DHQ: Digital Humanities Quarterly

Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure
2009
Volume 3 Number 1

Citation in Classical Studies

Neel Smith <dnsmith_dot_neel_at_gmail_dot_com>, College of the Holy Cross

Abstract

Citation practice reflects a model of a scholarly domain. This paper first considers traditional citation practice in the humanities as a description of our subjects of study. It then describes work at the Center for Hellenic Studies on an architecture for digital scholarship that is explicitly based on this model, and proposes a machine-actionable but technologically independent notation for citing texts, the Canonical Text Services URN.

Overview

For the past ten years, the Stoa Consortium has provided a hothouse for projects aimed at developing appropriate digital forms for scholarly work.[1] This endeavor is sometimes described as “translating” traditional print work into digital formats, but the metaphor of “translation” mischaracterizes the Stoa's accomplishment since it suggests that the print form is an established ideal to be more or less effectively replicated in a new medium. The work that the Stoa Consortium has supported is both more difficult and more profoundly significant: it asks first whether we can model our scholarly work *independently* of any particular technological infrastructure; and then, in second place, whether we can realize that model in a digital implementation. As a very small token in recognition of Ross Scaife's work on the Stoa, I would like to summarize some key results of a working group I have led at the Center for Hellenic Studies addressing precisely these problems as they apply to scholarly citation.[2] While my illustrations are drawn from the study of the Greek and Roman world, the following conclusions apply to the humanities more broadly: 1

- Citation is a form of ontology: how we cite the objects we study identifies and describes the material of our domain independently of any technology.
- We work most often either with two types of objects: discrete objects cited by unique identifiers, and texts in a notional hierarchy similar to the Functional Requirements for Bibliographic Records
- In a networked digital library, citation in a machine-actionable syntax with defined semantics underlies an architecture of services.

These conclusions have little meaning if they are not worked out in a real implementation, so I will describe in some detail the design decisions we have made in the creating a digital system for scholarly citation at the Center for Hellenic Studies. Since the range of technologies may include topics unfamiliar to some DHQ readers, I append a glossary of technical terms at the end of the paper.

Changing Technologies and the Fate of Homer's Commentators

A survey of information technology projects in the humanities over the past quarter of a century would vividly illustrate how difficult it is to conceptualize our work without simply following the rut that some technology has worn for us. I will leave it to my readers to select their own favorite (or least favorite) examples of projects conceived in the narrow terms of a particular technology: whether projects that unthinkingly reproduce the visual appearance of print publications in digital form (“shovel ware”), or projects that subordinate scholarly work to the requirements of an inadequate and rigid digital format. Instead, I would point out that the dangers of adapting scholarly work to an alternative information 2 environment are not new. One telling example from he history of classical scholarship is the changing treatment of scholarship on the Homeric poems.

In the Hellenistic and early Roman periods, the most learned literary critics of the ancient world — scholars like Aristarchus of Samos — composed extensive commentaries on the Homeric poems as free-standing works. Lemmata, or catch words, introducing each note provided an explicit link to the texts they commented on, but commentaries in the form of scrolls must have been awkward to use, requiring the reader to advance both the text and commentary to the appropriate passage. 3

Small wonder that in the transition from the scroll or volumen to the book form of the codex, later copyists saw an opportunity to create a much more convenient reading environment. Not only could the sheets of the codex be more easily turned to a particular passage: the wide areas frequently left as margins of each page in the codex were an inviting space for copying related commentary. By the time of the earliest surviving manuscript copy of the *Iliad*, we see a rich hypertext with multiple kinds of scholarly material surrounding the Homeric text.[3] 4

This improved reading environment came at a high price, however: the independent manuscripts of Aristarchus and others ceased to be copied. The individual comments, or scholia, that were placed in a given codex might only be a selection of an original commentary, and were indistinguishably merged with selections from other commentators. The form of the marginal scholia created a new experience of reading while ensuring the loss of much of the material the scholia were based on.[4] 5

