


# TEI Beyond XML – Digital Scholarly Editions as Provenance Knowledge Graphs

Andreas Kuczera   
University of Applied Sciences (THM)  
Gießen, Germany

## Abstract

This paper proposes to detach TEI semantics – a widely accepted standard for the description of textual phenomena – from its hierarchical XML framework in order to integrate its descriptive structures into a digital scholarly edition (DSE) of Hildegard von Bingen’s *Liber epistolarum* based on a knowledge graph enriched with provenance information.

*To which problem is digitization the solution?*  
(Nassehi, 2019)<sup>1</sup>

## 1 Introduction

The search for origins is a quintessential human activity. Scholars in the humanities – especially historians – engage in this activity by examining cultural artefacts such as texts, objects, and images. In so doing, they make use

---

 *Creative Commons License Attribution 4.0 International (CC BY 4.0).*

In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at <http://ceur-ws.org>

This long paper is based on research presented at “Graph Technologies in the Humanities 2019” (January 18-19, Academy of Sciences and Literature | Mainz, Germany).

<sup>1</sup>Translation by the author

of consensus-based techniques and methods, which enable a common understanding of their findings once they are published, for example, in the form of a critical edition. And yet these scholarly standards are themselves subject to a constant process of alteration and development: historical research is now increasingly permeated by digitization and the possibilities that come with it. Digital technology has significantly extended the methodological repertoire of researchers in the humanities, and the fact that scholars are no longer bound to paper as a medium means that a rich variety of new interpretative approaches have emerged – a state of affairs that is clearly at odds with Barbara Bordalejo’s contentious assertion that “there is no such thing as digital scholarly editing” (Bordalejo, 2018, pp 24).

The majority of today’s digital scholarly editions (DSEs) use the Text Encoding Initiative (TEI) standard in combination with XML and its inherent hierarchies, as it is widely considered to be “a well-documented format for archival long-term preservation” which allows researchers to describe “a large number of textual phenomena in general ways” (Cummings, 2018). But research data in general, and the data of DSEs in particular, is highly connected and far from easy to express within a hierarchy (Witt, 2018, pp 222-223); a point to which I will return in Section 4. James Cummings certainly does not exaggerate when he states that the notion that “XML has difficulty with overlapping hierarchies is not, in itself, strictly a myth” (Cummings, 2018, pp i70). And things become even more complex if we begin to include the divergent perspectives of researchers concerning the transcription and edition of a given text. One crucial step towards addressing this issue is to move from the text-as-document paradigm toward what Zundert and Robinson refer to as the text-as-work paradigm (van Zundert, 2016, p. 103-104). But as I will argue in what follows, we ought in fact to go one step further by fully recognizing that researchers in their various roles as transcribers, editors, annotators, and users are themselves a key factor in the system of textual editing. Such an approach follows Niklas Luhmann’s observation that

[t]he inclusion of the observer and the instruments of observation in the objects of observation themselves is a specific characteristic of universal theories.<sup>2</sup>

It goes without saying that this essay is not about presenting a universal theory – the point is that we would be well advised to think of editors and users as integral parts of an interconnected system. Everything contributors do, all their observations and decisions, become part of the DSE as a *work* (Kuczera and Kasper, 2019). Jeffrey C. Witt has already suggested to conceive of DSEs

---

<sup>2</sup>(Luhmann, 1987, 164) (translation by the author).

as multipartite networks (Witt, 2018), which is essentially a way of describing a graph. For Witt, however, researchers themselves do not form part of this network: on the textual level, he models “each text as an Ordered Hierarchy of Content Objects (OHCO)” (Witt, 2018, p. 231).<sup>3</sup>

What I would like to propose instead is to model a DSE as a provenance knowledge graph which contains the entire critical apparatus, one or more transcription(s) and, if applicable, details concerning the relationships among them, as well as information on the origin of every statement. Describing textual phenomena (including the actions of the editor or editors) by means of TEI semantics has the advantage of maintaining semantic interoperability, as TEI is the established standard in the field. Moreover, TEI renders the connection between researchers and their work transparent. Made available in the form of a provenance knowledge graph, this crucial information in effect turns research data into a collection of subjective decisions made by researchers – it is then up to individual users to decide how much they trust these decisions based on the expertise and academic profile of the scholar(s) in question. To manage this highly connected trove of research data, a labeled property graph (LPG) database can be used. With this groundwork in place, the next step will be to connect individual knowledge graphs either in part or in their entirety to a broader system of concurring and/or diverging statements and interpretations.

## 2 The Rise of Connected Research Data

To better understand the desideratum articulated in this paper, let us consider the process of digitization in the field of medieval history, a development that has taken place in at least two distinct stages.

### 2.1 Image Digitization

The first stage, which lasted until the end of the 1990s, was characterized by a strong focus on image digitization. In Germany, one important protagonist in the field was the dMGH project, in the course of which the volumes of the *Monumenta Germaniae Historica* (MGH) were scanned, saved as image files, and made accessible on the internet (Sahle and Vogeler, 2013).

By way of example, Figure 1 shows the scan of a page from the MGH with a transcription of a charter of Emperor Frederick Barbarossa. In most cases, these early attempts at digitization did not allow for text to be copied out of the images, but they were still a step in the right direction: a large amount of research material was made available to researchers even if the paper copies were absent from their library.

---

<sup>3</sup>On OHCO, see DeRose et al. (1990).

## 389.

Friedrich verkündet dem Klerus und Volk von Genf, er habe der Klage des Bischofs Arducius von Genf wegen der Entfremdung der Regalien seiner Kirche durch Herzog (Berthold) von Zähringen stattgegeben; er erklärt die Maßnahmen des Herzogs für ungültig und die Kirche von Genf für reichsunmittelbar und befiehlt, dem Bischof Gehorsam zu leisten.

(1162 September 7, Saint-Jean-de-Lozne).

Original im Staatsarchiv zu Genf (A).

Spon, *Hist. de Genève* 2, 33 n° 9 aus A = (Christin) *Diss. sur l'abbaye de Saint-Claude* 103. — Rivoire - v. Berchem, *Rechtsquellen des Kantons Genf* 1, 18 n° 12 aus A. — 10  
Böhmer *Reg.* 2468. — *Reg. genevois* 103 n° 389. — *Stumpf Reg.* 3969.

