DHQ: Digital Humanities Quarterly
2014
Volume 8 Number 2

Canonical References in Electronic Texts: Rationale and Best Practices

Joel Kalvesmaki <kalvesmaki_at_gmail_dot_com>, Dumbarton Oaks

Abstract

Systems of canonical references, whereby segments of written works are sequentially labeled with numbers or letters to facilitate cross-referencing, are widely used but seldom studied, undeservedly so. Canonical numbers are complex interpretive mechanisms with a great deal of potential for anyone editing and using electronic texts. In this essay I consider the rationale for and nature of canonical reference systems, to recommend principles to consider when deploying them in digital projects. After briefly reviewing the history of canonical references I note how they have been used so far, emphasizing the advances made by Canonical Text Services (CTS). I argue that the practical and theoretical problems that remain unaddressed require engagement with descriptions of how textual scholarship works and how notional literary works relate to the artefacts that carry them (using Functional Requirements for Bibliographic Records, FRBR). By correlating a theory of canonical reference numbers with those two models — editorial workflow and creative works — I offer key principles that should be addressed when planning, writing, and using digital projects.

The rationale of canonical references

Canonical references — e.g., Homer, *Iliad* 1.1; John 3:16; Sun Zu, *The Art of War* 3.6, *Hamlet*, Act III Scene 2 line 1895 — allow one to point to and discuss specific parts of literary works without having to stipulate a particular edition or version. Such numbers arose in earnest in the Renaissance as a way to place firm anchors in a fluctuating sea of texts, and have ever since been a critical means of cross-referencing. In the electronic age the textual sea has grown larger and more turbulent, so the anchors provided by canonical numbers are arguably more important now than ever, especially for scholarship. 1

For example, formal linguistic research into translations, paraphrases, and other acts of textual reuse depends upon the alignment of large corpora of texts. The preparatory stages for so-called bitext alignment are notoriously time-consuming [Tiedemann 2011]. There exist good stochastic methods to synchronize large units (paragraphs, sentences), but correcting the low level of errors still takes work. And by the time the bitexts have been aligned on this rough level, the end result is, in essence, a canonical numbering system, usually invented ad hoc and incorporated into other projects only after custom conversion. 2

Canonical reference systems also continue to facilitate everyday discourse about familiar literature. A Web publisher can install a javascript library such as reftagger.com's to automatically add widgets and hyperlinks to inline, naturally phrased canonical references that appear in individual pages. A writer doesn't need to do any special coding when quoting, and readers, through hover and click events, can consult the work that has been cited, in the language or version of choice. Such javascript libraries, which are always restricted to a particular textual corpus, would be impossible without the shared convention of canonical numbers. 3

A shared universal protocol for canonical numbers in electronic texts — some protocol readable by both computers and humans — would bring enormous benefit to linguists, digital humanists, and the public at large. But the challenges in finding such a protocol, as I will explain, are numerous and significant. Rather than tackling those challenges directly, I step back in this article to explore the history, theory, and assumptions that underlie canonical numbers. Briefly setting 4

them in historical context, I trace recent efforts to make them viably interoperable in digital projects. Initiatives such as Canonical Text Services have made critical advances, but important theoretical and practical questions remain. I argue that these problems are best understood against two backdrops: the process editors follow in their textual scholarship, and the model library cataloguers have used to describe creative works (Functional Requirements for Bibliographic Records, FRBR). I argue that a comparison shows exactly where each of the two major classes of canonical numbers — visual and semantic — are best suited. Finally, I detail principles that can help digital humanists make the most of canonical numbers in electronic texts. Those recommendations can be summarized here:

- Innovations in the number and specificity of canonical numbering systems have historically capitalized on new technology. That trend should be encouraged, not hindered. 5
- Both writers and users of textual data should declare in a computer-parsable manner the numbering schemes they have adopted, the type of each number, and any related normalizations applied to the texts.
- Numbering schemes both point to texts and interpret them. The first function requires stable and familiar numbers; the latter, the freedom to bring new insight into text structures. Digital projects should accommodate both impulses whenever possible.
- The distinction between visual and semantic numbering should always be kept in mind when creating data models and creating and using data. Neither type of number is inherently superior. The suitability of a particular system depends upon the assumptions and purposes of editors and researchers. Specifically:
    - Semantic numbering schemes are a heuristic ideally suited for knowledge, meaning, and insight; visual ones, for text-bearing objects.
    - Semantic numbering schemes are easier to apply across multiple versions of a work than visual ones are.
    - Visual numbering systems are easier and less controversial to fine-tune with increased specificity than semantic ones are.

My goal in this article is not to establish a definitive approach to employing canonical numbers. The problems, which I think remain numerous and complex, deserve sustained reflection. This essay is best seen as a catalyst to that reflection, and to experimentation in different practical solutions. 6

The complex tradition of canonical numbers

In antiquity, cross-referencing and text segmenting were two unrelated activities. The former was served by quotations, paraphrases, allusions, and other types of textual reuse, a flexible system that allowed (as it still does) authors a great deal of control over the quality, precision, and character of their references. Text segmentation, however, began as a way not to provide anchors for reference but to shape the semantic contours of the text. Large works would be divided into books or poems, e.g., the *Iliad* and the *Histories* of Herodotus, and smaller units were marked with individual symbols (e.g., ※, ⁖) that identified paragraph-sized units [de Moor and Korpel 2007], [Porter 2007]. Such segmentation, which relied upon marks that could not be sequenced, signaled the borders of logical or semantic units and furnished scribes or writers the means by which to shape their readers' interpretation of the text [Korpel and Oesch 2005], [de Hoop, Korpel, and Porter 2009].[1] 7

The use of numbers or letters, instead of symbols, so that sub-book units could be uniquely and sequentially labeled, is not attested until the Hellenistic/late antique period, with the numbering of Greek lyric poetry [Higbie 2010], the subsequent imitative numbering of the ancient Greek translation of the Hebrew Psalter [Yarchin forthcoming], and the numbered division of the four New Testament gospels into *titloi* and *kephalaia*, a structure that became an integral part of the canon tables invented by Eusebius of Caesarea (ca. 260–339/40) [Wallraff 2013, 6–7], [Zola 2012]. Eusebius's technique for numbering the parts of the four gospels could not have existed without the invention of the codex, a book format much easier than a scroll was for looking up references. The function was important to Eusebius's clientele, because the traditional methods of cross-reference — direct and indirect quotation — were not helpful for clergy who needed to quickly and conveniently find the Gospel passages designated for reading on a particular day. So the beginnings of what we know now as canonical reference systems emerged within the confluence of, among other 8

things, new needs and new technology.[2]

9 Segmenting and numbering works developed only sporadically and piecemeal through the late antique and early medieval period, gaining traction only later and in specific locales, particularly thirteenth-century Paris [van Banning 2007]. This late refinement of numbering schemes was again facilitated (but not determined) by social and technological developments such as the transition from scriptoria to a highly specialized bookmaking industry, which was eventually succeeded by the first printing shops [Rouse and Rouse 2009]. In this case the enrichment of numbering systems was motivated by pedagogy and scholarship, not liturgy. University scholars and students who wished to discuss and compare a text and its translation, e.g., a Hebrew work translated into Latin, required a rough method of alignment, and the best coordinating mechanism was to number the text segments and promulgate the system through the managed production and distribution of books [Moore 1893].[3] Broad consensus about numbering schemes, and the creation of complex hierarchies of small textual parts (e.g., subchapters) occurred only in the modern era, with the printing press, which helped stabilize clear, fixed reference systems [Zola 2012]. Long columns of a printed book could be marked by intercolumnar letters, and individual page or chapter lines could be numbered, each copy guaranteed to be identical.