Recent print publications of the scholia have compounded the problem. The potential of the printed edition is to make the contents of the unique manuscripts accessible to readers who are unable to view the manuscripts themselves. To accomplish this, the two works often considered standard references for the Homeric scholia to the *Iliad* today adopt a similar approach.[5] The editors select scholia from multiple manuscripts, and group them by line of the *Iliad*. Like the medieval copyist, they have created a new instrument that is more convenient for the particular task of line-by-line reading: as the medieval manuscript obviated the need to juxtapose different manuscripts, so the modern edition obviates the need to juxtapose print publications of scholia from different manuscripts. But the set of scholia in a given manuscript constitutes a unique commentary in its own right; to understand them, we need to be able read them together, as a group, and preferably in the context of the surrounding material assembled in the manuscript, such as the specific version of the text they comment on. This is impossible with the new print collections: the evidence of the individual manuscripts has disappeared from our print record just as the evidence of the Hellenistic commentaries disappeared from the medieval manuscript tradition.[6] 6

In developing new scholarly instruments for the codex manuscript of Homer, and for the printed edition of the Homeric scholia, innovators recognized how different technologies offered the possibility for different forms of reading. Both the manuscript adorned with scholia and the printed line-by-line collection of scholia privilege a single form of reading, however, and efface essential features of the sources they draw on. This was not due to any hostility towards their source material. The anonymous scribes who read the ancient commentators on Homer revered them, and cite them as authorities; Erbse and van Thiel are equally scrupulous in citing the specific manuscripts they have consulted. But in spite of their obvious, and self-professed, respect for their sources, both the medieval copyist and modern editors chose to work in a form that threatened the survival of the sources they admired. 7

Citation as a Heuristic

The scholiasts' citations of authorities and the modern editors' citations of manuscripts describe a model of study grounded in sources. The gap between this model and the effects of their work suggests that in the struggle to fashion a new and improved scholarly tool, they lost sight of some of the most central objects of their study. Today, we routinely confront examples of how new technologies can effect new forms of reading — that is, more generically, new ways of discovering, visualizing and manipulating information — and we risk falling into similar pitfalls. If we want to pursue the implicit agenda of the Stoa Consortium and define a technologically independent model of scholarly work, then considering our citation practice can serve as a practical first step. *What* we cite gives us an inventory of the objects we 8 study; *how* we cite an object suggests how we view its structure. Taken together, this probably comes as close to an abstract description of a given subject area or domain as any aspect of scholarly discourse in the humanities. Citation is a form of ontology.

Of course, in most scholarly publications in Classics, an enormous proportion of citations are to other scholarly publications. But scholarly publications are representations of an argument. They in turn depend ultimately on references to more fundamental objects of study. While the length of this chain of dependencies may obscure the relation of argument to source material (and while part of the appeal of digital publications is certainly the possibility of following such a series of dependencies automatically to its ultimate sources), if we want to use citation practice as a guide to modelling the objects we study, we will begin with the most elemental objects, on which others depend. 9

The following section of this paper begins with *what* we cite. I describe a simple model for identifying fundamental objects classicists work with, and how the Center for Hellenic Studies' Technical Working Group has attempted to implement that model. The next section deals with *how* we cite that material, and discusses the technical design of citation schemes satisfying our model. In the final section of the paper, I briefly discuss the relation of a digital citation system to the design of a networked architecture for humanistic scholarship. 10

Identification: What We Cite

Discrete objects. A citation must, before anything else, *identify* some object we are studying. For many categories of material, classicists or other humanists study are concerned with sets of discrete objects. Physical artifacts such as coins, manuscripts, or stone sculptures are clear examples: each object is physically distinct, and we need to be able to identify each physical object uniquely. But historically attested phenomena can equally clearly constitute sets of distinct objects: in a prosopography of Athenian citizens, every individual is unique and needs to be uniquely identified, although many persons may share the same name.[7] For any unique discrete object, our model of citation practice will require a means of ensuring that objects can be consistently and uniquely identified. 11

Within a single project citing a set of discrete objects, it should be straightforward to implement a requirement that objects have unique identifiers: relational database systems can enforce constraints for unique values of a field, XML documents can be designed to require unique identifying attributes on elements, etc. But how do we further guarantee that unambiguous identifiers will not conflict when they are disseminated across the internet? 12

This is directly analogous to a problem that the XML community faced a decade ago. When data from different schemas are commingled on the internet, how can automated systems determine the data structure that an element belongs to? If elements from a Dublin Core document, an XHTML document and a TEI document appear together, for example, how can we disambiguate an element named title that could belong to any of the three types of document? 13