Wie die in den Anmerkungen angeführten orthographischen Eigentümlichkeiten be-  
weisen, von einem Romanen mündlich, der auch als Diktator anzusehen ist und für  
die Formulierung des letzten Satzes D. 388 als Vorlage benützt hat; die Anklänge  
wurden durch *Petitsatz* gekennzeichnet. Die Datierung ergibt sich aus D. 388. 15

(+) § Fredericus dei gracia Romanorum imperator et semper augustus clero et populo Gebennensis ecclesie gratiam suam et omne bonum. § Noverit vestra dilectio vestraque universitas, quod venerabilem episcopum vestrum Arducium ad presentiam excellentie nostre et principum nostrorum venientem tamquam dilectum et honorabilem curie nostre principem solita inperali<sup>10</sup> mansuetudine et honorificentia suscepimus et eius querimoniis, quas pro alienatione regalium ecclesie sue a duce facta de Ceringe proposuit, nostras augustales aures pio favore accomodavimus. Facta enim eiusdem ducis, que com<sup>11</sup> Amedeo comite inisse dicitur, modis omnibus prohibemus et inperali<sup>12</sup> actoritate<sup>13</sup> revocamus in irritum. Nolumus enim, ut unquam etiam volente episcopo eiusdem civitatis comes vel aliqua alia persona medius possessor inter nos et ecclesiam Gebennensem existat, quia tali facto et iusticia obviat et ratio contradicit. Vestre igitur universitati dilectum episcopum Arducium principem nostrum cum plenitudine graciae nostre remittimus monentes et qua debemus actoritate<sup>14</sup> precipientes, ut ei tamquam<sup>15</sup> domino et patri<sup>16</sup> vestro per omnia obediatis.

(SP. D.)<sup>17</sup>

80

Figure 1: Scan of MGH DDFI.2, p. 260 with a charter of Emperor Frederick Barbarossa (Source: dMGH [https://www.dmgh.de/mgh\\_dd\\_f\\_i\\_2/index.htm#page/260/mode/1up](https://www.dmgh.de/mgh_dd_f_i_2/index.htm#page/260/mode/1up))

## 2.2 Full Text Digitization

Around the turn of the millennium, this first stage evolved into a phase of full text digitization with projects such as Regesta Imperii Online (Schulz, 2017). Having been personally involved in this project, I can vividly remember the discussions about whether image digitization or full text digitization should be used. One major argument advanced by the proponents of image digitization was that optical character recognition (OCR) was still in its infancy and highly error-prone. We addressed this issue by linking every full text item on our website to a scan of the corresponding book page in the *Regesta Imperii*, which gave users direct access to the material that was being digitized and allowed them to identify inaccuracies.

Figure 2: Full text version of the charter from Figure 1 (Source: [http://www.regesta-imperii.de/id/1162-09-07\\_2\\_0\\_4\\_2\\_2\\_587\\_1145](http://www.regesta-imperii.de/id/1162-09-07_2_0_4_2_2_587_1145))

With the advent of full text digitization, large-scale computer-based text retrieval from historical documents became a possibility, and this major im-

provement brought with it entirely new ways of scholarly exploration.

## 2.3 Entities in Focus

Today, we are facing the next important step: it is now time to focus on the *entities* in the text.

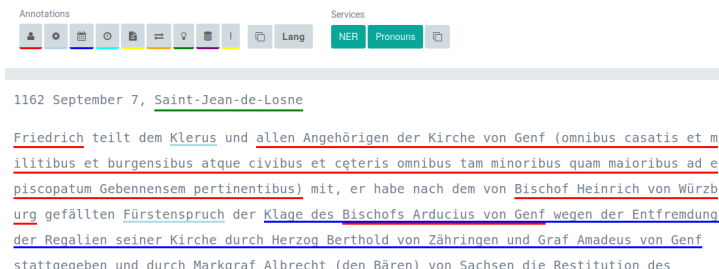


Figure 3: Fol. 341r of the *Liber epistolarum*

Identifying, annotating, and connecting these entities with data from authority files like GND or Wikidata enables the interconnection of individual research projects. It also becomes possible to model scholarly interpretations and the various steps of the research process in machine-readable statements – Section ?? will describe this process by example of a project dedicated to Hildegard von Bingen’s correspondence.

## 3 “Myths and Misconceptions about the TEI” – Thoughts From an Expert

In a recent article, Cummings (2018) shared his thoughts on what he perceives to be widespread myths or misconceptions concerning the TEI.

### 3.1 “XML is broken or dead”

The first of these myths is that “the TEI is XML (and XML is broken or dead).” As Cummings points out,

[t]he TEI Guidelines were first expressed in SGML as a markup language and only as of TEI P4 moved to recommending XML, but even this recommendation may change in the future (Cummings, 2018, i59).

With SGML, there had been no problems with overlapping markup – only with the shift to XML did overlap create the need for various workarounds.<sup>4</sup> On the other hand, however, the use of XML gave access to its entire ecosystem, and XML was the rising star in the field of markup at the time. Yet there is no reason why this arrangement has to be permanent:

<sup>4</sup><https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html>

[A]s new languages, technologies, and methodologies for text encoding emerge in future, the TEI Guidelines may move to them or include them as one of a set of ways to serialize digital text, so as long as they meet the basic requirements for easy long-term preservation, expressiveness, validation, integration, and mass adoption that is seen with XML. (Cummings, 2018, p. i59).

And this is precisely where our DSE of Hildegard von Bingen’s letters comes into play: its core principle is to employ TEI semantics without the hierarchical structure of XML.

### 3.2 The Future of XML as a Format for Text Encoding

At this juncture, I would like to share a personal observation regarding the future of XML. When we started using the format in our project *Regesta Imperii Online* in the early 2000s (Rübsamen and Kuczera, 2006), the project involved only comparatively basic annotation, so that any of the large number of freely available XML editors in plain XML mode was up to the task. Nowadays, many edition projects in the digital humanities employ the commercial software *OxygenXML*, often in combination with the virtual research environment *ediarum*,<sup>5</sup> which provides customizable GUI features. The reason for this is really quite simple: today’s annotation structures are often very complex, but *OxygenXML*’s author mode makes the intricate XML elements editable while conveniently hiding them from the user’s view.

