Journal of Data Mining and Digital Humanities

EPIsciences overlay journals

# Transparency and Discovery: Using a Text-Image Network to Study Manuscripts and Text Transmission

Jeffrey C. Witt[1*]

1 Loyola University Maryland, USA

*Corresponding author: Jeffrey C. Witt jcwitt@loyola.edu

**Abstract**

This article describes desired use cases of manuscript-text interactions that could positively influence future scholarship. These examples illustrate the potential for greater transparency between critical editions and their data sources (i.e., manuscript witnesses). Further, these use cases demonstrate how the possibility of a connected environment could facilitate the automated discovery of related manuscripts at granular levels (i.e. identifying connections at precise coordinates regions) through the discovery of related texts. In the second section, the article walks through the data modelling and engineering steps required to achieve these uses cases. Finally, the third section provides examples to illustrate how this future is already a fledging reality.

**keywords**

manuscripts; network graphs; digital editions

## INTRODUCTION

As the theme of this volume is the future of manuscript studies, I will first describe some desired use cases of manuscript-text interactions that could positively influence future scholarship. These examples will illustrate the potential for greater transparency between critical editions and their data sources (i.e., manuscript witnesses).[1] Further, these use cases will demonstrate the possibility of a connected environment that could facilitate the automated discovery of related

[1] The goal of transparency in the creation of a critical edition is frequently identified as a core aspiration. See [Bordalejo, 2021, section 68] who writes: "[W]hen I talk about a critical text, I am not referring to the reconstruction of a lost archetype, but to the construction of a new, well-informed text that can help readers understand the relationships between extant witnesses; a text that functions as a gateway to the others." This aspiration is echoed by [Huskey 2022] and [Sahle 2010]. However, without easy access to manuscript witnesses at granular levels, this desired transparency will remain difficult. This article intends to describe some of the engineering steps required to achieve this kind of immediate and granular access.

manuscripts at granular levels (i.e. identifying connections at precise coordinates regions) through the discovery of related texts.

These use cases focus on a new future interplay between manuscript and text at new levels of granularity. If this kind of transparency and discovery were enacted at scale, the future of manuscript studies would have at its disposable a connected web of distributed resources that a researcher could traverse at the touch of a button in ways that meaningfully impact their scholarship.

In the second section, I will walk through some of the data modelling and engineering steps required to achieve these uses cases.

Finally, the third section will provide examples to illustrate that this future is already a fledging reality. Thus, it is not a lack of technology holding back this future, but the challenge of encouraging the mass adoption of these new ways of publishing the textual data manifested in the manuscripts of the past.

## 1. DESIRED USE CASES

I imagine a future world in which a researcher, upon beginning at a point within a manuscript—perhaps a specific paragraph, citation or, even a specific word on a page—is immediately alerted to equally granular objects of related interest. And more than being merely "alerted", the researcher should be offered a way to view these related objects without having to follow a link or move to a new environment, regardless of where the object resides, whether physical or digital. In short, I imagine a researcher who would be able to begin at any given coordinate space within the image of a manuscript page and instantly be able to "move"— without actually physically moving — in at least three directions.[2]

1) First, while requiring no new information, the researcher should be able to call up the precise parallel coordinate regions containing the same textual objects in any sister witnesses. That is, if the user is focused on a particular paragraph manifesting itself in a particular manuscript, they should be able to view it alongside the coordinate regions of all other manuscripts of this text which contain this same paragraph. This movement is illustrated below in Figure 1.

[2] The word "movement" will be employed repeatedly in what follows to capture the dynamism of what is possible. At the same time, we want to avoid giving the impression that we are talking about the normal kind of movement that one experiences on the web, that of being given a link and being asked to follow that link to a new website and a new environment where one must re-contextualize themselves to their new surroundings. This kind of movement is a virtual re-enactment of movement within the book paradigm. Instead, I mean to speak of a kind of movement in which resources of interests, regardless of where they exist on the web, can be "called-up" and meaningfully contextualized in the user's current environment. Such is the early vision of [Bush, 1945] and [Nelson, 1981]. The idea of "movement" used here is meant to echo that vision.

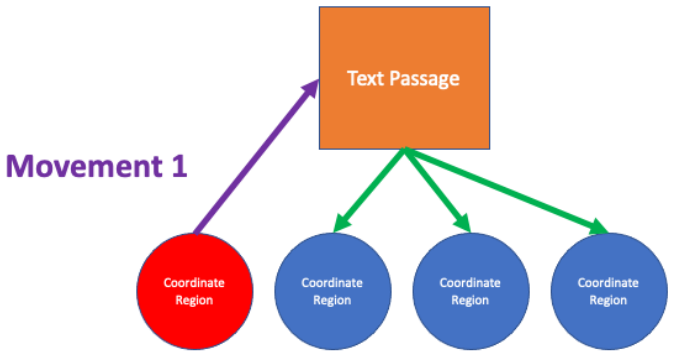



Figure 1.

2) Secondly, the researcher should be able to move backward from this paragraph to any material or formal mediating sources used within the passage. If, as seen in Figure 2 below, text X ("Text Passage") is quoting text Y ("Source Passage"), one should be able to move from the quotation in text X to the precise passage within text Y. And from here, they should be able to move to the coordinate regions of each manuscript containing this original text passage.

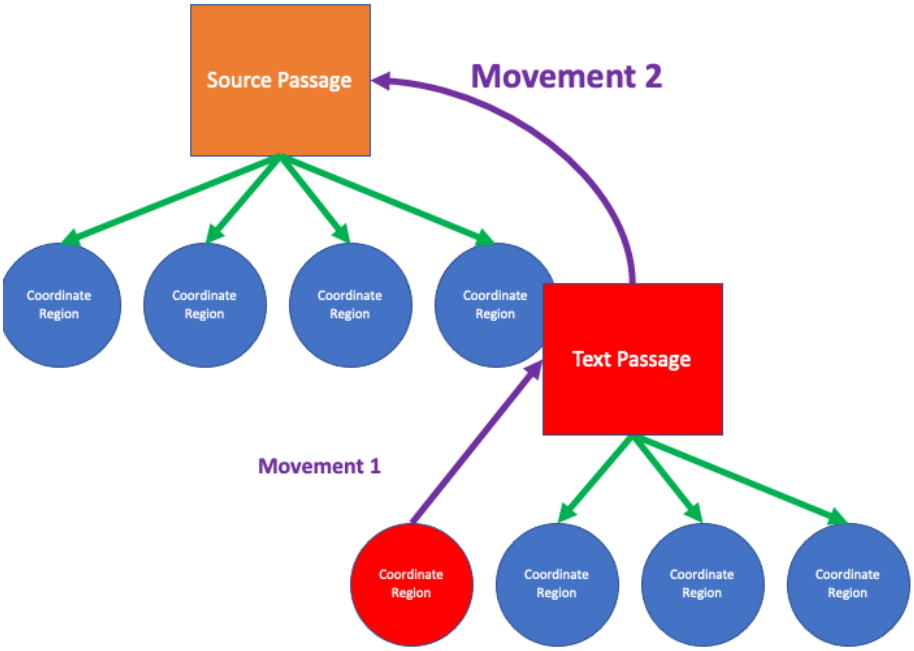



Figure 2.

3) Third, a researcher should be able to move forward toward to the specific passages in any texts that quote, cite, or reference the original text passage contained within the starting coordinate region. For example, as seen in Figure 3 below, if text X ("Text Passage") is quoted by text Z ("Quoting Passages"), the researcher should have instant access to all future quotations of this passage. And once again, they should be able to move from these quoting passages to any coordinate region of any manuscript containing the quoting passages.

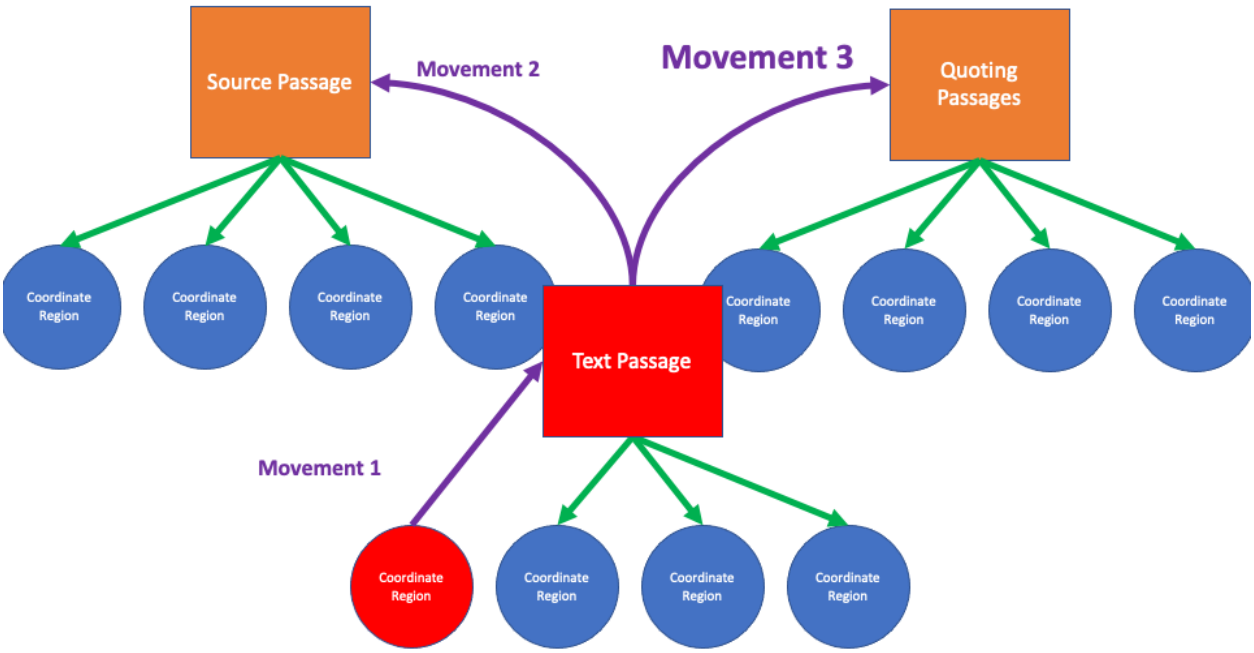



Figure 3.

4) Finally, the researcher's starting "target" should never be fixed. That means this tri-fold movement should be repeatable with respect to any text passage or coordinate region. In this way, the source passage can become the target text passage, and then we can move from here to its sources and its influences. The interconnected web that makes this possible is illustrated below in Figure 4.

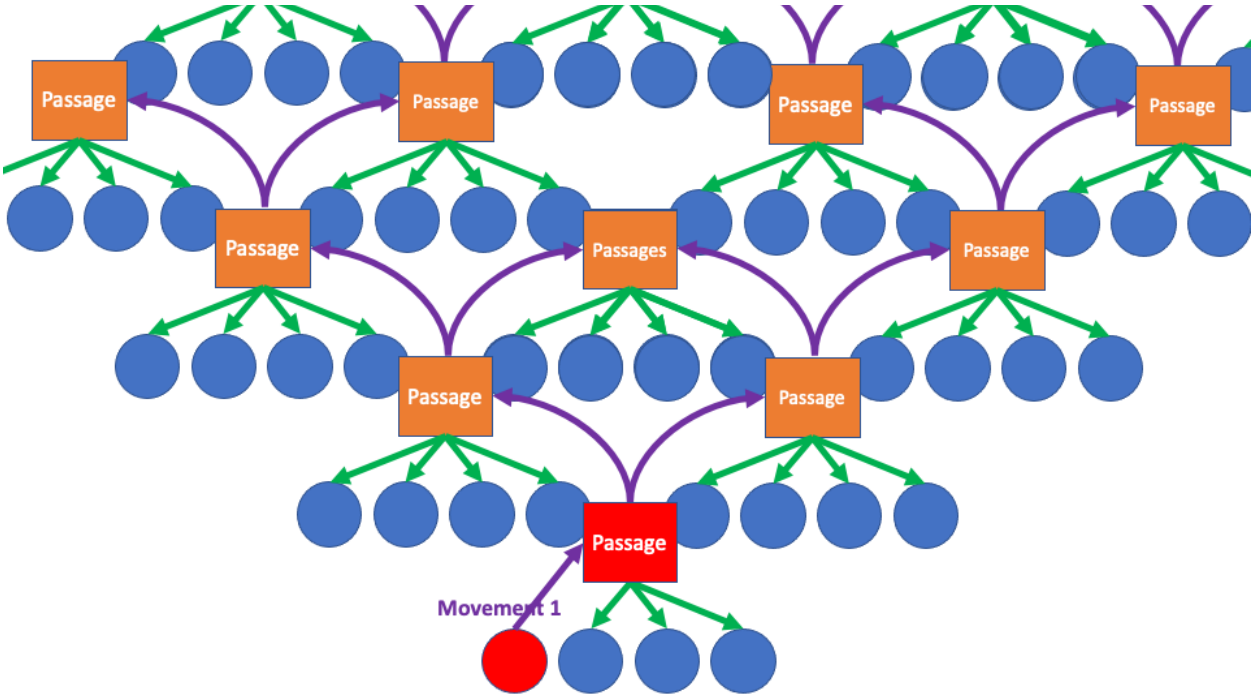



Figure 4.

The kind of discoverability, transparency, and instant access described above will not only create new possibilities for source criticism and reception history but will also let us expand our idea of what a manuscript witness is.

For example, with such an interconnected web, every quotation of a Patristic text in the medieval tradition will suddenly be identifiable. Subsequently, the witness of each quoting passage will be accessible as a witness to the original passage. This new accessibility should allow us to view the transmission of texts and their mutation through time with new clarity. Here are two brief examples of why this could be useful.

The following passage is a highly influential quotation from Augustine's *De Trinitate*, book 14, used repeatedly throughout the Middle Ages. This quotation discusses two kinds of knowledge, *sapientia* and *scientia*. Though Augustine never used the word "theology," the latter ter — *scientia* — was later understood to identify theology as a science and became a foundational source text in the debate over the scientific status of theology.[3]

> "alii datur sermo sapientiae, alii sermo scientiae", ista definitio diuidenda est ut rerum diuinarum scientia sapientia proprie nuncupetur, humanarum autem proprie scientiae nomen obtineat, de qua uolumine tertio decimo disputaui, non utique quidquid sciri ab homine potest in rebus humanis ubi plurimum superuacaneae uanitatis et noxiae curiositatis est huic scientiae tribuens, sed illud tantummodo quo

[3] See for example [Brown 1991] and [Witt, 2013].

> fides saluberrima quae ad ueram beatitudinem ducit gignitur, nutritur, defenditur, roboratur. [Augustinus, *De Trinitate*, Book 14]

With respect to this quotation, one may ask: at what point and why do later interpreters identify the discipline of theology with what Augustine here calls "science" or with the discipline that Augustine states is concerned "with human things, but only in so far as they grow, nurture, defend, and encourage the faith." One could have just as easily thought that the discipline of theology would better align with the "wisdom" that Augustine says is concerned with divine things. At what point was this interpretation made?

How can we study this reception scientifically rather than anecdotally? To do so, we would need to be able to collect, view, inspect, verify, and compare all future instances of this quotation in every manuscript. With this kind of access we could observe how the quotation has transformed over time. And with these comparisons we could identify the reasons for such a change (whether they be intentional, the result of scribal changes, or some combination of both).

Consider another famous example from Augustine's "Against the letter of Manichaeus" which illustrates the same point. In this famous text, Augustine declares that he would not believe the Gospel unless the authority of the Catholic Church "co-moved" him to do so.

> Ego vero Evangelio non crederem, nisi me catholicae Ecclesiae commoveret auctoritas. [Augustine, *Contra epistolam Manichaei quam vocant fundmanenti*].

According to [Heiko Oberman, 1963, p. 385], in some late medieval uses of this quotation the word "*commovere*…is replaced by *compellere* and interpreted as *approbare*, a term which suggests more duality between Scripture and Church than the phrasing of Augustine's text itself."

