folio might have ROIs defined that treat large regions—the main text-block, commentary text-blocks, illustrations—and very small regions—graphemes, punctuation. One ROI can overlap another, or many, as when a region defines a "poetic line" on the manuscript, while other regions identify individual words, and another identifies a large stain.

Similarly, annotation of geo-spatial data is infinitely flexible and granular, from the centimeter-scale mapping of a botanical garden to analysis that groups Roman amphitheaters scattered across the Mediterranean World.

1.2.2 Citation-Objects

Working with analyses of texts is more difficult.

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

3 πολλὰς δ᾽ ἰφθίμους ψυχὰς Ἄϊδι προΐαψεν
5 οἰωνοῖσί τε πᾶσι, Διὸς δ᾽ ἐτελείετο βουλή,
6 ἐξ οὗ δὴ τὰ πρῶτα διαστήτην ἐρίσαντε
7 Ἀτρεΐδης τε ἄναξ ἀνδρῶν καὶ δῖος Ἀχιλλεύς.

This is a passage of an ancient Greek text, which we can identify precisely and *declaratively* with a citation: *Iliad* 1.1–1.7. We can use a CTS-URN[2], which is both canonical and machine actionable to identify it:

```
urn:cts:greekLit:tlg0012.tlg001.persGrk:1.1-1.7
```

$$\underbrace{\underbrace{\mathit{urn} : \mathit{cts} : \mathit{greekLit}}_{\text{namespaces}} : \underbrace{\mathit{tlg}0012}_{\text{Homeric Poetry}} . \underbrace{\mathit{tlg}001}_{\text{Iliad}} . \underbrace{\mathit{persGrk}}_{\text{edition}} : \underbrace{1.1-1.7}_{\text{citation}}}_{\text{cts-urn}}$$

A citation resolves to a text, which may contain *mixed content*, markup describing the text. Here is the markup for line 4 of Book 2, from a transription of a particular manuscript of the *Iliad*.

```
urn:cts:greekLit:tlg0012.tlg001.msA:2.4

<l n="4">τιμήσ<choice><sic>η</sic><corr>ει</corr><choice>, ὀλέσῃ δὲ πολέας
ἐπὶ νηυσὶν Ἀχαιῶν.</l>
```

The citation is precise and explicit. The markup of the text is appropriate, too, in that it *documents* the Greek text. That is, it (a) captures the citation scheme, and (b) asserts the editorial status of the Greek text. In this case, the manuscript presents two different endings for the verb, "he might honor": -ῃ and -ει.

1.2.3 Analysis

A human being, reading texts, will inevitably engage in a number of simultaneous acts of analysis. A sophisticated reader, experienced in Greek epic poetry, will, without much conscious thought, analyze the text in the following ways:

- Lexical tokens: each word; its morphology; its complex lexicography.
- Named entities: some words are names: Achilles, Zeus. Some are complex, pointing to more than one person: "Son-of-Peleus".
- Syntactical units: phrases, clauses, sentences.
- Formulaic units: "Son-of-Peleus-Achilles", "Son-of-Atreus-Lord-of-Men", "Godlike-Achilles".
- Poetic lines: a fundamental structure of this text, and how we cite it.
- Poetic half-lines: a fundamental building-block of dactylic hexameter.
- Metrical feet: dactyls and spondees, themselves made up of…
- Syllables.

Iliad 1.1–1.7 includes seven citable units, according to the canonical scheme of citation for this text. The seven constitute a single sentence. But beyond that, things get complicated:

- First noun-phrase: μῆνιν… οὐλομένην (“destructive wrath”)
- First clause: μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος | οὐλομένην, (“Sing, goddess, of the destructive wrath of Achilles, son of Peleus”)
- Named Entity: Πηληϊάδεω Ἀχιλῆος (“Son-of-Peleus Achilles”)
- Named Entity?: Πηληϊάδεω (implies someone named “Peleus”?)
- First metrical foot: μῆνιν ἄ…
- Second metrical foot: …ειδε θε…
- First grapheme in *Iliad* 1.1: μ (a single character)
- First grapheme in *Iliad* 1.7 on the Venetus A manuscript: ἐξ (a ligature of two characters, and a diacritical mark)