There is a good reason why, in the broader field of software technology, XML is employed for purposes such as data exchange and the structuring of data in configuration files, but *not* for the sophisticated annotation of texts. Of course, publishers do use XML for their books, but these texts are nowhere near as deeply annotated as the ones that we are dealing with in the digital humanities today.

In fact, one could argue that the TEI community is in real danger of hitting a dead end, unless viable alternatives to XML are found in a timely fashion.

### 3.3 “XML (and TEI) cannot handle overlapping hierarchies”

Another myth discussed by Cummings is that “XML (and TEI) cannot handle overlapping hierarchies” (Cummings, 2018, p. i70-i71). Clearly, this is a bit of an overstatement: the TEI community has developed several mechanisms to deal with the issue of overlapping markup – at least to a certain extent.<sup>6</sup> But as the number of annotation hierarchies grows, these strategies

---

<sup>5</sup><https://www.bbaw.de/bbaw-digital/telota/forschungsprojekte-und-software/ediarum>

<sup>6</sup><https://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html>

do run into increasing problems. In light of this, we could rephrase Cummings’ statement as follows: “XML (and TEI) cannot handle *a sufficient number of overlapping hierarchies without complicated and ultimately inadequate workarounds.*” In some projects, a single annotation hierarchy may well be all that is needed – but being able to manage more of them should the necessity arise gives researchers considerably more flexibility.

To give an example, our graph-based DSE environment Codex (Kuczera and Neill, 2019) contains regions of text with up to 6 layers of annotation:

- Layout (page breaks, columns, alignment, etc.)
- Style (highlighted text, etc.)
- Entities (persons, places, concepts, etc.)
- Syntax
- Morphology
- Language

Customized annotation layers can easily be added to this list by the user. The substantial benefits of flexible, multidimensional annotation hierarchies in a DSE will be explained in detail in the following section.

## 4 *Hildegaph*: TEI without Hierarchies

In March 2020, a project based on the idea of using TEI semantics without the accompanying XML hierarchy was inaugurated under the title *The Book of Letters of Hildegard von Bingen. Genesis – Structure – Composition.*<sup>7</sup>

### 4.1 The Sources

The transmission history of Hildegard von Bingen’s (1098–1179) letters has taken many twists and turns. Within this complex and convoluted story, the so-called *Riesen-Codex* [‘giant codex’]<sup>8</sup> – a book of letters (*Liber epistolarum*) which consists of brief epistolary texts arranged to form a cohesive theological whole (see Figure 4 – the beginning of each letter is marked with larger characters and red ink) – assumes a particularly prominent position. The reason for this can be explained by two separate, albeit closely related, aspects of its reception history: first, both medieval and modern audiences are unanimous in their verdict that the *Liber epistolarum* is of equal importance to Hildegard’s works of visionary theology; second, the *Riesen-Codex* can lay claim to the special status of a last hand edition, as it was compiled by Hildegard’s staff from the entirety of her correspondence during her own lifetime and in accordance with her wishes.

---

<sup>7</sup>The project is funded by the Deutsche Forschungsgemeinschaft (DFG) <https://gepris.dfg.de/gepris/projekt/429863245?language=en>

<sup>8</sup>[https://tudigit.ulb.tu-darmstadt.de/show/Hild\\_R\\_Riesencodex](https://tudigit.ulb.tu-darmstadt.de/show/Hild_R_Riesencodex)

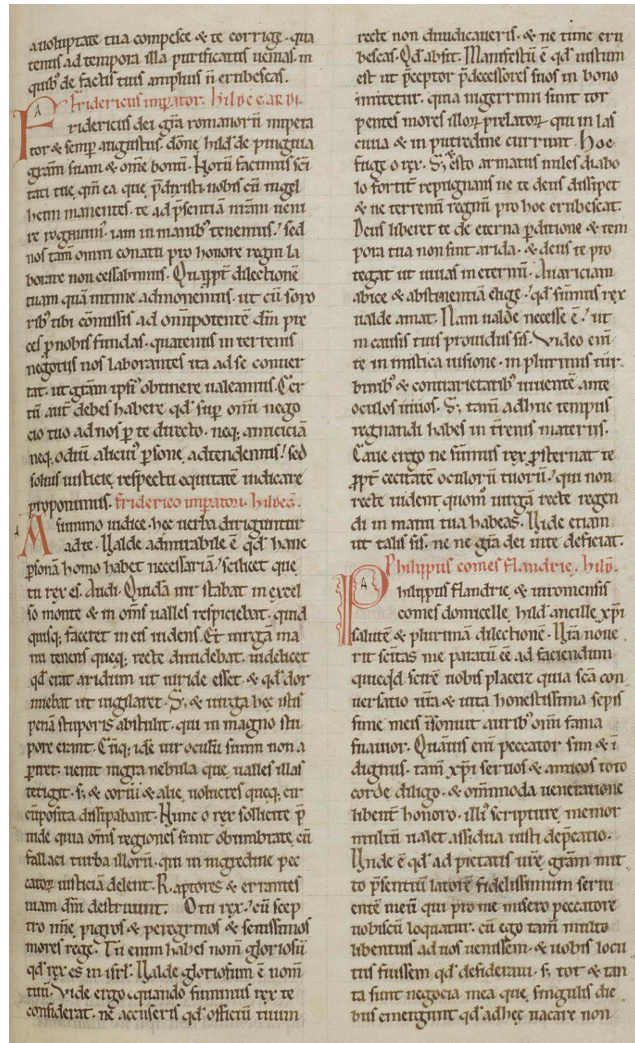


Figure 4: Fol. 341r of the *Liber epistolarum* Wiesbaden, Hochschul- und Landesbibliothek RheinMain (ehemals: Wiesbaden, Hessische Landesbibliothek) Hs 2 („Riesenkodex“) urn:nbn:de:hebis:43-972.



