CREATING CTS COLLECTIONS

Dorothy PORTER
William DU CASSE
Jerzy W. JAROMCZYK
Neal MOORE
Ross SCAIFE

University of Kentucky

Jack MITCHELL

Stanford University

Session Contact: Dorothy Porter,
Collaboratory for Research in Computing
for Humanities, University of Kentuck**y**

The Classical Text Services protocol [CTS] provides the means for coordinating and integrating XML-encoded documents on a single subject. Collaborative efforts among the Stoa Consortium and the Collaboratory for Research in Computing for Humanities at the University of Kentucky and Harvard's Center for Hellenic Studies have led to the development of several projects that seek to take advantage of CTS to provide comprehensive access to classical literary collections. Among these projects, some focus only on the textual aspects of the source materials, and the project content consists only of XML files (encoded according to the TEI Guidelines). Other projects are more elaborate and seek to link source texts to the physical artifacts which contain them, and these projects consist of TEI-XML files and digital image files.

CTS provides the means for organizing, referencing, and querying classical texts. Developed by classicists Chris Blackwell (Furman University) and Neel Smith (College of the Holy Cross), the aim of the Classical Text Services protocol is to define a network service enabling use of a distributed collection of texts according to notions that are traditional among classicists. The CTS adopts and extends the hierarchical scheme of bibliographic entities defined by the OCLC's and IFLA's Functional Requirements for Bibliographic Records, or FRBR [http://www.oclc.org/research/projects/frbr/default.htm].

FRBR describes bibliographic records in terms of a hierarchy of Works, each of which is realized through one or more Expressions, realized in turn through one or more Manifestations, realized through one or more Items. CTS implements this hierarchy using the traditional terms Work, Edition or Translation, and Exemplar, while extending the hierarchy upwards, grouping Works under a notional entity called "TextGroup'" (corresponding to authors, in the case of literary texts, or any other traditional and useful corpus, such as "Attica" for inscriptions, or "Berlin" for a published corpus of papyri). CTS also extends FRBR's hierarchy downwards, allowing identification and abstraction of citeable chunks of text (Homer, *Iliad* Book 1, Line 123), or ranges of citeable chunks (Hom.~ *Il.* 1.123-2.22). The CTS protocol allows sharing of information about texts at any level of the conceptual hierarchy, and allows retrieval of sections of an identified text at any hierarchical level supported by its scheme of citation.

In this session, we will describe the Classical Text Services protocol and explain how editors can use it to organize and query texts. We will also explain how we use CTS in one text-focused project, the Neo-Latin Colloquia project, and how we are expanding the CTS for use in the image-based Venetus A project. We end our session with a demonstration of the CTS Implementation Tool (CTS-IT), and introduce the prototype of the Network Tool for Collaborative Electronic Editing over the Internet (NeT-CEE), a tool that builds on CTS to provide support for large-scale editing projects requiring the special talents of geographically distributed individuals.

Creating a CTS Text Collection: The Neo-Latin Colloquia Project

William DU CASSE, Ross SCAIFE
University of Kentucky

Scholastic *colloquia* are didactic works from the 15th and 16th centuries designed to teach Latin to younger students through interactive listening and speaking. Accordingly, many of the *colloquia* deal with

the everyday life of schoolboys, enabling them to speak in ordinary situations using proper Latin. These dialogues, generally short and written with many idiomatic constructions, were simple enough for beginners who already possessed a basic knowledge of Latin grammar. Mastery of *sermo quotidianus* ("daily conversation") was and remains an excellent means of reaching the stage where one thinks directly in Latin rather than translating from a native language. Moreover, students who learn idiomatic Latin then read classical texts with greater facility, and their own written style improves. Writers such as Petrus Mosellanus (early sixteenth century) assert that *colloquia* offered the means for the learner to master the cultivated but familiar speech found in Cicero's letters or Terence's plays, but applied to subject matter and thoughts never treated by Cicero or Terence. Erasmus thought his contemporaries could learn the best Latin style for contemporary use by reading the best authors, but also imitating the Latin conversation of those who "spoke just as the best authors wrote."

Colloquia scholastica form a genre of largely unexplored texts that reveal much about the pedagogical practice that supported the continuing use of Latin as Europe's universal language for the educated into the Early Modern era, when it was no longer anyone's native tongue. The perpetuation of a stable language ("dead" in the argot of the linguists) based on texts only, supported by no vernacular usage, for so many centuries represents a significant and potentially illuminating linguistic phenomenon. The *colloquia* offer an untapped source from which we can learn more about the history of pedagogy during the rise of western humanism. In addition, the *colloquia* provide plentiful insights into social history. Designed to promote the use of spoken Latin for discussion of daily affairs, they reveal a great deal about the conditions and customs of life at the time when they were produced, especially about scholars, teachers, and students in and out of schools or universities, but also about conditions among the citizens, merchants, and tradesmen at large.