It is surely an interesting suggestion that highlights a single word's significance. But is Oberman correct? Do the uses of *compellere* really indicate an intentional deviation or could this be a transmission error? Are the scribes using either word following a tradition, or are they inventing something new? Are they anomalies, or are they part of a larger tradition? If these are scribal interpolations, where and when do they originate? And finally, how can we make the answers to these questions "transparent"? Is it possible to make the primary source material that stands behind these answers accessible for repeated consultation and review at the touch of a button? Can we make it possible for the research process that produced these claims to be easily repeatable and verifiable? Again, how can we study this scientifically rather than simply relying on the anecdotal evidence pulled from what Oberman has chosen to read and remember?

To these examples of known use cases, we should equally stress the power of the connected network to discover and draw attention to previously unknown passages and coordinate regions of interest. All active connections reported by editors can easily be inverted to show points of connections that no single editor knows individually, but which are collectively apparent.[4] Moreover, text passages that have received high degrees of re-use can be automatically discovered

[4] For a description of this inversion see [Christensen, Witt, Zahnd, 2021, p. 52].

and recommended for consultation through machine assisted methods — and by extension the corresponding manuscript coordinate regions can be detected. A few suggestive examples of such detection will be highlighted below in part 3.

## 2. DATA MODELING THE CONNECTED WEB

The question before us is how to achieve the uses cases described above. More precisely, how can we use textual connections to connect granular coordinates regions within large manuscripts distributed all over the world.

Creating this network of connections is difficult because it requires us to think differently about how we have traditionally discovered and located parts of a manuscript.

I call the traditional approach the "Russian Doll Discovery" approach. In other contexts, we might call it "location-based lookups." Location-based lookups work by starting at the largest known container and then subsequently dividing that container until one gets as close as possible to the desired granular object. In this approach, we must know the name of a long series of containers to find a small granular passage of interest in a manuscript.

In the case of manuscripts, this identification of containers requires us to find:

1) The institution that holds the containing codex.
2) The codex within the institution.
3) A further digital surrogate, such as a [IIIF] manifest.
4) The page or IIIF canvas within the digital surrogate.
5) Finally, a very precise coordinate region on a large and complicated page.

Worse yet, a traditional citation that lists out the requisite containers provides no mechanism for "saving" these discovery steps. It is, in short, a zero-sum game. The retrieval path to view a particular place in a particular manuscript does not reduce the retrieval time that it takes a second person to view the same coordinate region within a manuscript. While the citation offers users a description of the path that needs to be traveled, each consumer must still re-travel this path to see the object of interest.

For instance, if a researcher makes a claim based on the evidence in a manuscript cited as "France, Paris, Bibliothèque nationale de France, Lat 3074, f. 13, side verso, column A, somewhere near the middle," to confirm the claim, one will have to repeat all the discovery steps that were already performed in order to see the object.

This is not just true in the book world. It remains true in our current digital world. In most cases, one must travel to the correct website, perform a custom search, open the object, and navigate to the correct digital page, and then find the granular space of interest. If someone was hoping to compare this granular space of interest to another manuscript, they would have to repeat this entire process a second time. The labor involved de-incentivizes such a comparison. If someone had the ambition to compare 50 or 100 such regions of interest, they would quickly

abandon their plans when they learned the steps involved. Furthermore, on the rare case that someone did commit themselves to such a Herculean task, few would be interested in repeating it, making the results—practically speaking—unrepeatable and therefore unverifiable.

Decades ago, this procedure was familiar in another everyday context, namely regarding geographic location. If I wanted to locate a store or park on a map, I would first have to know the country it was contained within, then the city, then the part of the city, finally the street, etc. Even if I found what I was looking for and then wanted to share it with the person next to me, I would have to pass them the map. They, in turn, would have to perform the same steps that I had just performed.

But we generally do not do it this way anymore. We have something called the Global Positioning System (GPS). A GPS system uses unique identifiers—abstract and distinct from any particular map—to identify highly granular geographic spaces. Today when I want to find a place, I provide the identifier of the place or resource whose containers are already known and saved. Upon supplying the coordinates, an application can automatically look up, request, and load all the necessary containers needed to present the user with a detailed view of the region of interest with any level of desired contextual information.

We need a GPS for textual resources within their respective material manifestations (e.g. manuscripts, codices, etc.).

But things get tricky here. There is currently on-going work to create unique manuscript identifiers.[5] The IIIF API is likewise pushing cultural heritage institutions to mint RDF IDs for each page in their codices, called "canvas IDs". The addition of a coordinate region to these IDs would seem to get us close to what is needed. But while this is a necessary step, it is not sufficient.

GPS coordinates have allowed us to overlay historical maps on top of each other despite vast differences in scale and aspect ratio.[6] This is possible because GPS coordinates are an abstraction from any given map image. After all, the pixel coordinates of a city on a given map have nothing to do with the pixel coordinates of that same city on another map. Similarly, a IIIF Canvas ID plus the coordinates of the target region in a particular manuscript do not yet offer us the necessary abstraction. They will not allow us to tie together parallel points.

We need something that transcends all manuscripts. In the case of maps, this is the actual earth's longitude and latitude. In manuscripts, it is the abstract textual hierarchy that is "manifested" in each manuscript.

The solution to this problem involves using diplomatic text transcriptions (which are structured around the text hierarchy but still record the line breaks of the material text being transcribed) to stand as a link between textual abstractions and image zones (i.e., distinct coordinate regions).

Starting from a diplomatic text transcription, we can use inference rules to generate text abstractions. We can then tie these abstractions to associated coordinate regions by using the transcription to identify the first and last lines on which the target passage appears within a given

[5] See [Cassin, 2018].
[6] See for example [David Rumsey Map Collection].

material manifestation (i.e., a manuscript). The coordinates of the region where the entire passage is manifested can then be computed by using the coordinates of the top boundary of the first line and the bottom boundary of the last line.

The above process is illustrated in the following sequence of images (Figures 5-12). In Figure 5, we start with a lonely diplomatic transcription of a small paragraph, which includes line breaks as part of its textual transcription.

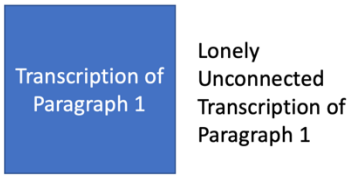



Figure 5.

These line breaks are automatically paired with the machine-captured line coordinates on a manuscript page. See Figure 6.

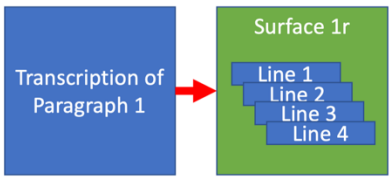



Figure 6.

Here, such a paring means that each paragraph is recorded as manifesting itself on a Zone (or Zones) via an "isOnZone" property. Furthermore, a given Zone may contain smaller Zones (such as line Zones). And from the coordinates of these line Zones, the bounding boxes of containing Zones can be constructed from the top boundary of the first child Zone and the bottom boundary of the last child Zone. An excerpt of this data model is visible below in Figure 7.

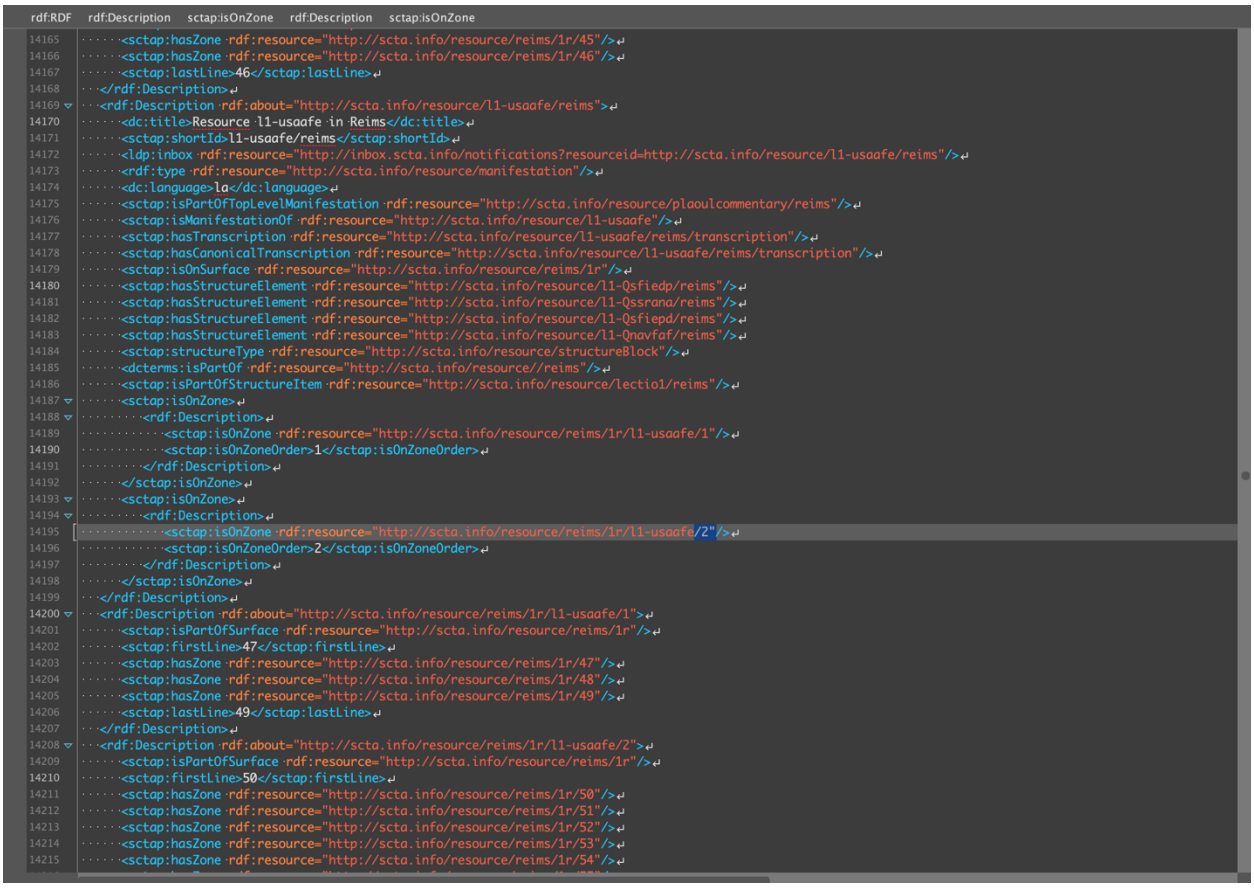



Figure 7.

The same transcription can also be used as a trigger to create resources representing abstract ideas about the text. Such abstract ideas roughly follow the FRBR ontology but are realized at all levels of the textual hierarchy.[7]

First, the existence of a paragraph transcription is reason enough to generate a Manifestation of the paragraph: i.e., the idea of this text passage at it appears in this specific edition. After all, if there is a transcription, there must be a Manifestation which is being transcribed. This is seen in Figure 8.

[7] For more on FRBR see [Noerr, 1998], [O'Neill, 2002], and [Bennet et al, 2003]. See also [Christensen et al, 2021] and [Witt, 2018].

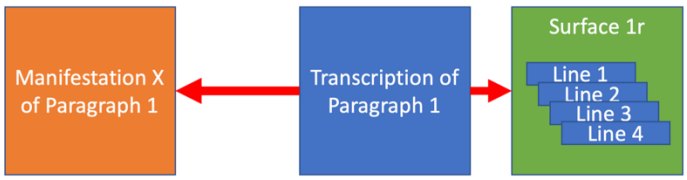



Figure 8.

Finally, a Manifestation becomes the trigger to generate an identifier for the Idea/Expression of a paragraph that may have many Manifestations. Again, if a Manifestation of this text exists, then the abstract content of this Manifestation must also exist. See Figure 9.

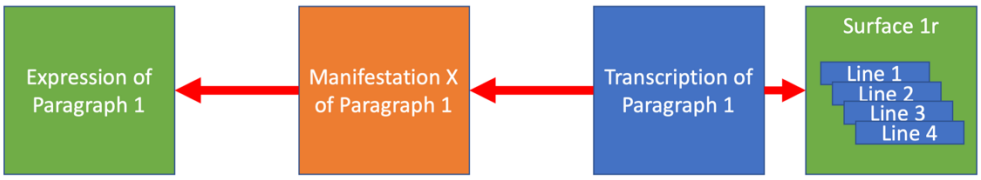



Figure 9.



Journal of Data Mining and Digital Humanities
**ISSN 2416-5999, an open-access journal**
http://jdmdh.episciences.org

Other Manifestations of the same Expression can then be built up in the same way. These can then be grouped and connected around their common text Idea/Expression. This concept is illustrated below in Figure 10.

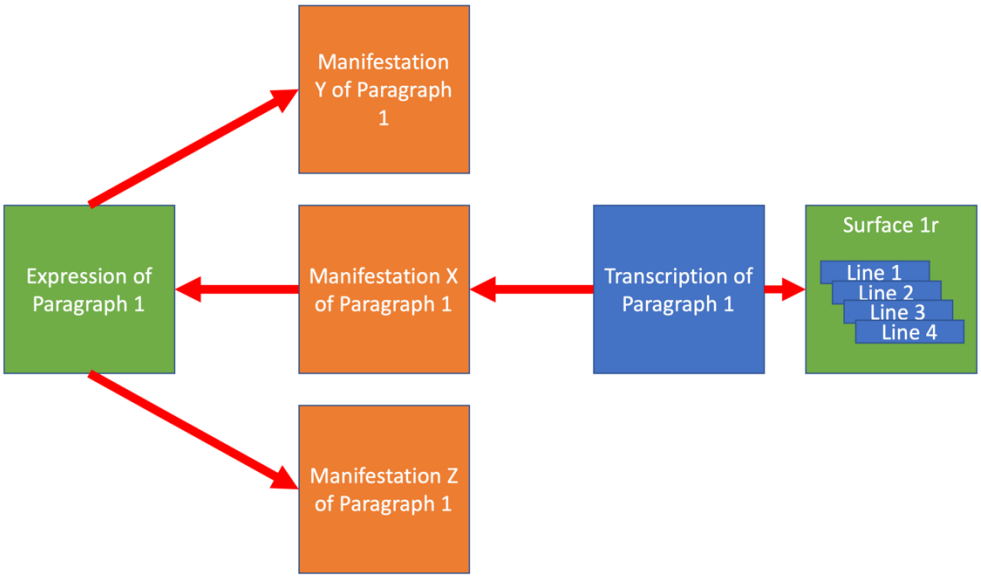



Figure 10.

Similarly, each of these Manifestations can lead us from a single Expression back to the respective coordinate region in each Manifestation, as seen in Figure 11.

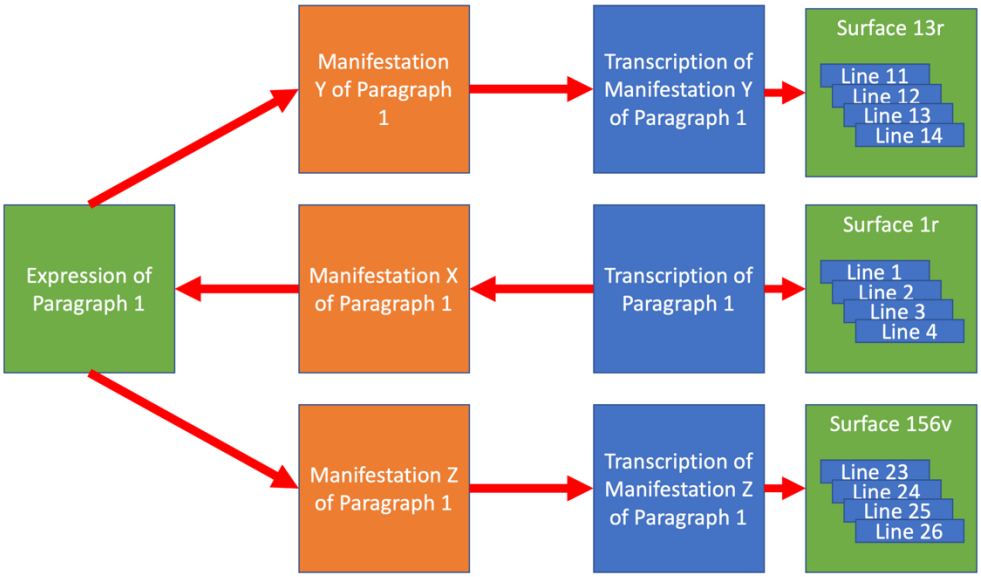



Figure 11.