Our project is the first to present the *Liber epistolarum* in the form of a digital scholarly edition. As opposed to the existing critical edition of Hildegard's letters (van Acker and Klaes-Hachmöller, 1993-2001), which seeks to reconstruct 'the correspondence that actually took place' while combining different stages of transmission, our focus lies on the final authorized form that Hildegard's letters assumed during her lifetime. Moreover, the individual letters found in the *Liber epistolarum* are not treated as mere historical witnesses, but rather as constituent parts of a deliberate and highly sophisticated theological-cum-literary composition.

Our edition of the *Liber epistolarum* is designed to be as media neutral as possible, allowing parts of it to be printed in book form. The changes that the text underwent over time can be traced by means of a graph model, in which the genesis of the individual letters – from the oldest known version to the form that appears in the *Liber epistolarum* – is modeled on the basis of the pertinent manuscripts. By way of example, Figure 5 shows the interdependencies of the various versions of letter #52 found in manuscripts Z, W, M, Wr, and R.

While information concerning the evolution of the *Liber epistolarum* over time is stored in a graph model, the texts of the letters themselves will be transcribed in a standoff property editor with the project name *hildegraph*.<sup>9</sup> For our purposes, standoff property (SPO) means that the texts are annotated on an index base, whereas TEI-based XML markup is mainly inline. The technical outline of SPO is explained in detail in (Kuczera and Neill, 2019).

The project began in April 2020 with a critical transcription of the text of the *Riesen-Codex*. As *hildegraph* was not yet operational at that point, the task was initially undertaken using an adapted version of the Leiden Conventions,<sup>10</sup> a system that employs various types and combinations of brackets to express textual phenomena in plain text. Currently, we are working on the transfer of these transcriptions into the *hildegraph* environment with the aid of TEI semantics.

As noted above, our DSE uses TEI semantics without XML hierarchies – in *hildegraph*, multiple annotation hierarchies can coexist in one system. Table 1 shows a preliminary list of annotation types based on TEI semantics. Here it is important to keep in mind that an annotation can be assigned to multiple semantic spaces – *Hildegraph* is capable of managing several indexing systems at once. For example, the first line in Table 1 shows that text between lines can be identified both by Leiden annotation *leiden/supralin-*

---

<sup>9</sup>*hildegraph* is derived from the Codex system described in (Kuczera and Neill, 2019) <https://www.hildegraph.org/>

<sup>10</sup>[https://en.wikipedia.org/wiki/Leiden\\_Conventions](https://en.wikipedia.org/wiki/Leiden_Conventions)



Explanation	Type	Description	TEI	Coding
<i>supra lineam</i>	text	between lines	<add place="above">	leiden / supralineam: orange text
<i>in margine</i>	text	margin note belonging to text	<add place="margin">	leiden / marginalia: purple text
<i>recensi manu</i>	extratext	additional text	<add place="margin">	leiden / additional-text: ZPA
start and end of column a-b	text	column	<cb>	leiden / column: EOC “//”
corr.	text	visible correction	<corr>	leiden / correction: red underline
original spelling (transcription)	interpr.	corrected to/sic!	<corr> and <sic>	leiden / sic: green underline
<i>in rasura</i> (text located in an place where prior text has been erased)	text	added on deleted text	<del @rend @reason><gap reason="rasure" unit="line" quantity="2"/> <add @place>	leiden / rewritten: yellow underline
text is in another line	text	transposition	<del cause="moved"> <supplied>	leiden / transposition: pink underline
erased text (erased, crossed out)	text	struck out or similar	<del rend="striked out">	leiden / striked-out: strike through line
rubricated text (words, letters)	text	rubrum	<emph rend="red">	leiden / emphasis: styled in red
resolution of abbreviations	interpr.	expansion	<expan> <abbr> </abbr><ex> </ex></expan>	leiden / expansion: styled in blue
empty space with missing text	text	gap	<gap reason="not readable">	leiden / gap: ZPA: underlined white space
start of line	text	line	<lb>	: EOL “/”

Table 1: List of annotation types

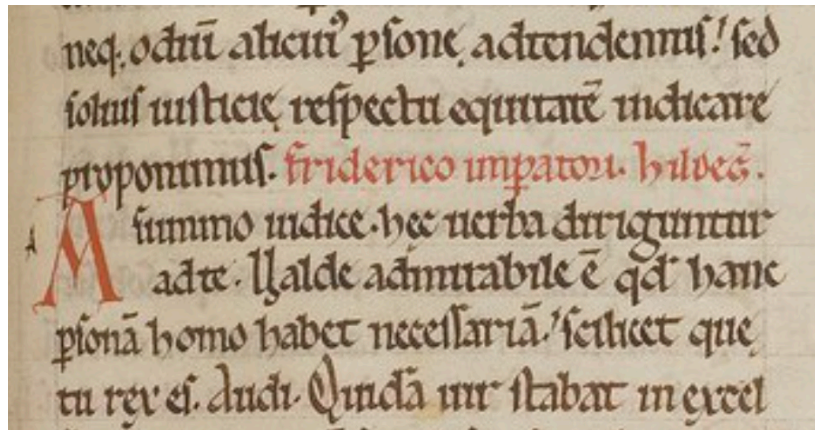


Figure 6: Part of fol. 341r with rubricated text

```
\#52\# [ru[friderico imp(er)atori hildeg(ardis).]ru] |
[ru[A]ru] summo iudice. hec uerba dirigitur |
ad te. Valde admirabile est q(uo)d hanc ||
p(er)sona(m) homo habet necessaria(m)! scilicet que |
tu rex es. Audi. Quida(m) uir stabat in excel
```

The manuscript's red, or rubricated, characters and the capitalized 'A' are a signal to the reader that a new letter is about to begin. In the transcription, these parts of the text are represented by square brackets and the siglum *ru*: [ru[A]ru] (for *Rubrum*), whereas the characters in round brackets spell out the abbreviations used by the original scribes. As this system of annotation unequivocally marks which start element belongs to which end element, overlapping markup is possible.

But what is the best way to represent rubricated text by means of TEI elements? Which of the several options provided by TEI is the most promising? In an effort to find an answer to this question, we reached out to two TEI experts.