Over the last three years, a group at the Stoa Consortium has begun to assemble the most comprehensive collection of neo-Latin *colloquia* availably anywhere in the world, in any medium. So far we have imposed the structural markup prescribed by the Text Encoding Initiative [TEI] in its "Base Tag Set for Drama" [http://www.tei-c.org/

release/doc/tei-p5-doc/html/DR.html] on over 650 *colloquia* containing over 620,000 words of text. We continue to build the collection through the addition of complex TEI markup, and we have enabled preliminary access to the collection online using the Classical Text Services protocol [CTS].

CTS allows us to access the Colloquia Collection on two levels. First, using CTS we can create a citation scheme specific to each set of *colloquia*, based on the organization of the dialogues themselves, and the internal structure of the individual dialogues. This structure is based mainly on the "Base Tag Set for Drama" as described in the TEI P5 Guidelines, so the typical citation scheme would follow the structure of a <div> for each individual *colloquia*, containing a <sp> for each speaker in turn, containing a <p> for the spoken text. CTS thus provides us with a way to easily cite (and thus to link to) any specific point in the *colloquia* texts.

In addition, CTS also enables us to create citations for the quotations (mainly from classical authors) that appear regularly in the *colloquia*. *Colloquia* originally served as a bridge to canonical literature. Including markup to identify references, both those specifically made by the *colloquia* authors, and those identified within the text by the modern editors, will illuminate how the *colloquia* authors used the classical texts, and which classical authors and works were most influential for early modern pedagogy. Encoding these bibliographical references will also enable us in many cases to point to full texts in the Perseus Digital Library [www.perseus.tufts.edu], where users may peruse the full context of the quotations or allusions. Using the TEI markup for Bibliographic Citations and References, we identify quotations and non-quotation references that the *colloquia* make to biblical and classical texts. For example, the author Pontanus quotes and refers to classical authors (and more recent ones) after each colloquium in his *Annotationes*, a section of notes following each colloquium.

In addition to bibliographic citations, we are explicitly marking **references to specific dates, people, and place names**. These references may be to historical events or people, or to events contemporary with the writing of the *colloquia*. Using the TEI tagset for Names and Dates [http://www.tei-c.org/release/doc/tei-p5-doc/html/ND.html], we mark instances where the *colloquia* name

specific people, places, dates, and events. In the short example below, the author Vives names both a specific person and a place. TEI not only enables us to mark names as they appear in the text, but also include a regularized version to simplify both searching and reading of the text (<orig> for the original version, <reg> for the regularized version).

<sp>

   <speaker>Mag.</speaker>

   <p>Ubi fecisti Latinae linguae tyrocinium? Nam non videris mihi prave institutus.</p>

</sp>

<sp>

   <speaker>Nep.</speaker>

   <p> <placeName> <choice><orig>Brugis </orig><reg>Bruges</reg></choice></placeName>, sub <persName> <choice><orig>Joanne Theodoro Nervio</orig><reg>Johannes Theodorus Nervius </reg></choice> </persName>.</p>

</sp>

<sp>

   <speaker>Mag.</speaker>

   <p>Viro diligenti, docto, probo. <placeName> <choice><orig>Brugae </orig><reg>Bruges </reg></choice></placeName> elegantissimae: nisi quod pereunt in dies vitio plebis profusissimae, dolendum </p>

</sp>

(*Colloquium* 7, in Ludovicus Vives, *Linguae Latinae Exercitatio Joan. Lodov. Vivis Valentini: libellus valde doctus, et elegans, nuncque primum in lucem editus, una cum rerum, et verborum memorabilium diligentissimo indice*. Parisiis: apud Joannem Foncher et Vivantium Gaultherot, 1538, p. 298.)

In this presentation, we will discuss the reasons for building CTS support into our edition, and give examples of the encoding we use to open up the *colloquia* to querying and organization through CTS.

Using CTS for Image-Based Electronic Editions: The Venetus A Project

Jack MITCHELL

Stanford University and Dorothy Porter, University of Kentucky

The Classical Text Services protocol [CTS] enables the integration of related texts – primary text transcriptions, different editions, translations, derived works, and annotations of any of these. The MultiText Homer project (MTH) is a difficult editing project because it seeks to connect all these various types of texts relating to the *Iliad*, starting with the oldest (and, arguably, the most important) manuscript of the text, Biblioteca Nazionale Marciana, Venice, Venetus A.