10 Ancient texts critically edited in the modern era are pinned to the scaffolding of numbering schemes that have come to be called canonical, much in the same special sense Eusebius used κανών [Wallraff 2013, 1–2]. That is, canonical numbers refer not to what we *ought* to read but to items in a *list*, i.e., a sequence, whether flat or hierarchical. That canonical list allows readers to point to items with precision, clarity, confidence, and convenience. These familiar and indispensable numbering schemes have become the basis for structuring the texts in digital corpora such as the Thesarus Linguae Graecae (TLG, an extensive database of ancient Greek literature) and the CLCLT/Cetedoc Library of Latin Texts (a similar database for Latin).

11 Despite their convenience, canonical numbers are not always simple. Many ancient and medieval works have attracted more than one system, each based on incommensurate logic. A reference to Arist., *Cat.* 5.2 (2a14–15) is commonly, and without contention, understood to refer doubly to a single passage in Aristotle's *Categories*. The first pair of numbers points to the medieval-inspired chapter 5, subchapter 2; the latter pair, to the modern milestones of the highly influential 1831 Bekker edition: page 2, first column, lines 14 and 15. The Bekker labels do not necessarily require one to consult the original nineteenth-century version. Any edition or translation that includes his numbers (very many do) could be used to find this passage. Both citation systems are widely used and highly regarded, since each serves a different purpose. The medieval system elegantly captures passages, *sententiae*, and logical segments; the Bekker numbers pinpoint specific lines, phrases, or words.

12 Canonical numbering systems are also fertile. They can change with new versions or editions. So, for example, the *Taktika* compiled by Leo VI in the tenth century has different paragraph numeration in the three editions published in the nineteenth, twentieth, and twenty-first centuries, each shaped by the distinct perspective of the editor.[4] Some alternative numbering schemes take hold and become standard; others face resistance. In some cases resistance or support for a particular system stems from assumptions about the origin and structure of the text. A good example is the Psalter. The most widely used numbering looks to the Hebrew tradition, in which Psalm numbers were not added until the tenth century at the earliest (and even then, in a system different from the modern one: [Yarchin forthcoming]). This numbering differs from that used in the oldest Greek translations, anachronistically but usefully called the Septuagint (LXX). Most of the Psalm numbers in the Hebrew and the LXX differ by one. Such differences might seem trivial today, but to many early Christians the number of a Psalm was a key to its interpretation. For example, Didymus the Blind (*Commentary on the Psalms*, 106.24-109.4) held that Psalm 50 LXX (51 Hebrew) was to be understood as a revelation that the number fifty, wherever it was found in scripture, symbolized repentance. According to Jerome (*Homilies on the Psalms* 5), Psalm 14 LXX (15 Hebrew) was to be read in concert with Exodus 12:6 and the fourteen-day half-cycle of the moon. Such interpretive readings corroborated how Christians understood themselves and how they characterized those outside the Church. Apologists as early as Justin Martyr (fl. 2nd c.) accused the Jews of having altered or dropped parts of the Bible that were faithfully preserved by the LXX (e.g., *Dial. Tryph.* 71–73).[5] Thus, the two systems sustain views about the ideal state of the Psalter, and the boundaries of individual Psalms (Hebrew 9 and 10 = LXX 9; Hebrew 147 = LXX 146 and 147). To assign a number to a Psalm is to take part, even if unwittingly or unwillingly, in old

Jewish-Christian debates about the priority of one text tradition over the other, and the authentic structure of the Psalter.

These are but a few of the difficulties found in the canonical numbering systems we have inherited. Many others could be adduced.[6] The problems affect technical decisions made in a digital humanities project. Shall one canonical numbering system be preferred over the other? Which one, and why? What support should be provided for alternatives? Should any be deprecated? Should canonical numbering be abandoned altogether? 13

How such questions are answered by a project may not suit everyone, including some of the editors and researchers who need to create and study that project's data. The assumptions and expectations scholars bring to the data they write and use greatly shape the practical ways canonical numbers are deployed in electronic texts. 14

Practical answers: TEI and Canonical Text Services

As has already been mentioned, canonical numbers were devised to help readers and writers refer to and find segments of literary works independent of specific editions or versions. In a digital environment, we would want the same function, and more. We should be able to make cross-references that can be understood unambiguously not just by humans but by computers, so that algorithms can cull the selected text from assorted other electronic versions, perhaps even versions unknown to the creator, and with minimal human intervention. 15

Furnishing an electronic text with machine-actionable canonical numbers is easy, at least on a project-by-project basis. The challenge is finding a shared, interoperable protocol. Such interoperability is in theory possible in Text Encoding Initiative (TEI) files, but documentation for `@cRef`, the lynchpin of the system, is relatively unspecific and tailored mainly to single-project use. The prime TEI example, that of Matthew 5:7, assumes conventions suited to a predetermined processing application [TEI Consortium 2.6.0, §16.2.6]. To my knowledge, no set of values for `@cRef` are shared across projects, despite the potential: `@cref` is bound to the datatype `data.text`, which permits unique, machine-readable names that could be parsed independent of any resolving algorithm. For example, using Matthew 5:7, one could theoretically encode in a TEI file `<ref cRef="http://example.org/NT/Mt/c5/v7" />` and, following the principles of linked open data, set up that URL to resolve HTTP requests with various serializations of Matthew 5:7 (e.g., in JSON-LD or another RDF format suitable for semantic web applications).[7] 16

The approach of using URL names for `@cRef` values is attractive, because it theoretically provides a way to refer to portions of texts by their canonical numbers, independent of any single TEI project's resolving algorithm or naming convention. And it would open up the legacy of all historical literature to the world of linked open data. But a `@cref`-based approach would be difficult to implement. The owner of `example.org` would need to configure the server to handle requests correctly and maintain that server, an extra set of duties that many potential TEI editors would prefer to avoid.[8] How should one specify which language or version to return? How would one define in a single URL a more complex range of text (e.g., Matthew 5:7-6:10a)? And what about permanence? What happens when the domain name changes ownership? 17

These difficulties have been addressed by Canonical Text Services (CTS), part of a larger protocol that coordinates collections, indexes, texts, and extensions (hence the name of the larger project, CITE).[9] The protocol depends upon a URN scheme (discussed below) designed for canonical numbers, with the central goal of allowing those who deploy servers to resolve CTS URNs so as to return to a client a string that corresponds to a specified canonical reference. The basic principles that govern CTS URNs have been outlined by Neel Smith, who has convincingly responded to doubts about whether text is an “ordered hierarchy of content objects” (OHCO) by arguing that the primary coin of the realm — citation numbers — demonstrably *are* such a hierarchy, and can and should be treated as such [Smith 2009, 26–32]. He does not claim that canonical numbers are the only way to structure texts in hierarchy, nor does he presume to say exactly what a text is, the question that vexed earlier scholars [DeRose et al.]. The end result is a system that works, one that reliably delivers text segments, because it focuses on one type of well-ordered hierarchical model. 18

CTS URNs, complex strings, form the heart of the protocol, and deserve detailed attention, since they rely upon a specific domain model of how texts are structured. Let us use an example, 19