One conceivable solution might have been to establish a centralized registry of data structures (schemas, DTDs), but that would have enormously burdened developers who frequently need to create or modify data structures with new DTDs or schemas. Instead, the World Wide Web Consortium defined XML namespaces as an easy and flexible way of qualifying XML structures with unambiguous references.[8] XML namespaces are defined in terms of the pre-existing system of Uniform Resource Identifiers (or URIs). URIs in turn piggy back on the internet's domain name system (discussed further below) that manages globally unique names. The result is that an XML element or attribute can be qualified by a namespace identifier that is guaranteed to be unique. A title element qualified by the namespace identifier http://purl.org/dc/elements/1.1/ unambiguously belongs to the Dublin Core Metadata structure, for example. 14

The Center for Hellenic Studies (CHS) Technical Working Group has followed the same reasoning to qualify references to specific objects (as opposed to data structures) with what we are calling “domain namespace identifiers,” or DNIDs.[9] Just as XML namespaces use URIs to qualify a reference to a schema, a “domain namespace identifier” can qualify an object identifier to guarantee that it will be globally unique. Since the CHS owns the domain name chs.harvard.edu, a CHS developer can build on this string to define namespace identifiers to refer uniquely to a set of objects. A data namespace like chs.harvard.edu/datans/images could be used to refer to a set of digital images, for example.[10] 15

Internally, the CHS can use any technology it chooses to guarantee that images are uniquely identified within this data set, and then associate the domain namespace identifier with a specific object's identifier when publishing the reference to the internet to ensure that it will be globally unique.[11]

16 **Hierarchical texts**. The ways scholars refer to the texts they read suggests a more complex, hierarchical identity than simple discrete objects. Our model of identifying texts needs to range from references to a poem like the *Iliad* as a conceptual work encompassing any version of the poem, to particular editions or translations, to even a single physical exemplar of a work. The library community has developed an abstract model known as the Functional Requirements for Bibliographic Records (FRBR) that describes the hierarchical nature of texts viewed as bibliographic entries, and in significant ways resembles the hierarchical identity seen in scholarly reference to texts.[12] The FRBR model posits four levels of hierarchy for cataloged works:[13]

FRBR term	Definition
work	“a distinct intellectual or artistic creation” (e.g., the *Iliad* as a conceptual entity)
expression	“the intellectual or artistic realization of a work” (e.g., a particular edition or translation of the *Iliad*)
manifestation	“the physical embodiment of an expression of a work” (e.g., a particular printing of a particular edition of the *Iliad*)
item	“a single exemplar of a manifestation” (e.g., a particular physical copy of a particular text of the *Iliad*)

Table 1.

FRBR provides a useful model that is largely congruent but not perfectly aligned with the ways classicists and other humanists cite texts. At the top of the hierarchy, classicists prefer to cite works as members of a larger group. These groupings are a taxonomical convenience, without any unifying semantics: a group might refer to authorship (“the works of Sophocles”), provenance (“the inscriptions from Aphrodisias”), or may represent a conventional grouping the semantics of which might even be disputed. “The Homeric poems,” for example, might be viewed by some as a category of author, and by others as a generic category. As purely conventional categories, the traditional groupings do not conflict with the librarians' proper concern to separate any kind of subject cataloging from the identification of a work. At the same time, the conventional groups provide a context for citing texts that classicists have found useful, and that permeates classical scholarship, so we will want to retain them in our model for identifying texts. Within these groups, the notional works classicists refer to correspond precisely to the notional works of the FRBR model.

17 In the center of the FRBR model, the distinction between expression and manifestation is not evident in classicists' citation practice. The difference between expression and manifestation matters to librarians responsible for physical holdings in a collection. Scholars citing texts focus instead on their semantic content. For their purposes, if two manifestations are so different that we need to distinguish them, they may as well be considered a new edition or translation.

18 Individual physical copies, on the other hand, may matter to the scholar as well as to the librarian, because finally the evidence for a given version rests in real, physical exemplars. If the only available or known exemplars are imperfectly preserved, scholarly citation may need to distinguish between evidence for a version from one copy versus another.

19 The following table summarizes the differences between the FRBR model and the model of texts suggested by our citation practice.

FRBR model	Classicists' citation practice
---	Grouping of texts (e.g., the Homeric poems)
work	= (notional) work (e.g., the *Iliad*)
expression	~ version (e.g., a citable edition or translation of the *Iliad* in any manifestation)
manifestation	---
item	= exemplar (e.g., a particular physical copy of a particular text of the *Iliad*)

Table 2.