Finally, from here—recalling the desired use cases described above—all quoted- and quoting-, referenced- and referencing-textual Expressions can be linked together. Once linked, it becomes possible to move from any coordinate space, not only to all sister witnesses, but to all passages wherein a text within a coordinate spaced is used, and from there to all the sister witnesses of this passage. This elaborate network can be viewed below in Figure 12.

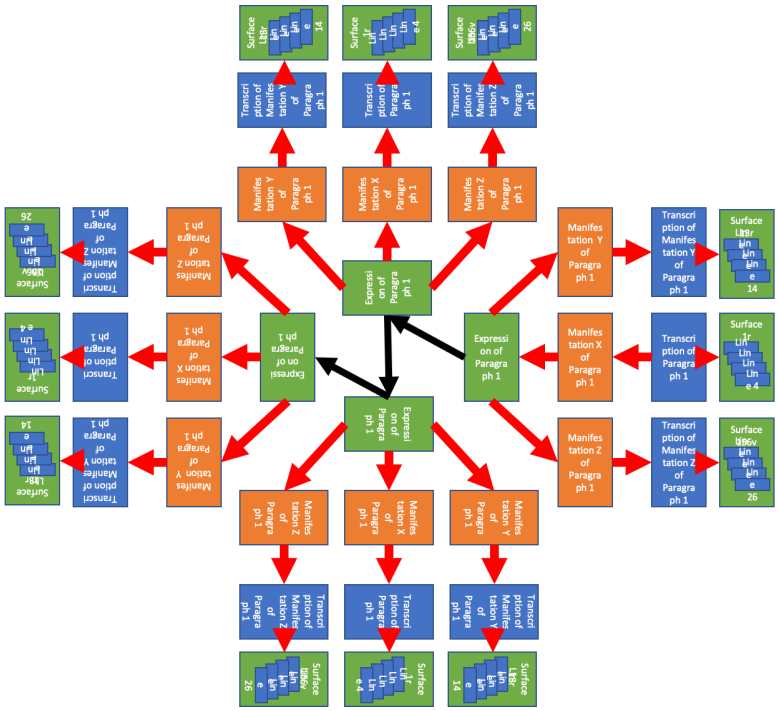


Figure 12.

## 3. REPRESENTATIVE EXAMPLES: THE FUTURE IS NEAR

In this third section, I show with a few extended demonstrations how the above network can facilitate the transparency and discovery.

First (3.1), I offer an example of how a researcher beginning from a coordinate space can be guided through the network described above in Figure 12 (which, in turn, was designed to enable us to traverse the network illustrated in Figure 4). Then (3.2), I will offer some examples of how basic but automated text analysis techniques can open new and unexpected paths through this same network, letting us discover cases of textual re-use. And because text and image regions are bound together, discoveries of such textual parallels automatically lead to discoveries of parallel image coordinate regions in previously isolated manuscripts and codices.

### 3.1 From Coordinate Space to Corpus

Let us imagine a researcher with a digital manuscript in front of them, not yet knowing what they are looking for. As they scroll through pages, different coordinates spaces on the page should be able to report connections, alerting readers to areas of possible interest. This process is akin to Google maps reporting further places of interest—in some way related to the initial target region of interest—and then offering links to further connections.

Here, in Figure 13, imagine that we are scrolling through a page in a manuscript owned by the Parker Library in Cambridge and served by its partner at Stanford University.

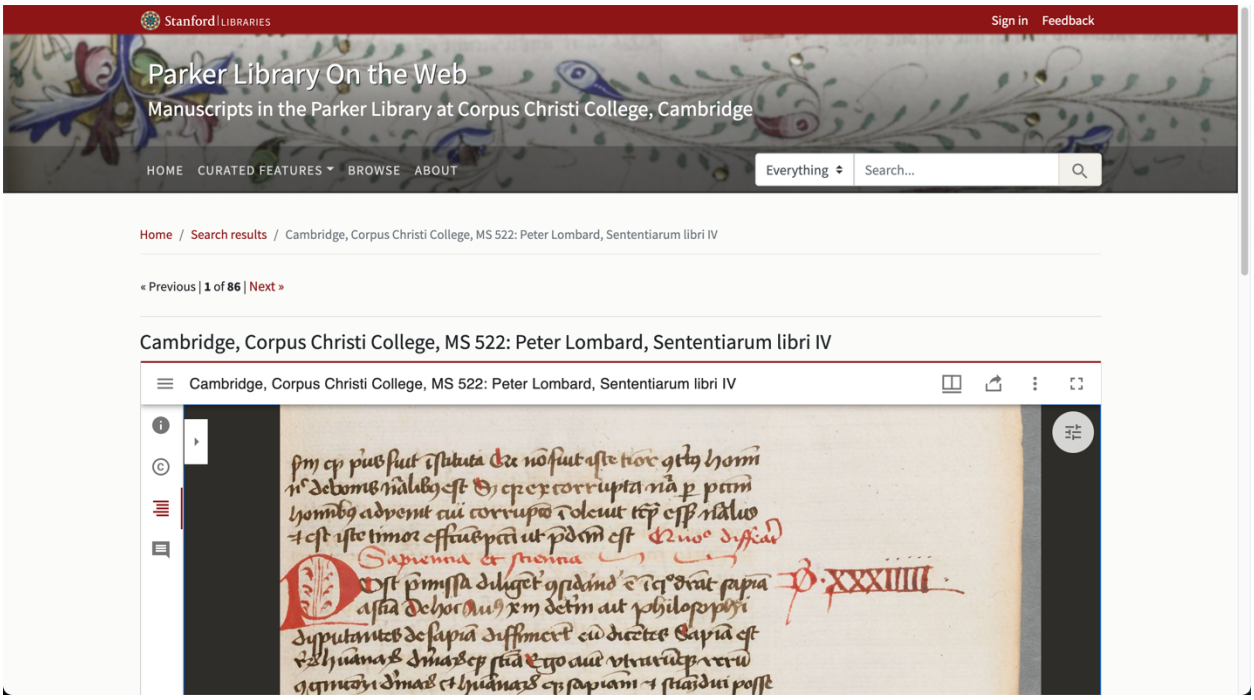



Figure 13.

At present, it remains unconnected. Although this page is from a manuscript containing Peter Lombard's *Sentences*[8] which has parallels in many manuscripts and contains numerous quotations of Patristic sources, the viewer at present cannot alert the user to any new pathways. This vacuum is unfortunate because the information exists. It just has not been connected.

Let us consider what would happen if the viewer knew about the existing data resources and took advantage of them. First, we would need the ID of the given page or the ID of a specific coordinate region. This is comparable to pointing to a very specific place on a map and saying: "tell me information about this place" or "show me this place in other maps."

Such IDs exist, but because people do not seem to be aware of the uses that can be made, they are not exploited. Here in Figure 14 we must go digging for the container IIIF Manifest because the viewer has decided not to show us the IIIF Canvas ID for this folio side.

[8] For an introduction to Lombard's *Sentences* see [Rosemann, 2007].

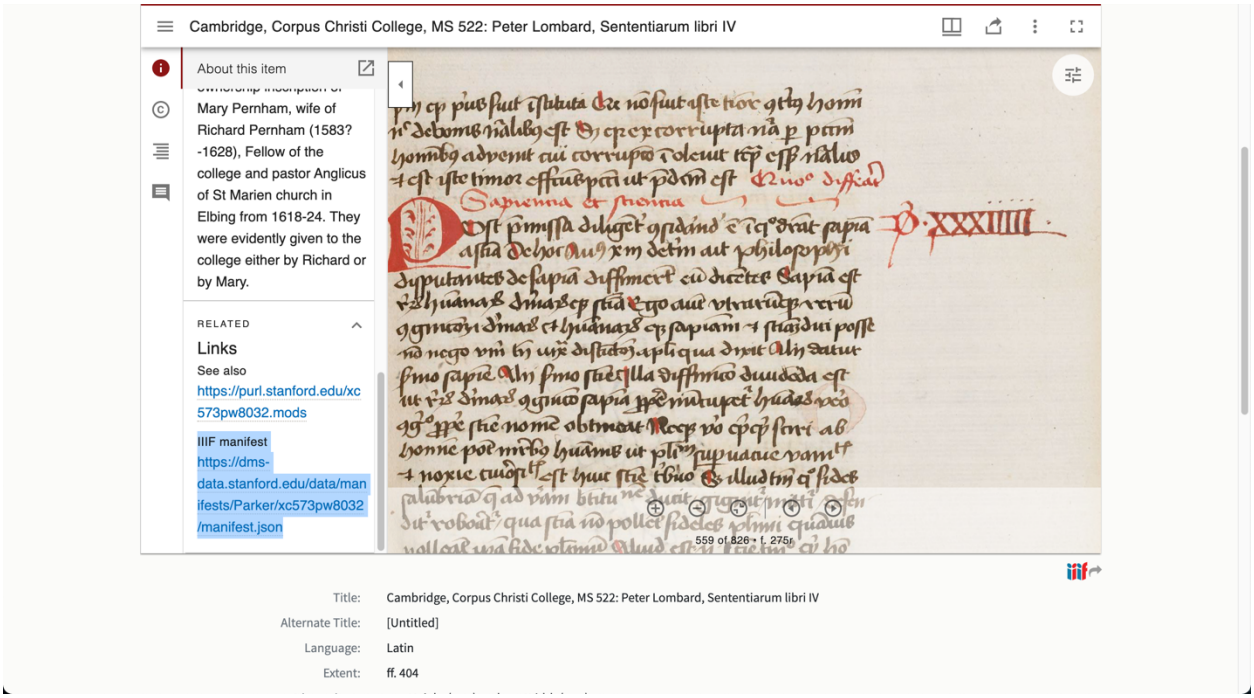



Figure 14.

Once we have the Manifest container, we can inspect it and find the Canvas ID, highlighted in Figure 15.


```
                    "width": 3432,
                    "service": {
                        "@id": "https://stacks.stanford.edu/image/iiif/xc573pw8032%252F522_274_V_TC_46",
                        "profile": "http://iiif.io/api/image/2/level1.json",
                        "@context": "http://iiif.io/api/image/2/context.json"
                    }
                },
                "on": "https://dms-data.stanford.edu/data/manifests/Parker/xc573pw8032/canvas/canvas-558"
            }
        ]
    },
    {
        "@id": "https://dms-data.stanford.edu/data/manifests/Parker/xc573pw8032/canvas/canvas-559",
        "@type": "sc:Canvas",
        "label": "f. 275r",
        "height": 5471,
        "width": 3303,
        "images": [
            {
                "@id": "https://dms-data.stanford.edu/data/manifests/Parker/xc573pw8032/imageanno/anno-559",
                "@type": "oa:Annotation",
                "motivation": "sc:painting",
                "resource": {
                    "@id": "https://stacks.stanford.edu/image/iiif/xc573pw8032/522_275_R_TC_46/full/full/0/default.jpg",
                    "@type": "dctypes:Image",
                    "format": "image/jpeg",
                    "height": 5471,
                    "width": 3303,
```


Figure 15.

Now that we have it, let us supply this ID to a viewer and that understands it and can report its connections.

In Figures 16 and 17, we see related textual regions in other manuscripts immediately detected and shown to the user without requiring the user to follow a link, leave the current interface, or adjust to a new environment.

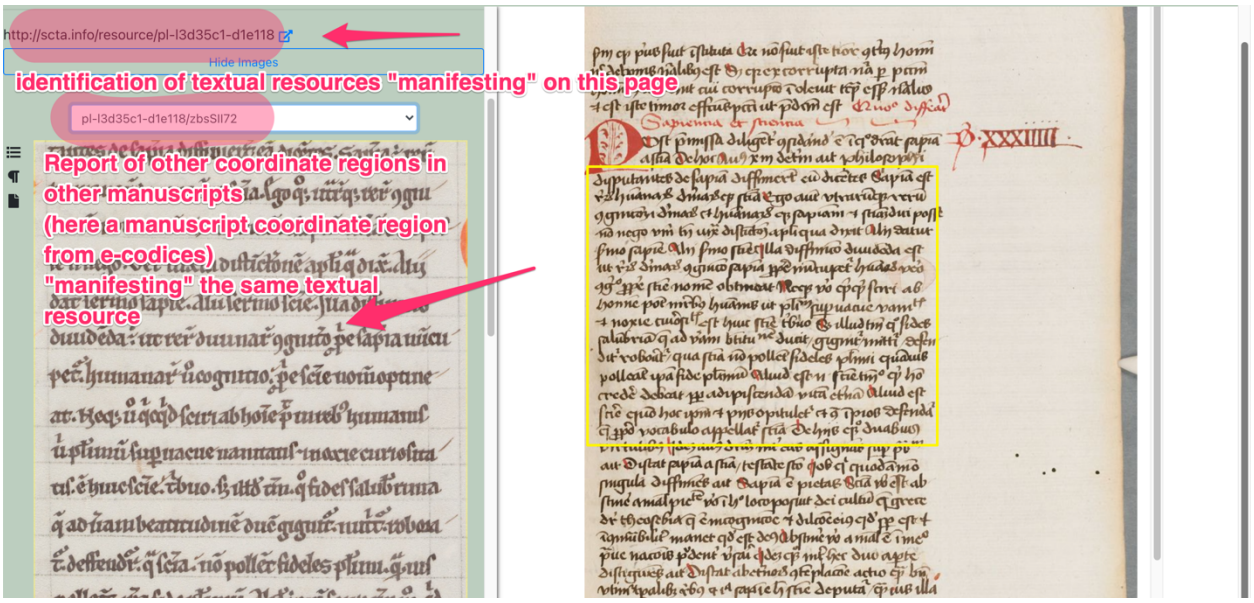



Figure 16.

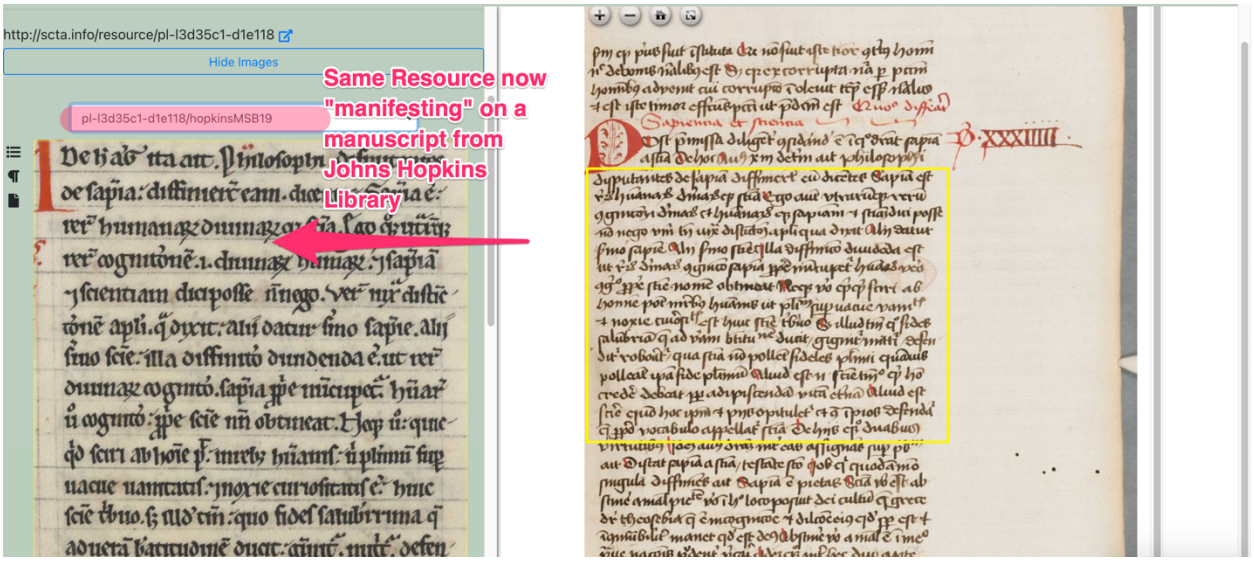



Figure 17.