`urn:cts:greekLit:tlg0012.tlg001:1.26`, one way in CTS to refer to Homer, *Iliad* 1.26. The URN has five major components.[10]

The first part, `urn`, declares that the text string is a Uniform Resource Name (URN). Defined in RFC 2141 and RFC 3986 (which defines Internationalized/Uniform Resource Identifiers [IRI/URI], the superset of URNs) URNs are administered by the Internet Assigned Numbers Authority (IANA).[11] At this writing, the second part of the URN, `cts`, is not in the IANA registry of URN namespaces, and governance of the namespace `cts` remains unclear.[12] 20

The third element in the URN, `greekLit`, designates the first part of the namespace-specific string (NSS), a part that CTS documentation has called a *naming authority* and a *namespace*, but which I will here call a *subnamespace*, mainly to distinguish it from the RFC 2141 use of *namespace*. For URNs that rely upon a registry, the namespace-governing body assigns subnamespaces to participating individuals or organizations and delegates to them the responsibility for assigning unique strings that follow the subnamespace. Publishers, for example, buy a block of ISBNs from their national ISBN registration agency and unilaterally assign each publication a single number. But a CTS subnamespace, at least for now, is an arbitrary string, perhaps coined by an individual or organization that deploys a service that is meant to be CTS URN-compliant [Smith 2009].[13] 21

The fourth part, `tlg0012.tlg001`, identifies the work, so is called the *work component*. The string follows an internal structure “analogous to the hierarchy of the Functional Requirements for Bibliographic Recordsr (FRBR).” [Smith and Blackwell 2012] (The role and import of FRBR is discussed at length below.) It can take up to four segments, delimited by full stops. In our example, which has only two segments, the first, `tlg0012`, designates a text group (writings attributed to Homer) and `tlg001` identifies the work (the *Iliad*).[14] Classicists will instantly (and correctly) associate `tlg` with the Thesaurus Linguae Graecae. But this association can be misleading, since one might assume that CTS URNs are wedded to the TLG cataloguing system. That would be a problem because TLG number assignments do not follow a FRBR-like principle, and would be inadequate for describing in a FRBR-compliant manner many ancient Greek texts [Berkowitz et al. 1990].[15] But in practice the construction of this fourth part of the URN is completely dependent upon the subnamespace owner, and so resultant strings are arbitrary. The subnamespace owner might have coined `Homer.Iliad` or `Ao7G34.p24b` with the same results, except that each has a different level of predictability and readability. The work identifier may take up to four segments. It could be simple (e.g., `Iliad`) or it could be extended to include a version segment, which may then be followed by an exemplar segment. Both segments are also arbitrary strings.[16] For example, `tlg0012.tlg001.grc01.msA` could be used to identify a particular version of the Iliad preserved in Greek in manuscript Venetus A. 22

The fifth part of the CTS-URN, called the *passage component*, specifies a node of text through canonical numbers. In our example, `1.26` designates book 1, line 26. A range of text may be indicated by two canonical numbers separated by a hyphen, e.g., `1.26–1.28`. The URN may be extended to refer to textual units even more refined by appending an “at” sign and a string and an optional sequence identifier. For example, `@ἐγὼ[1]`, may be appended to indicate the fifth word of *Iliad* 1.26 [Smith and Blackwell 2012]. Although this fifth element is theoretically just as arbitrary as the rest of the URN, in practice it is meant to be the least, because it trusts that people will want to use familiar canonical numbering schemes, simplified to ASCII letters and the Arabic numerals 0 through 9 (i.e., Roman numerals or Greek alphabetic numerals must be converted ahead of time).[17] For a given work, this fifth part must follow only a single canonical numbering system. But, theoretically, that text could be encoded multiple times with different work identifiers, each instance using a different numbering system. CTS provides no mechanism to declare explicitly which canonical system is being used. 23

CTS and CITE were developed to produce results, to create a reliable and predictable way to store and serve canonical texts. Early implementations suggest that CTS URNs are an important advance in facilitating technology-independent referencing. But the current version of the URN specification does not address important aspects of canonical references, distinguishable between theoretical concerns and practical ones. I momentarily suspend the first, to address practical issues. What principles should guide decisions in CTS implementation? CTS seems designed to be at least somewhat decentralized, without sacrificing interoperability. If one CTS project wishes to interact with another, how will 24

designers anticipate the conventions adopted by those projects to resolve complexities in canonical numbering schemes? If two CTS subnamespace owners mint URNs that describe the same work and they legitimately disagree about which numbering scheme to adopt, because each serves a different purpose, how can those differences be resolved? How would a translation of Aristotle in Arabic be divided by Bekker numbers, since a strict implementation would require overlapping and interlocked spans of text, i.e., a departure from a hierarchical model? What about a version of a work that is so disjoint with another that the received canonical numbering system is inadequate? How would one go about implementing a new one? Such practical questions are best approached by first considering how canonical numbers enter the process of textual scholarship.

Canonical numbering and the process of textual scholarship

25 Numbering parts of texts has traditionally been just one scholarly activity among many. Different texts demand different tasks, all of which can be placed somewhere on the following unidirectional sequence, consciously generalized here for the sake of argument.[18]

26 Before any activity at all, one must start with real, physical objects that bear written language: manuscripts, papyri, coins, seals, printed books, emails, etc. (I take digital files here to be material objects). Many of these are copies, of course, and scholars frequently hypothesize exemplars and archetypes. But even this hypothetical exercise, this urge to get to older but lost versions or to flesh out texts that are thought not to survive (e.g., Aristotle's treatise on comedy), must begin with the relevant writings that *do* survive. Our abstract constructions are always tethered to material artefacts.

27 From the basis of existing artefacts, text-critical work engages primarily with the oldest text-bearing objects, be they ancient unica, modern books, digital images of artefacts that no longer survive, or what have you. Scholars, inspecting and reading the objects themselves (e.g., books, codexes, seals, coins), under various conditions, engage directly with them, an experience mediated through vision, apparatus (e.g., magnifying glasses, computer screens), and mental processes.[19]

28 The work may then turn to creating surrogates of that original object, surrogates that replicate, describe, or interpret the visual features of the objects. Some surrogates are mechanically generated, e.g., printed photos, lithographs, squeezes, or digital images (2-D or 3-D scans or photographs), captured at various times, by various means, and under various conditions (e.g., of light). Other surrogates are derivative human artefacts created through a synthesis of judgment, interpretation, and skill, but intended to imitate the original object: line drawings, sketches, and other graphic-oriented interpretations (e.g., restored sketches of material too difficult to see in photographs). In every case, these surrogates are created to facilitate careful study and discussion of the original objects or the writing they carry, whether as an adjunct (e.g., an ultraviolet photograph of a papyrus fragment, to supplement other images) or as a replacement (e.g., a good, clean photocopy of a book, which requires the scholar to consult the prototype only when the photocopy is illegible).

29 Textual work then steps from graphically centered activity into transcription. Transcriptions fall along a continuum, with fidelity to the display of text on one side and concern for meaning on the other. On the first side are those transcriptions that frequently serve diplomatic editions. They respect graphic elements of the writing on the object or its surrogates, e.g., ligatures, punctuation, word space, line breaks, or letter position. On the other side are normalized transcriptions, where graphical features are resolved or converted. Ligatures might be broken into constituent letters, abbreviations spelled out, letterforms standardized, formats homogenized, or new punctuation and word space introduced. Visual cues on the object (size or color of letters, spacing) may nudge a transcriber to adopt a rudimentary form of semantic interpretation. Every diplomatic transcription requires at least some amount of normalization. And each normalized transcription follows a set of conventions, some used subconsciously, that adhere to readers' expectations for clarity and meaning. An object (or its surrogates) might undergo multiple transcriptions, but each one can be placed somewhere on this spectrum of graphic to semantic, diplomatic to normalized.