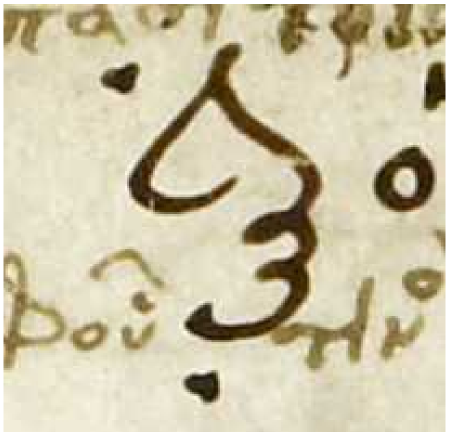


Figure 2: The word ἐξ, at *Iliad* (MS A 12-*recto*) 1.7: one, two, or three tokens, depending on the analysis.

Most of the above examples, however, cannot be cited precisely using the canonical scheme of citation. The first half-line—μῆνιν ἄειδε θεὰ—falls within 1.1, but is not the same as 1.1. The first syntactical clause—μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος | οὐλομένην—includes all of 1.1, and the first word of 1.2. There is a noun-phrase, the direct object of the verb ἄειδε, that includes the first word of 1.1 and the first word of 1.2, but nothing in between.

If we are to realize the potential of digital libraries, we need to be able to work with analyses like these *declaratively*. Possible analyses are limitless and complementary; some will cross citation-boundaries; some will be analyses of non-contiguous text. It is impractical to expect the *documentary markup* of a digital edition (e.g. TEI-XML) to serve for analysis as well.

1.2.4 Tokenization(s)

We could *add to* the canonical citation scheme a further level, making it Book, Line, Word, tokenizing the text. Thus our first syntactical clause—μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,—could be cited as 1.1.1–1.2.1 (Book 1, line 1, word 1, **through** Book 1, line 2, word 1).

But we would quickly find this limiting. The first metrical foot, a dactyl, includes the first word of 1.1 and the first syllable of the second word: μῆνιν ἄ....

We could tokenize by character, of course, so "μῆνιν ἄ" would be *Iliad* 1.1.1–1.1.7.

In all of these examples, we need to *declare* some combination of the citation hierarchy and the content. The CTS-URN specification allows us to add *subreferences*, by which our metrical-foot example could be expressed as "1.1@μ–1.1@α", or more precisely (since there might be more than one *mu* and more than one *alpha* in a line, "1.1@μῆνιν[1]–1.1@α[1]", that is, "1.1, the first instance of the string μῆνιν, through 1.1, the first instance of the string α."[3]

CTS-URNs with subreferences are an important start, but they are not sufficient.

> `τιμήσ<choice><sic>ῃ</sic><corr>ει</corr><choice>, ὀλέσῃ δὲ πολέας ἐπὶ νηυσὶν Ἀχαιῶν.` —*Iliad* 2.4 (Venetus A)

This line of a transcription of the *Iliad*, 2.4, as it appears on the Venetus A manuscript, is marked up to show that the scribe offered two alternative endings for the verb "he might honor": τιμήσῃ and τιμήσει.

What is the *content* here? If we want to cite "the two parallel verbs", and we cite "... 2.4@τιμήσῃ[1]-2.4@ὀλέσῃ[1]", as proposed above, the textual content of the electronic edition (the concatenation of the text-nodes in an XML document) would give us: τιμήσῃει, ὀλέσῃ. This does not make any sense.

And how would we cite our noun-phrase—μῆνιν ... οὐλομένην? 1.1@μῆνιν[1]–1.2@οὐλομένην[1] would include all the words in between the noun μῆνιν and the participle οὐλομένην. "1.1@μῆνιν[1] and 1.2@οὐλομένην[1]" is not a citation but two citations.

And so on. There is no single scheme of citation that can possibly serve the kinds of analysis that scholars employ every day.