While we want to allow scholars to refer easily to the wide range of objects we identify with simple unique identifiers, our corpus of primary texts is comparatively stable. For this body of material, the overhead of a more centrally coordinated registry system would be worth considering

The internet's domain name system (DNS) provides an example of how a system of hierarchical unique identifiers can function across a global network, and closely parallels what we need in citing primary texts. Key points are 20

- a tightly controlled set of top-level servers keep track of who has been granted control of some particular domain
- registrars of particular domains can add further extensions to the hierarchy
- a simple notation with dot-separated elements identifies objects in the hierarchy

So in DNS, a top-level server for .org addresses tracks the assignment of names like stoa.org to organizations like the Stoa Consortium. The registrar of the Stoa consortium can extend the hierarchy, and these entries are represented with extended names like www.stoa.org.

We want a system that would assign responsibility for registering identifiers for texts to appropriate organizations or projects, analogous to the internet's top-level domains. The Stoa Consortium, for example, might be a logical registrar for works of Latin literature, while the Aphrodisias project would be an obvious choice as registrar for the inscriptions from Aphrodisias. Within these domains, registrars could extend identifiers to any hierarchical level of a text's identification, from text group down to individual exemplar. In the following section, I will further develop a unified notational scheme for citing works and passages within works, but at this stage, note that we want a simple textual notation that can be extended to each hierarchical level of a text's identification. 21

As with DNS, this kind of delegated system of authority requires consensus among all participants. In a field like classics, “participants” really means the active projects that are disseminating digital texts — a comparatively small community, and probably a less fractious setting for trying to develop consensus than traditional professional organizations. The slow, hard work of building a consensus would be more than repaid when unambiguous hierarchical identifiers could be used for frequently cited texts. I will return in the following section to some potential rewards of a coordinated registry system for referring to texts. 22

How We Cite Objects

Simple vs. continuous citation. Whether we identify an object by a unique ID qualified by a data namespace, or by a hierarchically structured identifier of the kind we use for texts, how we choose to cite the identified object reflects our understanding of its structure. In the simplest case, scholars' citation may do no more than identify the object. In fact, this simple form of citation is extremely common when we refer to discrete objects that can be identified with unique identifiers. A numismatics publication, for example, might cite a coin using nothing more than a unique identifier, even when the discussion supported by that citation alludes to specific properties such as the weight of the coin or the location of the mint where it was struck. Is this minimal form of citation due to scholarly laziness? I would argue to the contrary that this can in fact be an advantageously light-weight form of citation. Different scholars will model any given type of object with different sets of properties corresponding to their particular research interests. For a study of mint 23 practices, the die axis of a coin and information about die linkages might be vital information; those might be irrelevant to a prosopographic study of names appearing in coin legends. By citing the coins simply as discrete objects, the numismatist and historian make it possible to recognize the identity of the objects they choose to model in radically different ways. Data models must correspond to specific research concerns: simple citation forms can serve to integrate different ways of modelling the same object.

24 In contrast to objects that are cited by simple identifiers, a handful of object types may be more precisely identified by citation pointing to some part of the object. Two familiar examples are geographic objects and images. In printed works, these may be "cited" with a visualization as a map or illustration, where a digital citation could more generically refer to coordinates within a continuous reference system (which could be visualized as a map or illustration). A geographic object such as a city may have a unique identifier in a collection of locations, but a reference in a geographic coordinate system could point to a particular section of that entity.[14] I am not aware of a standard convention for citing sections of images, but, like geographic objects, a digital image with a unique identifier could be cited with coordinate reference pointing to part of the image. At the Center for Hellenic Studies, we have used simple rectangular coordinates (top, left, bottom, right) to refer to a "region of interest" in a digital image. These coordinates are expressed in percentage units so that the citation is easily applied regardless of the scale of reproduction of the image. Clipping the cited section or highlighting it on the full image with alpha compositing are possible displays in a digital environment.

25 *Citation of texts*. Text citations combine identification in the FRBR-like hierarchy discussed above with a reference expressed in a different kind of continuous "coordinate system." Other humanists have fallen into the trap of citing texts by the accidental physical unit of the page, with the unfortunate consequence that the citation is valid only for a specific manifestation of the work, since references to pages cannot be applied to other printings, much less to other expressions such as translations into different languages. Classicists, together with scholars in Biblical studies, have generally recognized the importance of a *logical* citation scheme. "Book 2, chapter 5" of Thucydides has the same meaning no matter whether it is applied to the notional work, or a specific item.

