Cataloging for a Billion Word Library of Greek and Latin

Gregory Crane, Bridget Almas,
Alison Babeu, Lisa Cerrato, Anna Krohn
Perseus Digital Library, Tufts University
Medford, MA
gregory.crane@tufts.edu

Frederik Baumgart, Monica Berti,
Greta Franzini, Simona Stoyanova
University of Leipzig
Leipzig, Germany
fbaumgardt@e-humanities.net

ABSTRACT

This paper reports work on a catalog that includes not only standard metadata but also a complete reference transcription for each work so that users can explicitly cite not only every version but also every word in every version of a work. The Functional Requirements for Bibliographic Records conceptual model (FRBR) allows us to move beyond printed books and to track the logical units within (and often across) printed books: works (e.g., the *Iliad*) and expressions (e.g., versions such as the 10th century Venetus A manuscript or Butler's English translation). The Canonical Text Services (CTS) Data Model builds upon FRBR, allowing us to cite each word in any version of a text and to do so by building upon established citation schemes inherited from print (e.g., the chapter/verse citation scheme in the Bible). This paper describes a concrete implementation of such a catalogue of 3,679 Greek and Latin works that includes FRBR inspired metadata and TEI XML transcriptions that were revised to facilitate implementing a CTS API. It also describes how all the different versions of a work can be serialized as variations on the reference version. The FRBR+CTS catalog provides data by which text re-use and alignment services can automatically detect different versions of and quotations from the reference text, aligning all discovered instances according to a canonical citation scheme.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Standards*

General Terms

Documentation, Design, Standardization

Keywords

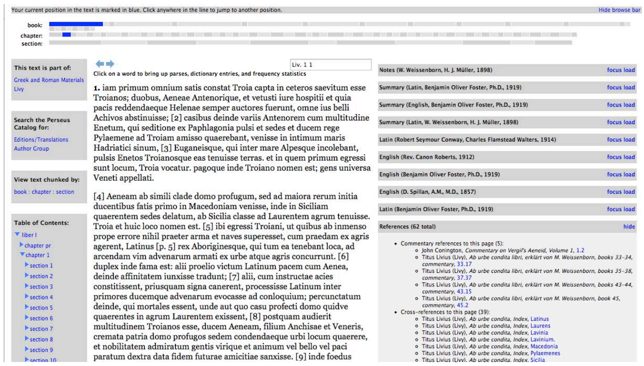
FRBR, CTS, metadata, catalogs

1. INTRODUCTION

In 2010, David Bamman identified among the first 1.2 million books downloaded from the Internet Archive, approximately 22,000 books that were primarily in Latin — a collection that contained more than 2.4 billion words. [3] By contrast, the commercial Thesaurus Linguae Graecae (TLG) Digital Library of Classical Greek and Brepols Series-A collections of Classical Latin collections claim 105 and 65 million words respectively.[1] The first survey of Latin alone from this first million plus books [3] thus created a collection of Latin that was an order of magnitude larger than the two largest commercial collections of Greek and Latin combined. The HathiTrust[2] lists 83,000 titles in Latin, of which 64,000 are in the public domain and available for extensive automatic processing by HathiTrust members. If proportions observed in the Internet Archive sample are true for the Hathi collections, more than 8 billion words of Greek and Latin are available in print books available in digital form.

Many users in many instances — perhaps most users in most instances — who are studying primary sources are not always interested in books. They are interested in the logical primary sources that may be published as parts of books and in all the information that they need to understand those sources. The standard reading environment of the Perseus Digital Library[3] provides a concrete, well-established, if still in some ways rudimentary, response to this need. Perseus has gradually developed digital collections since 1987, with a particular focus upon Greek and Latin; it serves an international audience, with the number of unique visitors in 2013 ranging from 237,000 during the summer to more than 420,000 during the fall semester.

A core function of Perseus has been to organize information relevant to a particular canonical chunk of a text. Perseus is designed to manage multiple versions of the same work.


Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.
Hide browse bar
book:
chapter:
section:
This text is part of:
Greek and Roman Materials
Livy
Search the Perseus Catalog for:
Editions/Translations
Author Group
View text chunked by:
book : chapter : section
Table of Contents:
liber I
chapter pr
chapter 1
section 1
section 2
section 3
section 4
section 5
section 6
section 7
section 8
section 9
section 10
Liv. 1 1
Click on a word to bring up parses, dictionary entries, and frequency statistics
Notes (W. Weissenborn, H. J. Müller, 1898)
Summary (Latin, Benjamin Oliver Foster, Ph.D., 1919)
Summary (English, Benjamin Oliver Foster, Ph.D., 1919)
Summary (Latin, W. Weissenborn, H. J. Müller, 1898)
Latin (Robert Seymour Conway, Charles Flamstead Walters, 1914)
English (Rev. Canon Roberts, 1912)
English (Benjamin Oliver Foster, Ph.D., 1919)
English (D. Spillan, A.M., M.D., 1857)
Latin (Benjamin Oliver Foster, Ph.D., 1919)
focus load
References (62 total)
hide

1. iam primum omnium satis constat Troia capta in ceteros saevitum esse Troianos; duobus, Aeneae Antenorique, et vetusti iure hospitii et quia pacis reddendaeque Helenae semper auctores fuerunt, omne ius belli Achivos abstinuisse; [2] casibus deinde variis Antenorem cum multitudine Enetum, qui seditione ex Paphlagonia pulsi et sedes et ducem rege Pylaemene ad Troiam amisso quaerebant, venisse in intimum maris Hadriatici sinum, [3] Euganeisque, qui inter mare Alpesque incolebant, pulsis Enetos Troianosque eas tenuisse terras. et in quem primum egressi sunt locum, Troia vocatur. pagoque inde Troiano nomen est; gens universa Veneti appellati.

[4] Aeneam ab simili clade domo profugum, sed ad maiora rerum initia ducentibus fatis primo in Macedoniam venisse, inde in Siciliam quaerentem sedes delatum, ab Sicilia classe ad Laurentem agrum tenuisse. Troia et huic loco nomen est. [5] ibi egressi Troiani, ut quibus ab inmenso prope errore nihil praeter arma et naves superesset, cum praedam ex agris agerent, Latinus [p. 5] rex Aboriginesque, qui tum ea tenebant loca, ad arcendam vim advenarum armati ex urbe atque agris concurrunt. [6] duplex inde fama est: alii proelio victum Latinum pacem cum Aenea, deinde affinitatem iunxisse tradunt; [7] alii, cum instructae acies constitissent, priusquam signa canerent, processisse Latinum inter primores ducemque advenarum evocasse ad conloquium; percunctatum deinde, qui mortales essent, unde aut quo casu profecti domo quidve quaerentes in agrum Laurentem exissent, [8] postquam audierit multitudinem Troianos esse, ducem Aeneam, filium Anchisae et Veneris, cremata patria domo profugos sedem condendaeque urbi locum quaerere, et nobilitatem admiratum gentis virique et animum vel bello vel paci paratum dextra data fidem futurae amicitiae sanxisse. [9] inde foedus

- Commentary references to this page (5):
    - John Conington, *Commentary on Vergil's Aeneid, Volume 1*, 1.2
    - Titus Livius (Livy), *Ab urbe condita libri, erklärt von M. Weissenborn, books 33-34, commentary*, 33.17
    - Titus Livius (Livy), *Ab urbe condita libri, erklärt von M. Weissenborn, books 35-38, commentary*, 37.37
    - Titus Livius (Livy), *Ab urbe condita libri, erklärt von M. Weissenborn, books 43-44, commentary*, 43.15
    - Titus Livius (Livy), *Ab urbe condita libri, erklärt von M. Weissenborn, book 45, commentary*, 45.2
- Cross-references to this page (39):
    - Titus Livius (Livy), *Ab urbe condita, Index*, Latinus
    - Titus Livius (Livy), *Ab urbe condita, Index*, Laurens
    - Titus Livius (Livy), *Ab urbe condita, Index*, Lavinia
    - Titus Livius (Livy), *Ab urbe condita, Index*, Lavinium.
    - Titus Livius (Livy), *Ab urbe condita, Index*, Macedonia
    - Titus Livius (Livy), *Ab urbe condita, Index*, Pylaemenes
    - Titus Livius (Livy), *Ab urbe condita, Index*, Sicilia