30 Letter or number labels may be applied to texts at any stage. They tend to come later rather than sooner. Numbers are

rarely applied to the original object, the numbers written by a modern curator in the corners of a manuscript's folios being a rare example. On surrogates such as digital images or line drawings, line or object numbers might be superimposed. And in transcriptions, numbering is all but inevitable. The further along the process, the more likely a scholar is to label and number the lines, columns, pages, paragraphs, gaps, or other visible textual units. This numbering nudges a transcription, even if very slightly, toward the normalizing side of the spectrum, since the editor makes judgments about the boundaries and contours of the writing.

That accounts for textual study of individual objects. Work may then turn to groups and collections of objects recognized as carrying similar or identical texts. Where sufficiently similar, transcriptions are collated, and aligned and compared where they are not. Well-collated transcriptions become the foundation for critical editions. And upon these transcriptions, collations, and editions, a whole series of other activities depend: translation, annotation, commentary, analysis, and interpretation (scholarly or not). This class of work normally addresses questions that go beyond the visual features of individual text-bearing objects. What is the plot? Does this sentence exhibit sarcasm? What significance does this claim have on assessing the price of whale oil in the nineteenth century? Such activities necessarily come at the end of a complex trajectory of textual scholarship. Our work is rooted in the visual, graphic experience of text-bearing objects, and, without ever abandoning those foundations, presses out to the semantic, conceptual, and ideal realm. 31

My description is, of course, a broad generalization. For any given text, steps could be omitted, repeated, or worked on concurrently. Scholars regularly step back and oscillate between reflecting on ideas and scrutinizing objects. And anything created within this process could itself become the starting point for a chain of scholarship to begin anew. My argument depends not upon specific paths taken but upon the overall thrust, which is better or poorly served at each stage by the two basic types of canonical numbering. Every canonical numbering scheme of a text serves one end of this process better than the other. Attending to the physical object, we number or label folios, pages, columns, sides, anonymous blocks, lines, glyphs, strokes, or Cartesian coordinates. On the semantic side, we label logical or conceptual components: books, poems, stanzas, chapters, entries, speeches, sentences, clauses, words, letters. Thus, Bekker numbers in Aristotle are object-oriented (page, column, line numbers that correspond to the 1831 manifestation) but the medieval references are semantic-oriented (chapter, subchapter). Even when those systems are repurposed (e.g., Bekker numbers applied to paragraphs), one type of numbering is better suited to the text-bearing object, the other to a text's meaning and import. 32

All taxonomies and structures respond to and serve the purposes of questions and assumptions scholars bring to their research. Canonical numbering systems are such a structure. Therefore an object-based canonical numbering scheme is ideally suited to serve questions or purposes that are focused on the original text-bearing object (or its surrogates) that is the foundation of that system. Similarly, semantic sequences help one set aside the variety of physical objects that carry versions of a text so one can engage questions of meaning, significance, and importance. Object-based canonical numbers nicely accommodate the need to refer to texts with specificity. Semantic-based ones tend to support discussions about how works are internally structured. The two have different functions; neither is inherently superior. But they are each governed by two different domains — physical and conceptual — whose relationship complicates any reference system that intends to bridge the two worlds. 33

Our example CTS URN for the manuscript A of *Iliad* 1.26 — `urn:cts:greekLit:tlg0012.tlg001.grc01.msA:1.26` — appeals to abstract entities everywhere except in the string `msA`. That a semantic numbering system (`1.26`) can connect to a physical object (`msA`) which can in turn be connected to three higher entities (the collection `tlg0012`, the work `tlg001`, and the version `grc01`) relies upon a particular view of texts, in this case a variation of a domain model called FRBR. 34

Canonical numbering and textual entities (Functional Requirements for Bibliographic Records [FRBR])

Just as scholarly activity can be described as a sequence, so textual entities, both physical and conceptual, can be — and have been — described in sequential or hierarchical models. The most thorough models are by those who most 35

often write and share metadata about texts: librarians. The Functional Requirements for Bibliographic Records (FRBR), highly regarded and influential in library cataloguing worldwide, is one of the best known domain models to describe material books and their conceptual archetypes.[20] First published in 1997 by the International Federation of Library Associations and Institutions, FRBR has been designed to help librarians create better data models, calatoguing rules, and metadata records.[21] Of the three groups stipulated by FRBR, only one, Group 1, which treats creative works, concerns us here.

36 FRBR describes Group 1 (creative artefacts) as a hierarchy of four entities, described top down as *work*, *expression*, *manifestation*, and *item* (sometimes called *WEMI* as a mnemonic). A *work* is “a distinct intellectual or artistic creation” but it is also “an abstract entity” [IFLA 2008, 17]. The *Iliad*, treated as a FRBR work, is to be found in no material object — it is an ideal, as is the second tier, *expression*, “the intellectual or artistic realization of a work.... [T]he specific intellectual or artistic form that a work takes each time it is ‘realized’.” Thus, Aristotle's *Categories* may be a work, but the lost Greek original, Andronicus's edition, Bekker's edition, and Bodéüs's translation would all be expressions of that single work. Works have a one-to-many relationship with expressions. Their metadata include descriptions of the creators, the subjects, dates of creation, and so forth.

37 On the third tier is *manifestation*, “the physical embodiment of an *expression* of a *work* ” [IFLA 2008, 21]. This is the first level at which the physical embodiment of a creative work can be described. Expressions and manifestations share a many-to-many relationship. Just as any expression may be embodied by any number of manifestations (e.g., multiple print runs of a book), so any physical object might have multiple creative works (e.g., an anthology). Metadata about manifestations usually include descriptions about who made the physical artefact, when, where, and so forth. Thus, Bekker's edition, its later reprints, and arguably the digital scans housed by Google or Archive.org are individual manifestations. Any specific instance of a manifestation is described by the fourth and final tier, the *item*: “a single exemplar of a *manifestation* ” [IFLA 2008, 24]. Examples of an item would be the copy of Bekker that sits on the shelf of my library, or the copy sold by a bookstore in New York. Manifestations and items form a pair, and like the corresponding upper pair (works and expressions), the first stands to the second in a one-to-many relationship.

38 The hierarchy of WEMI follows the sequence of the creative process. A writer (or artist or musician — FRBR covers any creative work) mentally conceives of a *work*, and decides to *express* it in a certain way. When the mental work becomes physical, that is, when the printing press inks the paper, the work is made *manifest*, and through distribution the public encounters individual *items* of the work. Each of the four stages is a necessary condition of the others.

39 The FRBR model, shaped as it is by specific priorities, assumptions, and needs, has come under sharp criticism, especially by librarians, best seen in the 2012 special issue of *Cataloging & Classification Quarterly* (vol. 50, nos. 5–7). One FRBR difficulty touches on canonical numbering systems, namely, how to interpret the definitions of the four Group 1 entities. For example, the lavishly illuminated manuscript the *Book of Kells*, as noted by Ian Fairclough, is not easily and unambiguously modeled in FRBR.[22] If treated as simply another biblical manuscript, the *Book of Kells* would be catalogued as the sole exemplar (item = manifestation) of the New Testament Gospels (the work) in Latin translation (the expression). And a digital project dependent upon this approach to FRBR would be inclined to use the New Testament's traditional canonical reference system. But if thought of as a creative work of art in its own right, the *Book of Kells* could be reasonably classified simultaneously as work, expression, manifestation, and item. To some this approach would be absurd; to others — especially those who need to catalogue photographic reprints of the *Book of Kells* as independent manifestations — it would be sensible. A digital project following this interpretation of FRBR would would be well suited to use principally folio numbers, columns, and cartesian coordinates.