26 In a classic article provocatively entitled "What is text, really?", DeRose, Durand, Mylonas and Renear suggested in 1990 that the best model for a representing a text digitally is an ordered hierarchy of content objects (OHCO).[15] Their suggestion was based on their extensive experience in the scholarly mark up of texts, especially within the Text Encoding Initiative, or TEI.[16] The TEI's markup guidelines had focused on many of the same kinds of content that tend to be used in our citation schemes (chapters and sections of prose works, or stanzas and lines of particular verse forms), so that it comes as no surprise that their OHCO so accurately describes the "coordinate systems" classicists use in the canonical citation of texts.

27 As early as 1993, however, Renear, Mylonas and Durand had backed away from the universality of their initial claim:[17] there is no *unique* logical hierarchy; the analytical perspective of the scholar might dictate different, overlapping or incompatible hierarchies. The editor of electronic texts therefore will have to apply different markup schemes for different purposes.

28 The 1993 revision to their original OHCO model contains important insights, but from the perspective of scholarly citation, the original OHCO thesis describes precisely how we cite texts. There is a single logical hierarchy for citation, and when we are interested in features of a text that are not aligned with the units of the citation scheme, we must nevertheless identify those features in terms of our citation scheme.

29 Renear et al. illustrate overlapping hierarchies with the example of sentences or speeches in a poetic work organized by metrical lines: the linguistic unit and the prosodic may not align, and the editor of the electronic text will be forced to privilege one hierarchy over the other, or devise a strategy for handling concurrent, overlapping hierarchies. Yet when classical scholars cite works by metrical line, they will not invent a new citation scheme to refer to a speech: they will express the citation in terms of lines no matter what feature they are analyzing. In book 6 of the *Iliad*, the Greek hero Diomedes addresses his opponent in battle, Glaukos, with the question, "Who are you? I don't think I have ever seen you on the field of battle before." The question begins with line 123, and wraps onto the first metrical foot of line 125. The speech unit and the metrical unit are incompatible, but for most purposes, this is of no consequence: the line is a

sufficiently precise pointer that we will simply refer to Diomedes' question with a reference like *Iliad* 6.123-6.125.

This is directly comparable to the approximation we accept when we cite other objects with a simple identifier: use of a common citation system allows us to cite without having to agree in detail on the underlying data model, since the data model will be dictated by the perspective of the individual scholar. For different purposes, we might view a coin as having one set of properties (die axis, weight) or another (attested personal names), but we can recognize a citation as referring to the same coin, regardless of the data model applied to it. Just so we might view a text as having one logical structure (syntactic unit) or another (prosodic unit), but we can recognize a span of lines in the *Iliad* as referring to the same passage without regard to the editor's views on the structure of the text. As scholarship increasingly relies on automated discovery of citations in digital material scattered across a global network, conventions keeping citation of familiar types of objects independent of assumptions about how those objects will be represented will become more and more valuable. 30

Texts and URNs. The hierarchical identification of texts along the lines of the FRBR model is becoming widespread in the library community, and scholars editing electronic texts have discussed the OHCO model for a decade and a half. It may seem surprising that the relevance of these models for citing digital texts has not been recognized, because for centuries classicists' canonical practice has been to identify texts in a notional hierarchy, and cite them in units of an ordered hierarchy of content objects. 31

Perhaps the complexity of citation expressed in prose has obscured the fact that canonical citations marry two hierarchies, one identifying the object, one describing its logical “coordinate system” for purposes of citation. Certainly, the variety in natural-language expressions for these ideas is an obstacle to machine recognition and action on our references. 32

To express these canonical references to texts concisely and unambiguously, the CHS Technical Working Group has defined a notation for canonical text citation. We chose to express these citations as Uniform Resource Names (URNs). URNs are “persistent, location-independent, resource identifiers” — precisely what a citation should be. The syntax of URNs (defined in RFC 2141) is “designed to make it easy to map other namespaces (which share the properties of URNs) into URN-space. Therefore, the URN syntax provides a means to encode character data in a form that can be sent in existing protocols, transcribed on most keyboards, etc.” [18] The syntax, too, is well adapted to the ways we use citations. Because we initially developed this notation for use in a network service called the Canonical Text Services, we refer to our URN notation as CTS URNs. 33

