Schriften des Instituts für Dokumentologie und Editorik — Band 12

Digital Scholarly Editions as Interfaces

edited by

Roman Bleier, Martina Bürgermeister, Helmut W. Klug,

Frederike Neuber, Gerlinde Schneider

2018

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 20. November 2018.

2018

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7481-0925-9
Einbandgestaltung: Julia Sorouri; Coverbild: “Desktop” von Franz Konrad (www.franzkonrad.com/gallery/desktop-2008-2010) .
Satz: Roman Bleier und LuaTEX

Digital Scholarly Editions and API Consuming Applications

Jeffrey C. Witt

Abstract

This article argues for the importance of a rigorous separation of the data of digital scholarly editions from the interfaces that display them. It is only when we are able to make such a separation, and in turn see our interfaces as API consuming applications, that we will be able to accommodate a plurality of innovative interfaces without redundancy and waste. As this paper will argue, the ability to a make such a separation requires a dramatic rethinking of the essence of a scholarly edition. The article first introduces in abstract what this rethinking looks like and then looks at one attempt to actualize this theory in the case of the medieval scholastic corpus. Through a number of examples of ongoing work, this article shows how such a paradigm shift enables the efficient construction of interfaces and that, because of this efficiency, a plurality of interfaces can be swiftly constructed in order to satisfy a wide variety of research interests.

1 Introduction

The committee for the *Digital Scholarly Editions as Interfaces* conference offered the following opening question to help frame the meeting's discussion and focus:[1]

> Plurality in representation is a core feature of digital scholarly editions. How do interfaces realize this plurality? Do we need different interfaces for different target audiences (i.e. scholars, digital humanists, students, public)?

What stands out as particularly important here is the question of how we can promote plurality without redundancy. The ability to create a plurality of interfaces for any given dataset is a major advantage that the digital medium affords us. This ability allows us to imagine, design, and produce interfaces created to address specific research questions or expose particular features of a dataset. Without this ability, we must be content with a single reading environment that will necessarily choose a particular presentation that privileges some concerns over others.

[1] Call for papers: informationsmodellierung.uni-graz.at/de/neuigkeiten/detail/article/call-for-papers-digital-scholarly-editions-as-interfaces/

Digital Scholarly Editions as Interfaces, edited by Roman Bleier, Martina Bürgermeister, Helmut W. Klug, Frederike Neuber, Gerlinde Schneider. Schriften des Instituts für Dokumentologie und Editorik 12. Books on Demand, 2018, 219–247.

This potential, however, is threatened by an underlying publication paradigm, adopted by the majority of digital scholarly edition projects, that results in wasteful redundancy. Digital scholarly edition projects in which data and interface are tightly coupled and data is not easily representable in rival displays pushes us back to the paradigm of the print medium where form and content are inseparable. In such cases, when innovative re-representations require massive data re-acquisition, plurality becomes prohibitively expensive and this expense forces us to be content with massive unrealized potential.

In my own experience, working on the corpus of medieval Latin philosophical and theological texts, this kind of redundancy is an acute problem. Because the corpus in question is so large and data acquisition is so laborious, every redundant act is a threat to the dream of one day being able to study this corpus as a whole and to explore that corpus through a plurality of illuminating interfaces.

Yet today, despite a clear interest in the promise of digital scholarly editions, unaffordable and unsustainable redundancies abound in attempts to take advantage of this new medium. A quick survey of websites devoted to a medieval scholastic text or author reveals an unnecessary duplication of technology stacks that deliver very basic and common user functionality.[2] Such sites typically include a frontend design that offers the user a predictable set of options, e.g. *view text*, *view bibliography*, *view about page* and so on. Yet despite this repetition in basic functionality, each group is setting up an entirely new technology stack. In creating this stack, each group is creating their own idiosyncratic way of connecting their front end to a private and siloed datastore. Accordingly, once this data is created, it can only be experienced via the specific interface to which it has been attached. Any attempt to create a new, rival interface would be stymied by lack of access to the private datastore or by the overwhelming task of re-acquiring this data and populating a new datastore. In the end, what we generally find is that isolated research groups are currently choosing the most inefficient way possible to make data available on the web, while leaving us with results that go "barely beyond" the capabilities of the printed page.[3]

In short, the problem, as it currently stands, is that energy and resources are being poured into the creation of mediocre websites that do basically the same thing, rather than allowing energy to be poured into common libraries and common interfaces for

---

[2] See for example the sites dedicated to the following medieval thinkers: Peter of Candia (Duba), Jacobus de Altavilla (Brinzei et al.), Peter Auriol (Duba et al.), John Mirecourt (Parodi and Caccia Dominioni), and Richard Rufus of Cornwall (Wood et al.). These are just a few representative examples among many. Each interface repeats a basic pattern of varying quality. Likewise, each site independently creates a very similar backend to support this limited functionality. Further, despite the fact that the information on one site is highly related to the information on the other sites, connections between sites or data sharing is impossible.

[3] The reference to going "barely beyond" is an explicit and conscious reference to the article by Joris van Zundert to be discussed below.

common presentations. Accordingly, the question that follows is: How do we stop making interfaces and digital scholarly editions that abound in redundancy? How do we, instead, prepare our data and editions in such a way that they enable the easy construction and maintenance of a plurality of interfaces that will, in turn, encourage the development of a plurality of exciting and revealing data presentations?

The solution to this problem of redundancy that prohibits plurality lies in a second question raised by the conference call for papers:

> Can we conceptualize machines as users? How can we include application programming interfaces (APIs) in the discussion on digital scholarly editions as interfaces?

As I will try to show in this article and case study, it is only when we are able to separate our idea of a digital edition from the interfaces that display such an edition, and in turn see our interfaces as API consuming applications, that we will be able to accommodate plurality without redundancy. But this, as we will see, involves a dramatic rethinking of the essence of a scholarly edition.

In what follows, I first introduce in abstract what this rethinking looks like and then focus on one attempt to actualize this theory in the case of the medieval scholastic corpus. From here, I will turn to show, through a number of examples of ongoing work, how this paradigm shift helps us solve the problem of redundancy noted above. Here, we will be able to see how this approach enables the efficient construction of interfaces and that, because of this efficiency, a plurality of interfaces can be swiftly constructed in order to satisfy a wide variety of research interests.

2 The *text-as-network* paradigm

The idea that current progress in digital textual editing has moved "barely beyond the book" is the theme of Joris van Zundert's enlightening paper of the same title that bemoans many of the grievances with the print paradigm aired above. He argues forcefully – with the help of Peter Robinson – that much of digital publication has only succeeded in reproducing this print paradigm in a new medium. He writes:

> Most digital scholarly editions, in fact, are all but literal translations of a book into a non-book-oriented medium. Peter Robinson, writing about the distinctions of text-as-work and text-as-document, argues that in the early days of digital editions – roughly until 2005 – scholars would privilege the text-as-work perspective, focusing on the potential of digital technology to express and support the properties of text that construct its meaning. In recent years, he continues, this trend has been exactly reversed. More recent

> digital scholarly editions harness the digital medium rather to represent the text-as-document – the faithful re-representation of a text according to its expression in the physical documents that carry it.…Robinson also notes that many collaborative transcription systems are designed to record text-as-document: not one of twenty-one tools listed in a survey by Ben Brumfield offers the possibility of recording text-as-work. Indeed it is far easier to point to examples of digital scholarly editions that are in essence metaphors of the book, or in other words: translations of a print text to the digital medium, apparently for no other reason than to fulfil the same role as the print text. (Zundert 103–104)

To do something more, we need a fundamental shift away from the *text-as-document* paradigm toward what Zundert and Robinson refer to as the *text-as-work* paradigm, or what I will refer to as the *text-as-network* paradigm.[4]

For those familiar with the history of the web, Tim Berners-Lee, and the recent advance of Linked Open Data (Berners-Lee *Data*), some of this critique should feel familiar. Berners-Lee introduced the notion of Linked Open Data (LOD) by pointing out that the modern web, despite all its success, fails to live up to its true potential. For most people, the current web is a web of *documents*, a web of connected documents to be sure, but just documents. Berners-Lee, with Christian Bizer and Tom Heath, writes:

> Despite the inarguable benefits the Web provides, until recently the same principles that enabled the Web of documents to flourish have not been applied to data. Traditionally, data published on the Web has been made available as raw dumps in formats such as CSV or XML, or marked up as HTML tables, sacrificing much of its structure and semantics. In the conventional hypertext Web, the nature of the relationship between two linked documents is implicit, as the data format, i.e. HTML, is not sufficiently expressive to enable individual entities described in a particular document to be connected by typed links to related entities. (Bizer 1)

Such a bland web, where data is intrinsically enmeshed with presentation (i.e. the hypertext markup language or HTML), makes it impossible for us to fully exploit the inferential capacities of modern computers. Berners-Lee's inspiration behind the proposal for Linked Open Data was the hope of creating a web of *data* (separated from any presentational form) standing behind the web of presentation-oriented *documents* that we regularly encounter on our computers screens. To Berners-Lee's proposal, we

---

[4] Zundert and Robinson are not the only ones rethinking the relationship of a text idea to its material instantiation, and for further discussions along similar lines see Sahle *Mediengebundenheit*; *Transmedialisierung*.

should also note the current and ongoing work of Sarven Capadisli to promote and actualize the notion of Linked Research and to escape the research isolation inherent in the *text-as-document* paradigm (Capadisli). This work, in turn, continues to pursue Ted Nelson's dream of a linked global corpus as described in *Literary Machines* and before him the idea of the *Memex* machine put forward by Vannevar Bush.