Figure 1: A text as viewed in the Perseus Digital Library

DATeCH 2014, May 19 - 20 2014, Madrid, Spain
Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-2588-2/14/05…$15.00.
http://dx.doi.org/10.1145/2595188.2595190

[1] http://www.tlg.uci.edu/;
http://www.brepols.net/publishers/pdf/Brepolis_LLT_EN.pdf

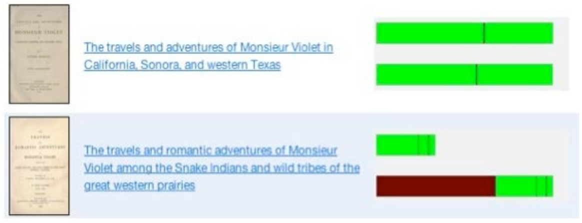
[2] http://www.hathitrust.org/

[3] http://www.perseus.tufts.edu/hopper/

The figure above presents an English translation for chapter 1, book 1 of Livy's *History of Rome*. A particular edition of the Latin is in the main field in the center while structural metadata appears on the left. The right hand column, however, illustrates an attempt to aggregate as much information as possible and aligns multiple editions, translations, an ancient summary of book 1 of Livy in Latin and in English, as well as a German commentary, and references to Livy 1.1 from various reference works.

The approach currently implemented in Perseus has at least two fundamental drawbacks. First, the results are not customized (adapted to user-specified parameters) or personalized (adapted to the needs of the user as inferred from prior behavior). The second drawback mitigates the effects of the first: the data aggregation depends upon hand-encoding of XML texts — a labor-intensive task that is inherently not scalable to millions of books, or even to the 90,000 digitized printed books that the HathiTrust lists as being in Greek and/or Latin. As we begin to aggregate this much information about Greek and Latin texts, simply reporting all known information will become less and less useful.

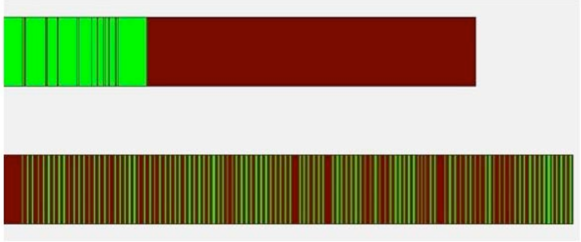
Fully automated methods can at least in part address both challenges. A variety of customization and personalization technologies exist and could be implemented. Methods also exist to automatically identify the chapter and verse style primary source citations by which students of Greek and Latin texts have cited most surviving texts for generations [9]. Methods also exist to identify multiple versions of, and even quotations from, a work [5, 6].


The travels and adventures of Monsieur Violet in California, Sonora, and western Texas
The travels and romantic adventures of Monsieur Violet among the Snake Indians and wild tribes of the great western prairies

In large scanning projects, such as those undertaken by Google or the Internet Archive in multiple libraries, a particular book or work may be scanned multiple times. In this visualization, aligned portions of two books are shown with green regions, which take up most of the text. Two of the editions of T*he travels and adventures of Monsieur Viole*t contain the entire work, while the third edition contains only the third volume. The green portion, therefore, of the last bar shows that it covers only the last third of the original work.

Figure 2: Duplicate detection by the Mining a Million Scanned Books Project

By looking for shared sequences of words and even letters, we can find many different versions of a text, even when other versions of the text have different editorial readings or when 20% of the characters are incorrectly transcribed. This builds upon extensive research in string matching from bio-informatics and other domains [8, 10, 11]. A transcribed TEI-XML[4] text is not simply a text that is useful by itself; it is also an extended query by which we can search very large collections. In a truly digital library, the text itself is part of the metadata by which we search and organize sources. Querying Google for three or more of the words in the quotation in Figure 4 generates thousands of documents that quote this part of the relevant passage (for example searching for "iactatus et alto" retrieves 27,800 hits).



Text alignment is also used for finding groups of texts whose structure corresponds in other ways, such as works published in different languages, or texts and their commentaries. Here, for instance, we see an automatically generated alignment between the Latin text of Vergil's *Aeneid* and a commentary. The first bar depicts the first eight books of the *Aeneid*. The green in this first bar indicates the aligned portions, from which we can tell that the commentary only deals with the first three books of the *Aeneid*. The second bar depicts the commentary. Its green portions are brief passages from the text of the Aeneid, and the intervening red bars are the commentary, which does not align.