One suggestion was to use the <emph> element<sup>11</sup> to highlight “words or phrases which are stressed or emphasized for linguistic or rhetorical effect,”<sup>12</sup> while the <rend> attribute could be employed to convey the information that the text's color is red. Here is what this approach would look like:

```
<emph rend="red">friderico imp(er)atori hildeg(ardis)</emph>
```

<sup>11</sup><https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-emph.html>

<sup>12</sup>The rhetorical aspect reminds me of Zundert's idea of a computational edition with performative texts (van Zundert, 2019).

The other expert proposed to use the <hi> element<sup>13</sup> to mark “a word or phrase as graphically distinct from the surrounding text, for reasons concerning which no claim is made,” a solution that would look like this:

```
<hi rend="colored">friderico imp(er)atori hildeg(ardis)</hi>
```

The <emph> element appeared to be a fitting choice to mark a section of text that had been highlighted for a specific purpose – but given that the red characters were graphically distinct from the surrounding text, a strong case could also be made for the <hi> element. Clearly, the rubricated text fulfilled at least two different roles: in the context of the overall layout of the page, the red characters mark the beginning of a new letter and thus serve a rhetorical and structural function, yet they also identify the sender and recipient of the letter in question. What, then, if we employed *both* elements to represent the distinctive red ink? Which element should come first and contain the other? And does this kind of containment even make sense?

In long discussions with various colleagues, no convincing arguments in support of the need for containment were put forward. There is simply no plausible need for it when it comes to accommodating different annotation layers like layout or rhetoric: as our *Hildeglyph* environment attests, all of these layers can be combined in various ways without the application of hierarchies.

### 4.3 “So what’s the *text*, then?”

The brief transcription from the *Liber epistolarum* discussed in the previous section contains several expansions of scribal abbreviations employed in the original text. The use of abbreviations in manuscripts was a very common practice in the Middle Ages and posed few obstacles to contemporary readers. A modern critical DSE, on the other hand, is expected to provide an expanded and normalized version of the text for convenience and ease of reference.

But which version of the text should be displayed in *Hildeglyph*’s plain text field? As a medievalist, I am inclined to argue that the version of the text that is as close as possible to the original should be shown in this prominent location; on the other hand, the expanded versions are much easier for casual users (and also for persons charged with maintaining the database) to read and understand. In the end, there are good arguments in favor of each of the two alternatives, and one of the major advantages of using SPO is that it does not force us to make a choice: as all versions can easily be converted

---

<sup>13</sup><https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-hi.html>

into one another, there is simply no need to decide once and for all whether the plain text to be indexed in SPO is the *original* or the *normalized* version – this decision can be made according to the specific requirements of the individual use case.

## 5 Digital Scholarly Editions as Provenance Knowledge Graphs

### 5.1 What is a Knowledge Graph?

If we continue along this line of thought, the whole DSE can be conceptualized as a provenance knowledge graph, in which every piece of information is stored together with information on where it comes from, who made which statement and when, etc.

Broader scholarly interest in knowledge graphs began to arise when Google discussed their own approach to the issue in a blog post, which essentially described an enhancement of their search engine through semantics without going into the technical details (Amit Singhal, 2012). Since then, a fair amount of research has been carried out in this area, notwithstanding the absence of a universally accepted definition of what constitutes a knowledge graph (Ehrlicher and Wöß, 2016).

At times, the term has simply been used as a synonym for ontology. According to Paulheim (2016),

[a] knowledge graph (i) mainly describes real world entities and their interrelations, organized in a graph, (ii) defines possible classes and relations of entities in a schema, (iii) allows for potentially interrelating arbitrary entities with each other and (iv) covers various topical domains.

For Ehrlicher and Wöß (2016),

[k]nowledge graphs are large networks of entities, their semantic types, properties, and relationships between entities,

whereas Pujara et al. (2013) point out that there are systems that

[u]se a variety of techniques to extract new knowledge, in the form of facts, from the web. These facts are interrelated, and hence, recently this extracted knowledge has been referred to as a knowledge graph.

Another common definition<sup>14</sup> is that a knowledge graph represents a collection of interlinked descriptions of entities (real-world objects, events, situ-

---

<sup>14</sup>See, for example, <https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-knowledge-graph>.

ations, or abstract concepts), while other scholars use the term to refer to any knowledge base modeled as a graph.

As if this confusion was not enough, none of these definitions say anything about the *technical* specifications of a knowledge graph: some explicitly mention RDF (Resource Description Framework) (Färber et al., 2016) and some suggest node properties (Ehrlicher and Wöß, 2016), but as of yet, no clear picture as to possible technical backgrounds of knowledge graph systems has emerged.

## 5.2 The Provenance of a Statement

The various concepts of knowledge graphs discussed in the previous section do have one thing in common: they treat information as objective truth. But in the field of the (digital) humanities, the ‘truth’ is always a matter of interpretation. The interpretative process begins the moment the very first characters of a text are transcribed. Each of the editor’s decisions is open to discussion and constitutes a subjective statement – and that is precisely the point where provenance comes into play. Once we begin to model a DSE as a knowledge graph that includes comprehensive provenance information, we end up with a huge amount of statements. Expressing all of this information in RDF would produce a huge and completely unmanageable graph, which is why we have opted to use labeled property graphs (LPGs) in our DSE.

## 5.3 LPG vs. RDF

RDF is a W3C standard for data exchange in the web that represents data as a graph, and this is the most important point of commonality it shares with LPGs. RDF structures information in triples in the form of subject-predicate-object, with the subject, predicate, and object being identified by Uniform Resource Identifiers (URI) (Barrasa, 2016).

The statement that Emperor Frederick Barbarossa was a human being would look like this:

(Emperor Frederick Barbarossa)-(INSTANCE\_OF)-(human)

Here is the same statement translated into URIs:

(<https://www.wikidata.org/wiki/Q79789>)  
(<https://www.wikidata.org/wiki/Property:P31>)  
(<https://www.wikidata.org/wiki/Q5>)

Adding his place of death – Weingarten – would involve another triple:

(<https://www.wikidata.org/wiki/Q79789>)  
(<https://www.wikidata.org/wiki/Property:P19>)  
(<https://www.wikidata.org/wiki/Q572427>)

In principle, RDF understands the world as a network of connected entities and literals. Its popularity surged with the rise of the Semantic Web<sup>15</sup>, which operates on the basic idea that users should publish data in structured formats with well defined semantics so that this data can be ‘understood’ by machines. Originally, this structured information was to be contained in RDF triple stores<sup>16</sup>, a vision that soon evolved to quad stores which added a named graph to each RDF triple. Today, the product of this evolution is commonly referred to as “semantic graph database” (Barrasa, 2016).