Journal of Data Mining and Digital Humanities
**ISSN 2416-5999, an open-access journal**
http://jdmdh.episciences.org

Nor are we confined to a single viewer. When the Mirador viewer is connected to the text-image network, we can supply a single ID of a granular text passage and in response see parallel coordinate regions of multiples witnesses automatically loaded within a Mirador instance (seen below in Figure 18).

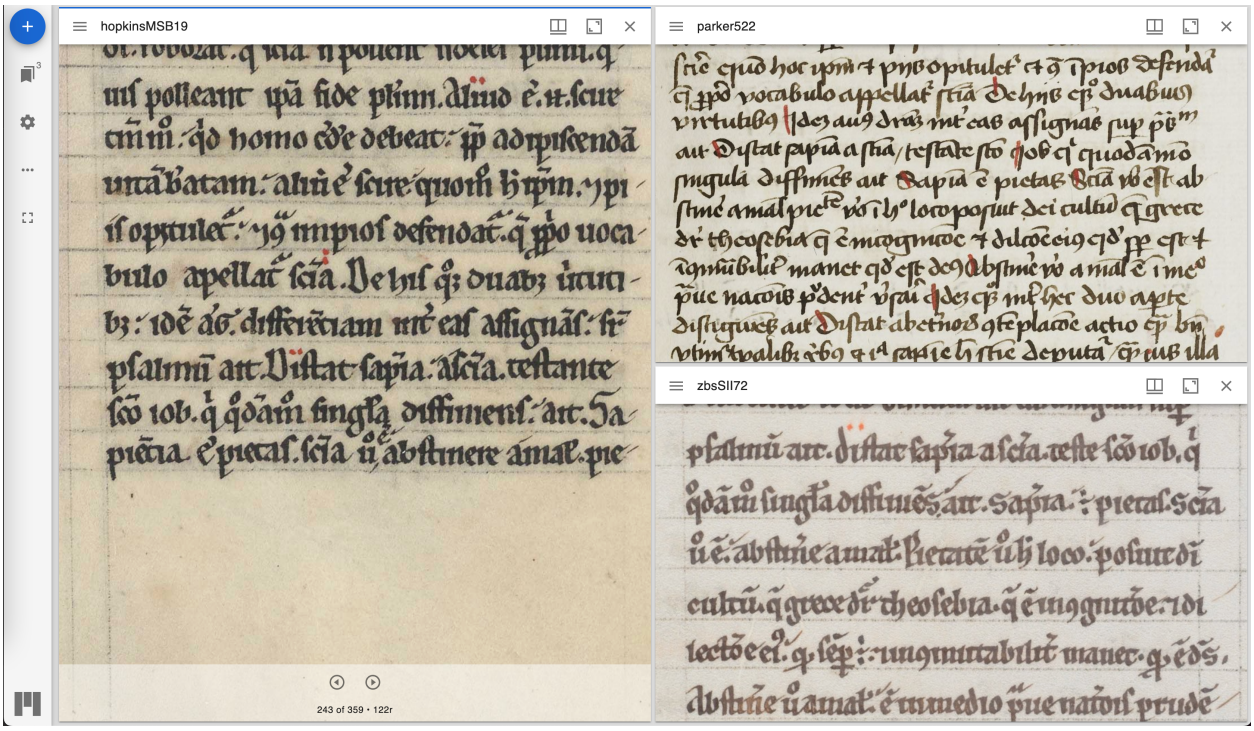



Figure 18.

Nor must it be paragraph; it can be any kind of granular resource, from a paragraph or a small quotation to a scientific image. Consider Figure 19 below.



Journal of Data Mining and Digital Humanities
**ISSN 2416-5999, an open-access journal**
http://jdmdh.episciences.org

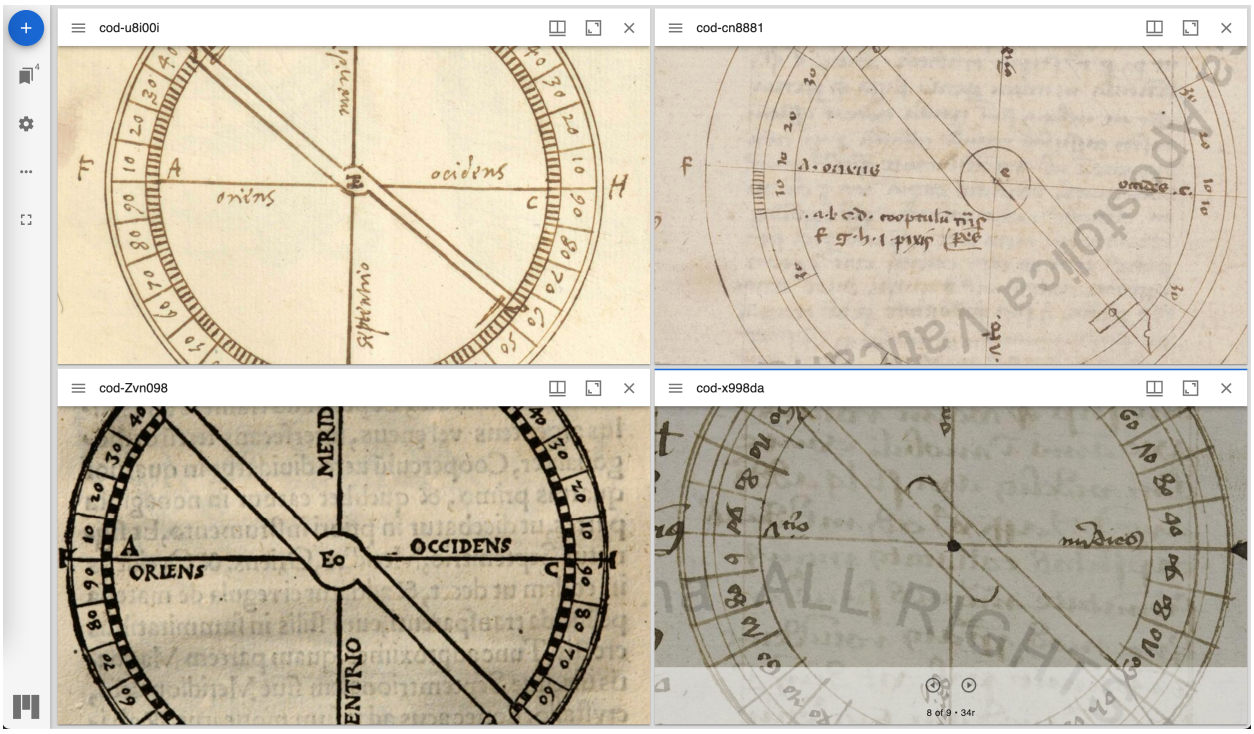



Figure 19.

In loading these three instances of a paragraph or the four instances of a scientific figure, notice what has happened. We have supplied one ID without knowledge of any of the many containers required for Mirador to present these paragraphs or figures. Instead, the network in Figure 12 can be traversed by simply supplying a single Expression ID. Then all the containing information needed to load the desired resources within a meaningful context (in this case, the Manifest IDs, Canvas IDs, and respective coordinates) can be automatically retrieved.

For the same reason, we never need to supply the traditional citation information (which is basically a list of containers). Instead, we can (as seen in Figures 20 and 21) supply the ID of the textual resource, and the let that resource report the traditional citation information.

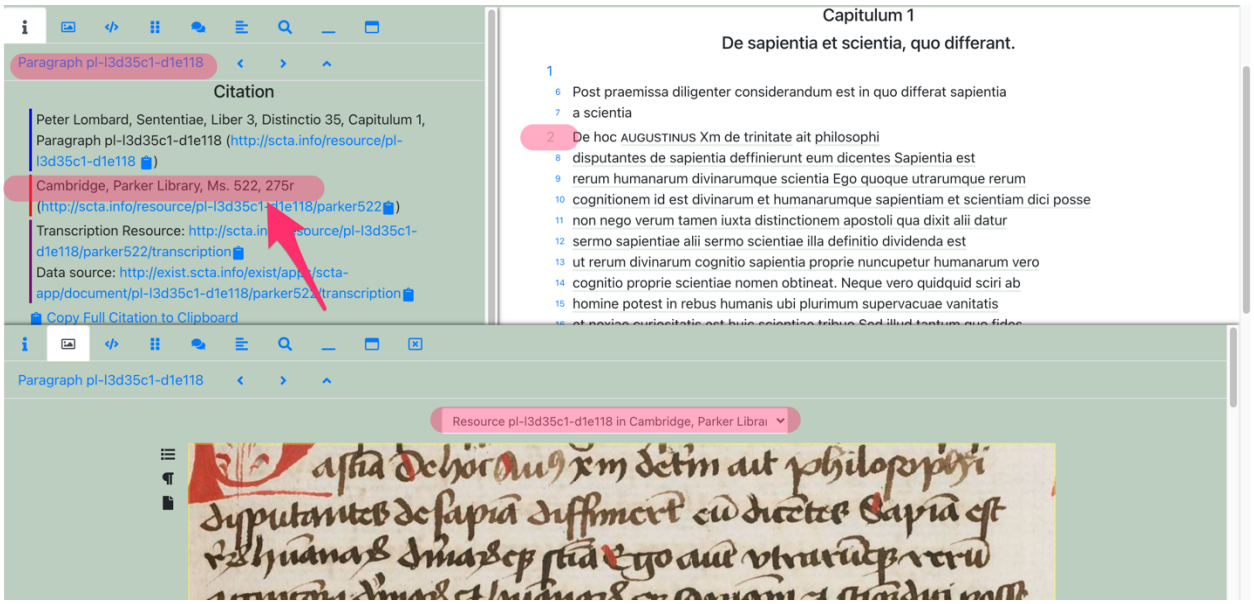



Figure 20.

Figure 21 shows that we can even highlight an arbitrary text fragment within a paragraph and be delivered both the traditional citation information and the image coordinates region containing only the arbitrary selected text.

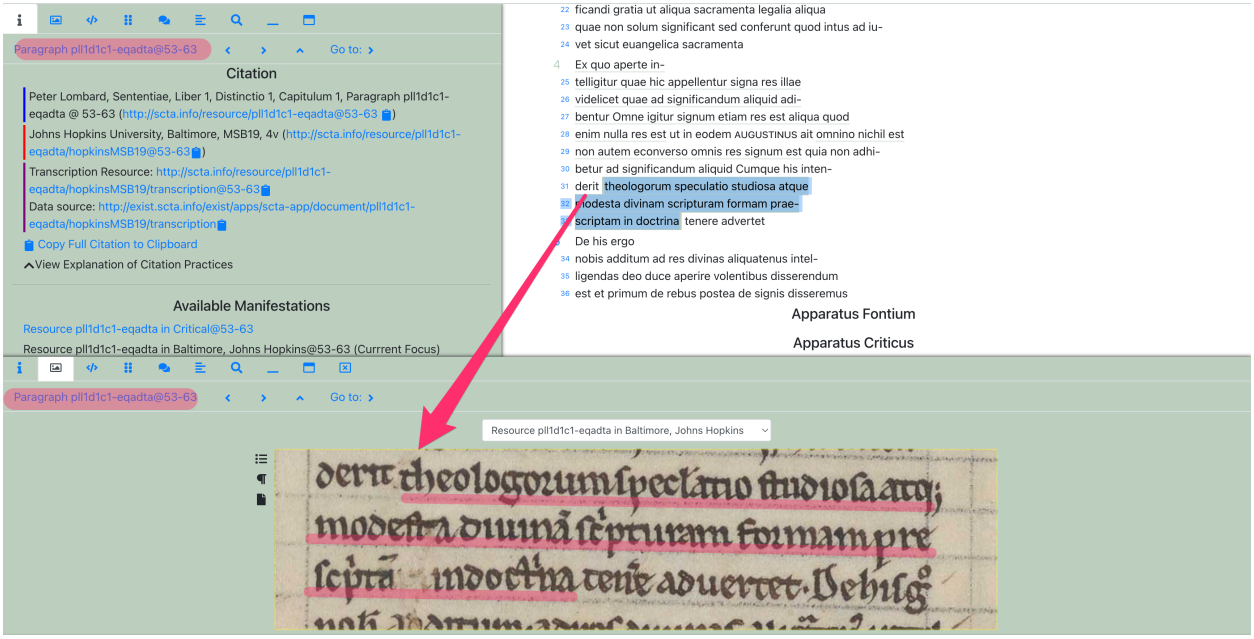



Figure 21.

The ease of consultation here should not be underestimated. For known manuscripts, the greatest obstacle to transparency, consultation, and verification is not some genuine impossibility, but the practical difficulty (in time and labor) of the actual lookup. Consultation will become effortless and therefore incentivized when this friction is reduced to almost nothing.

At this point, we have seen how "movement 1" from our above use case has been accomplished. Starting from a specific coordinate region, we instantly discovered and accessed the coordinate regions of parallel manifestations of a textual phenomenon.

Let us consider movements 2 and 3. Recall that from any coordinate region, we wanted to be able to travel upward to the textual resource manifested in this space and then ask for information specifically about its sources and its influences. From there, we wanted to move down again to consult the manuscript witnesses to these granular text passages.

In Figure 22, we can see that our focus is on the "target passage" in yellow, which reports known and marked sources (a bible quotation and a quotation from Augustine).

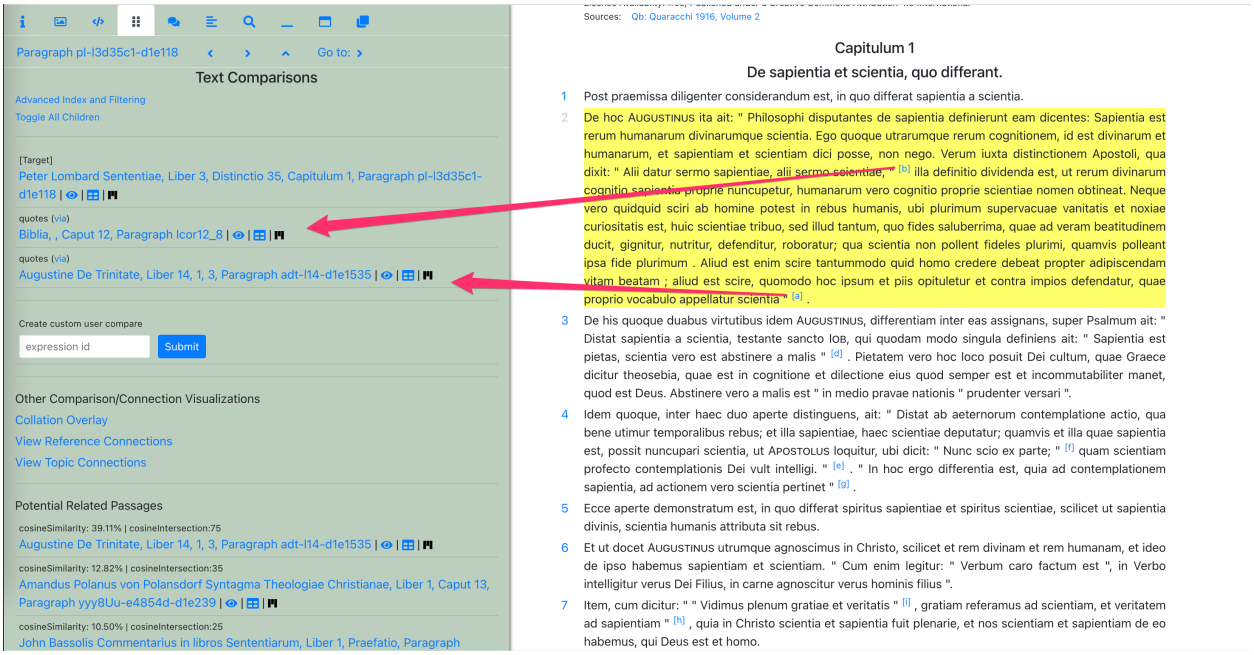



Figure 22.