40 All other problems aside, FRBR's architecture has salient points of comparison with the process of textual scholarship (described above) that touch on canonical numbers. Some differences should be immediately evident. For example, textual work begins with a physical object and moves up; the FRBR model begins at the opposite end, and moves down.[23] The two models describe two different kinds of activities, each with a distinct set of necessary conditions. FRBR focuses on the creative process that leads from the mind of the writer to the hands and eyes of the readers. An account of the workflow of textual scholars describes how they take a preexisting artefact and incorporate it into their

own creative work. Consequently, a digital project that approaches its texts with the FRBR perspective is likely to start with semantic canonical numbering as a foundation upon which to superimpose any other visual ones. But a project that takes a viewpoint grounded in the sequence of textual work is likely to adopt visual numbers as a starting point. Once again, neither approach is inherently superior. But in any given project one should always be aware which type of system has been given primacy.

41 A second obvious difference between editorial work and FRBR's model is in the latter's differentiation between *manifestation* and *item*. Scholars dealing with texts that survive primarily on premodern text-bearing objects will nearly always be dealing fundamentally with unica perhaps best thought of as item-manifestations. No manuscript or carving is identical to another, regardless of how closely a copy imitates its model. This distinction then is of little consequence to some canonical numbering systems, but of great importance to those that must include both manifestations and items as distinct categories, for example, projects that compare inscriptions on coins or seal impressions (among the few ancient artefacts that can be described in terms of manifestations) or that compare different print impressions of a book. Any canonical numbering convention that disallows manifestations or requires them runs the risk of excluding certain classes of texts.

42 Some points of comparison between FRBR and scholarly practice are tempting to construe as differences, but those differences are only superficial. The most salient for canonical numbering pertains to what CTS calls text groups or collections, e.g., the collected works of Homer. This class is treated as a conscious departure from FRBR [Smith 2009, 19–21]. But that deviation may be more apparent than real since, in fact, FRBR allows works to encompass, and to be encompassed by, other works, so-called aggregating works. For example, the Bible, the New Testament, the Gospels, Matthew, and the Lord's Prayer can all be legitimately classified as FRBR works [IFLA 2008, 29].[24] That complexity makes it impossible to introduce a FRBR Group 1 entity higher than work, even if it were desirable.[25] Deliberating over how works relate to collections, or works relate to works, importantly affects how one thinks about canonical numbering systems. Canonical references, as noted above, follow a hierarchical model, which can be represented as a tree. Each node in a hierarchical tree is the child of one node and no more, and it does not overlap with its sibling nodes. But a hierarchical model cannot fully describe works or collections. Any work may incorporate multiple works or collections or be incorporated by multiple works or collections, and “sibling” works frequently overlap. So the relationship of works to each other — singly or in groups — cannot be described in a tree hierarchy. One needs not a hierarchical model but a network one, to express sets and many-to-many relations. Thus, any protocol that is meant to deal with the relations that hold between textual parts and wholes cannot depend exclusively upon a hierarchical model, which will suit only canonical numbers of individual predetermined works. Alternatively, if a protocol chooses for practical reasons to adopt a hierarchical model not only for canonical references but for their superstructures, then the documentation for that protocol should explain how users should convert complex work-work or work-collection relations into a simpler hierarchical model.

43 Now some similiarities between FRBR and editorial work habits, particularly those that touch on canonical numbering systems. One pertains to the distinction between conceptual and material realms. As might already be apparent, object-based numbering schemes (e.g., page numbers) correspond to the lower half of the FRBR model (item and manifestation); semantic-based ones, e.g., paragraphs, parallel the upper half (expression and work). The bottom half of FRBR Group 1 entities is comparatively straightforward, implemented by library cataloguers without a great deal of confusion. This corresponds to the relatively clear and unambiguous use of object-based canonical numbering systems. An editor needs merely to number the pages and lines of any transcription, which is rather uncontroversial to do if the writing is clear. A reference to the smallest unit in an object-based numbering system, e.g., folio 3v, left column, 7th line, 5th glyph, can be written, read, and cited with little potential for confusion — low entropy as it were.

44 But the upper half is oftentimes not well understood, and precision can be problematic [Peponakis 2012]. I have already pointed out how FRBR has not attempted to formalize a mechanism to describe how one work relates to another. To that should be added the uncertainty about the boundary line between expression and work. By FRBR definition, translations are individual expressions of a single work. But clearly some translations or adaptations take a life of their own, generating other translations or becoming independently culturally significant (e.g., Plautus's plays, or free

adaptations of Shakespeare). The transition point, where an expression might be legitimately described as a work in its own, is noted but not defined by the FRBR model [IFLA 2008, 293.2.1]. The discernment a cataloguer must exercise to distinguish a work from an expression resembles the judicious choices a textual scholar makes in devising a semantically oriented system of canonical numbers. To divide a text and attach numbers to sentences and paragraphs requires interpretive discernment of a sort quite different than the one that labels lines and pages of an object. Segmentation of the former kind relies upon the contestable interpretation of the boundaries of meaning, much like the work-expression and work-work boundaries that must be determined by library cataloguers. Some libraries might disagree with how the Library of Congress classifies a work, and they would be within their rights to catalogue it as they see fit, especially if the decision is in service of the libraries' specialized missions. So too in textual scholarship. If the modern editor of a previously edited text faults the interpretation that motivated the received semantically oriented canonical numbering system, and if the editor cannot adapt it without doing violence to the interpretation of the text, then that editor is entirely justified in renumbering. Of course this runs the risk of introducing confusion. But it is up to the editor to think critically about how a work should be interpreted and to determine whether the risk is justified. And it is up to readers and other editors to decide whether or not the new system deserves acceptance. Our reasonable desire for a uniform convention should not always prevail. All semantic canonical numbering schemes are interpretations, and all interpretations merit critical reflection and revision.

Recommendations for using canonical numbers in digital projects

45 Historically, refinements in canonical citation have capitalized on technological innovation. Chapter numbers in prose came with the codex and served the Church. Subchapter and verse numbers came with the book industry and served both the Church and the academy. We now have the opportunity to refine to an unprecedented level our existing cross-reference systems, and to introduce varieties of new ones, to serve a much wider swathe of society. Architects of digital projects should do all they can to support that trend, not curtail it.

46 Anyone designing and populating digital projects should keep in mind the two types of canonical numbering systems, and make decisions accordingly. This is not to say that the two types cannot or should not be combined (e.g., Bekker numbers to identify paragraphs in Aristotle's *Categories*). There may be situations where a mixture would be appropriate. But project architects should be on the lookout for cases where a fusion of the two might entail confusion or invalid data.

47 New or enriched numbering systems should be encouraged, not discouraged. New systems should be coordinated with predecessors. Provision should be made, when possible, for tables that allow one canonical numbering system to be converted to another. This would permit scholars not only to structure and therefore interpret their corpora in new ways, but to look afresh at texts through obsolescent numbering schemes that may be objects of study in their own right. Consider, for example, an edition of a classical text coded to be read not only in its modern segmentation but in an earlier one, so that a scholar specializing in sixteenth-century culture might study how the classical legacy was then being interpreted.