In LPGs, each node and edge not only has a unique and distinctive ID, but also a set of key-value pairs (or properties) that characterize it. Our example of Emperor Frederick Barbarossa and his place of death could be expressed like this:

```
(e:Entity{type:'human', wikidataId:'Q79789'})  
-[r:PLACE_OF_DEATH {wikidataId:'P19'}]->  
(p:Place {label:'Weingarten', wikidataId:'Q572427'});
```

When comparing RDF triples with an LPG, it is important to keep in mind that in the latter, nodes and relationships have an internal structure. In RDF, on the other hand, a triple is composed of two nodes connected by an edge (subject-predicate-object); the subject and the relationship are each identified by a URI, and the object can be another node or a literal, so that neither nodes nor relationships have an internal structure – they are merely unique labels. It is evident from this that an RDF graph could easily reach ten times the size of an LPG containing the same amount of information.

Another important difference is that RDF does not uniquely identify instances of relationships of the same type, nor does it allow instances of relationships to be qualified. In an LPG, the information is stored in the graph structure and in the internal structure of nodes and relationships. In RDF, all of this must be expressed in simple RDF triples.

In light of this, *Hildegraph* uses an LPG to store all of the information contained in the DSE as a statement, which allows us to explore and compare diverse (and potentially competing) interpretations. The provenance information – who made what statement when and where – is stored in the properties. Here, the versioning of graphs as discussed in Martina Bürgermeister’s contribution to this volume plays an important role.

---

<sup>15</sup> <https://www.scientificamerican.com/article/the-semantic-web/>

<sup>16</sup> <https://en.wikipedia.org/wiki/Triplestore>



## 5.4 Manuscript Structures in the Graph

The physical structure of the manuscript (and its relationship to its digital descendants) can also be modeled as a graph. In Figure 7, the main manuscript R is represented by the image in the upper part of the picture. This node is then connected to the folio nodes, which correspond to the individual folios that make up the manuscript, and which are connected by `IS_FOLLOWED_BY` edges that model the order of the folios within the manuscript. This part of the graph – or in other words, this subgraph – contains the information about the physical structure of the manuscript. One example of the usefulness of this information is a manuscript in which the order of the folios has been changed at some point – in a graph, both the original order and the new order can easily be modeled.

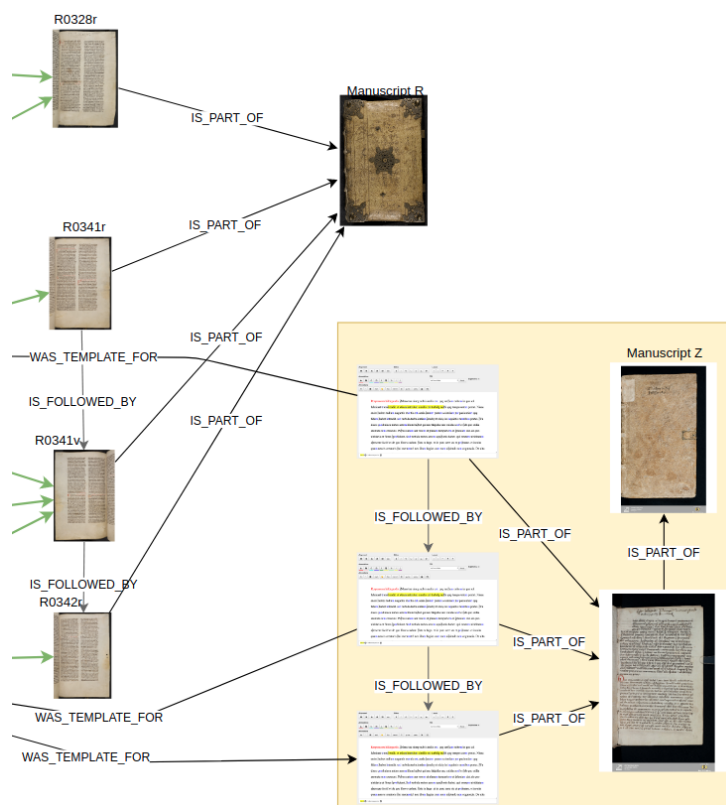


Figure 7: Physical structure of the manuscript as represented by the graph

## 5.5 Transcription as Connected Parts of Text

A TEI/XML-based transcription aims at expressing the layout structure of a manuscript with inline markup in one XML document. In *Hildes-graph*, every distinct unit of text is assigned its own SPO node. The re-

relationships between the different text blocks are represented by means of IS\_NEIGHBOUR\_OF edges, which encode the visual impression of vicinity in the graph. In addition, the individual passages of text are linked to a corresponding image of the folio in question.

Figure 8 shows two lines of text (*initium libri Epistolarum et orationum Sanctae Hildegardis*) added by a later hand in the upper margin of fol. 328r.

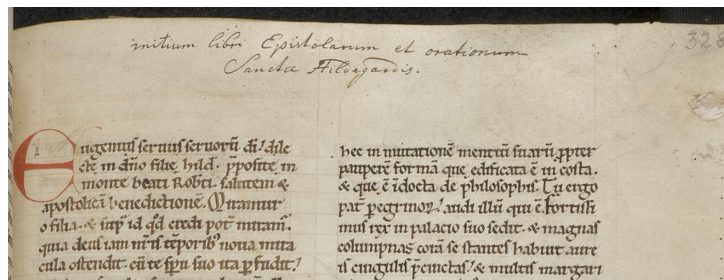


Figure 8: Addendum by a later hand (fol. 328r.)