The values used to identify texts in CTS URN citations must be publicly documented, ideally in a DNS-like distributed registry system. In order to begin developing software using the CTS URN notation immediately, the CHS Technical Working Group is maintaining a hierarchical registry of identifiers for works of ancient Greek literature, and a registry of other CTS registries covering other domains (such as the Stoa Consortium for works of Latin literature). 34

Syntax of a CTS URN

URNs always begin with the string urn, followed by a protocol identifier. We propose the identifier cts for our protocol. CTS URNs are composed of up to four further top-level elements, separated by colons. They are 35

1. a CTS domain (required). This is a string identifying the “top level registry” where text identifiers can be looked up or resolved. We use the string greekLit to refer to the CHS registry of works of ancient Greek literature.
2. a work identifier (required): a FRBR-like hierarchical identifier, with individual components separated by dots; the entire string can be resolved in the CTS domain named in item 1. Where possible, the greekLit registry uses values derived from the Thesaurus Lingue Gracae's Canon; the Homer poems are tlg0012, and within that group, the *Iliad* is tlg001, so a reference to the Iliad as a notional work would be expressed as tlg0012.tlg001.
3. passage reference (optional; if absent, the URN refers to an entire work): a string referring to the canonical citation scheme of the cited work, with different levels of the citation hierarchy separated by dots. Line 123

> of book 6 of the Iliad would be represented as 6.123. Spans are indicated by starting and ending points, separated by a hyphen, so lines 123 to 125 of book 6 would be 6.123-6.125.

The general structure of a CTS URN is therefore `urn:cts:DOMAIN:WORK:PASSAGE?` and a full CTS URN citing lines 123-125 of Iliad 6 would be `urn:cts:greekLit:tlg0012.tlg001:6.123-6.125` Because the URN explicitly indicates the hierarchy of both the work and the citation, applications can choose to interpret the reference at whatever level they consider appropriate. A URN might refer to a specific English translation, but an application could ignore the more specific components of either the work or citation hierarchy to apply it to a Greek edition of the *Iliad*.

An example application. A simple example can illustrate some advantages of using a notation with specified syntax and semantics.[19] Google's Base application allows account holders to upload content associated with a given URL, and optionally including key words. Although Google clearly sees this as a way to expose sales catalogs, it could be used more generally to expose database contents to Google's search engine. 36

The CHS has a small registry of scholarly publications that it maintains on line, including Gregory Nagy's works *Best of the Achaeans* (urn:cts:chs:nagy.bofa:) and *Pindar's Homer* (urn:cts:chs:nagy.ph:). These are organized in a logical citation scheme of chapters and sections. I exported each section (the leaf node in the citation tree), and associated it with its proper URN. Searching Google base for a cluster of terms like “Pindar's Homer Nagy lyric” ranks as the top match a site offering *Pindar's Homer* for sale (and mentioning “lyric” in its abstract of the book). Searching for “urn:cts:chs:nagy.ph: lyric” finds only content matching that URN as well as the term “lyric”. Because the URN structure is hierarchical, searching for “urn:cts:chs:nagy. Achilles” finds all occurrences of Achilles in the textgroup “nagy” in the chs domain, ranked according to Google's search algorithm. Because the CTS URN expresses the semantics of hierarchical text citation in a simple flat string, we can use its precision when to limit results when searching Google Base for fuzzier terms. 37

More generally, the simple string of a CTS URN is well suited to passing around the internet as a precise form of machine-actionable citation. The links associated with the Google Base entries for the Nagy books, for example, pass the URN as a parameter to a text browsing application so that from a search of Google Base, a reader can pass directly to a continuous browser through the full text. 38

Beyond Citation: Architecture

Source citation is just one part of scholarly publication, and conventions for citing resources digitally must be viewed as part of a larger architectural design. I have previously argued that when the digital library is the global internet, the natural architecture for scholarly publications is a hierarchy of services.[20] This service tier can be further subdivided into its own hierarchy of higher-order services that depend on more fundamental services, the most basic of which are identification and retrieval. An analytical “diff” service, for example, describing differences between the same passage in two versions of a text, could be built on top of a service that retrieved passages by canonical reference. 39