In addition to the text of the *Iliad*, Venetus A includes numerous commentaries called *scholia*, which serve to describe particular aspects of the main text. The textual variants that are preserved in the scholarly commentary of the Venetus A allow us to recover some of the multiformity that was lost in the process of text fixation that gave final or near-final shape to these two monumental oral poems. Variants also give us valuable insights into the process of oral composition-in-performance. The Homer *scholia* help us reconstruct the broad diachronic dimensions inherent in the evolution of Homeric textual traditions. Different Homeric textual traditions may have been definitive at different historical moments, but no single Homeric textual version can be deemed definitive beyond its own historical context. The *scholia* to the Venetus A manuscript of the *Iliad* preserve a treasury of ancient variants and allow us the opportunity to consider many possible texts at many different stages of transmission. Through the *scholia*, we can recover both a more accurate and a more accessible picture of the fluidity inherent in the Homeric tradition, especially during the earliest stages of the text.

The *scholia*, though vital for building an understanding of the *Iliad*, have never been completely edited, or even transcribed. The reason for this may be obvious given a cursory glance at the manuscript: their appearance on the page is incredibly complex. See Figure 1 below for one small example that shows a representative section of Venetus A.

Figure 1: A slice from Venetus A, folio 111v, showing (from left to right): Marginal numbers in the Greek alpha-numeric system; marginal scholia (main collection); illuminated capital, showing the first letter in a new book; inter-marginal scholia; book title; main text of the Iliad, including interlinear notes.

The layers of text evident throughout Venetus A are perhaps best described by Thomas W. Allen in an article of 1898, the last paragraph of an article concerning the history of the creation of Venetus A [Allen]. We include it here in its entirety:

> To recapitulate the history of the MS. which we have now reconstructed; the sheets, numbered and ruled, were given out to be written. The scribe who received them wrote the text and the principal scholia in the places ruled to contain them; during the act of writing he made corrections from time to time both in the text and the scholia. This done, he apparently began the book again and wrote in the irregular space left between the scholia and the text, and between the lines of the text, other shorter scholia in a different type of hand. He took advantage of this opportunity to correct in an exhaustive manner the text he had written; he added and altered breathings, accents and apostrophes, added and corrected critical signs, and wrote above or in the inner margin corrections of words. The book, thus complete in substance, was given to the original scribe who had numbered the quires and ruled the lines; he compared it throughout with the archetype and noted on the edge of the page differences; sometimes he accompanied these with a mark to call attention; he added lines left out, and omitted scholia either in the ruled margin or the intermediate space. In a few places he explicitly refers to his authority to defend himself from corrections already made in the text by, as it would seem, the first hand on his second round. Lastly, a third person reviewed in detail the suggestions of the reviser; deleted a great number of them in favour of the reading in the text, and in other cases substituted a correction of his own. He added likewise omitted scholia and remarks of a general nature upon the context. This excessive carefulness in the preparation of the book is further seen in the numbering of the similes, the quantitative marks, and the supplements of the elisions.

For a successful edition, we need software and encoding support that will enable the editors to separate out and edit these multiple layers of text, and changes made to the manuscript over time, while specifying the relationships among those layers.

To ensure that the TEI-XML files are accessible to CTS querying, we need to ensure that the various manuscript texts (main text and *scholia*) are encoded clearly, while at the same time maintaining links among the texts and between the texts and the manuscript images. We are currently storing main text and *scholia* for each of the 24 books of the Iliad, plus introduction, in separate XML files – 50 files in all. The various *scholia* are encoded as separate divisions, one division each for marginal *scholia* (Am), intermarginal *scholia* (Aim), interior marginal *scholia* (Aint), interlinear *scholia* (Ail), book subscription (Asub), and book metrical introduction (Amet) – six divisions total. Within the divisions, groups of *scholia* that comment upon or provide alternate readings for a single lemma are grouped together as numbered segments alongside the lemma. The following example shows the encoding for the first set of marginal *scholia* in Book 4:

```
<div type="marginal" n="Am.4">

    <p n="1">

        <seg type="lemma">Lemma for first marginal
        scholia, Book 4</seg>
```

<seg type="schol" n="1">First marginal scholion, Book 4</seg>

<seg type="schol" n="2">Second marginal scholion, Book 4</seg>

</p>

The "n" attributes provide the means by which CTS can access the *scholia* texts and build citations either from the lowest level – all the marginal *scholia* (Am – all marginal *scholia*)– to within the individual *scholion* (Am.4.1.0 – the lemma for the first marginal *scholia* of Book 4).

In addition to the XML files we have 641 image files, one for each side of the manuscript folios that survive and contain text. We are currently using the File Section and Structural Map functions of the Metadata Encoding and Transmission Standard [METS] to build indices that associate image files to the corresponding areas in the main text file, and that will associate areas of the image files with the corresponding TEI-encoded *scholia*. For this presentation, we will describe the complex textual and physical organization of the Venetus A manuscript including display of representative folios containing all of the six types of *scholia*. We will illustrate how CTS along with TEI and METS provides a robust scheme for organizing and accessing a complex image-based editing project.