The path towards a true scientific and holistic understanding of the scholastic corpus requires a similar paradigm shift. But it should be repeated once more that this shift has very little to do with a simple shift from the printed page to a website. This is merely a shift from one medium to another within the same paradigm. The shift in question involves a radical reconsideration of the essence of a text.

During the course of several years working on semantically encoded text editions, my own conception of what I am doing as a textual editor has dramatically shifted. Instead of seeing myself as engaged in the task of creating pages of text, I have come to see my primary task as one of identifying discrete data points and then documenting both the data types of these data points and the relationships between them. The editing of a paragraph, therefore, is not just the creation of a visible paragraph, but the recognition of a node that is the third child of the second section that has three preceding siblings and two following siblings. And of course, this tree structure is only one kind of networked representation. The same nodes can be re-used in other networks that can explore other kinds of relationships. For example, it is possible to recognize that the content of this node is a commentary on the content of another node in another commentary, which itself is a commentary on another text. In this way, a network can not only track the position of a node within the historical linear text, but also track the position of nodes within a common discussion taking place in multiple texts over centuries. As Berners-Lee remarked in his TED Talk on Linked Data: "Data is relationships". Thus, as a textual editor, I am first and foremost in the business of identifying and describing relationships between identifiable text parts.

It should be kept in mind that most of our modern computer applications work directly against such a realization. The common skeuomorphism of showing the "page" as we write in Microsoft Word is only the most obvious example of how our digital applications are still quietly and subtly forcing us to think about a text as a "thing" that lives on a page rather than as a network of related but highly diverse data types. Zundert describes this same phenomenon as "paradigmatic regression" exemplified in the metaphors used by GUI applications. He writes:

> In order to help the user understand a new target domain or a new paradigm, it is expressed by way of a conceptual domain or a paradigm that is already known to the user. An obvious example is the metaphor of the desktop, which was used to communicate the functions of the PC to as broad an audience as possible. The only trouble is that such metaphors are necessarily incomplete

> as they conceal both the good and the bad of the deeper computation model. (Zundert 86)

Given the pressure exerted by most document composition software, it will require an extra intentional effort to begin to think differently. This reconceptualization will only occur if we force ourselves to think about texts as data first, divorced from the page, book, or any visual representation of this data. The shift away from the *text-as-document* to *text-as-network* requires a revolution in how we think about our texts, our subsequent publication of this data, and the material artifacts (manuscripts, printed books, and digital display applications) that exists as temporary carriers of this data.

For example, the publication of a text edition should not be identified with the publication of a book or website or anything that a reader will encounter directly, precisely because this kind of presentation is already a derivative "representation" of the underlying network of relationships.[5] Again, as Zundert notes, this demand seems to directly contradict current common practice:

> Current reality, however, is very different. In textual scholarship, Internet nodes are mostly placeholders that point via a URL to a digital document or to a digital edition as a whole, as a data silo. The edition of the Van Gogh letters, for instance, sits at the node identified by http://www.vangoghletters.org/vg/ as a fully integrated and monolithic pile of edited text from letters; the pile includes comments, annotations, translations and so on. The finest granularity presented to the network of the web is at the level of the individual letter (e.g. http://vangoghletters.org/vg/letters/let043/letter.html). Even that URL identifies a compound object, that is, a meaningful set of multiple scholarly objects: two facsimiles, a transcribed text, annotations, bound together by an interface that…represents an editorial argument about what constitutes the digital scholarly edition of this particular letter. According to this argument, there is no need to address the transcription, the facsimile, a particular annotation, in isolation. Most of the digital scholarly editions on the Web are expressed similarly. It is hardly better than a network of nodes in which each node represents a particular edition that is offered as a PDF. This situation renders it impossible to address texts (and thus editions) beyond their graphical interface in ways compatible with a hypertext model. (Zundert 101)

In contrast to this reality, it is imperative that the publication of a digital scholarly edition should coincide, first and foremost, with the publication of a granular dataset

[5] On this point, see also Sahle's notion of "transmedialization" (see Sahle *Mediengebundenheit).*

that reveals the underlying logic of the text network and is made accessible according to the best practices of the field and Linked Open Data. Thus, contrary to the practice of the Van Gogh edition as described by Zundert, resource identification should extend much further than to a web*page* – emphasis is placed on the word page to note the "paradigmatic regression" to be fought against. As the web*page* is itself just a unique compilation of a number of distinct resources, each of these resources should be identified and published as distinct nodes along with the relationships between these nodes. Each component – at the very least the idea of a paragraph, a division, the transcription of a paragraph, the image of a paragraph, etc. – should be a de-referenceable node that exists independent of any particular interface. Only with this kind of separation can the interface *page* referenced by Zundert be seen as a unique argument about how best to visualize the network of relationships between these text nodes. Moreover, only this separation will allow rival interfaces – a welcome and positive plurality – to make counter arguments about the proper representation of the logic inherent in the text network.

3 Editing and publishing via the *text-as-network* paradigm

The *Scholastic Commentaries and Texts Archive* (SCTA) (Witt et al.) is one example of an attempt to actualize the aspirations outlined above.[6] The medieval scholastic tradition is a rich, vibrant, and highly influential corpus of philosophical and theological material. As noted above, the corpus is enormous, complex, and interconnected in complicated and fascinating ways. By connecting editorial work on this corpus into a global network of data, we will ultimately be able to gain a holistic perspective on the entire corpus. A prime example of the kinds of texts that belong to the scholastic tradition are the medieval commentaries on the twelfth-century book, the *Sentences* of Peter Lombard, known as *Sentences* commentaries.[7] Lombard's *Sentences* quickly became the preeminent theological and philosophical textbook in the high and late Middle Ages. Many of the greatest intellectuals of the Middle Ages wrote commentaries on this common textbook. Consequently, these commentaries constitute an enormous corpus that serves as a critical witness to the history of medieval philosophy and theology. Today, we know of approximately 1,000 such commentaries written from the 12th to the 16th century, each typically ranging from 1,000 to 3,000 pages in modern printed form. Further, *Sentences* commentaries only begin to scratch the surface of the wider tradition. To these should be added biblical commentaries, commentaries on Aristotle, *summae*, quodlibetal questions, logical treatises, and many other types. The SCTA is an attempt to publish this corpus

[6] For a discussion of the early inspiration behind the SCTA, see Witt *Sentences*.
[7] For an introduction to Lombard's *Sentences*, see Rosemann.

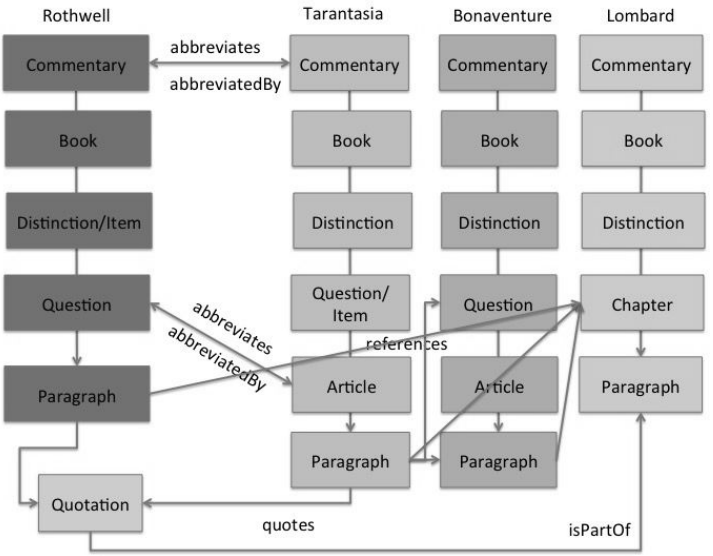

Figure 1: Example of Commentary Text Hierarchy and Relationships.

first and foremost as a data network that, when published according to an open and documented set of standards, can be endlessly re-purposed for an infinite number of purposes.

A critical part of this data organization and publication lies in the development of field standard data models that can make this data accessible in predictable ways to data consuming applications. In the context of the scholastic corpus, we face two major challenges when trying to model it in a presentation agnostic way.

First, we face the problem common to almost any critical project: namely, describing its generative history from its inception to its modern reception. This history abounds with distinct but highly related resources that an interface needs to be able to navigate in order to allow a user in turn to navigate massive amounts of data in an intelligible and citable way. We need to be able to identify and navigate between manuscript versions as well as early modern and contemporary printings of the same text.