48 Numbering systems should be maximally human readable without sacrificing machine parsability, and they should be useful independent of any specific project or algorithm. Although not every project will find CTS URNs suitable, their model should be studied. For example, the use of the token specifier `#ἐγὼ[1]` discussed above may make no sense for a particular project that wishes to specify a fragment independent of a particular language. But the CTS convention demonstrates an important way to attain new levels of specificity in our canonical numbers. Along these lines I have argued that object-based numbering schemes offer more precision than semantic-based ones do, because the former are not as likely to be confused or contested. But that is not to say that granularity in semantic numbering systems are impossible; rather, specificity in each of the two types requires different strategies.

49 For object-based numbering systems, there is enormous potential for clear, unambiguous, and precise (or precisely fuzzy) cross-referencing. With existing technology one can point not just to pages and lines but to shapes, blotches, white space, or any spatial region. Thus, project managers do well to accommodate descriptions of specific rectangular regions of an image. They do better to support polygonal regions. Even better are scalable vector graphics, which can

identify the most erratically shaped regions in a flat digital image.[26] Although there are challenges in coordinating and synchronizing the description of a single region across multiple digital surrogates of the same object, that difficulty is best addressed not in the numbering scheme but in how the surrogates are coordinated. Scholars who are editing and labeling texts should not have to negotiate that answer.

50 Semantic numbering schemes also have enormous potential for specificity. In theory, we could label not just paragraphs and sentences, but semantic units as small as words, letters, and accents. For example, one could coin Arist., *Cat.* 5.2.9.2 to refer to the 9th word, 2nd letter of Aristotle's (*Categories*) , chapter 5 subchapter 2, i.e., the ρ in the word πρώτως (“firstly”). But editors and linguists may legitimately differ on how to define semantic units, and even then on how to interpret those definitions. At what point does a paragraph or sentence begin and end? What is a word? Does a contraction count as one or two? Should a ligature be counted as one or two letters? Should combining characters (U+0360..036F) be counted individually or not? How should punctuation factor in any word or character counts? [Manning and Schütze 1999, 124–136], [Zwicky and Pullum 1983] The problem is twofold, because such commitments must be made both by those who create transcriptions and by those who use them. If people working on the same transcription are in disagreement or confusion about the interpretation of a new numbering system, then it cannot be reliably shared. But I believe the challenge merits consideration, not despair. If the assumptions and definitions that underlie incommensurate semantic numbering schemes could be unambiguously declared, then it should be straightforward to develop concordance tables, much like the ones proposed above for larger semantic units. The main challenge is to find ways to allow those who are preparing a transcription the means to declare, in a machine-readable way, what normalization techniques they have imposed upon a text. And those who wish to use transcriptions should be provided a similar mechanism, to declare how they have defined and interpreted semantic units such as sentences, words, and characters. Both mechanisms — one for editors and one for researchers, interdependent — would be enormously useful. At the same time, if it can be demonstrated that certain kinds of declarations are limited, difficult, impossible to make, or impossible to coordinate with other declarations, then the philosophical implications for linguistics, computing, and logic would be significant, making the effort all the more worthwhile.

51 Almost every text lends itself to numerous types of milestones that could be used for canonical numbering systems. Beyond the simple, straightforward examples of Bible chapters and verses, the systems can get quite complex and can be prone to ambiguity. To mitigate confusion on the part of both humans and computers, digital projects should require canonical numbers to be typed, or they should be clear about what default typology has been assumed, so each number or letter clearly corresponds to a particular unit. For example, an encoding of Arist., *Cat.* 5.2 (2a14–15), which cites five different measures, should specify which label corresponds to what kind of unit. I defer for now a discussion of how to do this effectively.[27]

52 Project managers may need to discourage the use of some numbering systems. For example, a project aligning a text and its translations would be ill-advised to use an object-based numbering system such as Bekker numbers. Such a project, focused as it is upon coordinating semantic units (words and phrases) across versions will have to deal with segmentation challenges both in the source text and in the translated text. For example, the πρώτως discussed above breaks midword and so spans two Bekker line numbers. And translations of the *Categories* will rarely respect the boundaries of Bekker line numbers. Fine-tuned bitext aligment based on visually based numbering systems would be difficult if not impossible.

53 On the other hand, some digital projects might best serve their editors and their researchers by supporting *only* object-oriented numbering schemes. For example, a project focusing on documentary editions or a project that focuses upon coordinating diplomatic transcriptions with their digital surrogates might prevent considerable confusion by compelling users and editors to work only with visual landmarks.

54 Or maybe not. It all depends upon the needs and expectations of project stakeholders. Managers should temper flexibility with realism, prioritizing support for canonical numbering systems in accord with the project's budget, staff, schedule, and goals. It is rarely easy to get resources to match our ambitions. But I hope my suggestions and reflections on the rationale and best practices for canonical number systems are helpful to those who are trying to calibrate that match.

Acknowledgements

For reading drafts of this article or for offering their constructive comments on random ideas I thank Ian Fairclough, Andrew Sulavik, John Hostage, Martin Wallraff, Karen Coyle, Matthew Beacom, and Dot Porter. If while describing phenomena in their fields of specialty I have erred, I alone am to blame.

Notes

[1] The best research on the topic has come from the Pericope Group, http://www.pericope.net/, in their eponymous monograph series.

[2] My position does not imply determinism. Although the technology was a necessary condition for innovations in canonical numbers, it was not a sufficient one (else Eusebius's tables would have been invented centuries earlier). I exclude from this narrative stichometry, another equally ancient counting technique in which lines on a scroll were tabulated to estimate scribal costs. There is no evidence that stichometric sums served any labeling or indexical function; the only number that mattered — the total — held no value outside any single scroll except as a rough estimate of how long a work was [Kennedy 2012], critiqued by [Gregory 2012].

[3] Eusebius's canon tables are the best-attested early example (Eusebius, *Life of Constantine* 4.36–37) but are also an anomaly that highlights the significance of much later developments. Although the tables were widely replicated across a host of New Testament manuscripts, they have widely divergent readings. The variety is so complex that a critical edition, a sine qua non for determining how, when, and where the tables were actually used, is being prepared only now, by Martin Wallraff.

[4] See [Haldon 2013, 455–466] for a concordance of chapter numbers assigned in the Patrologia Graeca, Vári's edition, and that of Dennis.

[5] Justin is also the earliest attested author to unambiguously and clearly use numbers for cross-reference. *Dial. Tryph.* 73.1.

[6] Other examples: (1) Text associated with some canonical numbers may be deprecated, leaving gaps in the numbering system, e.g., roughly forty verses of the New Testament since the edition of Stephanus. [Zola 2012] (2) Some competing systems are fused together in hybrid form, e.g., chapter numbers from later systems combined with line numbers of earlier ones. [Zola 2012, 249 note 30]. (3) Some canonical numbers designate spans of text with vague or ambiguous beginning and end points, e.g., a number of intercolumnar letter labels in the Patrologia Graeca and Patrologia Latina. (4) Texts from fragmentary authors take at least two sets of canonical numbers: the fragment numbers of critically edited anthologies and the canonical numbering schemes used in the original texts that quote them.

[7] HTTP: hypertext transfer protocol. URL: uniform resource locator. JSON-LD: JavaScript Object Notation for Linked Data. RDF: resource description framework.