1.3 Analytical Exemplars

Our approach is to create a new text, derived from an Edition (or Translation) that expresses a particular analysis. We call these "Analytical Exemplars". They are subordinate to and specifically dependent on the Edition from which they derive. The Exemplar inherits the citation-structure of the Edition. The Exemplar may *extend* the Edition's citation hierarchy to an additional level of depth.

("Exemplar" has always been part of the CTS bibliographic hierarchy of: `text-group` → `work` → `edition/translation` → `exemplar`.[3][4])

While all of our Editions and Translations begin life as TEI-XML, our Analytical Exemplars are created as tabular data. There is no reason these Exemplars could not be re-expressed as TEI-XML, but we have as yet see no reason to do so. Like our Editions and Translations, the

Exemplars are further processed into RDF statements for serving via the SPARQL endpoint that feeds our CTS service.

1.3.1 Data Defining an Analytical Exemplar

We create an Analytical Exemplar, derived from a specific version (Edition or Translation), by capturing the following data, initially in a plain-text table, and (after processing) as RDF statements:

- **Analyzed Text** This is a CTS-URN, with or without a substring, which may be range, identifying the passage of text analyzed in the Edition. If the text in question is an XML text contained mixed content, the 'text' here includes **the concatenation of all text-nodes in a citation unit**.

- **Analysis Record** This is a CITE-URN identifying **uniquely** the pairing of analysis+text.

- **Analysis** This is a CITE-URN pointing to the analysis being attached to a text. It *may* be identical to the analysis record,

  - When the *analysis* is unique (*e.g.* "The first clause of the *Iliad* in the 'msA' edition."), then the `Analysis Record` (a URN) and the `Analysis` (a URN) may be **identical**.
  - When the *analysis* is not unique (*e.g.* "verb", or "dactyl"), the `Analysis Record` (a URN) and the `Analysis` (a URN) must be **different**.
  - The *analysis* URN points to an object to which any desired metadata may be attached.

- **Analytical Exemplar URN** This is a CTS-URN used to construct an "analytical exemplar", which is a text derived from the version identified by the Analyzed Text CTS-URN, with one additional level of citation-hierarchy, each of whose leaf-nodes is an analysis, identified by the **Analysis URN** (above). The Analytical Exemplar, when processed into the OHCO2 data model, will act like any other CTS text. The *text content* of each leaf node is…

- **Text-Content** This identifies the *text-content* of the leaf-nodes of the analytical exemplar.

1.3.2 The Result

We have the original *edition* of the text, with its canonical scheme of citation. *E.g.* The Homeric *Iliad*, edition of the Venetus A, which begins with 1.1:

```
urn:cts:greekLit:tlg0012.tlg001.msA:1.1 = <l n='1'>μῆνιν ἄειδε θεὰ Πηληϊάδεω
Ἀχιλῆος</l>
```

We have an *analytical exemplar* derived from the *edition*. *E.g.* The Homeric *Iliad*, edition of the Venetus A, exemplar tokenized by word.

$$\underbrace{\underbrace{\mathit{urn}:\mathit{cts}:\mathit{greekLit}}_{\text{namespaces}}:\underbrace{\mathit{tlg}0012}_{\text{Homeric Poetry}}.\underbrace{\mathit{tlg}001}_{\text{Iliad}}.\underbrace{\mathit{persGrk}}_{\text{edition}}.\underbrace{\mathit{wt}}_{\text{exemplar}}:\underbrace{1.1.1}_{\text{“μῆνιν”}}}_{\text{cts-urn}}$$

So, `urn:cts:greekLit:tlg0012.tlg001.msA.wt:1.1.1` has *text content* μῆνιν. It is *aligned with* `urn:cts:greekLit:tlg0012.tlg001.msA:1.1@μῆνιν[1]`. It is *analyzed by* `urn:cite:hmt:iliadLexMSA.1`, a CITE-Object which might tell us that this object is a "noun", "feminine", "accustive", "singular", from the lemma "μῆνις", or even that it is the direct object of the sentence.

We can navigate the *exemplar* as we navigate the *edition*, and we can likewise identify or retrieve its citation-units at any level of granularity by URN reference.

Since the *exemplar* is aligned to the textual content of the *edition*, and all other *exemplars* derived from this *edition* are as well, we have implicit alignment across any analyses that anyone produces for this edition of the text.