Second, we face a problem perhaps more unique to scholastic philosophy and theology. This is the problem that scholastic texts constitute a highly intertextual corpus of non-linear texts. That is to say, every text is in some way making reference to discrete parts of other texts. As researchers, we need the ability both to display the traditional historical hierarchy, but also the flexibility to dynamically construct new

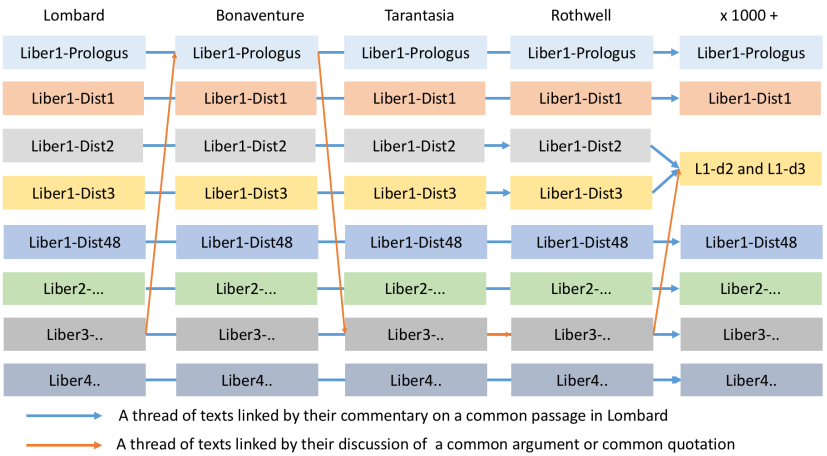

Figure 2: Example of Multiple Text Hierarchies in the Commentary Tradition.

hierarchies based on unique citation patterns or investigative queries. Below are just two examples that illustrate this latter need.

Figure 1 illustrates the fact that each commentary itself is a composition of an elaborate text hierarchy. Each hierarchy contains within itself hundreds if not thousands of relations to other text hierarchies. Further, these relations do not necessarily run through parallel units of the different commentaries' text hierarchies (e.g. from paragraph to paragraph), but can zigzag and crisscross from one level in one commentary to a completely different level in another commentary (e.g. a small paragraph in one commentary can reference a much larger distinction in another commentary or a medium sized article – containing multiple paragraphs – can be abbreviated by a single paragraph in another commentary).

Figure 2 illustrates the fact that because many of these texts exist within a larger commentary tradition, researchers often have an interest only in particular sections of any given commentary: e.g. a small identifiable section that discusses a common theme or argument throughout the history of the commentary tradition. Thus, we need the ability to categorize granular text units within a text's larger hierarchy as belonging to a thematic discussion. In this way, researchers can easily request, and systems can construct, new text hierarchies created from selecting particular text units from the entire corpus and arranging these units in a manner best suited to the research question at hand.

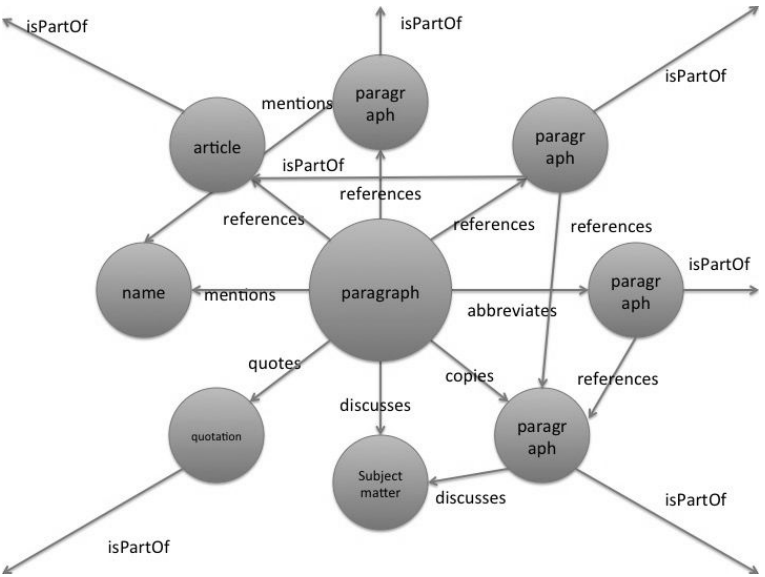

Figure 3: Simplified Example of a Text Network.

To solve these challenges, and in order to create a truly critical medieval scholastic corpus, the SCTA aims, first and foremost, not to publish a website, but to publish a dataset as RDF triples, such that every connected concept in the corpus has a de-referenceable ID through which it can be annotated with subsequent properties or annotations and linked to other resources. The result is something like the purposively simplified web seen in figure 3.

To solve the first challenge of modeling the generative history of our corpus, we have designed a model based off the Functional Requirements for Bibliographic Reference (Working Group on FRBR/CRM Dialogue) model that has been significantly modified and extended.[8]

[8] For more information on FRBR, see Noerr et al. as well as Bennet et al. This approach shares similarities with, and indeed is partially inspired by, the model underlying the service known as Canonical Text Service (CTS), currently employed by the Homer Multitext (Blackwell and Smith) and Perseus. It is also an area of active development, evidenced by the ongoing efforts to update CTS to what is currently being called Distributed Text Services (DTS) (Almas et al.). Because there is so much active thinking and theorizing currently underway, the SCTA ontology offers us a chance to deviate slightly from existing models in order to meet the unique demands of the scholastic corpus. In the future, the SCTA will be working, wherever possible, to align this ontology with the recommendations of the CIDOC-CRM FRBRoo model, while preserving its ability to adapt the model to meet the unique demands of the scholastic corpus. For a short write-up explaining a few of the reasons for our deviation from the CTS

The basic and most fundamental classes within this model are the following:

- WorkGroup
- Work
  E.g. the idea of Moby Dick
- Expression
  E.g. the idea of Melville's Expression (as opposed to a screen play Expression)
- Manifestation
  E.g. the idea of the 1959 edition of Moby Dick
- Item
  E.g. One physical copy of the 1959 edition in a particular library
- Transcription
  E.g. the idea of a digital transcription of the 1959 edition of Moby Dick; includes properties like hasXML, hasJson, hasPlainText, hasHtml

Very briefly, we begin with a concept of WorkGroups that can contain other WorkGroups. But WorkGroups can also contain what FRBR calls Expressions which get us closest to the idea of the book or text we are generally familiar with. This might be the Expression of *Moby Dick* that Melville wrote or the Commentary on the *Sentences* that Thomas Aquinas wrote.

Expressions can then have Manifestations, which are roughly equivalent to the idea of various editions that have survived. A Manifestation might be the idea of a manuscript, or the Venice 1505 printing of a text, or the 1959 version of *Moby Dick*. A Manifestation has instances called Items and these Items live in physical spaces like a library.

To the FRBR model, we also add the concept of a Transcription. A Transcription is the idea of a digital representation of any Manifestation. It is not yet a file or file format, but a Transcription alone can take properties like hasXML which points to wherever the XML serialization of this transcription exists on the web, accessible via the HTTP protocol.

Finally, while the Manifestation is not yet something we can take a picture of, it is the idea of something physical. Thus, we can also create IDs for the idea of a "Manifestation Surface", that is, the idea of folio 1 recto, the idea of folio 2 verso. These Surfaces can then be connected to the "Item Surfaces" belonging to an actual physical codex possessed by a library. It is from here that we can make important connections to related web resources in a different but related model. This is the concept of a *IIIF Canvas* defined by the *International Image Interoperability Framework* (IIIF) API (IIIF Consortium). This API is widely adopted by world libraries for making images of

vocabulary and re-adoption of the FRBR vocabulary, see Witt (*Modelling*). This post originated as part of the discussion within the DTS working group about various modeling ontologies and became an argument for why the SCTA sees FRBR, with some modification, as an ideal modeling vocabulary.