Much of the energy of the CHS Technical Working Group has been focused on defining and implementing network services in support of scholarly applications, in part because we recognize the commonplace that while end-user applications are short-lived, thoughtfully designed services upon which end-user applications can be built can have much longer lives. I would suggest that we can extend our digital architecture one tier deeper to include a “reference tier” that is more fundamental than the service level. This relationship is summarized in the following table. 40

Level	Function	Technological design	Life expectancy
3. Application tier	End-user applications	Draw content from network services	Short life
2.(b) Service tier	Higher-order services for manipulation and analysis	APIs that depend in turn on standard retrieval APIs	Like retrieval APIs
2.(a) Service tier	Identification and retrieval	APIs that work in terms of immutable citations	Long life
1. Reference tier	Citation	Syntax and values fixed in machine-actionable form	Immutable

Table 3.

Because citations should be immutable — the concept “book 6, lines 123-125 of the *Iliad* ” should have a fixed meaning, and remain valid — it is important for us to design digital expressions that will be both immediately practical, and likely to remain sustainable for the indefinite future. Both the application of data namespaces to qualify unique identifiers and the use of CTS URNs for references to texts should satisfy those conditions.

Conclusions

To study, as humanists do, historically unique products of culture — works of literature, historical events, artifacts of material culture — is an extraordinarily complex undertaking. I suspect that for most humanists that complexity is part of the fascination of their work, a reflection of the richness that gives meaning to the object of their study. We are not merely comfortable with the complexity of our material, we revel in it. Most often we are less familiar with an idea that seems natural to the computer scientist: that any degree of complexity can be constructed from the composition of simpler elements. 41

Yet the way we cite sources suggests that at some level we intuit this. We tend to cite a provocatively large proportion of the material we study either as simple objects (expressible as a unique identifier qualified by a data namespace), or as a continuous reference within a hierarchically identified text (expressible as a CTS URN). It is worth considering how much of our work could be modelled in information systems that rest ultimately on foundations laid with these two simple shapes of building blocks. 42

Glossary of Technical Terms and Abbreviations

- **CTS:** Canonical Text Services. A network service developed at the Center for Hellenic Studies for identifying and retrieving texts by canonical reference. (See http://chs75.harvard.edu/projects/diginc/techpub/cts or http://katoptron.holycross.edu/cocoon/diginc/techpub/cts)
- **CTS URN:** Canonical Text Services URN. A notation with defined syntax and semantics for citing passages of text in a technology-independent but machine-actionable form. (See http://chs75.harvard.edu/projects/diginc/techpub/cts-urn or http://katoptron.holycross.edu/cocoon/diginc/techpub/cts-urn.)
- **DC metadata:** Dublin Core metadata. Basic metadata such as authorship, and summary of contents, expressed in the standards developed by the Dublin Core Metadata Initiative. (See http://dublincore.org/)
- **DNID:** Domain Namespace Identifier. A notation using internet domain names to guarantee a unique context for object identifiers. (See http://www.dnid-community.org/)
- **FRBR:** Functional Requirements for Bibliographic Records. A conceptual model for cataloging library resources. FRBR's “group 1 entities” describe a hierarchical model for texts, from a notional work down to a

single specific item. (Summary in Wikipedia with links to technical documents: http://en.wikipedia.org/wiki/FRBR.)

- **XML namespace:** A “simple method for qualifying element and attribute names used in Extensible Markup Language documents by associating them with namespaces identified by URI references.” (See http://www.w3.org/TR/REC-xml-names/.)

Notes

[1] This paper was initially submitted to honor ten years of Ross Scaife's work on the Stoa. It was, in that context, a meager sign of respect for a valued colleague. Now the final submission must be part of a collection in memoriam, where it cannot begin to stand for the loss to our field, much less to friends and family.

[2] It is an especially small token, since Ross himself contributed directly to the work at the Center for Hellenic Studies described here. I am grateful to the Center for Hellenic Studies, its director, Gregory Nagy, and director of publications, Lenny Muellner, for support of this work; among the many contributors listed at http://chs75.harvard.edu/projects/diginc/who, I would especially like to take this opportunity to thank Chris Blackwell and Gabe Weaver for their continuous and insufficiently recognized contributions to the CHS Technical Working Group, as well as for their help in clarifying my own thoughts on the topics addressed here.

[3] The Greek MS Z. 458 (= 841) in the Biblioteca Nazionale Marciana, in Venice. For an introduction with further references see [Nagy 2004, 4–24]; [Dickey 2007, 18–23].