1.4 Examples

The example above is so simple as to seem pointless: 1.1@μῆνιν[1] in the Edition is aligned to 1.1.1 in the Exmplar, with text-content "μῆνιν". Below, we give some examples of more complex or problematic kinds of analysis that this approach makes possible.

1.4.1 Lexical Tokens

The easiest case would be a traditional tokenization by lexical entities. This is a straightforward tokenization by word, allowing us to attach metadata to word-tokens.

> μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος — *Iliad* 1.1

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1@μῆνιν[1]`
Sequence	`1`
Analysis Record	`urn:cite:hmt:lexTokens.1`
Analysis	`urn:cite:hmt:lexTokens.1`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.lexTokens:1.1.1`
Text-Content	`μῆνιν`

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1@ἄειδε[1]`
Sequence	`2`
Analysis Record	`urn:cite:hmt:lexTokens.2`
Analysis	`urn:cite:hmt:lexTokens.2`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.lexTokens:1.1.2`
Text-Content	`ἄειδε`

Field	Value
Analyzed Text	urn:cts:greekLit:tlg0012.tlg001.msA:1.1@θεὰ[1]
Sequence	3
Analysis Record	urn:cite:hmt:lexTokens.3
Analysis	urn:cite:hmt:lexTokens.3
Analytical Exemplar URN	urn:cts:greekLit:tlg0012.tlg001.msA.lexTokens:1.1.3
Text-Content	θεὰ

1.4.2 Markup Problems

Even a simple "tokenization by word" becomes difficult when a text has complex editorial markup. A "lexical-token-exemplar" might choose to ignore editorial markup, but because its tokens would still be aligned to the Edition, the editorial status of any given token—unclear, supplied, *vel sim.*—could be determined. But for this analysis the text-content would simply be strings of Greek. The description of the analytical exemplar expresses the principles for its construction.

μῆν<unclear>ιν ἄει</unclear>δε θεὰ Πηληϊάδεω Ἀχιλῆος — *Iliad* 1.1

Field	Value
Analyzed Text	urn:cts:greekLit:tlg0012.tlg001.msN:1.1@μῆν[1]-1.1@ιν[1]
Sequence	1
Analysis Record	urn:cite:hmt:lexTokens.1
Analysis	urn:cite:hmt:lexTokens.1
Analytical Exemplar URN	urn:cts:greekLit:tlg0012.tlg001.msN.lexTokens:1.1.1
Text-Content	μῆνιν

Field	Value
Analyzed Text	urn:cts:greekLit:tlg0012.tlg001.msN:1.1@ἄει[1]-1.1@δε[1]
Sequence	2
Analysis Record	urn:cite:hmt:lexTokens.2
Analysis	urn:cite:hmt:lexTokens.2
Analytical Exemplar URN	urn:cts:greekLit:tlg0012.tlg001.msN.lexTokens:1.1.2
Text-Content	ἄειδε

Field	Value
Analyzed Text	urn:cts:greekLit:tlg0012.tlg001.msN:1.1@θεὰ[1]
Sequence	3
Analysis Record	urn:cite:hmt:lexTokens.3
Analysis	urn:cite:hmt:lexTokens.3
Analytical Exemplar URN	urn:cts:greekLit:tlg0012.tlg001.msN.lexTokens:1.1.3
Text-Content	θεὰ

1.4.3 Metrical Feet

A different tokenization, and a different analytical exemplar. This one captures metrical feet, which cross word-boundaries. The "Analysis" would be a urn identifying the kind of foot (*dactyl* or *spondee*, in this case).

> μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος — *Iliad* 1.1

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1@μῆνιν[1]-1.1@ἄ[1]`
Sequence	`1`
Analysis Record	`urn:cite:hmt:metricalAnalysis.1`
Analysis	`urn:cite:hmt:meter.dactyl`
Analytical Exemplar urn	`urn:cts:greekLit:tlg0012.tlg001.msA.feet:1.1.1`
Text-Content	`μῆνιν ἄ`

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1@ειδε[1]-1.1@θε[1]`
Sequence	`2`
Analysis Record	`urn:cite:hmt:metricalAnalysis.2`
Analysis	`urn:cite:hmt:meter.dactyl`
Analytical Exemplar urn	`urn:cts:greekLit:tlg0012.tlg001.msA.feet:1.1.2`
Text-Content	`ειδε θε`

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0012.tlg001.msA:1.1@ὰ[1]-1.1@Πη[1]`
Sequence	`3`
Analysis Record	`urn:cite:hmt:metricalAnalysis.3`
Analysis	`urn:cite:hmt:meter.spondee`
Analytical Exemplar urn	`urn:cts:greekLit:tlg0012.tlg001.msA.feet:1.1.3`
Text-Content	`ὰ Πη`

1.4.4 Syntax Problem

For analyzing syntax, it is common to separate certain words, so for οὔτε, the οὔ is treated as an adverb, and the τε as a coordinator. One approach as been to edit the text by splitting those words into two. But breaking up Greek words in an Edition, merely to serve a single kind of analysis, is not ideal. This approach lets us keep the Greek intact, while analyzing things like οὔτε according to its parts.

> ἵν᾽ οὔτε φωνὴν οὔτε του μορφὴν βροτῶν — Aeschylus, *PV* 21

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0085.tlg003:21@οὔτε[1]`
Sequence	`N`
Analysis Record	`urn:cite:fu:pvSyntax.45`
Analysis	`urn:cite:fu:pvSyntax.45`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0085.tlg003.synTok:21.2`
Text-Content	`οὔ`

Field	Value
Analyzed Text	`urn:cts:greekLit:tlg0085.tlg003:21@οὔτε[1]`
Sequence	`N+1`
Analysis Record	`urn:cite:fu:pvSyntax.46`
Analysis	`urn:cite:fu:pvSyntax.46`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0085.tlg003.synTok:21.3`
Text-Content	`τε`

1.4.5 Clauses

> 1 μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος
> 2 οὐλομένην, ἣ μυρί᾽ Ἀχαιοῖς ἄλγε᾽ ἔθηκε,
> 3 πολλὰς δ᾽ ἰφθίμους ψυχὰς Ἄϊδι προΐαψεν
> …
>
> — *Iliad* 1.1–1.3

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.

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`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.1.1`
Text-Content	`μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,`

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`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0012.tlg001.msA.clauses:1.2.1`
Text-Content	`μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος οὐλομένην,`

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

This example requires some discussion. 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*.[1])

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, say:

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

If we were to navigate our Edition and the derived Exemplar via a CTS service, the following URNs would return the following text-content:

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,

Exemplar-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,

If we were to submit a `getNextUrn` request to the CTS Service, we would get the following results:

[1] An example where the *analysis* and the *analysis record* would have different URNs might be an analysis of personal names. We might choose to analyze "Πηληϊάδεω" and "Ἀχιλῆος" individually. Each would have a unique *analysis record*, but each would bye *analyzed* the same CITE-URN, identifying an entity that is "Achilles, son of Peleus, hero of the Trojan War in Homeric Epic."

Input URN	Result of `getNextUrn`
`urn:cts:…msA.clauses:1.1.1`	next = `urn:cts:…msA.clauses:1.2.2`
`urn:cts:…msA.clauses:1.2.1`	next = `urn:cts:…msA.clauses:1.2.2`
`urn:cts:…msA.clauses:1.1`	next = `urn:cts:…msA.clauses:1.2`

1.5 Non-contiguous Text

> 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."

Colored text indicates "text reuse".

Field	Value
Sequence	`N`
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]`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0007.tlg012.perseus-grc1.histfrag:26.3.1`
Text-Content	`ὑπὸ τοῦ Μελίσσου καὶ Περικλέα αὐτὸν ἡττηθῆναι ναυμαχοῦντα πρότερον`

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-content** 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 not damaged our Edition, but we can present our analysis of quotation *as we choose*, and attach commentary, *vel sim.*, to the object pointed to by the Analysis URN.