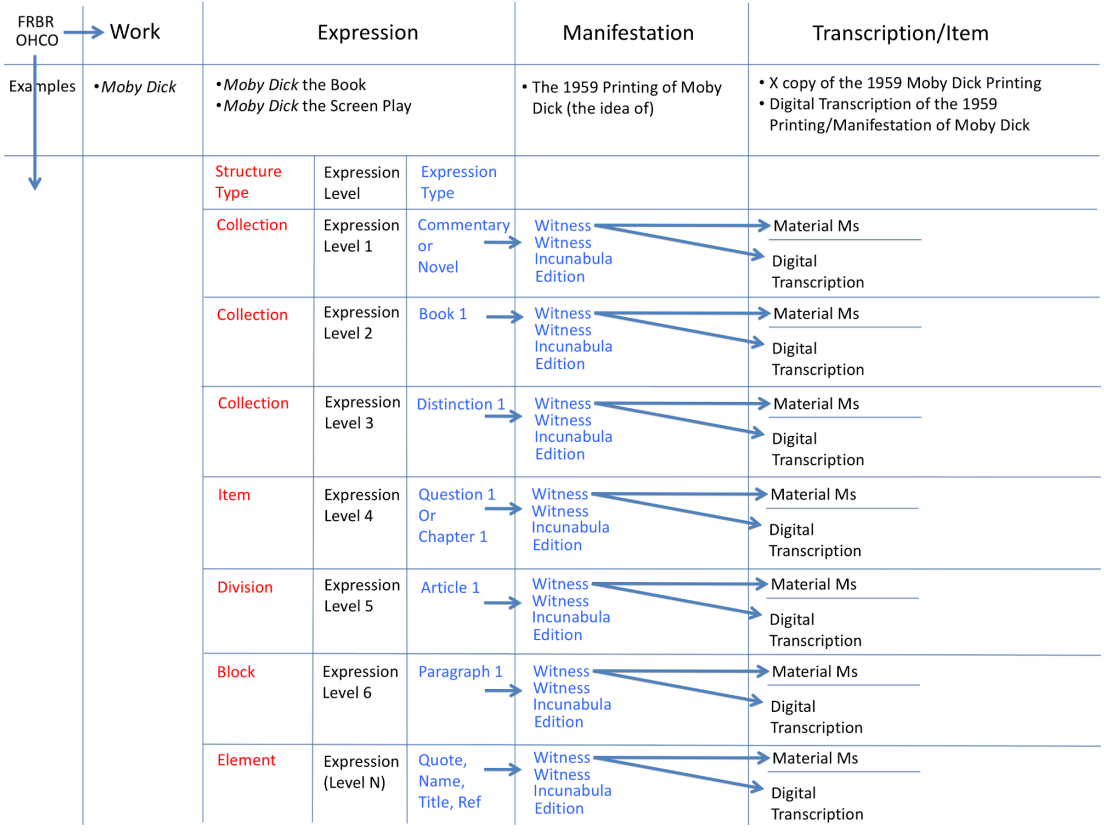

Figure 4: Text Network Matrix.

cultural heritage resources (in this case, manuscript folio images) available on the web.[9]

The basic classes used to connect out from the idea of a Manifestation Surface to the concept of a IIIF Canvas are as follows:

- Manifestation Surface
  E.g. the idea of page 1 in the 1959 edition of Moby Dick
- Item Surface
  E.g. the physical page 1 in a particular copy of the 1959 edition of Moby Dick
- IIIF Canvas
- IIIF Image Annotation
  E.g. images taken of the physical page 1 in a particular copy

The SCTA's connection to the IIIF model and API means that a consuming application will be able to seamlessly move from a fragment of text to a related image hosted somewhere completely different on the web.

This, however, only models the generative history of our corpus, it does not help us with the tangled web of internal and external connections within the deep hierarchy of any given Expression. To meet this challenge, we break down our corpus vertically as well as horizontally. In other words, we conceive of each text as an *Ordered Hierarchy of Content Objects (OHCO)*,[10] and accordingly create resource IDs for every division within the document hierarchy down to the individual paragraph and quotation level. Each resource, no matter how small within the hierarchy, gets further linked to the FRBR model as an Expression, Manifestation, Item, Transcription, etc. And each level of this hierarchy can be further annotated, so that, for example, we can identify any point in the hierarchy as an instance of a "Prologue". With these annotations, we can create new paths that slice through the corpus, taking only a cross section of a relevant section from each commentary or text. The result, as seen in figure 4, is a complicated matrix of relationships.

A similar matrix exists for Surfaces and their connections to each part of the text as seen below in figure 5. In this matrix, one can see that we have IDs not just for the image facsimile or the IIIF Canvas, but an ID for the Manifestation Surface, which itself is not something that one can take a picture of because it is still just an idea. From here we can link out to the actual physical Surface (an Item Surface) found in a particular Item (e.g. a single printed book). While, in the case of manuscripts, this relationship will be one-to-one, the conceptual separation of Manifestation Surface and Item Surface is particularly important when dealing with incunabula and printed books. In such cases, the Item Surface, which is a particular realization of the Manifestation

[9] For a lengthier discussion of the model used for connecting Surfaces, Canvases and Zones to text Manifestations, see Witt *Surfaces*.

[10] For an early discussion of OHCO, see DeRose.

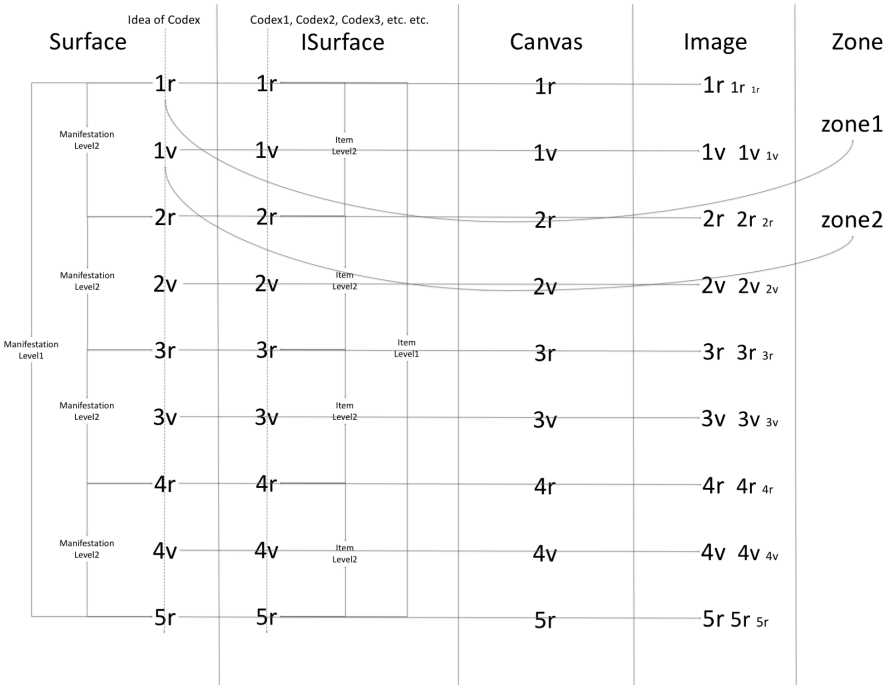

Figure 5: Text Network Surface and IIIF Canvas Matrix.

Surface, may have received unique written marginal notes, and we need to be able to distinguish this Surface from other Item Surfaces that realize the same Manifestation Surface. In such a case, we want to be able to link out to the IIIF Canvas ID for each unique Item Surface while still being able to group all of these Item Surfaces under the idea of a given Manifestation Surface. The matrix in figure 5 further shows how these Manifestation Surfaces can be identified with various parts of the Manifestation text hierarchy. Such connections ultimately allow interfaces to navigate between two overlapping hierarchies: the conceptual hierarchy of the text (e.g. books, chapters, sections, paragraphs, etc.) and the material hierarchy of the historical material carriers of these texts (e.g. codex, quire, bi-folio, folio, recto-verso, column, etc.)

Finally, with these models established, the SCTA can automatically construct the actual dataset by simply standardizing much of the work that critical editors are already doing. Even when an editor is focused simply on preparing a print edition, they are still engaged in precisely the kind of data creation needed to construct this data set, such as identifying manuscript witnesses, identifying structural units, and

developing transcriptions for each of these units. When these basic pieces of raw data are encoded according to a common field standard, such as a customization of TEI like the LombardPress-Schema (which the SCTA currently supports), we can automatically construct the data set described above. The result of this standardization means that the SCTA build script can crawl this data and construct the text network according to the models outlined above. The output of this build script at the present time is a database of more than eleven million triples and a fully indexed, searchable corpus of over ten million words.

4 Fulfilled promises: development efficiency and interface plurality

Because the SCTA publishes its texts first as a connected data network rather than as a text tightly bound to any presentational interface, such as a page, it is publicly available to anyone and any client that knows how to consume the API or query the public SPARQL endpoint. This allows the data to be efficiently reused in a number of different ways.

4.1 Common libraries

The first place we can see the payoff of this kind of data-first publication is in the newfound ability to build common code libraries that can be reused by interface developers to quickly navigate and query the resulting dataset.

One such library is the lbp.rb library. In keeping with the overall emphasis on decoupling distinct components, it is important to note that this library is a conceptually distinct component from the SCTA dataset. In other words, it is possible for many different libraries to be written that aim to consume this dataset in various ways. Another library might be written to access the network in a different way for a different purpose. It is for this reason that we do not refer to this as the SCTA ruby library, but the "lbp" (LombardPress) ".rb" (ruby) library, which means this is a particular library designed by the LombardPress project to make use of SCTA data, implemented in the ruby programming language. Other groups interested in alternative languages can develop different libraries for other purposes and in alternative programming languages (such as Python, R, or javascript). In this way, we promote plurality. Yet, we also combat redundancy because, if the lbp.rb library works for a programmer's present purpose, she has no need of re-writing the library. She can simply adopt the existing library and move on to the next programming task, saving a considerable amount of time in the process.