In the following image, Figure 23, we can see how we have refined the granularity of our focus even more. Now we are targeting a specific quotation by Peter Lombard of a specific Bible verse, namely I Corinthians 12:8. Without leaving the present viewer, we can view that quotation in all its sister witnesses (in this case in the Parker Library manuscript, an E-Codices manuscript, and a Johns Hopkins manuscript).

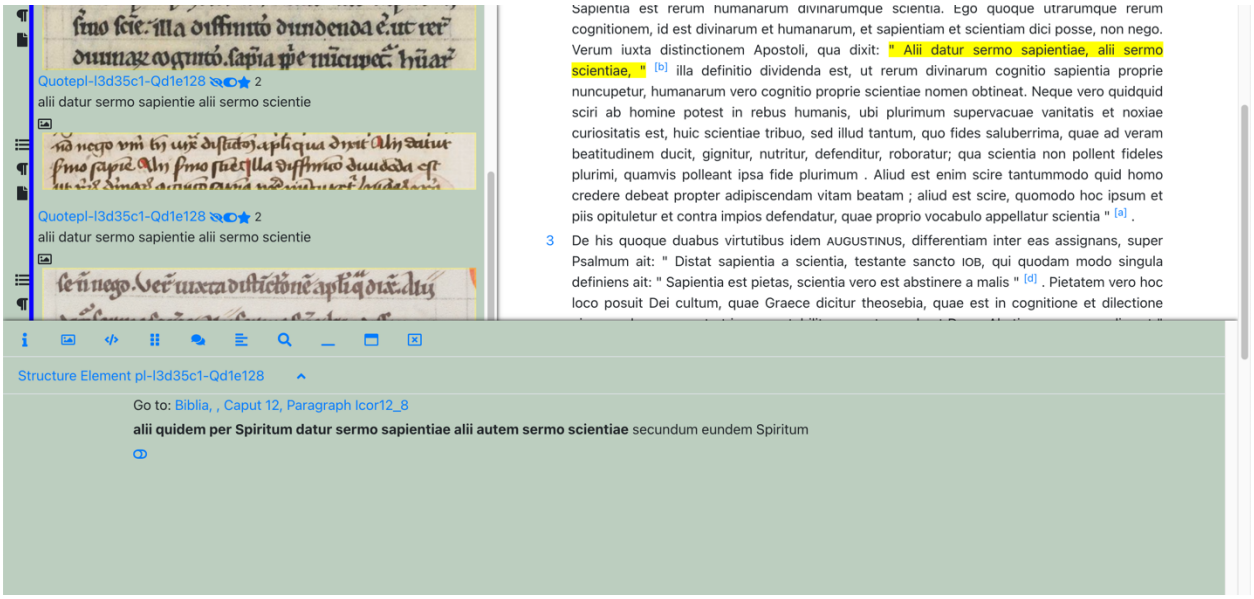



Figure 23.

Below, in Figure 24, we have changed our focus again to the quotation by Peter Lombard of Augustine's *De Trinitate,* book 14 (this is the critical passage noted at the outset containing the discussion of wisdom and science).

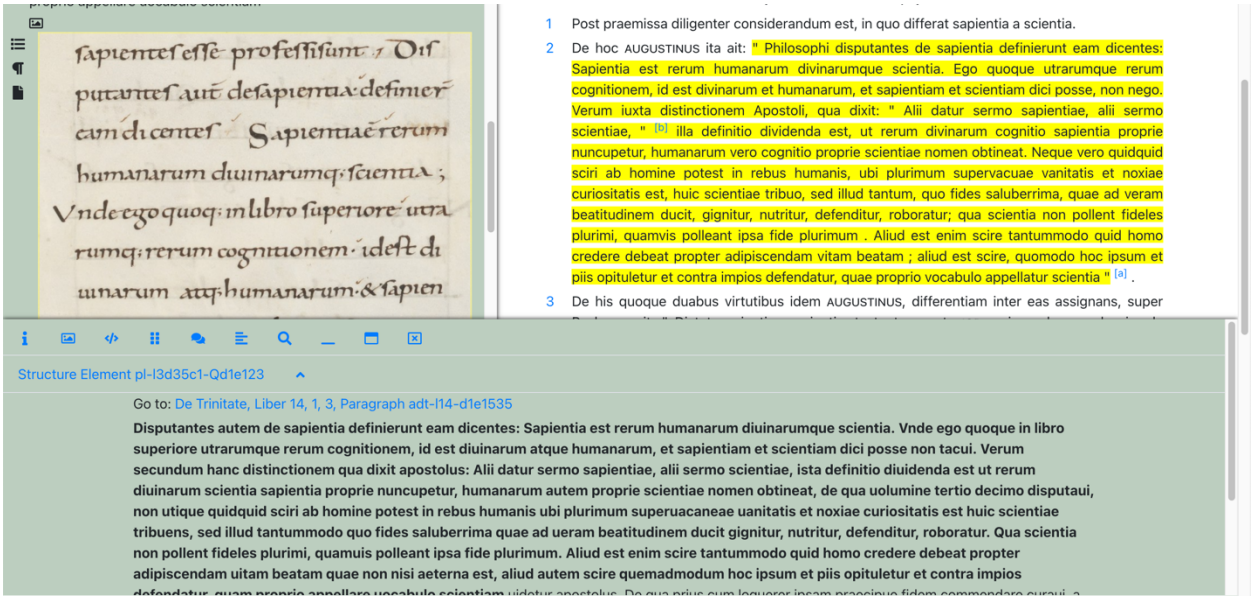



Figure 24.

From here we can move from the Lombard text to the Augustine passage he has quoted. In this, case, instead of looking at sister witnesses of the quoting text (Lombard's text), we can turn to



Journal of Data Mining and Digital Humanities http://jdmdh.episciences.org
**ISSN 2416-5999, an open-access journal**

inspect the Augustine text itself and its manuscript witnesses. Thus, in Figure 24 above, we are no longer looking at manuscripts containing Peter Lombard's *Sentences*, but we are looking at a witnesses to Augustine's *De Trinitate* at the precise coordinate region in the manuscript that Lombard is quoting. This is the second movement that we discussed earlier: moving to a source passage and then out to its manuscript witnesses.

In moving to the Augustine text, we can finally illustrate the third kind of movement discussed above: moving from a text passage, not to its sources, but to its influences (i.e., all passages in future texts that have quoted or referred to the target passage in some way) and from these passages to their manuscript sources.

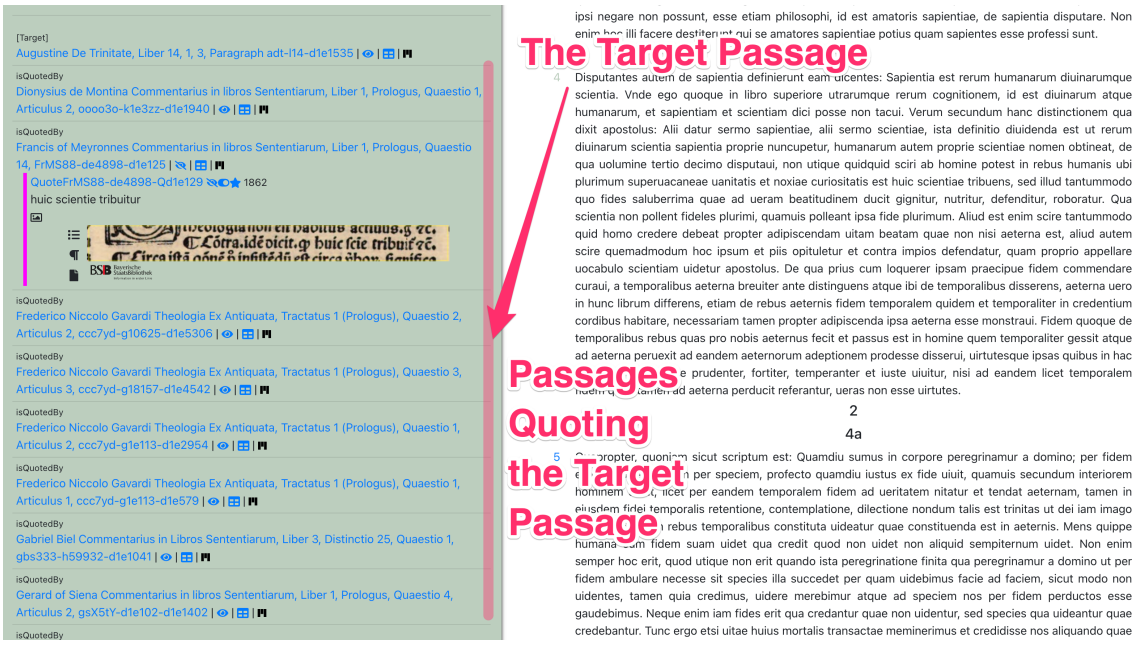



Figure 25.

As we can see here in Figure 25, each quoting passage has been inverted into an "isQuotedBy" property. This inversion lets us aggregate all the scattered passages quoting the Augustine target passage. [9] In turn, we have access to a comparable transcription and manuscript/codex image of each quoting passage. Each data layer can be shown without navigating away from the target passage, even though text and images are served from various libraries throughout the globe.

Finally, the ability to aggregate all these quotations together and the transcription of each witness allows us to expand our notion of a witness. No longer are we just talking about the original

[9] By "all the scattered passages," we mean all the passages that have been marked-up according to field standards, published in a machine accessible form, and indexed by the *Scholastic Commentaries and Texts Archive* [SCTA]. The SCTA's focus is on building an aggregated corpus of scholastic texts. Thus, each time a new text is published in this way, the SCTA index expands and the results seen above in Figure 25 are updated and improved.

witnesses to a text, but we can add the witnesses of any passage quoting the original text. In Figures 26–28 below, we can see that the witnesses of each quoting instance can be added to an ever-expanding collation table.

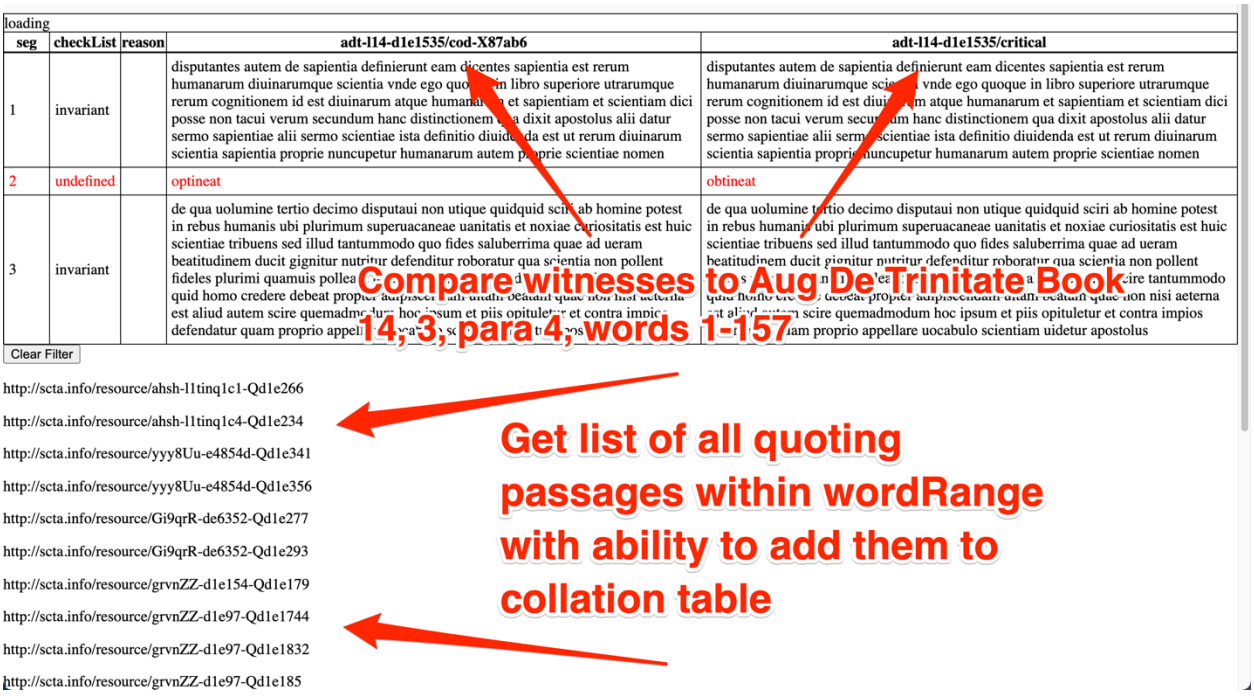



Figure 26.

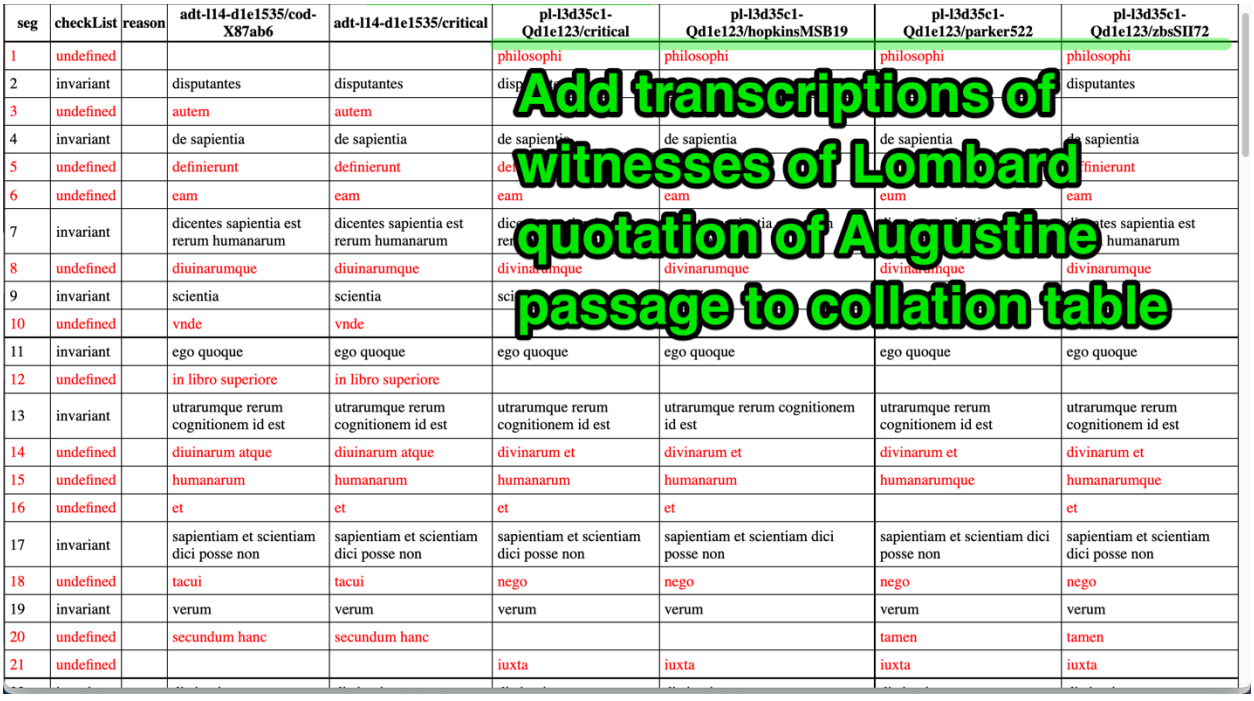



Figure 27.