Tools for Building CTS Projects: CTS-IT and NeT-CEE

Jerzy W. JAROMCZYK, Neal MOORE
University of Kentucky

For building both text editions and image-based editions, we have extended the CTS reference implementation [http://chs75.harvard.edu/projects/diginc/code/ctswebapp] to create the CTS Implementation Tool (CTS-IT). CTS-IT works with TEI files stored in an eXist XML database, providing a user-friendly interface through which an editor can upload TEI-encoded files, sort the files into groups and build a citation scheme to enable CTS querying of the data. As described in the CTS specification [http://chs75.harvard.edu/projects/diginc/specs/cts], information about the edition files is stored in the CTS Text Inventory file (TextInventory.xml), which provides an index of the files housed in the server, along with important metadata about those files.

The CTS-IT Upload function incorporates CTS metadata into the eXist upload capabilities, creating the relevant sections of the Text Inventory when storing documents in the database. CTS-IT uses the eXist API, the CTS library [http://chs75.harvard.edu/projects/diginc/code/ctslib], and the file upload functionality from Apache Commons. Functionality includes:

· Create a new textgroup in eXist and assign the textgroup a unique ID. Both the textgroup name and ID can be updated (changed) later.

· Add new texts to the textgroup

· Separate IDs for editions and translations

· Automatically validates the TEI files against DTDs or Schemas that are stored online, if they are declared in the file.

· Automatically names the uploaded file as an online node, named according to : textgroupID_projectName_projectID

· During upload, given input from the editor, CTS-IT assigns a citation scheme based on the structure of the individual files.

· Cross-check – make sure that there is a file for everything in the Text Inventory, and that every listing in the Inventory has a match in the database.

· Set eXist as a block inside the Cocoon publishing framework, so we can build on-the-fly pages based on the information stored in the database.

However, the CTS-IT provides only basic support for text-only editions. Though it is a good start, we have started development of a new tool, the Network Tool for Collaborative Electronic Editing over the Internet (NeT-CEE), which will vastly expand the functionality of CTS, allowing an editor to relate XML text files with digital images of the physical objects on which the texts are founded. NeT-CEE will allow scholars to build editions encompassing text, images, and annotations using the Extensible Markup Language (XML), the de facto standard for encoding electronic editions in the

humanities, complying with the Text Encoding Initiative (TEI) standards for markup. NeT-CEE will also support overlapping XML markup, which will occur when (for example) an edition includes markup to describe a word in a text, and that word on the page is broken between two lines.

NeT-CEE will be oriented towards collaborative electronic edition projects that bring together a number of scholars with diverse skills and interests. NeT-CEE will implement a distributed editing framework, with access control and version management systems which will allow several different editors to collaborate on an edition with different levels of access, and without fear that one editor might inadvertently overwrite another's work. Finally, since NeT-CEE will be accessible through a regular web browser it will encourage collaborative work among individuals who are geographically dispersed, and may encourage electronic editing by those many accomplished humanities scholars who are familiar with a browser view but who may be put off by regular XML editing software.Software that enables image-based collaborative editing will be applicable to countless manuscripts and papyri that have their own complex textual organizations. We will provide the means to cite not only the primary text of a document, but also the array of marginal notes and annotations that accompany it (as in Venetus A). Likewise, manuscripts of Euclid and Plato include not only marginal commentaries but tables and figures to which we wish to provide access. In addition, many historical scientific texts are already available though the CTS protocol and collaborative image-based editions could be compiled by importing new manuscript images into the existing CTS compliant texts.

We anticipate that NeT-CEE will foster the creation of scholarly works by forging partnerships between individuals and institutions, enabling them to share resources, both physical resources (in the form of texts and images) and intellectual (in the form of subject knowledge and editing experience). Because we will release NeT-CEE under an Open-Source license, it will especially promote cooperation among smaller institutions that might not have the resources to purchase expensive software. NeT-CEE will be a significant resource for scholars, but also for teachers and students, potentially encouraging collaborative projects between K-12 schools in different regions of the United States (or, indeed, around the World).

In this presentation, we will demonstrate the functionality of the CTS-IT and the prototype NeT-CEE.

[Allen] T. W. Allen. "On the Composition of Some Greek Manuscripts." Journal of Philology, vol. XXVI, pp. 161-181, 1898.

[CTS] Classical Text Services Protocol. <http://chs75.harvard.edu/projects/diginc/techpub/cts>

[METS] Metadata Encoding and Transmission Standard <http://www.loc.gov/standards/mets/>

[TEI] TEI P5 Guidelines for Electronic Text Encoding and Interchange, edited by C.M. Sperberg-McQueen and Lou Burnard. Revised and re-edited by Syd Bauman and Lou Burnard. January 2005.