It is precisely this aspect of re-usability that makes many of the interfaces discussed below possible. The lbp.rb library is used repeatedly in many of these interfaces.

Because each interface can adopt this common library, developers can avoid redundant coding tasks and develop the specific interface more efficiently.

4.2 Scta.info

Scta.info is a specific kind of interface designed to visualize the logical connections that constitute the text network. In keeping with the paradigm shift described above, this visualization could be considered the real heart of the digital scholarly edition. Scta.info is a very simple interface that visualizes the logic of the text network in tables. This interface is designed for a particular purpose. Its primary purpose is to be used by subsequent interface designers in order to become familiar with the logic of the network so that they can exploit this logic when they build different interfaces.

4.3 LombardPress-Web

The LombardPress-Web Application is an interface that makes extensive use of the lbp.rb library and is primarily designed to show critical editions of scholastic texts within the context of diplomatic transcriptions and manuscript images. It is also the primary and flagship application designed to show how independent clients can make use of the SCTA SPARQL endpoint, the lbp.rb library, as well as the IIIF API to display texts and images that are distributed throughout the web.

The LombardPress-Web application is a "dumb client", and it is important that this should be kept in mind. By "dumb" I mean simply that the application is completely agnostic to the data it displays. No text files or image files reside (or better, are *siloed*) on the server of this application. Nothing particular to this set or genre of texts is part of the hard code. Rather, it has been designed to understand a particular API and data model. Accordingly, any project that publishes their texts as a data network following the schema outlined above could reuse this client to view their data. The key take-away from this design is that not only is the data reusable in other clients, but the client, when de-coupled from a particular dataset, becomes reusable for a variety of datasets.

A quick tour of the LombardPress-Web application will illustrate some of the ways a client can exploit the logic of the text network.

Multiple text hierarchies

In figure 6, we see the display of the traditional text hierarchy for a given commentary. The client parses the URL for the RDF ID of the commentary in question, and then queries the SCTA SPARQL endpoint for the information it needs to display a basic table of contents.

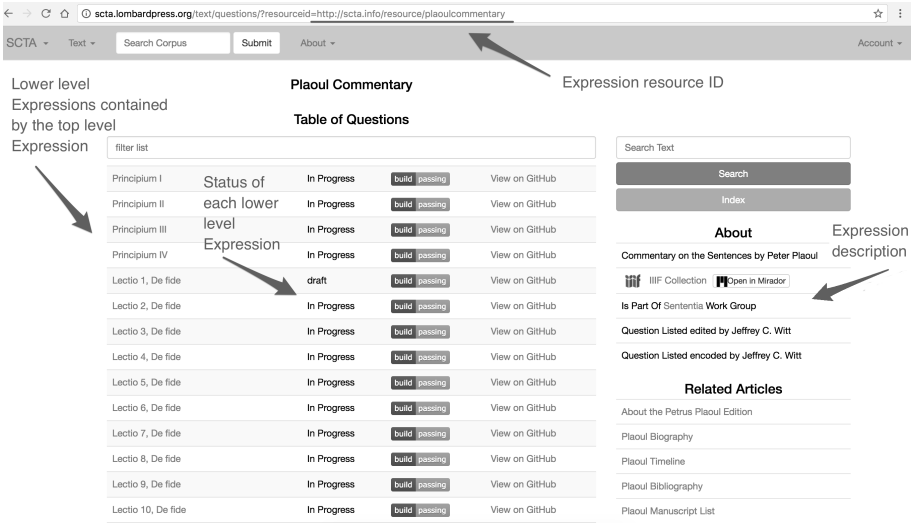

Figure 6: Traditional Text Hierarchy Display.

Figure 7 illustrates the way in which we can use this same interface to display non-traditional, non-linear hierarchies. In this case, the RDF ID given as a query parameter no longer corresponds to a Work or Expression, but to a specific category of a text part (what we call an ExpressionType) that occurs repeatedly within all commentaries of this type. In this example, the ID refers to the prologue of book 1 of all commentaries on the *Sentences* of Peter Lombard. Generally, each commentary contains several questions that fall within this type of text part and they all discuss themes of faith, theology, and science. When using this ID, the client creates a new text hierarchy and the user sees a new text that is effectively the result of cutting out sections of hundreds of texts and then arranging those sections according to a specified order (such as date or author name). In this way, a user has, in seconds, isolated a focused historical discussion that took place over centuries and until now was hidden within a multitude of much larger texts.

Negotiating multiple text manifestations

When a text part is selected, the client uses the SCTA database to locate an XML serialization of this section somewhere in the wider web, retrieves it, and displays it for the user in a fairly traditional way, as a document with an *apparatus fontium* and *apparatus criticus.* This visualization can be seen in figure 8.

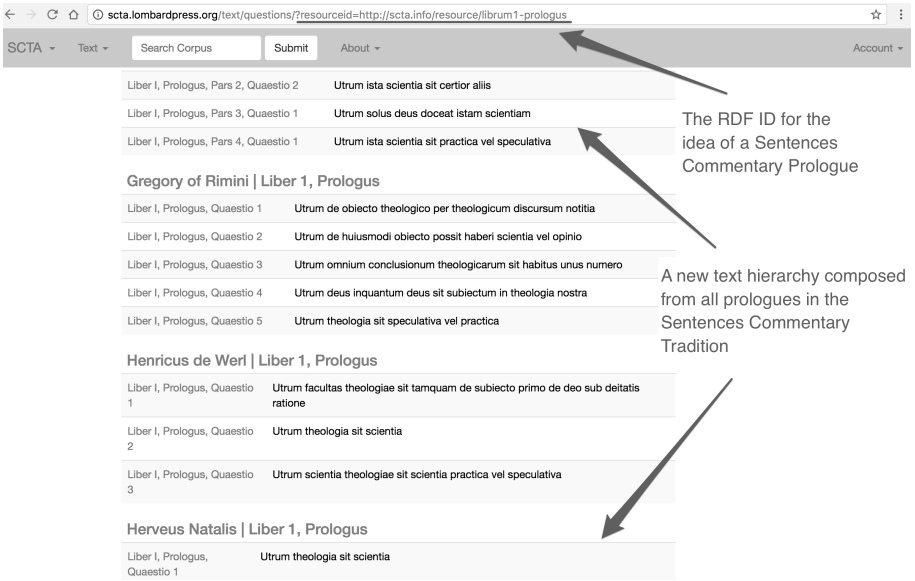

Figure 7: Dynamically Constructed Non-Traditional Text Hierarchy Display.

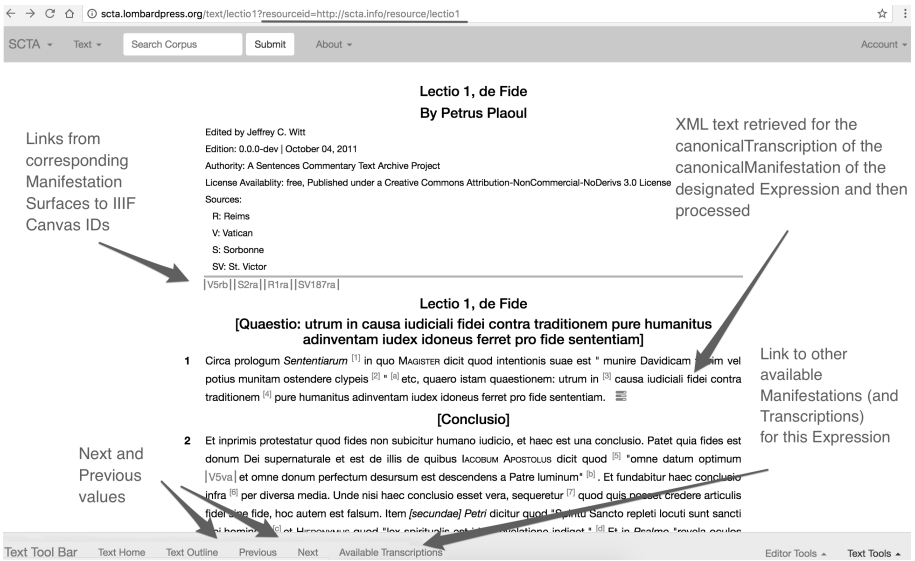

Figure 8: Critical Text Display.

But even the traditional display in figure 8 is a good example of how the client is using the known network of relations between the various Manifestations of a given Expression to offer the user a default view of the text. In this example, the user only gives the client the ID for the abstract Expression of the Work. Without any prior knowledge of the text, the client uses the properties on this Expression to find the canonical or default Manifestation (from available Manifestations) of an Expression and then looks for the canonical or default Transcription of the canonical Manifestation.