Figure 3: Partial duplicate detection by the Mining a Million Scanned Books Project

Figures 2 and 3 show how work done by the Mining a Million Scanned Books Project[5] can detect full and partial duplicates of a work in a very large collection. The work presented here allows such a system not only to detect where a work appears in part or in full but also to align those different versions and quotations to a conventional citation scheme.

Automatic text reuse detection cannot by itself create a structured visualization such as the one offered by Perseus, because the Perseus visualization is organized around standard chapter/verse, book/line etc. citation schemes. We need a transcription of at least one reference edition where the citation scheme has been encoded.

```
<div1 type="Book" n="1">
<milestone ed="p" n="1" unit="card"/>
<l n=”1”>Arma virumque cano, Troiae qui primus ab oris</l>
<l n=”2”>Italiam, fato profugus, Laviniaque venit</l>
<l n=”3”>litora, multum ille et terris iactatus et alto</l>
```

Figure 4: Encoded text from Virgil's *Aeneid*

Once we have a text encoded as in Figure 4, we know that the words "Arma virumque cano, Troiae qui primus ab oris" constitute line 2 of book 1 of a particular edition of the *Aeneid*. This citation+string becomes a key that we can use to scour large collections for quotations and alternate editions [2]. Because we can use n-grams of characters as well as words, we can identify many different versions of a work — for instance only about 10% of the words in a complicated text such as Aeschylus's *Suppliant Women* will differ from edition to edition. A single reference edition is all that is needed to identify many quotations and almost all duplicate editions. The reference edition does not have to be the most up-to-date edition and it can even include residual data entry errors.

In the rest of this paper, we describe two foundational data structures: (1) FRBR inspired metadata by which we can track many different versions of a work (whether or not we have a full transcriptions) and (2) the TEI-XML reference transcriptions used to align multiple versions of a work.

2. THE CANONICAL TEXT SERVICES DATA MODEL

The Canonical Text Services (CTS) Protocol is a specification that "defines a network service for identifying texts and for retrieving fragments of texts by canonical reference expressed as

[4] http://www.tei-c.org/index.xml

[5] http://ciir.cs.umass.edu/research/massivedata/

CTS-URNs."[6] The CTS data model extends the Functional Requirements for Bibliographic Records (FRBR) hierarchy so that it can address every word in every version of a Work. This data model includes the following entities:

Textgroups are defined as "traditional, convenient groupings of texts such as 'authors' for literary works, or corpus collections for epigraphic or papyrological texts"[7] and include unique identifiers but also support multiple titles (to support multi-lingual collections). The use of this concept allows us to rationalize the fact that classicists cite Thucydides (a particular author) and the *Greek Anthology* (a Byzantine collection with poems by many authors) in much the same way (Thuc. 1.22.1 vs. Anth. 3.22.2).

As defined by the official FRBR guidelines[8], a Work is "a distinct intellectual or artistic creation." For example, Plato's Allegory of the Cave occurs in his work, *The Republic*.

We use these entities to develop CTS compliant URNs[9]. As a part of the CTS and CITE Architecture[10], these URNs "provide the permanent canonical references" on which CTS relies "in order to identify or retrieve passages of text." An example may prove illustrative. The text group Homer has the URN urn:cts:greekLit:tlg0012; the work *Iliad* has the URN urn:cts:greekLit:tlg0012.tlg001; and lastly, the edition of the *Iliad*, published in 1931 and edited by Thomas W. Allen, has the URN urn:cts:greekLit:tlg0012.tlg001.perseus-grcX1.

3. THE PERSEUS CATALOG

The Perseus Catalog[11] is an attempt to provide systematic catalog access to at least one online edition of every major Greek and Latin author (both surviving and fragmentary) from antiquity to 600 CE. To do so, the Perseus Catalog uses the CTS/FRBR data model to represent different expressions (primarily manuscript witnesses, scholarly editions and translations into other languages) of particular works. The Perseus Catalog provides breadth of coverage, including many editions for which we have no scanned page images, much less curated TEI XML.