While one editor might be content merely on the attributed paraphrase, another might want to analyze this text of Plutarch by promoting the quotation to direct speech. The *text content* of the Exemplar is a matter for editorial judgement. That editor's analysis would look like this:

Field	Value
Sequence	`N`
Analysis Record	`urn:cite:histfragNormal: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]`
Analytical Exemplar URN	`urn:cts:greekLit:tlg0007.tlg012.perseus-grc1.histfragNormal:26.3.1`
Text-Content	`ὑπὸ τοῦ Μελίσσου καὶ Περικλῆς αὐτὸς ἡττήθη ναυμαχῶν πρότερον`

1.5.1 The “Analysis-Object”

The **Analysis URN** may exist only to give a unique identifier to the analysis, or it may point to a CITE object with various fields. A CITE Object record for the example above might look like this:

URN	`urn:cite:histfrag:arist.577`
Type	“Quotation”
Genre	“Prose”
Source	“Aristotle”
Auth	M. Berti
Date	???
Notes	“…”

1.6 Generating this Data & Processing it into CITE Collections and CTS Texts

There are no generic analyses. Every specific analysis of each text is going to be unique. Any project that has undertaken even the simplest kind of tokenization knows how quickly it becomes necessary to make editorial decisions. For the *Homer Multitext* and work on editions of Aeschylus at Furman University, we have scripts that generate specific tokenizations. For the paleographic work on the *Homer Multitext* we rely on human editors to define characters, glyphs, abbreviations, and so forth, on our Homeric manuscripts. Some analyzes can be generated from elements in a TEI-XML text (our personal-names analyses for the *HMT* texts is one example).

Generally, there are ways to automated parts of the process, such as generating analysys-URNs in sequence for a table of analyses. We indend to supplement our CTS utilities along the lines that Bridget Almas has already demonstrated extremely effectively in SOSOL, to make it easier to select passages of “analyzed text” from an Edition.

Each of the examples above can be represented by a tab- or comma-delimited text file. This can then be processed to generate a CITE collection and the necessary RDF to include the Analytical Exemplar in a CTS library.

We are working on incorporating these scripts to turn ORCA records into fully processed CITE and CTS data. These will be integrated into our CITE Manager utility: https://github.com/cite-architecture/citemgr.