If there is no critical text, the client will use the SCTA metadata to simply default to the best diplomatic transcription currently available. This ability to display whatever version of the text is currently available, while having the flexibility to modify or replace the canonical Manifestation when a new and better Transcription becomes available, has been tremendously useful for expanding the corpus quickly with what is available while also making room for improved quality over time. Because we have an ID for everything, we can take transcription contributions from everyone, no matter if they are just working on a single manuscript or a tiny section of a much larger text. While it may be extremely difficult for a novice or student to construct a perfect critical text, they may be able to produce a usable diplomatic transcription. Because we have a place for this diplomatic transcription, we can include it and make it available for use without having to wait for a perfect critical edition to be completed. For the time being, this usable edition will help provide search results and enable discovery. The ability to discover topics and themes in a text is the key to generating more interest. Later, as interests grow because of the early availability of these usable editions, other diplomatic transcriptions can be added and, eventually, a superior critical transcription can replace it as the canonical Manifestation.

Interacting with granular text units

Because we have IDs for every level of the document hierarchy, we can also use the network of relations (between Expression, Manifestation, and Transcription) to create functionality at granular levels. This can be seen, for example, in figure 9, where, for every paragraph Expression, we can request on-demand collations from the available Transcriptions of available Manifestations. When the request for a collation is made by the user, the client queries the SCTA endpoint for available Manifestations and populates two lists: a drop-down list for the base text and a drop-down list for the text version to be compared. When the selection is made, the client requests the canonical Transcriptions for the selected Manifestations, runs the collation algorithm, and returns visualized results.

Similar functionality is possible with respect to the available digital facsimiles for a given paragraph Expression. Figure 10 shows just one way a user can experience such

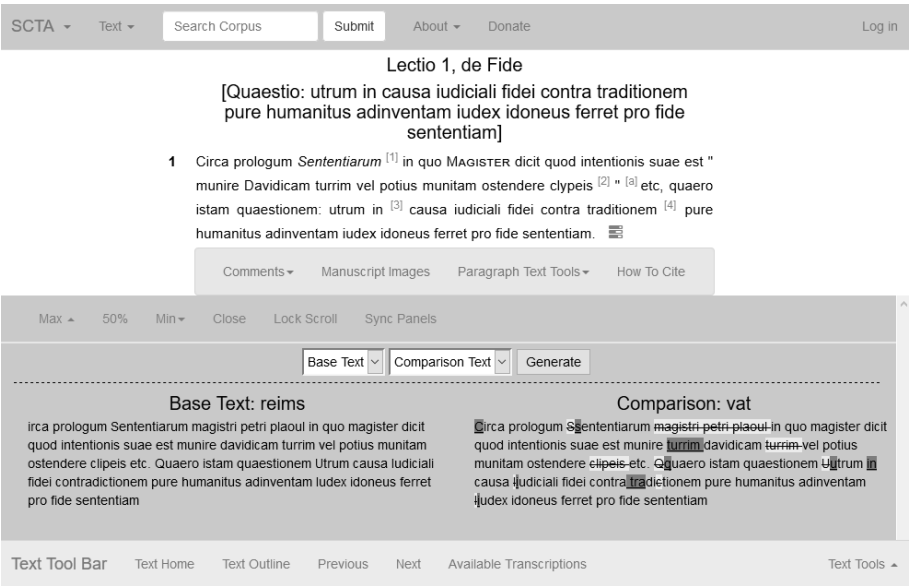

Figure 9: On-Demand Collation for Text Parts.

images. When a user asks to view the images for a given paragraph, a list is created – again with data from the SCTA database – of available Manifestations. When a particular Manifestation is selected, the SCTA is queried for which Zones belong to this Manifestation of this paragraph and which Surfaces these Zones fall on. From these Manifestation Surfaces, the query reaches out to the canonical Item Surface (or ISurface) and from here a connection to the IIIF Canvases minted by the holding library can be made. Finally, the IIIF Canvas leads to a URL at the hosting library where the actual image can be requested. Using the coordinates for the Zones in question and the IIIF Image API, only specific coordinate regions of the image are requested from the library server in question. The results of this image query are presented to the user as the image of one Manifestation of the target paragraph as seen in figure 10. If the user selects a different Manifestation, this query is repeated and a new image is requested from a different library server.[11]

[11] Other kinds of granular functionality can be added as well, such as commenting and annotating text sections. See my post on how we are attempting to aggregate distributed discussions of a common resource using LinkedData Notifications (Social Web Working Group) (Witt *Linking*).

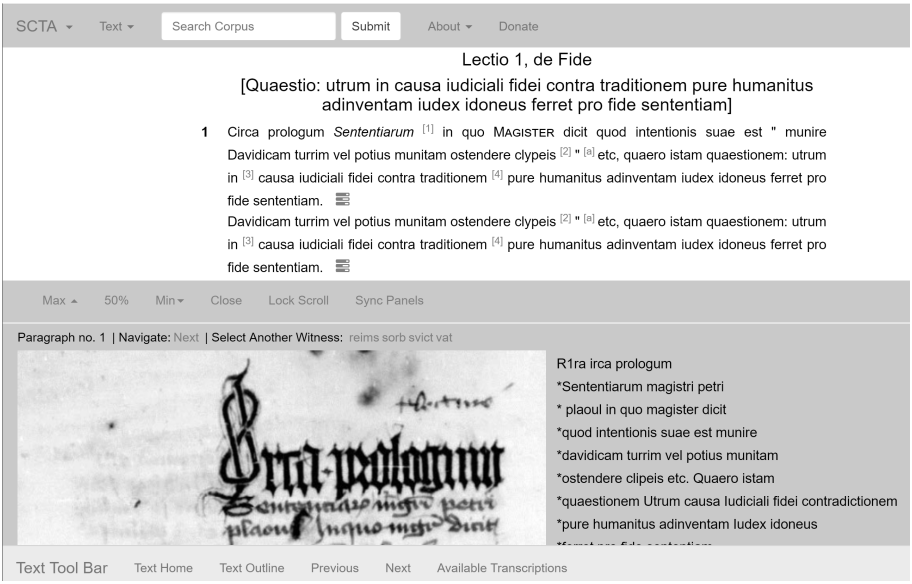

Figure 10: On-Demand Image Display for Text Parts.

Crawling the network

Finally, the client can exploit the many kinds of interconnections between texts and crawl the corpus in a non-linear way. Figure 11 offers a simple example of much grander possibilities. In this example, the user is reading the target passage of a given commentary. Upon requesting information about this passage – which is actually a request by the client interface to the SCTA database for more information about this resource – the client alerts the user to the fact that this small paragraph is actually an abbreviation of another text part in an entirely different commentary. The user can request to see this other text part without navigating away from the target text and can now read the abbreviating paragraph in the context of the text part it is abbreviating. This same kind of comparison, or contextual reading, can be extended to other kinds of text relations, such as *copies*, *references*, *quotes*, *discussions*, or, more generically, *isRelatedTo* as well as the inversion of these relations such as *isCopiedBy*, *isReferencedBy*, *isQuotedBy*, or *isDiscussedBy*.

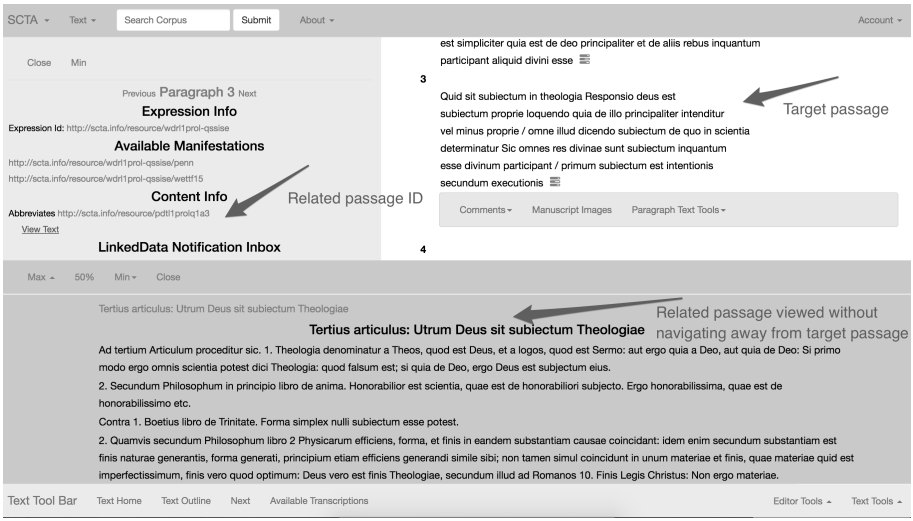

Figure 11: Navigation of Related Text Hierarchies.

Mirador

The LombardPress-Web client discussed above is an example of an interface designed for a particular purpose. For those people interested in displaying their data for such a purpose, we have severely reduced redundancy because there is now no need to reconstruct such an interface. Their focus, instead, should be on preparing their data according to the field standards and recommended data models so that it can be automatically displayed by a reusable client.