Still a work in progress, the catalog currently includes 3,679 individual works (2,522 Greek and 1,247 Latin), with over 11,000 links to online versions of these works (6,419 in Google Books, 5,098 to the Internet Archive, 593 to the HathiTrust). The Perseus interface now includes links to the Perseus Catalog from the main navigation bar, and also from within the majority of texts in the Greco-Roman collection.

The current Perseus Catalog of Greek and Latin was first conceived of in 2005 as a "FRBRized" catalog for the Perseus Digital Library's online collection of Greek and Latin texts [7]. This eventually grew into what became known as the "FRBR-Inspired catalog" [1] for a growing collection of digitized Greek and Latin books (both being produced in-house at Perseus and in the Open Content Alliance).

This collection made use of the MODS[12] and MADS[13] standards developed by the U.S. Library of Congress and was intended to provide cataloged access to at least one version of every surviving major Greek and Latin author from antiquity. The catalog is currently being re-conceptualized as the basis for the Open Greek and Latin Project and is also part of both the Billion Word Library and the Reinventing Humanities Publication[14], projects both supported by the European Social Fund and hosted at the University of Leipzig. Open Greek and Latin is but one component of the recently-announced Open Philology Project launched by the Humboldt Chair of Digital Humanities at Leipzig.[15]

In the Perseus Catalog, the user will find record identifiers such as:

urn:cts:greekLit:tlg0284.tlg052.perseus-grc1[16]

urn:cts:latinLit:phi0474.phi052.opp-lat1[17]

urn:cts:latinLit:stoa0255.stoa004[18]

These Canonical Text Service (CTS) -URN identifiers used in the Perseus Catalog reference identifiers for authors and works from the TLG and Packard Humanities Institute (PHI)[19] canons as well as in some cases the Stoa Latin Text Inventory. These identifiers were used because they have domain-specific meaning for members of the classical community and provide a semantic cue as to the author or work being referenced. They do not indicate that a specific edition from the TLG or PHI canon is being referenced.

In addition, over the course of cataloging in the last seven years, many works have been discovered without identifiers in any of these canons, including the following types:

- Anonymous works that could not be identified reliably or did not have a work identifier (this applies in particular to a number of smaller Latin poems in anthologies).
- Works by later classical Latin authors (due to the relatively early end date of the PHI and the sparser coverage of the Stoa inventory).
- Works by authors about whom nothing was reliably known, often not even the correct form of their name.
- Fragmentary works.

[6] http://www.homermultitext.org/hmt-doc/cite/cts-urn-overview.html

[7] http://www.homermultitext.org/hmt-doc/cite/texts/ctsoverview.html

[8] http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[9] http://en.wikipedia.org/wiki/Uniform_resource_name

[10] http://www.homermultitext.org/hmt-doc/cite/index.html

[11] http://catalog.perseus.org

[12] http://www.loc.gov/standards/mods/

[13] http://www.loc.gov/standards/mads/

[14] http://sites.tufts.edu/perseusupdates/2013/05/02/reinventing-humanities-publication-project-receives-e1-1-million-grant-from-the-saxon-ministry-of-science-and-european-social-fund/

[15] http://sites.tufts.edu/perseusupdates/2013/04/04/the-open-philology-project-and-humboldt-chair-of-digital-humanities-at-leipzig/

[16] http://catalog.perseus.org/catalog/urn:cts:greekLit:tlg0284.tlg052.perseus-grc1

[17] http://catalog.perseus.org/catalog/urn:cts:latinLit:phi0474.phi052.opp-lat1

[18] http://catalog.perseus.org/catalog/urn:cts:latinLit:stoa0255.stoa004

[19] http://packhum.org/

- Works by authors later determined to be fictitious (pseudo-authors or names simply attributed to a work).

During the first few years of cataloging, the basic procedure was to simply create a catalog record for a work with no identifier and label it as such. Starting in 2011, when the catalog first became available online through an eXtensible catalog[20] implementation, the importance of unrepeated identifiers to ultimately (and ideally) support the aggregation and discovery of all uniquely cataloged works became increasingly clear. Because of the expandable nature of the Stoa Registry of Latin Literature, identifiers have been created both for Latin works that had either 1) previously been cataloged and had no PHI or existing Stoa ids or 2) for newly cataloged Latin works without any identifiers. For Greek works that were not found in the TLG canon (a much smaller number), a basic pattern of tlg-author name has been used as a placeholder in the MADS and MODS files until such time as a more formal system of identifier creation for fragmentary and fictitious authors is decided upon (that will likely make use of both CTS and CITE Collections). Currently, if a work does not have a unique identifier, it cannot be found within the current catalog interface.