With XML/TEI, this text would be contained in the main body of the letter in the XML document. In *Hildeglyph*, the added text receives its own SPO node, and the two SPO nodes are connected with an IS\_NEIGHBOR\_OF edge (Figure 11). While textual information is thus stored separately from its visual arrangement on the folio page with all the benefits such an approach entails, the combination of text and layout can easily be examined if and when this is needed. Figure 9 shows the entire data model of *Hildeglyph*.

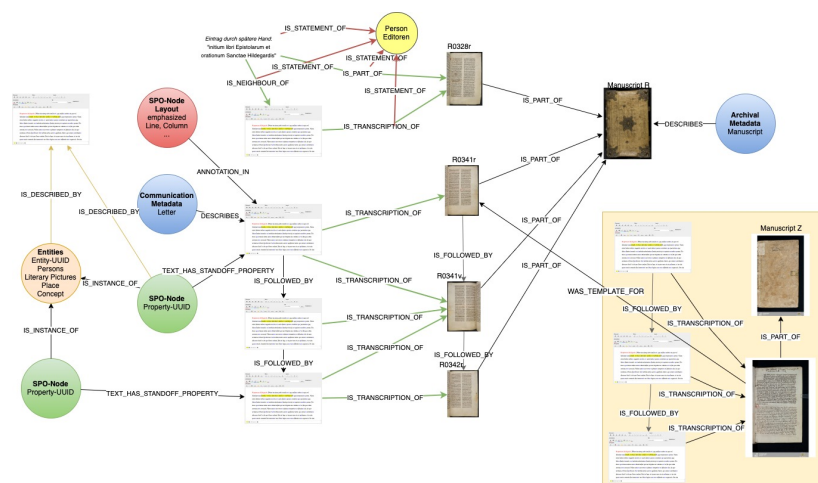


Figure 9: Data model for the digital scholarly edition of the *Liber epistolarum*

## 5.6 Text as a Graph

Given the continuing lack of technical solutions for managing text directly as a graph (Kuczera, 2016), we developed our own set of standoff properties (Kuczera and Neill, 2019) based on Desmond Schmidt's ideas (Schmidt, 2016, 63-69).

Since SPOs are index based, one must select a base text for indexation. In practice, however, every version of a text can be used as base text because they can all be converted into one another. Indexing the base text makes every character of the text addressable – they are strung out like pearls in a long row, forming a chain of nodes in the graph which is given order by the direction of the text. All additional information is then connected to these indexed characters in a process that builds a bridge between the more transcription-related sphere of the text and the predominantly semantic and interpretative sphere of the graph. In this regard, *Hildegraph* goes well beyond Witt's above quoted proposition to model text as an ordered hierarchy of content objects (OHCO).

## 5.7 Transcription with Annotations of Annotations

Another SPO is created whenever a user adds an annotation, which can then be annotated again (most likely by another user) with yet another SPO, and so on. With standoff properties, every annotation is stored together with a Globally Unique Identifier (GUID) and can be traced back to the user who added it (Kuczera and Neill, 2019). From this perspective, annotations can be seen as a statements by a certain user – as users add these statements to the base text (which is itself a statement), and as these statements are in turn annotated by other users, the resulting knowledge graph continues to grow.

Figure 10 shows the subgraph concerned with transcription and interpretation. The manuscript consists of letters from and to Hildegard. These letters are assigned one SPO node each, which are connected with `IS_PART_OF` edges to the corresponding folio nodes. A letter can belong to one or more folio pages, and may have a predecessor or a successor connected with `IS_FOLLOWED_BY` edges (See Figure 9). The red, blue, and green circles on the right represent metadata, layout, and semantic content of the letters.

## 5.8 Annotations as Individual Statements

Figure 11 shows how provenance is stored. Every information is connected to a node which represents the user who created the statement.<sup>17</sup> In our example, the user *Andreas Kuczera* has transcribed a margin note. This

<sup>17</sup>Using TEI semantics, this information can be stored with `<respStmt>` (<https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-resp.html>), and `<revisionDesc>`.

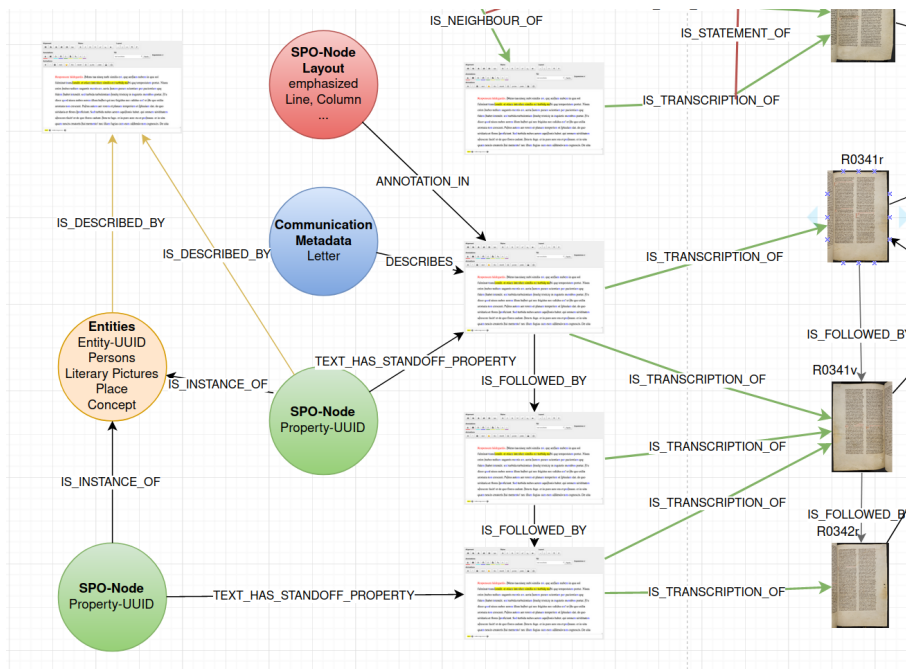


Figure 10: Transcriptions with annotations, entities, described by other text nodes and metadata.