However, while this text-focused interface might be the preferred interface for some research questions and activities, it might not be the best interface for other concerns. There may be other research needs where the images of the manuscript witness should be front and center. If creating an alternative viewer required that the dataset be re-produced and that a new storage solution be created, this would be prohibitively expensive. But if the data can be reused, it becomes easy and trivial to offer alternative views.

This is precisely the case in the combination of the generic IIIF compliant viewer, called Mirador, with the SCTA dataset. Mirador, like LombardPress-Web, is a "dumb client" that is designed to understand a particular API, namely the IIIF API. Unlike the LombardPress-Web application, which is text-focused, Mirador focuses on images, or a digital representation of the material artifact, and views the text as a kind of annotation. If the scholastic corpus data were somehow welded to the LombardPress-

Figure 12: Mirador and IIIF Collection Display.

Web application, automatic reuse of this data for display in the Mirador viewer would be nearly impossible. However, because the SCTA is a separate public data service and the LombardPress-Web interface is just another dumb client, we can easily repurpose the same data to be exploited by a different interface. Figures 12, 13, and 14 offer an illustration of this data reuse.

Figure 12 shows how IIIF collections can be dynamically constructed from the SCTA concept of WorkGroups, Works, Expressions, and Manifestations – the very same data used by the LombardPress-Web application described above. Upon entering the interface, a top level IIIF collection is created for every author in the collection. Each author collection includes a sub-collection that corresponds to every Expression written by that author. When a user selects a particular author, a custom IIIF manifest is created for every Manifestation for every Expression attributed to that author. Each IIIF manifest includes the IIIF canvases for the folios or pages of the manuscript or book that correspond to Manifestations for every Expression attributed to the author in question. When a user selects a particular Expression, they see in return only IIIF manifests that correspond to the Manifestations for this particular Expression.[12]

Figure 13 shows us how the data used to create a basic table of contents for a text in figure 6 can be reused to create a table of contents for a specific manuscript.[13] Figure 14 illustrates how the same diplomatic transcriptions used to create on-demand

[12] For a more detailed discussion of this kind of reuse of SCTA data, see Witt *Manifest*s.

[13] Another example of the reuse of this structural metadata by a completely separate client can be seen in the *Sentences Commentary Catalogue* (RCS) maintained by Ueli Zahnd at the University of Basel. For a write-up and description of this reuse, see Witt *Dataset*.

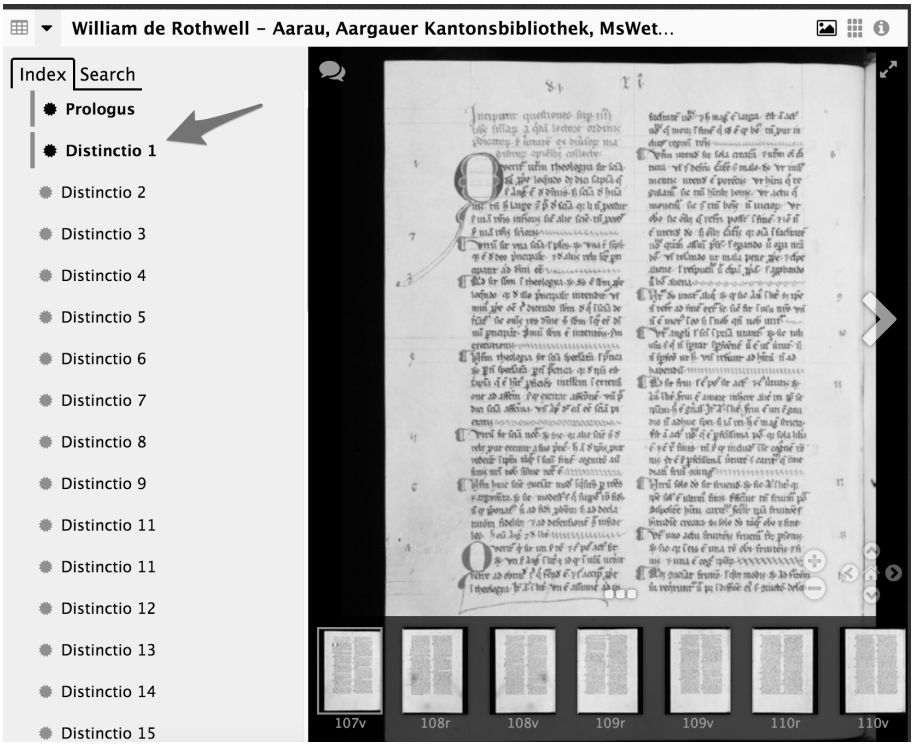

Figure 13: Mirador and Reuse of Text Hierarchy Data.

collations in the LombardPress-Web application seen in figure 9 can be used as the data for the navigation of a manuscript via search results in the Mirador application.

Once more, what it is critical here is the fact that none of this data has been recreated. It is identical to the data seen in the other viewer, it has simply been repurposed by a different interface. The data being produced for one visualization, because it has been decoupled and published separately from this visualization, is all we need to quickly and efficiently build rich manuscript viewing environments.

LombardPress-Print

Finally, because buy-in to the importance of the shift from *text-as-document* to *text-as-network* is so often hindered by those who remain exclusively interested in a codex manifestation of their editorial work, we need to do more to explode the false binary

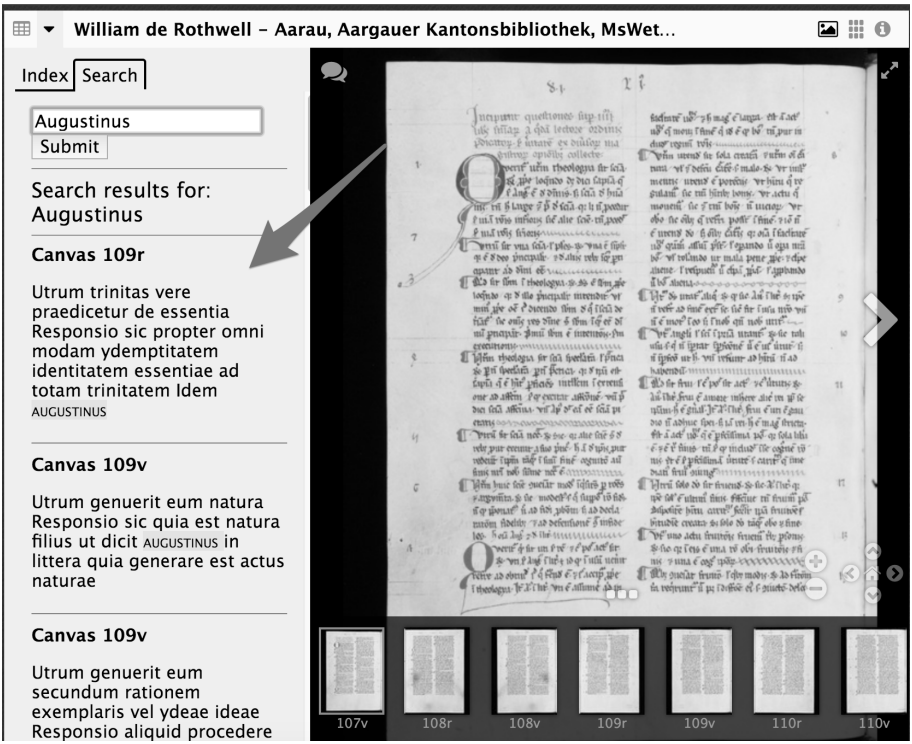

Figure 14: Mirador and Reuse of Transcriptions as Manuscript Search Results.

between creating a digital edition or creating a physical book. This is a false choice. It is imperative that we understand that a printed book is just another interface that can and should be seen as separable from the underlying data-network.

Thus, I have made a prototype command line tool that allows a user to select any ID from the SCTA database and run a PDF conversion with a single command. This tool, once again leveraging the power of the common lbp.rb library, is able to access a text fragment of any part of the corpus from the cloud (that is, distributed anywhere on the web) and deliver a camera-ready print visualization. As seen in figure 15, a user does not need a prior knowledge of where on the web the source XML file is. She needs only the RDF ID of the text fragment in question. The command line tool will take this ID, crawl the text network, discover the source file, perform the conversion, and return the print-ready output.