Journal of Data Mining and Digital Humanities
**ISSN 2416-5999, an open-access journal**
http://jdmdh.episciences.org

Here in Figure 28, we can see how the word “scientia theologiae” makes its way into the tradition.


| seg | checkList | reason | adt-l14-d1e1535/cod-X87ab6 | adt-l14-d1e1535/critical | l1-Qstendr/critical | l1-Qstendr/reims | l1-Qstendr/sorb | l1-Qstendr/svict | pg-b1q1-Qd1e4272/critical | pl-l3d35c1-Qd1e123/critical | pl-l3d35c1-Qd1e123/hopkinsMSB19 | pl-l3d35-Qd1e123/par |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | undefined | | | | | | | | | philosophi | philosophi | philosophi |
| 2 | undefined | | disputantes | disputantes | | | | | | disputantes | disputantes | disputantes |
| 3 | undefined | | autem | autem | | | | | | | | |
| 4 | undefined | | de sapientia | de sapientia | | | | | | de sapientia | de sapientia | de sapientia |
| 5 | undefined | | definierunt | definierunt | | | | | | definierunt | deffinierunt | deffinierunt |
| 6 | undefined | | eam | eam | | | | | | eam | eam | eum |
| 7 | undefined | | dicentes sapientia est rerum humanarum | dicentes sapientia est rerum humanarum | | | | | | dicentes sapientia est rerum humanarum | dicentes sapientia est rerum humanarum | dicentes sapi est rerum humanarum |
| 8 | undefined | | diuinarumque | diuinarumque | | | | | | divinarumque | divinarumque | divinarumqu |
| 9 | invariant | | scientia | scientia | scientia | scientia | scientia | scientia | scientia | scientia | scientia | scientia |
| 10 | undefined | | vnde | vnde | theologiae | theologicae | theologiae | theologiae | scilicet theologia | | | |
| 11 | undefined | | ego quoque | ego quoque | | | | | | ego quoque | ego quoque | ego quoque |
| 12 | undefined | | in libro superiore | in libro superiore | | | | | | | | |
| 13 | undefined | | utrarumque rerum cognitionem id est | utrarumque rerum cognitionem id est | | | | | | utrarumque rerum cognitionem id est | utrarumque rerum cognitionem id est | utrarumque r cognitionem |
| 14 | undefined | | diuinarum atque | diuinarum atque | | | | | | divinarum et | divinarum et | divinarum et |
| 15 | undefined | | humanarum | humanarum | | | | | | humanarum | humanarum | humanarumq |
| 16 | undefined | | et | et | et | et | et | et | | et | et | |

Figure 28.

In Figure 29’s collation table, we can see the famous quotation about “not believing the Gospel unless moved to do so by the Church,” and we can see through the addition of a quoting witness that the word “compelleret” (as noted by Oberman) is creeping in and competing with “commoveret”. But we can also see that this is not the only variation. We also see instances of “moveret” and “approbanti”.



Journal of Data Mining and Digital Humanities
**ISSN 2416-5999, an open-access journal**
http://jdmdh.episciences.org

| | | | AgCMfu-sdf83a-d1e185 | l76-Qd1e1315/critical | l76-Qd1e1315/critical | ooij7j-d1e108-Qd1e1353/cod-nHHjHd | ppr-l76-Qd1e1315/erlang | ppr-l76-Qd1e1315/pal | yyy8Uu-e12439-Qd1e930/cod-MkL9py | yyy8Uu-e58209-Qd1e1387/cod-MkL9py |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | undefined | | ego | euangelio | euangelio | euangelio | | | ego | |
| 2 | undefined | | vero | | | | | | | |
| 3 | undefined | | evangelio | | | | evangelio | evangelio | | evangelio |
| 4 | invariant | | non | non | non | non | non | non | non | non |
| 5 | undefined | | crederem | crederem | crederem | crederem | crederem | credam | crederem | crederem |
| 6 | undefined | | | | | | | | evangelio | |
| 7 | invariant | | nisi | nisi | nisi | nisi | nisi | nisi | nisi | nisi |
| 8 | undefined | | | | | crederem | | | | |
| 9 | undefined | | me | | | | | | me | me |
| 10 | undefined | | | auctoritas | auctoritas | | auctoritas | auctoritas | | |
| 11 | undefined | | | | | | | | | ecclesi catholica moveret autoritas |
| 12 | undefined | | | | | | | | ecclesiae | |
| 13 | undefined | | catholicae | | | | | | catholicae | |
| 14 | undefined | | ecclesiae | ecclesiae | ecclesiae | ecclesiae | ecclesiae | ecclesiae | | |
| 15 | undefined | | | | | | | | autorita | |
| 16 | undefined | | | | | | me | me | | |
| 17 | undefined | | commoveret | | | approbanti | | | commoveret | |
| 18 | undefined | | auctoritas | | | | | | | |
| 19 | undefined | | | | | | conpelleret | conpelleret | | |
| 20 | undefined | | | hoc | ad hoc | | ad hoc | ad hoc | | |
| 21 | undefined | | | me compelleret | me compelleret | | | | | |

A collation of future quotations of: Augustine, Contra epistolam Manichaei quam vocant fundamenti, Liber 1, 5, 6, Paragraph AgCMfu-sdf83a-d1e185 @ 139-149... ...starts to show how other words are beginning to replace "commoveret"... ...words like "compelleret", "approbanti", and simply "moveret"

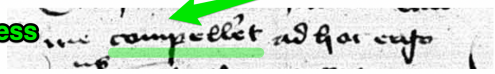
Figure 29.

Once again, each data point is fully transparent. Clicking on any word in the collation table for any witness, as seen below in Figure 30, can call up the precise coordinate region in the respective manuscript where the word occurs.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | undefined | | ego | euangelio | euangelio | euangelio | | | ego | |
| 2 | undefined | | vero | | | | | | | |
| 3 | undefined | | evangelio | | | | evangelio | evangelio | | evangelio |
| 4 | invariant | | non | non | non | non | non | non | non | non |
| 5 | undefined | | crederem | crederem | crederem | crederem | crederem | credam | crederem | crederem |
| 6 | undefined | | | | | | | | evangelio | |
| 7 | invariant | | nisi | nisi | nisi | nisi | nisi | nisi | nisi | nisi |
| 8 | undefined | | | | | crederem | | | | |
| 9 | undefined | | me | | | | | | me | me |
| 10 | undefined | | | auctoritas | auctoritas | | auctoritas | auctoritas | | |
| 11 | undefined | | | | | | | | | ecclesi catholica moveret autoritas |
| 12 | undefined | | | | | | | | ecclesiae | |
| 13 | undefined | | catholicae | | | | | | catholicae | |
| 14 | undefined | | ecclesiae | ecclesiae | ecclesiae | ecclesiae | ecclesiae | ecclesiae | | |
| 15 | undefined | | | | | | | | autorita | |
| 16 | undefined | | | | | | me | me | | |
| 17 | undefined | | commoveret | | | approbanti | | | commoveret | |
| 18 | undefined | | auctoritas | | | | | | | |
| 19 | undefined | | | | | | conpelleret | | | |
| 20 | undefined | | | ad hoc | ad hoc | | ad hoc | | | |
| 21 | undefined | | | me | me | | | | | |
| 22 | undefined | | | compelleret | compelleret | | | compelleret | | |
| 23 | undefined | | | | | | | ad hoc | | |

Selected Segment to Inspect

Immediate and Transparent Access to Transcription Source



141r(281),a - line: 30: me conpelleret ad hoc ergo

view in Ibpweb

Figure 30.

### 3.2 From Text Analysis to New Connections

In the examples above, the pathways between quoted and quoting texts result from the editorial work that examines a text and identifies source targets. This is by and large the traditional work of constructing an *apparatus fontium*. The only difference—and it is a powerful difference—is that the result of tracking down these sources is recorded as a machine-actionable link rather than as a footnote confined to the bottom of a page. The machine-actionable nature of this link lets us invert and automatically create an index of influences rather than a mere index of sources.

However, this traditional work can be assisted by various forms of automated analysis that can recommend sources to an editor as probable cases of passage re-use. In short, the automated analysis can itself recommend pathways through the network. These pathways can be provided directly to end users or can be used by an editor to detect and confirm cases of passage re-use. In either case, the full network is at the user's disposal. While textual re-use detection analyzes textual similarity, the embeddedness of these text passages within a web of links to corresponding coordinate regions lets the automated process connect not only texts but also parallel coordinate regions within scattered manuscripts across the globe.

The method of similarity detection used in the following examples is computationally simple. In this case, the corpus of scholastic texts constructed and maintained by the *Scholastic Commentaries and Texts Archive* [SCTA] is crawled and indexed into a list of n-grams of size 4. Each n-gram takes a property of "isFoundIn" and the object of this property is any paragraph passage that contains the n-gram in question. Passage similarity is then determined through an n-gram intersection threshold. In this case, the threshold was set at 6. This means that any two passages that share 6 unique n-grams of size 4 are considered similar. This threshold can be adjusted in order to fine-tune results.

Using the corpus metadata and textual hierarchies, we can then request all the paragraph passages within a given range of text and then ask for all "similar" passages. The results can be visualized as a kind of similarity matrix useful for consultation, inspection, and verifiability. To illustrate the results of this approach we can look at two examples. The first example (3.2.1) is small and requires only a brief explanation. The second example (3.2.2) is more involved and requires a more extended explanation.

#### 3.2.1 Example 1.

Staying focused on the important passage from Augustine that discusses "sapientia" and "scientia", we can query for all paragraph passages within Augustine's *De Trinitate*, book 14 and then query for all passages that have been identified as potentially similar. The results are visualized below in Figure 31. Along the x-axis there is a column for every paragraph in book 14 of Augustine's *de Trinitate*. Along the y-axis, we list every paragraph (in chronological order) in the SCTA corpus that is deemed "similar" to a paragraph in the Augustine text. The blue cells within the chart indicate which paragraph along the y-axis is similar to which paragraph of the Augustine text in the x-axis.

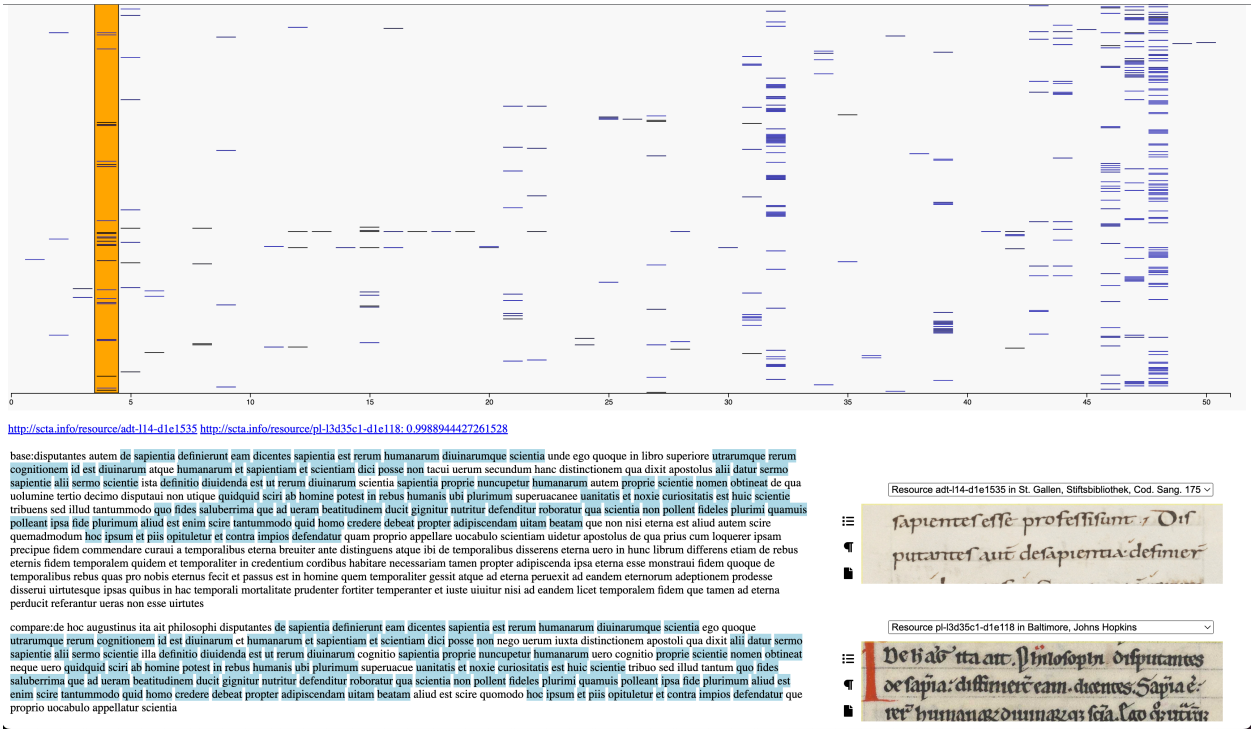



Figure 31.

The similarity matrix seen in Figure 31 shows several vertical columns indicating high textual re-use of text fragments within specific paragraphs in Augustine's *De Trinitate*, book 14. Beyond merely showing us these patterns of re-use, each computer recommended pairing is fully transparent. The underlying data layers the computer uses to report similarity can be immediately inspected. Clicking on a given match reveals the "base" and "compare" passages with n-gram intersections highlighted. Finally, the corresponding manuscript coordinate regions for each passage appear in consort, allowing the end user full transparency of all data layers used to generate the similarity recommendation.

### 3.2.2 Example 2

Example 2 uses the same technique but offers us a real-world example of how the text-image network provides genuine and meaningful discoveries. Appreciating such discoveries requires a more extended explanation.

In 2008, Venicio Marcolino [Marcolino, 2008] published an article attempting to show the dependence of the *Sentences* Commentary by the 14th-century Augustinian Peter Gracilis on the *Sentences* Commentary of an earlier Augustinian, John Hiltalingen of Basel. Hiltalingen was a Bachelor of Theology in the 1360s and became a Master of Theology in 1371.[10] Gracilis wrote (or

[10] See [Zumkeller, 1964, p. 231].

one might argue "compiled from sources") his own commentary around the mid-1370s[11] and became licensed to teach in 1381 and a Master of Theology by at least 1385.[12]

The result of Marcolino's article is roughly 30 pages of tables showing where he has found parallels between the earlier text of Hiltalingen and that of Gracilis. It is a remarkable achievement considering that neither text was available in a printed critical edition at the time. As best he could, Marcolino painstakingly observed the text in each manuscript and recorded the parallels that he found.

Unfortunately, this highly repetitious approach (scanning one paragraph, checking for parallels in the other, and then moving on to the next) was a painstakingly manual process for Marcolino, who worked page by page, line by line. His achievement was partially possible because of Gracilis's previously known dependence on Hiltalingen. [Damasus Trapp, 1956, p. 254], on initially recognizing the dependence in 1956, wrote:

> Petrus Gracilis...followed not only the footsteps but the very phrases of Hiltalingen in a way so deceptive that it does not cast the best light on Gracilis. He read *secundum Hiltalingen* without ever mentioning him. Only by a *lucky coincidence* [emphasis mine] was I enabled to "unmask" Gracilis' dubious literary honesty.

While withholding judgment about whether Trapp is correct to accuse Gracilis of "dubious literary honesty," Trapp's admission that his discovery was the result of a "lucky coincidence" speaks directly to the anecdotal nature of this approach and why a text-image network that can be used to systematically detect parallels is desirable. Without thinking systematically about the scholastic corpus as data and the discovery methods that could be applied to that data at scale, we are left with the unscientific methodologies of "coincidence" and "luck."

Marcolino's detections were likewise helped by the systematic way that Gracilis tended to follow Hiltalingen. In nearly all cases, Gracilis followed him distinction by distinction, question by question.[13] Thus, once the alignment of questions was found, Marcolino could use his understanding of the structure of a *Sentences* Commentary (i.e., the organization around topic-based distinctions) to begin at a particular distinction in Gracilis and then move to the parallel distinction in Hiltalingen where text re-use was expected to be highly likely.