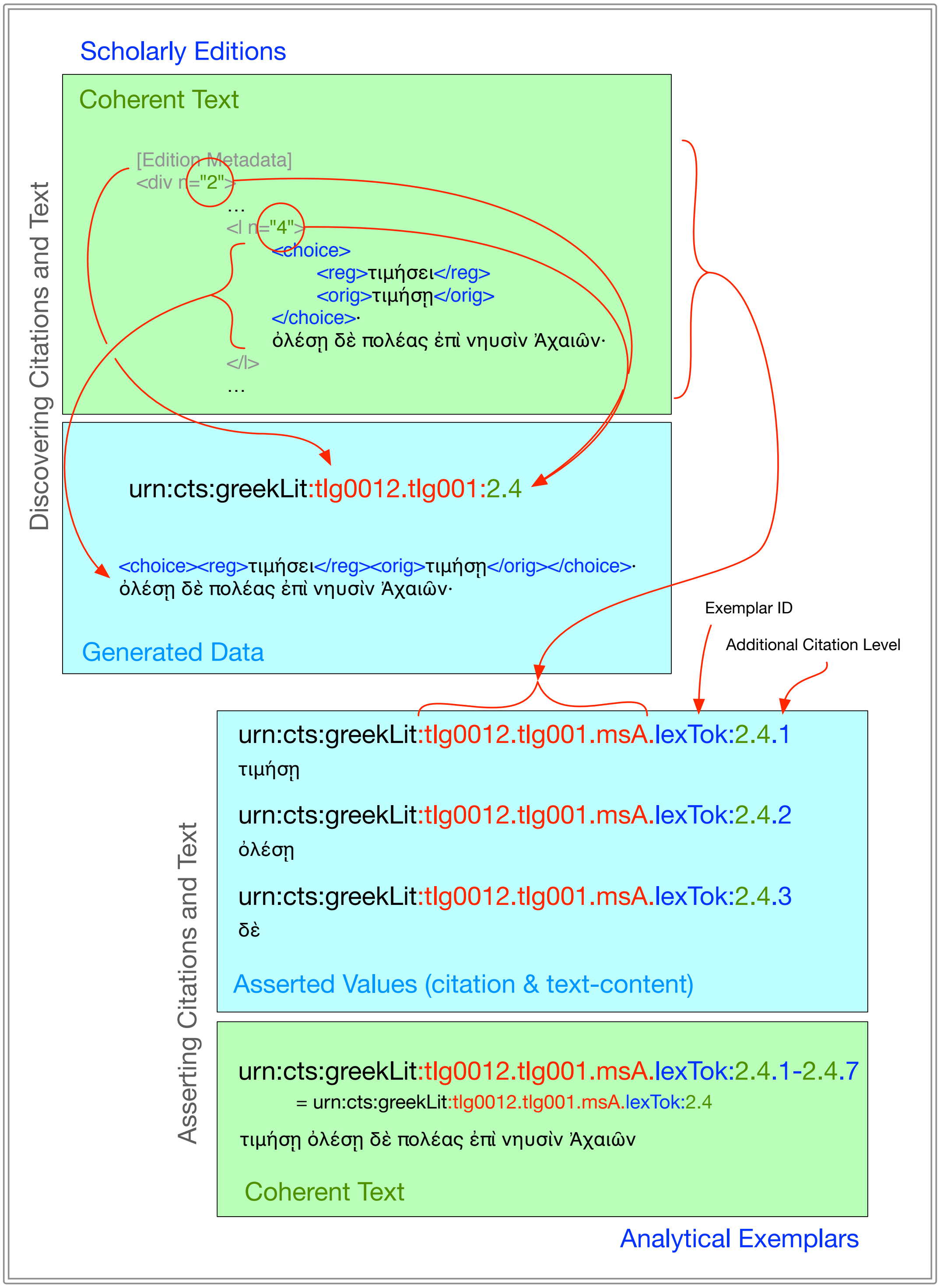

Scholarly Editions
Discovering Citations and Text
Coherent Text
[Edition Metadata]
<div n="2">
…
<l n="4">
<choice>
<reg>τιμήσει</reg>
<orig>τιμήσῃ</orig>
</choice>·
ὀλέσῃ δὲ πολέας ἐπὶ νηυσὶν Ἀχαιῶν·
</l>
…
urn:cts:greekLit:tlg0012.tlg001:2.4
<choice><reg>τιμήσει</reg><orig>τιμήσῃ</orig></choice>·
ὀλέσῃ δὲ πολέας ἐπὶ νηυσὶν Ἀχαιῶν·
Generated Data
Exemplar ID
Additional Citation Level
Asserting Citations and Text
urn:cts:greekLit:tlg0012.tlg001.msA.lexTok:2.4.1
τιμήσῃ
urn:cts:greekLit:tlg0012.tlg001.msA.lexTok:2.4.2
ὀλέσῃ
urn:cts:greekLit:tlg0012.tlg001.msA.lexTok:2.4.3
δὲ
Asserted Values (citation & text-content)
urn:cts:greekLit:tlg0012.tlg001.msA.lexTok:2.4.1-2.4.7
= urn:cts:greekLit:tlg0012.tlg001.msA.lexTok:2.4
τιμήσῃ ὀλέσῃ δὲ πολέας ἐπὶ νηυσὶν Ἀχαιῶν
Coherent Text
Analytical Exemplars

Figure 3: Complementary models of creating a “text”: (a) discovering citation-values and associated text in an XML file; (b) asserting citation-values and assigning text-content to them.

References

[1] D. N. Smith and G. Weaver, "Applying domain knowledge from structured citation formats to text and data mining: Examples using the CITE architecture," *Text Mining Services*, p. 129, 2009.

[2] C. Blackwell and D. Smith, "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* (L. Muellner, ed.), Washington, DC: The Center for Hellenic Studies of Harvard University, 2012.

[3] C. Blackwell and N. Smith, "CTS URN specification 2.0," Feb. 2014.

[4] C. Blackwell and N. Smith, "CTS protocol specification," Feb. 2014.