Fragmentary authors and authors of small surviving texts have been particularly challenging due to their often inconsistent treatment in the traditional canons. For example, the *TLG Canon of Greek Authors and Works* (3rd edition)[21] assigns unique identifiers to many fragmentary historians and to the individual epigrammatists found in the *Greek Anthology*. In the online canon, however, searching on the identifiers for epigrammatists yields no results as all the individual epigrams are now found under the identifier 7000.001 (*Anthologia Graeca*) and the user needs to know the number of the book and individually numbered epigram to find a specific epigram by an author. Although the individual Greek epigrammatists are not individually searchable (although their texts are still extant), a user can, nonetheless, individually search for fragmentary historians (although technically their work only exists as part of other surviving works) thus leading to double results [4]. Thus, the Perseus Catalog has made use of the last printed TLG canon and utilized the identifiers for both fragmentary authors and other authors such as epigrammatists.

Another identifier issue occurs when a single group identifier is used to identify works that are often individually referenced in published editions, such as the *Lives* of Nepos (phi588.1, stoa0210-stoa003) and Suetonius (phi1348.1, stoa0268-stoa006). In some cases, such as the *Dialogi* of Seneca the Younger, the works have individual Stoa identifiers (stoa0255-stoa004, stoa0255-stoa006 to stoa0255-stoa014), but only a single PHI identifier (1017.12). In the cases of both Nepos and Suetonius, the Perseus Catalog has followed an earlier solution developed for the Perseus Digital Library of creating unique identifiers called Abstract Bibliographic Object (ABOs) in order to uniquely reference each of the individual lives (for example, urn:cts:latinLit:phi0588.abo002 for the life *Themistocles* by Nepos). For the *Dialog*i of Seneca and in several other instances we have chosen to use the Stoa rather than the PHI identifier as the default work identifier in order to support the most granular level of work identification possible within the catalog.

Another problematic case is when a single, top-level work identifier is used for a work attributed to multiple traditional (often dubious) authors, all of whom have authority records. This is the case with the authors in the *Scriptores Historiae Augustae* or *Historia Augusta*, which has the top-level identifier phi2331, with various sub level identifiers for the individual work titles. This can create data aggregation challenges when attempting to support both searching for a textgroup such as the *Scriptores Historiae Augustae*, while also attempting to preserve the ability to search for the traditional individual author names.

4. CTS-COMPLIANT EPIDOC TEI XML REFERENCE TRANSCRIPTIONS

Where the FRBR catalog provides breadth of coverage by tracking many different versions of a work, the collection of reference transcriptions provides depth. The collection attempts to provide at least one curated transcription for every work so that this transcription can be used as a query to locate, and a framework around which to organize, many other versions and quotations of the same work in potentially very large collections.

In order to implement this we decided to make the canonical citation scheme the primary hierarchical structure in our TEI XML. CTS describes a protocol and the backend data can be stored in many different ways — the Perseus Hopper developed by David A. Smith in the 1990s could already extract multiple overlapping hierarchies from XML texts. But we made a decision to structure the TEI XML texts so as to facilitate third parties who wished to integrate them into a CTS compliant environment. This transformation required a substantial restructuring: citation boundaries often occurred in the middle of other hierarchies (most often speeches or extended quotations).

We have also begun to review the tagging in our Greek and Latin texts to make them more consistent with each other and to bring these TEI XML documents more consistent with each other and to make the collection as a whole more consistent with the epiDoc[22] subset of TEI XML tags developed by classicists, originally for inscriptions and papyri. At present we have draft revised TEI XML P5 versions for 2,232 of 2,770 Greek and Latin primary texts in the Perseus collection. These can now be represented in a serialized form where each word has a unique identifier.