Nevertheless, despite his achievement, Marcolino would likely admit (and his article does admit in the form of numerous question marks indicating places of likely but as of yet undetected textual reuse) that the results could be more complete and precise. First, the article format of presenting his data creates ambiguity. The format limits how precisely he could communicate the parallels he saw. Thus, he settled on asserting parallels between sections (using section headings as navigation markers), even though many of these sections have several paragraphs by both

[11] [Marcolino, 2008, p. 497] suggests 1377 as a correction to the older dating of 1375/1376 found in Zumkellar, 1964, p. 236] and [Sullivan, 2004, p. 196], who follows Zumkellar.
[12] See [Sullivan, 2004, p. 196].
[13] According to [Marcolino, 2008, p. 500], Gracilis substantially follows Hiltalingen in 35 out of 40 questions in book one of his *Sentences* Commentary.

authors, only some of which show dependence. For example, in the case of Gracilis's third conclusion and Hiltalingen's fourth conclusion (highlighted in Figure 32 below), the alignment gives the impression that the passages in Gracilis are mere copies of the corresponding passages in Hiltalingen.

**Quaestio 1** (12r-15r): *Utrum quis viator scientifice sine dono speciali possit assentire theologice omni vero supernaturali.*

| | |
|---|---|
| Quod sic: 1°-4° | Q. 1 princ. 1, 2, 4, 5 |
| In oppositum: 1°-2° | Q. 1 Ad oppositum: 1°, 3° |
| Prima conclusio | Facinus, 1 prol. q. 1 prop. 1 |
| Corollarium 1-3 | Facinus, 1 prol. q. 1 prop. 1 corol. 1-3 |
| Secunda conclusio | Facinus, 1 prol. q. 2 prop. 1-2 |
| Corollarium 1-2 | Facinus, 1 prol. q. 2 prop. 1-2 corol. |
| Corollarium 3 | ? |
| Tertia conclusio | Q. 1 a. 3 concl. 4 |
| Corollarium 1-3 | Q. 1 a. 3 concl. 4 corol. 1-3 |
| Contra conclusionem 1 : 1°-5° | Facinus, 1 prol. q. 1 contra prop.: 1°-5° |
| : 6° | Q. 1 a. 4 contra concl. 3: 2° |
| Contra conclusionem 2 et corol.: 7°-8° | Q. 1 a. 4 dub. 3 |
| Contra conclusionem 3 : 9° | Q. 1 a. 4 contra concl. 4: 1° |
| : 10°-11° | ? |
| Ad argumenta contra conclusionem 1: 1°-5° | Facinus, 1 prol. q. 1 ad. obiect. |
| : 6° | Q. 1 a. 5 ad obiect. contra concl. 3: 2° |
| Ad arg. contra conclusionem 2 et corol.: 7°-8° | Q. 1 a. 5 dub. 3 |
| Ad argumenta contra conclusionem 3: 9° | Q. 1 a. 5 ad obiect. contra concl. 4:1° |
| : 10°-11° | ? |
| Ad rationes principales: 1°-4° | Q. 1 ad princ. 1, 2, 4, 5 |

Figure 32.

However, if we look closely at a few examples of the parallel between Gracilis's third conclusion and the parallel passages in Hiltalingen (see Figures 33 and 34), we can clearly see that they are not mere copies. They are in fact highly selective.



Journal of Data Mining and Digital Humanities http://jdmdh.episciences.org
**ISSN 2416-5999, an open-access journal**

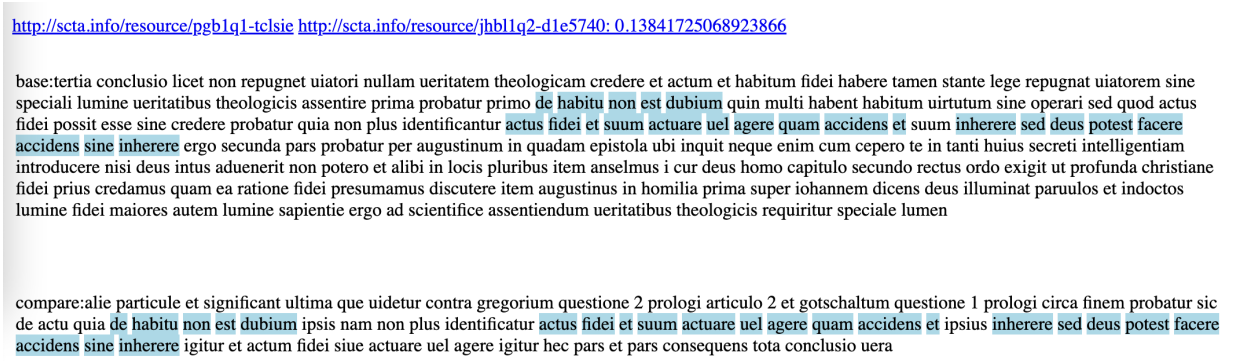



Figure 33.

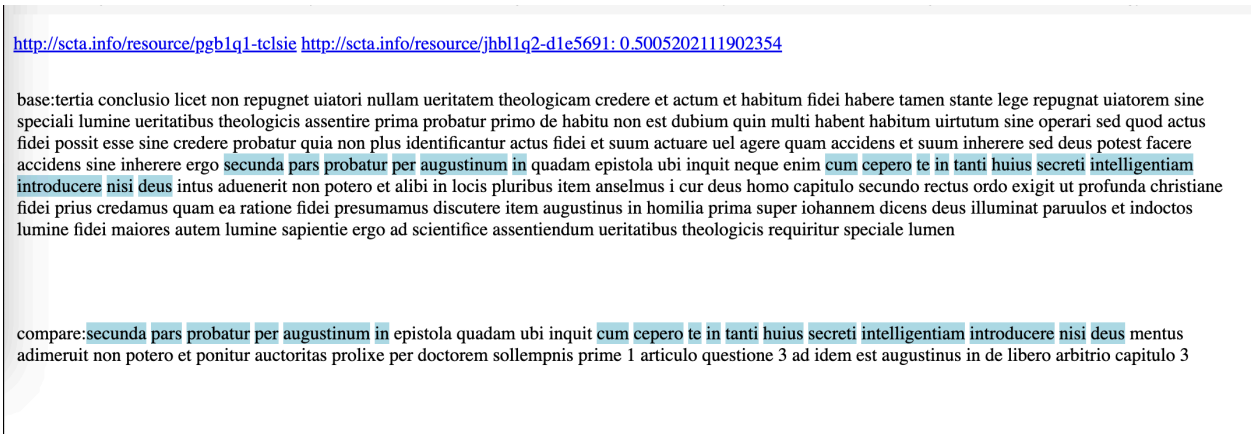



Figure 34.

The passages are dependent, to be sure. However, Marcolino can only point us to this general dependence; he cannot show us with granularity the highly selective way Gracilis has re-used phrases and sentences while at the same ignoring large sections of Hiltalingen's text and intermingling his own ideas.

Second, beyond the ambiguity of asserted parallels, question marks permeate his tables at junctures where he believes Gracilis is compiling from other sources, but for which he has not found another source. See Figure 35.

**Quaestio 25** (77r-79r): *Utrum summum independens referatur ad contingens.*
*Utrum trinum simplex invariabile referatur ad contingens seu producibile.*
*Utrum per productum possit formaliter inferri Deum ad creata vere realiter referri.*

| | |
|---|---|
| Quod non: 1°-2° | ? |
| : 3° | Q. 25 princ. 3 |
| In oppositum | Q. 25 In oppositum |
| Prima conclusio | ? |
| Corollarium 1-3 | ? |
| Secunda conclusio | ? |
| Corollarium 1-3 | ? |
| Tertia conclusio | Q. 25 concl. 3 |
| Corollarium 1-3 | ? |
| Ex dictis sequitur: 1° | ? |
| Contra conclusionem 1: 1°-3° | ? |
| Contra corollarium 1: 4°-5° | ? |
| Contra corollarium 2: 6° | ? |
| Contra conclusionem 2: 1°-5° | ? |
| Contra conclusionem 3: 1°-2° | ? |
| Ad argumenta contra conclusionem 1: 1°-3° | ? |
| Ad arg. contra corollarium 1: 4°-5° | ? |
| Ad arg. contra corollarium 2: 6° | ? |
| Ad argumenta contra conclusionem 2: 1°-5° | ? |
| Ad argumenta contra conclusionem 3: 1°-2° | ? |

Figure 35.

Third and finally, at some junctures—for example at questions 28, 30, 35 and question 38 (see Figure 36 below)—he offers no parallels at all.


Journal of Data Mining and Digital Humanities http://jdmdh.episciences.org
**ISSN 2416-5999, an open-access journal**

**Quaestio 37** (110v-114v): *Utrum per divinae voluntatis vigorem ordinentur et regulentur omnia ad universi decorem.*
*Utrum per divinae voluntatis rectitudinem ordinentur omnia ad universi pulchritudinem.*

| | |
|---|---|
| Quod non: 1° | ? |
| : 2°-3° | Q. 33 princ. 1, 3 |
| In oppositum | Q. 33 In oppositum |
| Prima conclusio | ? |
| Corollarium 1-5 | ? |
| Secunda conclusio | Q. 33 concl. 2 |
| Corollarium 1-3 | Q. 33 concl. 1 corol. 1-3 |
| Tertia conclusio | Q. 33 concl. 3 |
| Corollarium 1-3 | Q. 33 concl. 3 corol. 1-3 |
| Contra conclusionem 1: 1°-4° | ? |
| Contra conclusionem 2: 1° | Q. 33 contra concl. 2 |
| : 2°-4° | Q. 33 contra concl. 1: 1°-3° |
| Contra corollaria: 5°-6° | Q. 33 contra concl. 1 corol. 1-2 |
| Contra conclusionem 3: 1°-6° | ? |
| Ad argumenta contra conclusionem 1: 1°-4° | ? |
| Ad argumenta contra conclusionem 2: 1° | Q. 33 ad obiect. contra concl. 2 |
| : 2°-4° | Q. 33 ad obiect. contra concl. 1: 1°-3° |
| Ad arg. contra corollaria: 5°-6° | Q. 33 ad obiect. contra concl. 1 corol. 1-2 |
| Ad argumenta contra conclusionem 3: 1°-6° | ? |
| Ad rationes in oppositum: 1° | ? |
| : 2°-3° | Q. 33 concl. 1 ad princ. 1, concl. 3 ad princ. 3 |

**Quaestio 38** (115r-v, 117r-119v): *Utrum divinum velle rerum immediatum productivum principium sit non esse earum causale principium positivum.*

**Quaestio 39** (120r-122v): *Utrum quodlibet bonum possibile aliud a prima bonitate sit in se ponibile ab increata voluntate.*

| | |
|---|---|
| Quod non: 1°-3° | Q. 34 princ.3, 2, 1 |
| In oppositum | Q. 34 In oppositum |
| Prima conclusio | Q. 34 concl. 3 |
| Corollarium 1-3 | Q. 34 concl. 3 corol. 1-3 |
| Secunda conclusio | Q. 34 concl. 2 |
| Corollarium 1-3 | ? |
| Tertia conclusio | Q. 34 concl. 1 |
| Corollarium 1-2 | Q. 34 concl. 1 corol. 1-2 |
| Corollarium 3 | ? |
| Contra conclusionem 1: 1°-5° | Q. 34 contra concl. 3 |
| Contra conclusionem 2: 1°-6° | Q. 34 contra concl. 2 |
| Contra conclusionem 3: 1° | Q. 34 contra concl. 1: 2° |
| Contra corollaria: 2°-3° | Q. 34 contra concl. 1 corol. 1-2 |
| : 4°-5° | ? |



Journal of Data Mining and Digital Humanities http://jdmdh.episciences.org
**ISSN 2416-5999, an open-access journal**

Figure 36.

The lack of stated parallels, set against the backdrop of many questions with identified parallels, leaves it ambiguous to the reader whether questions like question 38 are unique questions in which Gracilis has abandoned his role as compiler and has suddenly made an original contribution or whether Marcolino simply has not yet found any parallels.

Marcolino's focus is no doubt on finding parallels within the texts he knows. This means, first and foremost, the work of Hiltalingen. But where he can, he identifies other parallels in other authors. For example, he has found parallel passages in parallel distinctions of other Augustinian writers such as Facinus de Ast[14] or Thomas de Argentina.[15] Given Gracilis's status as a member of the Augustinian Order and the predictable structure of a *Sentences* commentary, these are logical and predictable places to look.

This, however, also shows us the limits of the method. As no individual can read the entire scholastic corpus, our collective endeavor to understand the inter-textuality and dependence/influence of scholastic texts will be guided by our current expectations. These expectations, in turn, will be derived from what we know. Marcolino's article (and the parallels he has found) shows us what he knows and expects while occluding the unexpected. Here, Marcolino understands Gracilis, the Augustinian, to be a compiler of Augustinian sources. Thus, he scours the Augustinian sources in the expected parts of their respective *Sentences* commentaries, and his results confirm his expectations.

But what would happen if we had access to the entire scholastic tradition as data transformed into a text-image network that a machine could analyze at a scale not achievable by a human being? Might such a corpus not only let us make Marcolino's results more precise and transparent but also show us the unexpected? Connecting texts with other texts we had not anticipated? And through these texts automatically connecting manuscripts regions that were previously isolated?

We can do this by using the simple n-gram similarity detection method described above. In Figure 37 below, we have arranged every paragraph within the Gracilis text along the top horizontal x-axis. Running down the vertical y-axis are any paragraphs detected within the SCTA corpus that meet the similarity criteria. Texts are arranged chronologically and then paragraphs within a containing text are arranged in sequential order. This simple arrangement reveals some clear patterns of dependence.

[14] See [Marcolino, 2008, p. 503].
[15] See [Marcolino, 2008, p. 508].

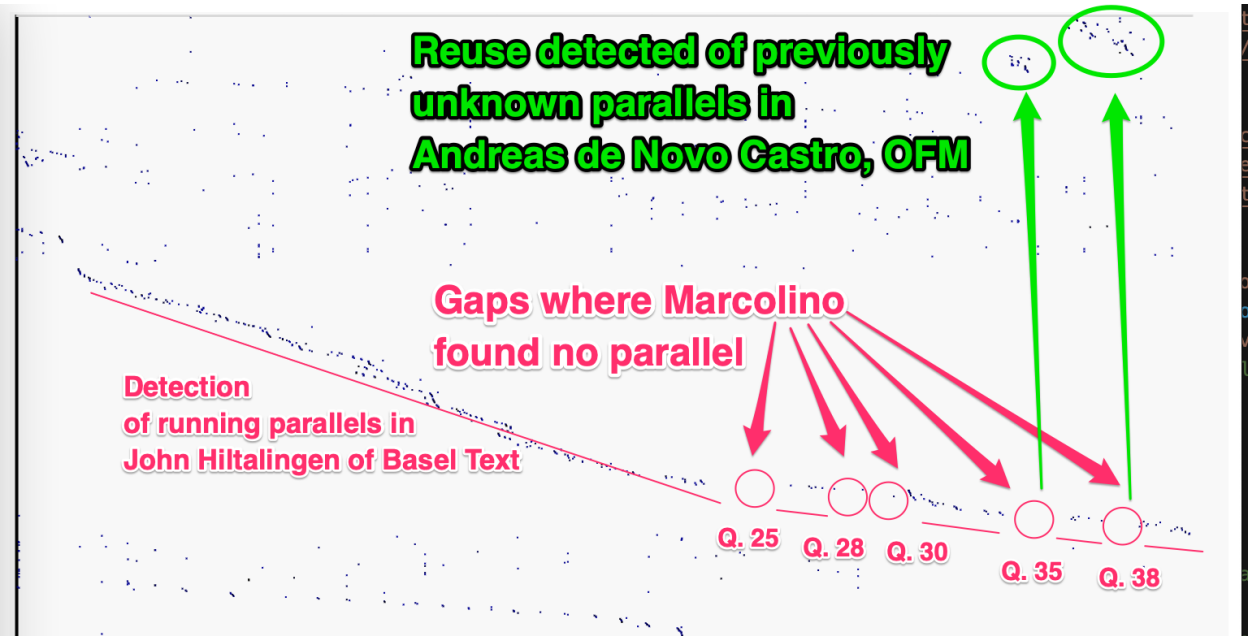



Figure 37.

Toward the middle of the graph, a strong clear blue line moving gradually downward can be detected. This line shows the habitual way that Gracilis has appropriated the text of John Hiltalingen of Basel throughout his own text. It moves diagonally because Gracilis consistently follows Hiltalingen in the same order. So, as we move one paragraph to the right in the Gracilis text, we generally move to the following paragraph in the Hiltalingen text, and thus move one step down in the vertical y-axis.