[8] A SPARQL endpoint has been valorized as an ideal implementation, but setting one up and maintaining it can pose high costs, evident in a 2013 survey that showed that only 52% of a set of public SPARQL endpoints were operating. See [Rogers 2013], which offers other cautionary points.

[9] Material in this section derives from a combination of documentation at www.homermultitext.org and my own experience setting up a CTS server. Over the course of 19-22 May 2013, I was a participant at a CTS workshop held at Furman University, Greenville, South Carolina. Project directors D. Neel Smith and Christopher Blackwell helped participants configure individual CTS servers on portable hard drives. On my CTS server I prepared and deployed around twenty Greek texts: the Greek Septuagint version of Genesis (Rahlfs edition) and numerous eighth- to tenth-century Byzantine Greek saints lives. Prepared in advance as TEI-compliant XML, the files were converted by the CTS server into RDF triples, segmented according to canonical number, and stored in Jena Fuseki as the basis for a SPARQL endpoint. HTTP requests routed through Apache fulfilled browser-based requests for data from the endpoint triplestore.

[10] My description differs slightly from the latest official specification, 2.0.rc.1, accessed 2014-04-17 from http://www.homermultitext.org/hmt-docs/specifications/ctsurn/

[11] RFC 2141: http://tools.ietf.org/html/rfc2141. RFC 3986: http://tools.ietf.org/html/rfc3986.

[12] See http://www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml. Like the administration of internet domain names (maintained by IANA's mother organization, the Internet Corporation for Assigned Names and Numbers [ICANN]), all IANA-approved URN schemes must define a mechanism for uniqueness and persistence. That persistence is to be guaranteed either with or without an actively administered registry. In the case of registry-dependent URNs, e.g., ISBNs and ISSNs, a namespace is granted to an international organization or coalition of organizations capable of delegating numbers, of maintaining the registry, and of safeguarding the uniqueness and stability of

namespace-specific strings. (The ideal does not always hold. It is well-known among librarians that ISBNs cannot always be trusted, as some publishers have been known to recycle them or poach on those belonging to other publishers.) Registry-independent URNs, e.g., UUIDs and tag URNs, must be defined precisely such that validly constructed strings are guaranteed to be unique and persistent. Homer MultiText, the chief user of CTS, claims “A CTS-URN...is *unique* and *immutable* ”" (emphasis in the original) [Dué et al. 2012]. And [Smith 2009, 21–22, 34] suggests the beginnings of a central registry, but there are still few details on governance and how persistence and uniqueness will be guaranteed.

[13] See previous note.

[14] CTS treats the first segment, called *collection* or *text group* as being outside the purview of FRBR; the second segment, *work*, however, is said to be equivalent to the FRBR definition. [Smith 2009, 19]

[15] The two-part, seven-digit (xxxx.yyy) TLG numbers are structured upon principles of practicality, not a consistent domain or data model. The goal of the TLG has been to put all of ancient Greek literature into a convenient, simple numbering scheme. Some inconsistency has been allowed in the taxonomy to minimize confusion and duplication. The end result is a useful key for text retrieval but not a FRBR-like model, which the architects of TLG never intended to provide. In fact, TLG violates key FRBR principles. A full TLG number might point not to a work (as defined in FRBR) but to an expression, a manifestation, or an item, sometimes many of them at once (see any TLG number assigned the title *fragmenta*, which collect assorted texts from known but otherwise lost works). A work or an expression could have multiple TLG numbers (see, e.g., Irenaeus, author 1447). Some TLG numbers refer to texts that, according to FRBR's definitions might be preferably treated as works of other languages (e.g., the books of the Septuagint). Comparable communities using Greek texts that overlap with TLG's catalog may classify ancient or medieval Greek works in a way incommensurate to TLG numbers. See, for example, the Dumbarton Oaks Hagiography Database (http://www.doaks.org/research/byzantine/resources/hagiography-database), e.g., the *Life of Theophanes the Confessor* (which corresponds only roughly to TLG 3153); Euodius's version of the *42 Martyrs of Amorion* (largely based on TLG 3083.007 but incorporating readings from TLG 5098.003); and about nineteen lives that are culled from individual works collected in an edition classified singly under TLG 4411. None of the observations in this note should taken as a criticism of the TLG. I mean only to show that it is impossible to rely on a registry system such as TLG if one needs to follow a consistent domain model of texts.

[16] The third segment, called *version*, corresponds to FRBR's term *expression*; the fourth segment, *exemplar*, corresponds to *item* in FRBR. FRBR's term *manifestation* has no correlate in the CTS URN scheme.

[17] By definition (RFC 2141) all URNs are restricted to the twenty-six Latin letters (upper U+0041..005A and lower U+0061..007B, which cannot be treated as equivalent) plus the numerals and nineteen punctuation signs (some of which are reserved). It is unclear what provision will be made for canonical numbering conventions that extend outside the basic Latin alphabet or require decisions on whether uppercase or lowercase Latin letters should be used. For example, Frankenberg's edition of the Syriac translation of works by Evagrius of Pontus introduces canonical numbers that use a manuscript-based double-Greek letter system, e.g., `fol. 2bα` to refer to the part in his edition that corresponds to folio 2, verso, 1st column of the manuscript he used. Of course this can be converted to a CTS URN scheme, but neither an editor nor a researcher will be able to predict whether such a number should be rendered `2b1`, `2ba`, `2bA`, `2B1`, `2Ba`, or `2BA`.

[18] Digital Research Infrastructure for the Arts and Humanities (DARIAH), Bamboo DiRT, and other partnering organizations are pursuing a thorough study of scholarly activity, which they plan to turn into an ontology (a formalized description, in a given domain, of concepts and their interrelationships) of digital research methods, suitable for the semantic web.

[19] I do not champion here any particular epistemology or view of representation, since I think any reasonable theory of how we perceive and know can be harmonized with my description of scholarly activity.

[20] A much more complex extension of FRBR is FRBRoo ("oo" = object oriented; http://www.cidoc-crm.org/), which synthesizes FRBR with an ontology of cultural artefacts developed in the museum world, CIDOC CRM. I focus here on FRBR because of its apparent conceptual elegance and strength, and because CTS has used it as a point of departure.

[21] Cataloguing rules: Resource Description and Analysis (RDA), an international standard that replaces the Anglo-American Cataloguing Rules (1978). http://www.rda-jsc.org. Data models: BIBFRAME, spearheaded by the Library of Congress and inspired both by RDA and FRBR, is intended to replace MARC 21 records.

[22] Email posts "The Book of Kells," FRBR listserv, 2003-09-29 and 2003-10-08.

[23] This is not the only way to construe FRBR. Matthew Beacom has argued that the WEMI model presupposes a kind of Platonism, and that following an Aristotelian approach (his preference), “turning the model upside down — flipping WEMI to IMEW — will make the model easier to

understand and to apply." Email post "FRIDAY, or, turning the bibliographic entities hierarchy (WEMI) upside down?", FRBR listserv, 2003-11-18,11:29. Under this approach to FRBR, one begins with written artefacts and tries to extrapolate types and archetypes, much as literary scholars do.