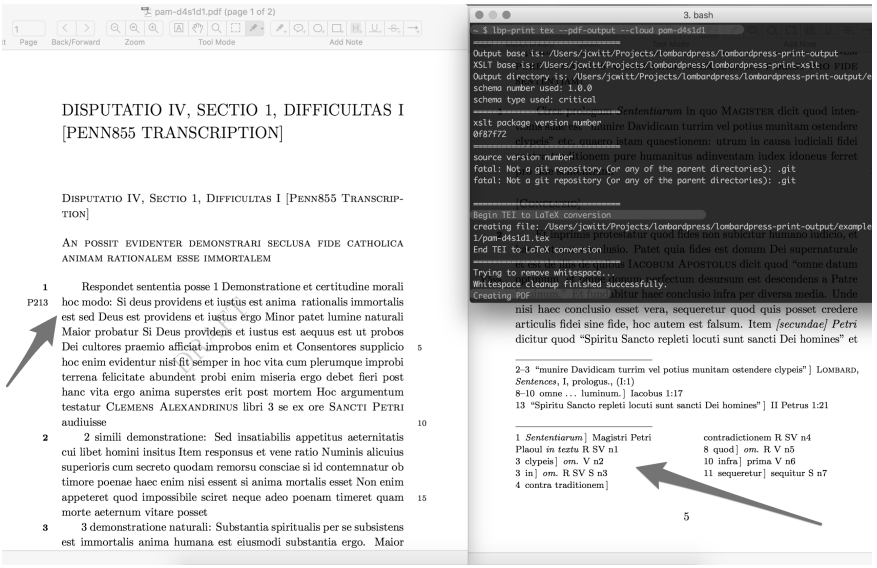

Figure 15: Data Reuse in Automatic Print Display Creation.

4.4 Conclusion

In closing, I return to the original question: how can we promote interface plurality without redundancy? This article is an attempt to show that the answer to this question is already technologically possible. The greater challenge is overcoming the social inertia of an academic culture that, despite protestations to the contrary, tends to be content with the existing paradigm: a paradigm that insists that a text is fundamentally a thing that belongs in a book and that any digital representation of this book should re-create this paradigm as closely as possible.

The hope behind the efforts described above is to offer enough early glimpses of what becomes possible when we understand our texts as networks of various interrelated data types so that scholars will begin to embrace this new way of thinking in mass. Such a transition will be difficult and will require scholars and researchers to learn new things rather than simply shrug off the tasks of data type recognition and encoding as someone else's job. But this mental adjustment is a non-negotiable. If we want to achieve results such as those described above on a massive scale, a simple transition from print documents to a web of documents is not enough. In such cases, we will end up pouring money and resources into digital environments that take us nowhere. Instead, we must see, understand, and publish our texts first as

ideas divorced from any material presentation. Only then will we be able to efficiently and cost-effectively pursue a plurality of innovative interfaces that truly advance the pursuit of historical knowledge.

Bibliography

Almas, Bridget, et al. "DTS: Distributed Text Services." Github, 2016, github.com/distributed-text-services. Accessed 18 June 2017.

Bennett, Rick, et al. "The Concept of a Work in Worldcat: An Application of Frbr." OCLC Office of Research, 2003, www.oclc.org/research/publications/archive/2003/lavoie_frbr.pdf.

Berners-Lee, Tim. "Linked Data." W3, 2009, www.w3.org/DesignIssues/LinkedData.html. Accessed 18 June 2017.

—. "The Next Web." TED, 2009, www.ted.com/talks/tim_berners_lee_on_the_next_web. Accessed 18 June 2017.

Bizer, Christian, et al. "Linked Data - The Story So Far." *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, 2009, pp. 1–22. doi:10.4018/jswis.2009081901.

Blackwell, Christopher, and Neel Smith. "CTS: Canonical Text Services." *Homer Multitext Project*, 2012, www.homermultitext.org/hmt-doc/cite/texts/ctsoverview.html. Accessed 18 June 2017.

—. *Homer Multitext Project,* 2012. www.homermultitext.org. Accessed 18 June 2017.

Brinzei, Monica, et al. *Jacobus de Altavilla.* Universitatea Babeș-Bolyai, 2014, www.jacobus-dealtavilla.ro/contact.html. Accessed 18 Aug 2017.

Bush, Vannevar. "As We May Think." *Atlantic Monthly* 176, July, 1945, pp. 101–108. www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881. Accessed 18 June 2017.

Capadisli, Sarven. "Where is Web Science? From 404 to 200." 2016, csarven.ca/web-science-from-404-to-200. Accessed 18 June 2017.

DeRose, Steven J., et al. "What is Text, Really?" *Journal of Computing in Higher Education*, vol. 1, no. 2, 1990, pp. 3–26.

Duba, William O. *Peter of Candia Homepage.* University Cyprus, 2004, www2.ucy.ac.cy/isa/Candia. Accessed 18 Aug 2017.

Duba, William O., et al. *The Peter Auriol Homepage.* 2017, www.peterauriol.net. Accessed 18 Aug 2017.

IIIF Consortium. *International Image Interoperability Framework.* iiif.io. Accessed 18 June 2017.

jeffreycwitt. "Lbp.rb: The Lbp.rb Ruby Library." Github, 2017, github.com/lombardpress/lbp.rb. Accessed 18 June 2017.

Linked Open Data. linkeddata.org. Accessed 27 Jan. 2017.

Linked Research. 4 July 2017. linkedresearch.org. Accessed 18 Aug. 2017.

LombardPress. lombardpress.org. Accessed 18 June 2017.

Mirador. 15 Sept 2016. projectmirador.org. Accessed 18 June 2017.

Nelson, Ted. *Literary Machines: The report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and freedom.* Mindful Press, 1981.

Noerr, Peter, et al. "User benefits from a new bibliographic model: follow-up of the IFLA functional requirements study." *64th IFLA general conference, IFLANET*, 22. April 2009, www.ifla.org/IV/ifla64/084-126e.htm. Accessed 18 June 2017.

O'Neill, Edward T. "FRBR: Application of the entity-relationship model to Humphry Clinker." OCLC Office of Research, 28 March 2002, www.oclc.org/research/publications/archive/2002/oneill_frbr22.pdf. Accessed 18 June 2017.

Parodi, Massimo, and Lucia Caccia Dominioni. *Giovanni di Mirecourt, Commento alle Sentenze (libro 1).* Università degli Studi di Milano, 27 Sept 2003, www.filosofia.unimi.it/mparodi/mirecourt/home.htm. Accessed 18 Aug. 2017.

Robinson, Peter. "Towards a Theory of Digital Editions." *Variants*, vol. 10, 2013, pp. 105–31.

Rosemann, Philipp W. *The Story of a Great Medieval Book*. University of Toronto Press, 2007.

Sahle, Patrick. *Digitale Editionsformen, Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Befunde, Theorie und Methodik*. Schriften des Instituts für Dokumentologie und Editorik, vol. 8, Books on Demand, 2013. urn:nbn:de:hbz:38-53523.

—. "Zwischen Mediengebundenheit und Transmedialisierung. Anmerkungen zum Verhältnis von Edition und Medien." *Editio, vol.* 24, 2010, pp. 23–36. doi:10.1515/edit.2010.004.

Social Web Working Group. "Linked Data Notifications." W3C, 2 May 2017, www.w3.org/TR/ldn. Accessed 18 June 2017.

Witt, Jeffrey. "Creating an Aggregated Dataset from Distributed Sources - a report from the 2016 Basel Meeting." *LombardPress Project Blog*, August 2016, lombardpress.org/2016/08/25/basel-workshop-report. Accessed 18 June 2017.

—. "Linking Research, the SCTA, LombardPress, and LinkedData Notifications." *LombardPress Project Blog*, January 2017, lombardpress.org/2017/01/24/linking-research. Accessed 18 June 2017.

—. "Creating Dynamic Custom IIIF Manifests and the Importance of Great Data." *LombardPress Project Blog*, October 2016, lombardpress.org/2016/10/22/dynamic-manifests-and-great-data. Accessed 18 June 2017.

—. "SCTA Modelling Proposal." *LombardPress Project Blog*, June 2016, lombardpress.org/2016/06/12/scta-modeling-proposal. Accessed 18 June 2017.

—. "The Sentences Commentary Text Archive: Laying the Foundation for the Analysis, Use, and Reuse of a Tradition." *Digital Humanities Quarterly*, vol. 10, no. 1, 2016, www.digitalhumanities.org/dhq/vol/10/1/000231/000231.html. Accessed 18 June 2017.

—. "Surfaces, Canvases, and Zones." *LombardPress Project Blog*, August 2016, lombardpress.org/2016/08/09/surfaces-canvases-and-zones. Accessed 18 June 2017.

Witt, Jeffrey et al. *Scholastic Commentaries and Texts Archive*. 2011, scta.info. Accessed 18 June 2017.

Wood, Rega, et al. *The Rufus Project*. Stanford University, 2017. rrp.stanford.edu. Accessed 18 Aug 2017.

Working Group on FRBR/CRM Dialogue. "Definition of FRBRoo: A Conceptual Model for Bibliographic Information in Object-Oriented Formalism." *IFLANET*, 2017, www.ifla.org/publications/node/11240. Accessed 18 June 2017.

Zahnd, Ueli. "RCS: Repertorium Commentariorum in Sententias Petri Lombardi." University Basel, rcs.philsem.unibas.ch. Accessed 27 Jan. 2017.

Zundert, Joris van. "Barely Beyond the Book?" *Digital Scholarly Editing: Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, OpenBook Publishers, 2016, pp. 83–106.