[4] Only one such work survives from the Hellenistic period as an independent commentary today: the commentary by Hipparchus of Nicaea on the (now lost) *Phaenomena* of Eudoxus, and the similarly entitled poem of Aratus. It is possible that the very fact that Hipparchus' work offers a comparative commentary on two distinct works may have made it less amenable to abstracting in the form of scholia to one or the other *Phaenomena*, and so may have helped preserve for us today the only surviving work by the Hellenistic period's greatest astronomer.

[5] Erbse (eight printed volumes); van Thiel (current version available as pdf documents, described as a “Proecdosis”” or preliminary release, to the “scientific publication,” evidently conceived of as a print publication.

[6] Fortunately, we still have the manuscripts, and so long as they are preserved, can hope to rectify this absence from the printed record. Nineteenth-century scholars took a different view of publishing scholia, and in the case of the Homeric commentaries, established a series of publications of scholia from a given manuscript. While this still divorces the commentary from the specific version of the text it explicates, this scheme at least maintains the coherence of the collected comments in a given manuscript. Compare the edition of the scholia to the Venetus A by Dindorf (1875), with the recent comparative editions by Erbse and van Thiel.

[7] Note that *evidence* for a historical individual may be ambiguous, and we may be unable to determine if we have evidence for one or more individuals of the same name; but the simple point made here is that our underlying model of historical personnages insists that each individual is historically unique.

[8] For the 2006 second edition of “Namespaces in XML 1.0,” see http://www.w3.org/TR/REC-xml-names/, superseding the previous recommendation from 1999.

[9] For current information on and further discussion of DNIDs, see http://www.dnid-community.org/.

[10] Although defining XML namespaces directly in terms of URIs makes the XML namespace definition admirably simple, it does twist the sense of a URI a bit. URIs must begin with a protocol identifier such as http://. This has no meaning when the URI string is applied to an XML namespace for the sole purpose of ensuring its uniqueness. To disambiguate object identifiers with data namespaces, we think it is preferable to omit the protocol component of the URI string, e.g., chs.harvard.edu/datans/images rather than http://chs.harvard.edu/datans/images.

[11] In any XML document, XML namespaces can be declared and bound to an abbreviated prefixusing the xmlns prefix on attributes. Software and data structures relying on data namespaces to disambiguate object identifiers could similarly work with abbreviated aliases for lengthy DNIDs.

[12] The formal description of the model is available from http://www.ifla.org/VII/s13/frbr/frbr.pdf For current information about FRBR, and ongoing activity in the very active FRBR community, see the FRBR blog at http://www.frbr.org/.

[13] These are the “group 1 entities” of the FRBR model.

[14] Geographic objects occupy an important place in historically grounded disciplines of the humanities, but I will not discuss them further in this paper, because other projects are already effectively addressing the problems of citing and using geographic resources. I have found the work of the Open Geospatial Consortium especially valuable in thinking about citing and using networked resources: see their web site at http://www.opengeospatial.org/

[15] Full text available from http://doi.acm.org/10.1145/264842.264843/.

[16] http://www.tei-c.org/

[17] "Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies," available from http://www.stg.brown.edu/resources/stg/monographs/ohco.html

[18] http://www.faqs.org/rfcs/rfc2141.html

[19] I would like to thank Gabe Weaver for sharing with me his work that suggested the application of CTS URNs to Google Base content.

[20] In "Digital publication for digital libraries," available from http://chs75.harvard.edu/projects/diginc/techpub/digitalpub.

Works Cited

Dickey 2007 Dickey, Eleanor. *Ancient Greek Scholarship: A Guide to Finding, Reading, and Understanding Scholia, Commentaries, Lexica, and Grammatical Treatises, from Their Beginnings to the Byzantine Period*. Oxford: Oxford University Press, 2007.

Dindorf 1875 Dindorf, Willhelm. *Scholia Graeca in Homeri Illiadem, Vol 1*. Oxford: Oxford University Press, 1875.

Nagy 2004 Nagy, Gregory. *Homer's Text and Language*. Champaign, IL: University of Illinois Press, 2004.

van Thiel van Thiel, Helmut. "Scholia D in Iliadem". Preliminary publication online from http://kups.ub.uni-koeln.de/volltexte/2006/1810/.