Given a text with citation scheme, we can now build a unique identifier for each word in that text The URN cts:latinLit:phi0914.phi0011.perseus-lat3:1.40 designates chapter 40 of book 1 of Livy's *History of Rome* in the Conway/Walters Oxford Classical Text edition of Livy books 1-5. The string "cum intentus in eum" within that chapter can be represented as cum@3, intentus@1, in@9, eum@1, se@4. Thus we uniquely identify the third instance of the word cum, the ninth instance of the word in, and the fourth instance of the word se in that version of that cited chunk of Livy.

We use this index (Table 1) rather than a serial number (e.g, word 1, word 2, … word n) to simplify representing variants in different editions. This allows us to represent different editions in a compact fashion. In the examples below, we see how the transcription of the edition in the Oxford Classical Text differs from the transcription of the edition in the Loeb Classical Library.

[20] http://www.extensiblecatalog.org/

[21] http://www.worldcat.org/oclc/20828572

[22] http://sourceforge.net/p/epidoc/wiki/Home/

Table 1: Index of words in Livy editions

OCT				Loeb	
1.41	confidere	1	Same	Confidere	1
1.41	**propediem**	**1**	**Sub**	**Prope**	**1**
1.41			**Insert**	**Diem**	**1**
1.41	Ipsum	1	Same	Ipsum	1
1.41	Eos	1	Same	Eos	1

5. TOWARDS LINKED DATA

The CTS data has been developed over several years and as such is currently at various levels of compatibility with current best practices for Linked Data.

Because the data in the catalog is essential to facilitating the current development efforts of the Perseus Digital Library and the new Open Philology Project, and we also believe it can be a valuable tool for others in the Digital Classics and Digital Humanities communities, we made a decision to release the data and the catalog interface to it before we could claim full compliance with Linked Data standards and before the revised TEI XML texts were ready. We are taking an incremental approach to compliance. This coincides with a larger effort of the Perseus Project to make all of its data available as proper Linked Data.

We have started by thinking carefully about the URIs that we are using to name and address the Perseus texts, catalog metadata, and other data objects from the Perseus Digital Library, ensuring that these URIs will be stable and properly dereferenceable.[23] Publishing and supporting these URIs was a core requirement for 1.0 Release of the catalog.

As of the 1.0 Release, URIs are used to name all Textgroups, Works, Editions and Translations in the catalog, and we have published alternate versions of these URIs for the HTML and Atom resource formats currently available for the catalog data. We have also linked the texts in the Perseus Digital Library to the Perseus Catalog. And finally we have linked the canonical data URIs for the Perseus texts and citations to the catalog via HTTP 303 redirects, so that if a text or citation addressed using a Perseus data URI is not yet available in the Perseus Digital Library, users of these URIs can be redirected to any bibliographic information available in the Perseus Catalog for the requested resource.

The next steps on the roadmap to linked data compliance will be:

- to release all the Perseus Catalog data as RDF triples, available via common RDF serialization formats, including RDF/XML and JSON-LD.
- to add RDF-A attributes to the HTML displays of the Perseus Catalog

6. MORE CTS REFERENCE EDITIONS

As of January 2014, Leipzig has data entry contracts on-going with two different vendors to create TEI XML transcriptions for all 55 public domain volumes of the *Corpus Scriptorum Ecclesiasticorum Latinorum* (CSEL) and the first 50 volumes of the Patrologia Latina series. Leipzig is also preparing a data entry contract based upon a new workflow for correction of OCR-generated Classical Greek.

7. CONCLUSION

The CTS data model provides a structure by which the Perseus Catalog and Perseus Collection of Greek and Latin TEI XML source documents can smoothly drive each other. We can identify a new edition of a Greek or Latin work in the Perseus Catalog and then create a TEI XML transcription with citation scheme. We can then use the TEI XML transcription as a query with which to search for many other editions and quotations of the work, returning not only matched text but the appropriate citations (e.g., this edition of Livy covers only books 1-5, this section of Athenaeus quotes that section of the *Iliad*). The matched text and citation data then feeds back into the Perseus Catalog (e.g., allowing us to specify precisely what chunks of a work appear in a given printed volume or on a given printed page).