Figure 37 also clearly shows some of the presumed gaps noted by Marcolino, for example, in questions 25, 28, 30, 35, and 38. These are places where Gracilis's habitual practice of following Hiltalingen temporarily stops, and where Marcolino has noted few or no parallels (e.g., see Figure 36 above).

However, Figure 37 also shows us something new. In addition to showing the gaps in dependence on Hiltalingen, it has also detected other clusters in other texts precisely at the moment where no parallel in Hiltalingen was detected. In the highlighted example, parallels to questions 35 and question 38 have been found in the earlier work of Andreas de Novo Castro, OFM, who was most likely composing his *Sentences* Commentary in the late 1350s, perhaps 1358-1359.[16] While upon inspection, the textual re-use is straightforward and clear, it is off Marcolino's radar

[16] For an insightful and thorough account of de Novo Castro's life and the surrounding scholarly debate, see [Idziak, J., 1997]. De Novo Castro's dates have been subject to much debate, but there seems to be growing consensus that he probably delivered his lectures on the *Sentences* in the late 1350s. In fact, evidence of Gracilis's textual re-use can even contribute positively to this debate, helping us to confirm that de Novo Castro must have written before Gracilis's composition in 1377. This would help definitively rule out the earlier suggestion of some scholars that de Novo Castro ought to be situated in the 15th or even 16th centuries. See [Idziak, 1997, p. 136] for these much earlier theories.

because he expected to find dependence on other Augustinian writers and de Novo Castro was a Franciscan. This reminds us once again of the limitations of the anecdotal approach and the need for a more exhaustive methodology. The anecdotal approach confirms what we already expect while overlooking what we do not.

Because parallels are arranged in sequential order, rough diagonal patterns indicate sustained textual re-use. Diagonal patterns occur when successive paragraphs in one text show a high degree of similarity in successive paragraphs in a matched text.

Once aware of this pattern, we can detect it not only visually but algorithmically by beginning at a given match point in a text and then asking if its previous paragraph matches an earlier paragraph in the same matched text. When this happens repeatedly (e.g., three or more times), we can score this as a high-probability match. Finally, we can take the sum of all of these high-quality matches within a text section (e.g., a *quaestio* or chapter). Where the sum for a given text exceeds a threshold point, in this example 10 or more times, this text becomes a very likely candidate for *sustained* textual re-use.

We can see this threshold at work, first, through the confirmation of the parallels in the Hiltalingen text that Marcolino already noted. Consider Figure 38.


Figure 38.

Figure 38 shows us that for Quaestion 12 of Peter Gracilis, small clusters in a few different texts have been detected. These likely indicate a common quotation or repeated technical phrases. However, the entry for "jhb-l1q13" points to 42 sequence clusters, far exceeding the clusters

detected in other texts. This text happens to be precisely the text for which Marcolino asserted the known parallel.

We can repeat this process for Question 35 and Question 38 for which Marcolino identified no known parallel. Consider Question 35 in Figure 39 below.


Figure 39.

The passage from de Novo Castro (D78nca-f12127) shows a very high number of similarity clusters just as we saw in the Hiltalingen parallel (Figure 38). And this detection matches the the visualized clusters for Question 35 seen in Figure 37.

Using the graph network of metadata, we can restrict the visualization to only show detected parallels for this question against the detected question of de Novo Castro. In this way, we can programmatically zoom in on the fragment of the visualization suggested by the algorithm. From here, we can call up textual parallels used by the computer to suggest the similarity and confirm it for ourselves. Finally, we can move outward from these granular text passages to the parallel coordinate regions in each manuscript where the connected texts are manifested in each manuscript. For Question 35, this zoomed-in and transparent perspective can be seen in Figure 40.

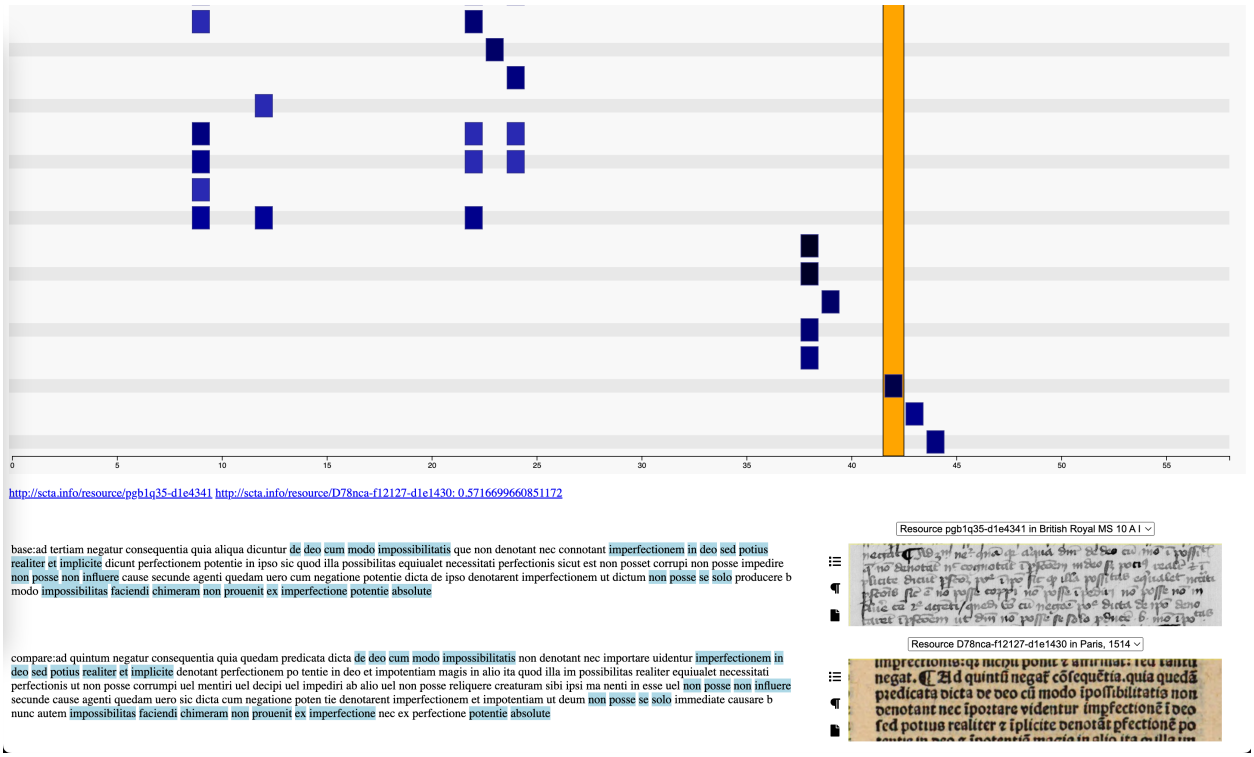



Figure 40.

This pattern of detection is once more clear in the case of Question 38. Figure 41 shows the significant re-use of passages by de Novo Castro algorithmically detected by the computer.


Figure 41.

By zooming in (seen in Figure 42), we can inspect and confirm.

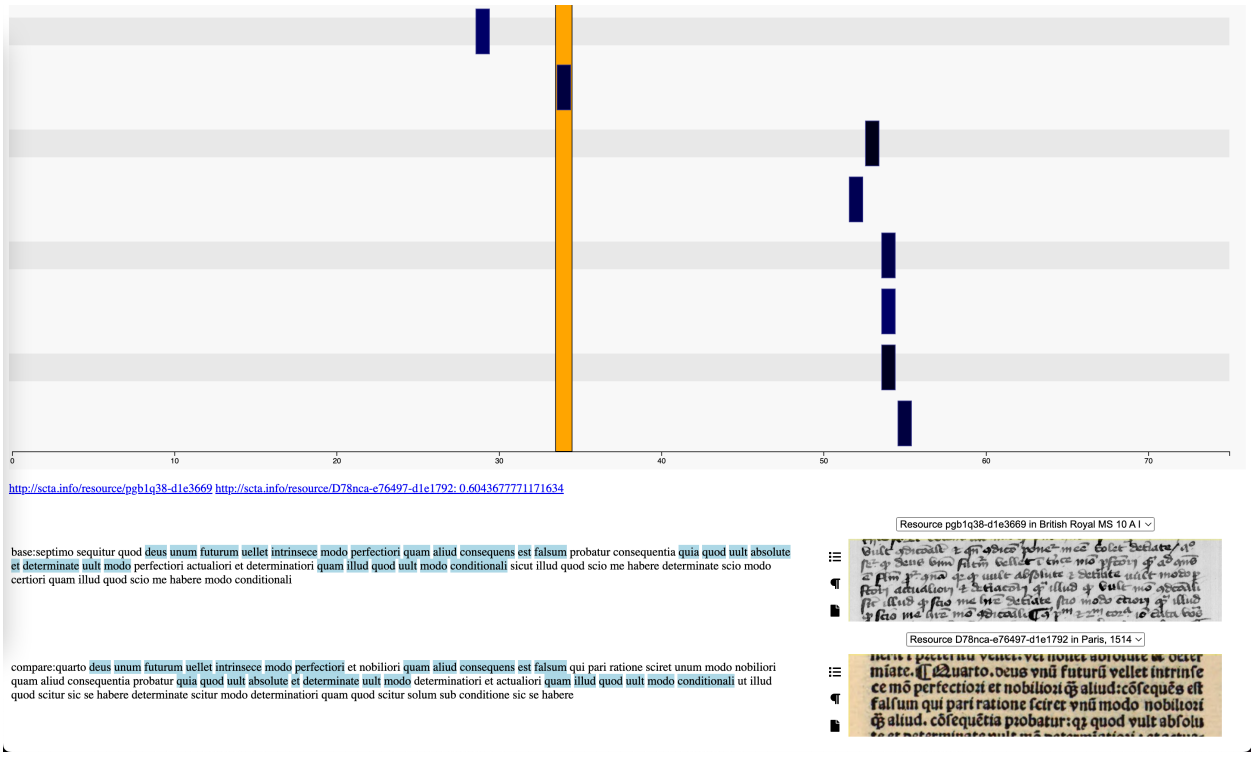



Figure 42.

To summarize, the text-image network lets the researcher move instantly from an abstract analytical level to seamlessly inspect every data layer that underlies these computational results.

## CONCLUSION

The list of passage re-use detection by the above method is growing and a full report is better left for another article. For now, these use-case studies are enough to highlight the exciting future for manuscript studies when texts and image data are interconnected in an elaborate web at granular levels.

It is also a reminder that even these computational constructed connections do not happen automatically. These outcomes depend on lower data management layers that have accepted the challenge of thinking about texts and images in new ways. Text and images alike must be released from the bounds of presentation, whether that be a digital or print presentation. Instead, they must be published first as atomized data resources that can be combined and re-combined into a thousand different forms of presentations, intricately customized to the needs of the researcher and their unique research questions.[17] Likewise, such combination and recombination require new ways of cataloging texts and their parts at levels of granularity. But this approach is foreign to

[17] These atomic and re-combinable units echo the vision seen Ted Nelson's idea of Xanalogical Storage in which new presentations of text can be programmatically constructed from a common "pool of data". See [Nelson, 1981, chapter 0, p. 0/6].

most traditional library practices. We will thus need a collective effort in the form of academic societies and archives to maintain these data models and to incentivize these new publishing practices. As we have seen in these examples, such an effort for the data considered here (i.e., medieval scholastic texts) is already underway in the form of the *Scholastic Commentaries and Texts Archive* [SCTA].With the proper support, the possibilities demonstrated here will only grow and, in turn, will enable a new and exciting future for manuscript and textual studies.

## REFERENCES


Augustinus. *De Trinitate*. Edited by Mayer, C. in *Corpus Augustinianum Gissense (CAG)*, Electronic edition. (Charlottesville, Virginia) 2000.

Augustine. *Contra epistolam Manichaei quam vocant fundmanenti*. Edited by Migne, J-P. in *Patrologia Latina* 1865;42.

Bennett, R., Lavoie, B., O'Neill, E. The Concept of a Work in WorldCat: An Application of FRBR. *Library Collections, Acquisitions and Technical Services* 2003;27(1).

Bordalejo, B. Well-Behaved Variants Seldom Make the Apparatus: Stemmata and Apparatus in Digital Research. *Digital Medievalist*. 2021;14(2). https://doi.org/10.16995/dm.8065.

Brown, S. Peter of Candia's Hundred-Year "History" of the Theologian's Role. *Medieval Philosophy and Theology*. 1991;1:156–190.

Bush, V. As We May Think. *The Atlantic*, 1945. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

Christensen, M., Witt, J., and Zahnd, U. Re-Conceiving the Christian Scholastic Corpus with the Scholastic Commentaries and Texts Archive. *Digital Humanities and Christianity*. Edited by Tim Hutchings and Claire Clivaz. De Gruyter (Berlin), 2021:47–76. https://doi.org/10.1515/9783110574043-003.

Huskey, S. The Visual [Re]Presentation of Textual Data in Traditional and Digital Critical Editions. *Magazén*. 2022;1. https://doi.org/10.30687/mag/2724-3923/2022/05/005.

IIIF. International Image Interoperability Framework. https://iiif.io.

Cassin, Matthieu. International Standard Manuscript Identifier. 2018 https://www.csmc.uni-hamburg.de/written-artefacts/working-groups/permanent-seminar/conference-contributions/files/cassin-pres.pdf.

David Rumsey Map Collection. https://www.davidrumsey.com/.

SCTA. Scholastic Commentaries and Texts Archive. https://scta.info.

Idziak, J. The Disputed Biography of Andreas de Novo Castro. In *Questions on an Ethics of Divine Commands*. Notre Dame Press (Notre Dame, IN), 1997:136-157.

Marcolino, V. Zum Abhängigkeitsverhältnis Der Sentenzenkommentare Der Augustinertheologen Petrus Gracilis († n. 1393) und Iohannes von Basel († 1392). *Analecta Augustiniana*. 2008;71:493–529.

Nelson, T. *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 (Sausalito, California), 1981.

Noerr, P. User Benefits from a New Bibliographic Model: Follow-up of the IFLA Functional Requirements Study. 64th IFLA general conference. 1998. http://www.ifla.org/IV/ifla64/084126e.htm.

Oberman, H. *The Harvest of Medieval Theology. Gabriel Biel and Late Medieval Nominalism.* Harvard University Press (Cambridge), 1963.

O'Neill, E. FRBR: Application of the Entity-Relationship Model to Humphry Clinker. *Library Resources and Technical Services* 2002;46(4). http://www.oclc.org/research/publications/archive/2002/oneill_frbr22.pdf.

Rosemann, P. *The Story of a Great Medieval Book: Peter Lombard's Sentences*. Broadview Pres (Toronto), 2007.

Sahle, P. Zwischen Mediengebundenheit und Transmedialisierung. *Editio*. 2010;24:23–36, https://doi.org/10.1515/9783110223163.0.23.

Sullivan, T. *Parisian Licentiates in Theology, A.D. 1373-1500: A Biographical Register: Vol. 1. The Religious Orders*. Brill (Leiden), 2004.

Trapp, D. Augustinian Theology of the 14th Century. *Augustiniana*. 1956;6:147–274.

Witt, J. Interpreting Augustine: On the Nature of 'Theological Knowledge' in the Fourteenth Century. *Studia Patristica*. 2013;69:359–71.

Witt, J. Tradition, Authority, and the Grounds for Belief in Late Fourteenth-Century Theology. *Augustine in Late Medieval Philosophy and Theology*. Edited by Slotemaker, J. and Witt, J., in *Studia Patristica.* 2017;87:147-160.

Witt, J. Digital Scholarly Editions and API Consuming Applications. *Digital Scholarly Editions as Interfaces*. 2018;12:219–47.

Zumkeller, A. Die Augustinerschule des Mittelalters: Vertreter Und Philosophisch-Theologische Lehre. *Analecta Augustiniana*. 1964;27:167–262.