[24] FRBR 3.3: "The structure of the model, however, permits us to represent aggregate and component entities in the same way as we would represent entities that are viewed as integral units. That is to say that from a logical perspective the entity work, for example, may represent an aggregate of individual works brought together by an editor or compiler in the form of an anthology, a set of individual monographs brought together by a publisher to form a series, or a collection of private papers organized by an archive as a single fond. By the same token, the entity work may represent an intellectually or artistically discrete component of a larger work, such as a chapter of a report, a segment of a map, an article in a journal, etc. For the purposes of the model, entities at the aggregate or component level operate in the same way as entities at the integral unit level; they are defined in the same terms, they share the same characteristics, and they are related to one another in the same way as entities at the integral unit level." Further work on aggregates is being conducted by an IFLA working group: http://www.ifla.org/node/923. For a critical reflection see [Žumer and O'Neill 2012].

[25] To differentiate *collection* from *work* in a domain model entails difficult questions, e.g., should collections of collections motivate yet another ontological layer of *metacollections*? The logic that compels one to avoid an infinite regress is the same that motivated FRBR architects to stick simply with *work*.

[26] The Open Annotation model offers an example of how to leverage SVG for a referential system. http://www.openannotation.org

[27] I am in the process developing a possible approach, as part of an XML encoding format for text alignment.

Works Cited

Berkowitz et al. 1990 Berkowitz, Luci, Karl A. Squitier and William Allen Johnson. *Thesaurus Linguae Graecae Canon of Greek Authors and Works*. New York: Oxford University Press, 1990.

DeRose et al. DeRose, Steven J., David G. Durand, Elli Mylonas and Allen H. Renear. "What Is Text, Really?". *SIGDOC Asterisk J. Comput. Doc* 21: 3 (1997), pp. 1-24.

Dué et al. 2012 Dué, Casey, D. Neel Smith and Christopher Blackwell. *A Gentle Introduction to CTS & CITE URNs*. *Homer Multitext Project*. 2012. http://www.homermultitext.org/hmt-doc/guides/urn-gentle-intro.html.

Gregory 2012 Gregory, Andrew. "Kennedy and Stichometry – Some Methodological Considerations". *Apeiron: A Journal for Ancient Philosophy and Science* 45: 2 (2012), pp. 157-179.

Haldon 2013 Haldon, John F. "A Critical Commentary on the Taktika of Leo VI". *Dumbarton Oaks Studies* 44 (2013).

Higbie 2010 Higbie, C. "Divide and Edit: A Brief History of Book Divisions". *Harvard Studies in Classical Philology* 105 (2010), pp. 1-31.

IFLA 2008 IFLA Study Group on the Functional Requirements for Bibliographic Records., and International Federation of Library Associations and Institutions. Section on Cataloguing. Standing Committee. *Functional Requirements for Bibliographic Records: Final Report. Revised*. München: K.G. Saur, 2008.

Kennedy 2012 Kennedy Jr., John Bernard. "Plato's Forms, Pythagorean Mathematics, and Stichometry". *Apeiron: A Journal for Ancient Philosophy and Science* 43: 1 (2011), pp. 1-32.

Korpel and Oesch 2005 Korpel, Marjo Christina Annette, and Josef M. Oesch, eds. *Layout Markers in Biblical Manuscripts and Ugaritic Tablets*. Assen: Koninklijke van Gorcum, 2005.

Manning and Schütze 1999 Manning, Christopher D., and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1999.

Moore 1893 Moore, G.F. "The Vulgate Chapters and Numbered Verses in the Hebrew Bible". *Journal of Biblical Literature* 12: 1 (1893), pp. 73-78.

Peponakis 2012 Peponakis, Manolis. "Conceptualizations of the Cataloging Object: A Critique on Current Perceptions of FRBR Group 1 Entities". *Cataloging and Classification Quarterly* 50: 5-7 (2012), pp. 587-602.

Porter 2007 Porter, Stanley E. "The Influence of Unit Delimitation on Reading and Use of Greek Manuscripts". In Marjo Christina Annette Korpel Josef M. Oesch and Stanley E. Porter, eds., *Method in Unit Delimitation*. Assen: Koninklijke Van Gorcum, 2007. pp. 44-60.

Rogers 2013 Rogers, Dave. *The Enduring Myth of the SPARQL Endpoint*. *Dave's Blog*. June 6 2013. http://daverog.wordpress.com/2013/06/04/the-enduring-myth-of-the-sparql-endpoint/.

Rouse and Rouse 2009 Rouse, Richard H., and Mary A. Rouse. *Manuscripts and Their Makers: Commercial Book Producers in Medieval Paris, 1200-1500*. Turnhout: Harvey Miller, 2000.

Smith 2009 Smith, Neel. “Citation in Classical Studies”. *Digital Humanities Quarterly* 3: 1 (2009).

Smith and Blackwell 2012 Smith, Neel, and C.W. Blackwell. “Four URLs, Limitless Apps: Separation of Concerns in the Homer Multitext Architecture”. In *Donum Natalicium Digitaliter Confectum Gregorio Nagy Septuagenario a Discipulis Collegis Familiaribus Oblatum: A Virtual Birthday Gift Presented to Gregory Nagy on Turning Seventy by His Students, Colleagues, and Friends*. Washington D.C.: Center for Hellenic Studies, 2012.

TEI Consortium 2.6.0 TEI Consortium. *TEI: P5 Guidelines*. 2014. http://www.tei-c.org/Guidelines/P5/.

Tiedemann 2011 Tiedemann, Jörg. *Bitext Alignment*. *Synthesis Lectures on Human Language Technologies*. Morgan and Claypool, 2011.

Wallraff 2013 Wallraff, Martin. “The Canon Tables of the Psalms: An Unknown Work of Eusebius of Caesarea”. *Dumbarton Oaks Papers* 67 (2013), pp. 1-14.

Yarchin forthcoming Yarchin, William. “Why Were the Psalms the First Bible Chapters to Be Numbered?” .

Zola 2012 Zola, Nicholas J. “Why Are There Verses Missing from My Bible? the Emergence of Verse Numbers in the New Testament”. *Restoration Quarterly* 54: 4 (2012), pp. 241-253.

Zwicky and Pullum 1983 Zwicky, Arnold M., and Geoffrey K. Pullum. “Cliticization Vs. Inflection: English N’T”. *Language* 59: 3 (1983), pp. 502-513.

de Bruin 2013 De Bruin, Win. *Isaiah 1-12 as Written and Read in Antiquity*. Sheffield: Sheffield Phoenix Press, 2013.

de Hoop, Korpel, and Porter 2009 De Hoop, Raymond, Marjo Christina Annette Korpel and Stanley E. Porter, eds. *The Impact of Unit Delimitation on Exegesis*. Boston: Brill, 2009.

de Moor and Korpel 2007 De Moor, J.C., and M.C.A. Korpel. “Paragraphing in a Tibero-Palestinian Manuscript of the Prophets and Writings”. In Marjo Christina Annette Korpel Josef M. Oesch and Stanley E. Porter, et al., eds., *Method in Unit Delimitation*. Assen: Koninklijke Van Gorcum, 2007. pp. 1-34.

van Banning 2007 Van Banning, SJ, and H.A. Joop. “Reflections upon the Chapter Divisions of Stephan Langton”. In Marjo Christina Annette Korpel Josef M. Oesch and Stanley E. Porter, et al., eds., *Method in Unit Delimitation*. Assen: Koninklijke Van Gorcum, 2007. pp. 141-146.

Žumer and O'Neill 2012 Žumer, Maja, and Edward T. O'Neill. “Modeling Aggregates in FRBR”. *Cataloging and Classification Quarterly* 50: 5-7 (2012), pp. 456-472.