The integrated collection of FRBR metadata and CTS compliant TEI XML Greek and Latin sources thus provides us with a framework that supports large scale and intensive analysis. The FRBR hierarchy allows us to organize many different versions of the same work. The CTS encoded transcriptions allow us to identify partial instances of a text (e.g., an edition of book 6 of the *Aeneid* or a quotation from chapter 42 of book 2 of Thucydides' *History of the Peloponnesian War*). The word by word serialization allows us to support precise annotations on particular words in particular editions. Overall, the FRBR/CTS data provides a foundation for big and deep data analysis that is essential to a mature digital infrastructure.

8. ACKNOWLEDGMENTS

The activities reported here received support at different times from the Mellon Foundation, the National Endowment for the Humanities, the Institute for Museum and Library Services, the Deutsche Forschungsgemeinschaft, the Alexander von Humboldt Foundation and the European Social Fund.

9. REFERENCES

[1] Babeu, A. 2008. *Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience.* White Paper. Mellon Foundation. http://www.perseus.tufts.edu/publications/PerseusFRBRExperiment.pdf

[2] Bamman, D., Babeu, A., and Crane, G. 2010. Transferring structural markup across translations using multilingual alignment and projection. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL '10)*. ACM, New York, NY, USA, 11-20. DOI= http://dx.doi.org/10.1145/1816123.1816126

[3] Bamman, D., and Smith, D. 2012. Extracting two thousand years of Latin from a million book library. *J. Comput. Cult. Herit*. 5, 1, Article 2 (April 2012), DOI=http://doi.acm.org/10.1145/2160165.2160167.

[23] The syntax for objects in the Perseus catalog is described here (http://sites.tufts.edu/perseuscatalog/documentation/user-guide/catalogdata-uris/), while the URI syntax for texts, citations and other data objects in the Perseus Digital Library can be found here (http://sites.tufts.edu/perseusupdates/beta-features/perseus-stable-uris/)

[4] Berti, M., Romanello, M., Babeu, A. and Crane, G. 2009. Collecting fragmentary authors in a digital library. In *JCDL '09: Proceedings of the 2009 joint international conference on Digital libraries*, New York, NY, USA, pp. 259-262. ACM. DOI=http://doi.acm.org/10.1145/1555400.1555442.

[5] Büchler, M., Crane, G., Moritz, M. and Babeu, A. 2012. Increasing recall for text re-use in historical documents to support research in the humanities theory and practice of digital libraries. Volume 7489 of *Lecture Notes in Computer Science*, Chapter 11, pp. 95-100. Berlin, Heidelberg: Springer Berlin / Heidelberg. DOI= http://dx.doi.org/10.1007/978-3-642-33290-6_11.

[6] Coffee, N., Koenig, J.P., Poornima, S., Forstall, C.W. Ossewaarde, R., and Jacobson, S.L. 2013. The Tesserae project: intertextual analysis of Latin poetry. *Literary and Linguistic Computing*, (2013) 28 (2): 221-228. DOI= http://dx.doi.org/10.1093/llc/fqs033

[7] Mimno, D., Crane, G., and Jones, A. 2005. Hierarchical catalog records: implementing a FRBR catalog. In *D-Lib Magazine*, 11 (10), 2005. http://www.dlib.org/dlib/october05/crane/10crane.html.

[8] Olsen, M., Horton, R. and Roe, G. 2010. Something borrowed: sequence alignment and the identification of similar passages in large text collections. *Digital Studies/Le champ numérique 2* (1). http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/190/235#d0e1039

[9] Romanello, M. Creating an annotated corpus for extracting canonical citations from Classics-related texts by using active annotation. *Computational Linguistics and Intelligent Text Processing, ser. Lecture Notes in Computer Science*, A. Gelbukh, Ed. Springer Berlin Heidelberg, 2013, vol. 7816, pp. 60-76. http://dx.doi.org/10.1007/978-3-642-37247-6_6.

[10] Smith, D. A., Manmatha, R., and Allan, J. 2011. Mining relational structure from millions of books: position paper. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing*, *BooksOnline '11*, New York, NY, USA, pp. 49-54. ACM. DOI= http://dx.doi.org/10.1145/2064058.2064069

[11] Yalniz, I. Z., Can, E.F. and Manmatha, R. 2011. Partial duplicate detection for large book collections. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, New York, NY, USA, pp. 469-474. ACM. http://dx.doi.org/10.1145/2063576.2063647