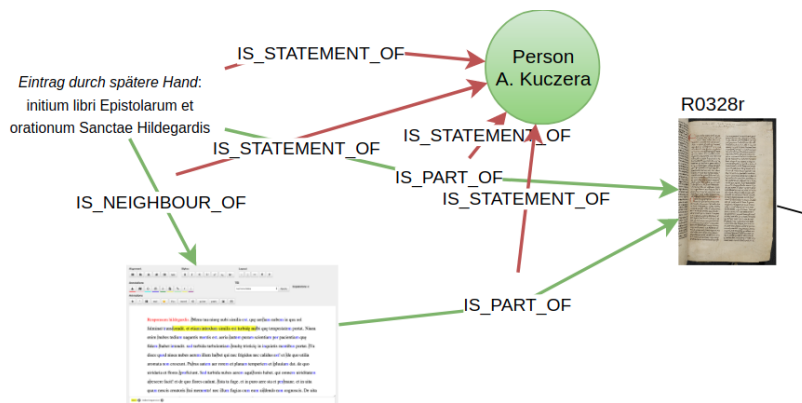


Figure 11: Every annotation can be traced to the user

transcription is not stored in the SPO node of the corresponding letter, but rather in a separate SPO node which is then connected with an `IS_NEIGHBOUR_OF` edge to a zero point annotation in the letter text – it is this separate storage of transcription and allocation that makes the modeling of multiple interpretations possible.

## 6 Conclusion

By way of conclusion, I would like to return to the brief epigraph of my essay: “To which problem is digitization the solution?” From my point of view, digitization enables researchers to publish their findings with a maximum of flexibility and transparency. Ideally, hierarchies should only be involved in this process when they are actually needed, and not because they are forced upon us by technological limitations. One of the fundamental properties of research data in the (digital) humanities is that it is highly connected, and I would argue that scholars should be granted the capacity to store every bit of information concerning these connections even if, for the time being, a standard or suitable ontology to express them might still be lacking. From a technical perspective, graph technologies can provide us with the capability to model multiple and multidimensional layers of information. TEI semantics could be another important piece of the puzzle, but their practical utility is dramatically reduced by the limitations of the XML hierarchies with which they are currently yoked together. As the *Hildegraph* environment shows, there is no reason why the problematic coupling of TEI semantics and XML should continue.

## References

- Amit Singhal (2012). Introducing the Knowledge Graph: Things, Not Strings. <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html>.
- Barrasa, J. (2016). RDF Triple Stores vs. Labeled Property Graphs: What’s the Difference? <https://neo4j.com/blog/rdf-triple-store-vs-labeled-property-graph-difference/>.
- Bordalejo, B. (2018). Digital Versus Analogue Textual Scholarship or The Revolution is Just in the Title. *Digital Philology: A Journal of Medieval Cultures*, 7(1):7–28, DOI: 10.1353/dph.2018.0001.
- Cummings, J. (2018). A World of Difference: Myths and Misconceptions About the TEI. *Digital Scholarship in the Humanities*, 34(1):i58–i79, DOI: 10.1093/lhc/fqy071.

- DeRose, S. J., Durand, D. G., Mylonas, E., and Renear, A. H. (1990). What Is Text, Really? *Journal of Computing in Higher Education*, 1(2):3–26, DOI: 10.1007/BF02941632.
- Ehrlicher, L. and Wöß, W. (2016). Towards a Definition of Knowledge Graphs. In Martin, M., Cuquet, M., and Folmer, E., editors, *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16)*, number 1695 in CEUR Workshop Proceedings. <http://ceur-ws.org/Vol-1695/>.
- Färber, M., Bartscherer, F., Menne, C., and Rettinger, A. (2016). Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. <http://www.semantic-web-journal.net/content/linked-data-quality-dbpedia-freebase-opencyc-wikidata-and-yago-0>.
- Kuczera, A. (2016). Digital Editions Beyond XML – Graph-Based Digital Editions. In Düring, M., Jatowt, A., Preiser-Kappeller, J., and van Den Bosch, A., editors, *Proceedings of the 3rd HistoInformatics Workshop on Computational History*, number 1632 in CEUR Workshop Proceedings, pages 37–46. <http://ceur-ws.org/Vol-1632/>.
- Kuczera, A. and Kasper, D. (2019). Modellierung von Zweifel – Vorbild TEI im Graphen. In Kuczera, A., Wübbena, T., and Kollatz, T., editors, *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, volume 4 Special Issue of *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/SB004\_003.
- Kuczera, A. and Neill, I. (2019). The Codex – An Atlas of Relations. In Kuczera, A., Wübbena, T., and Kollatz, T., editors, *Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten*, volume 4 Special Issue of *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/sb004\_008.
- Luhmann, N. (1987). *Archimedes und wir : Interviews*. Number 143 in Internationaler Merve-Diskurs. Merve, Berlin.
- Nassehi, A. (2019). *Muster: Theorie der digitalen Gesellschaft*. C.H.Beck, München, 3rd edition.
- Paulheim, H. (2016). Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semantic Web*, 8(3):489–508, DOI: 10.3233/SW-160218.

- Pujara, J., Miao, H., Getoor, L., and Cohen, W. (2013). Knowledge Graph Identification. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., et al., editors, *Advanced Information Systems Engineering*, number 7908 in Lecture Notes in Computer Science, pages 542–557. Springer, Berlin/Heidelberg, DOI: 10.1007/978-3-642-41335-3\_34.
- Rübsamen, D. and Kuczera, A. (2006). Verborgen, vergessen, verloren? Perspektiven der Quellenerschließung durch die digitalen 'Regesta Imperii'. In Hering, R., Sarnowsky, J., Schäfer, C., and Schäfer, U., editors, *Forschung in der digitalen Welt. Sicherung, Erschließung und Aufbereitung von Wissensbeständen*, number 20 in Veröffentlichungen aus dem Staatsarchiv der Freien und Hansestadt Hamburg, pages 109–124. Hamburg University Press, DOI: 10.15460/HUP.STAHH.20.77.
- Sahle, P. and Vogeler, G. (2013). Digital Monumenta Germaniae Historica (dMGH). *Digital Philology: A Journal of Medieval Cultures*, 2(1):135–139, DOI: 10.1353/dph.2013.0006.
- Schmidt, D. A. (2016). Using Standoff Properties for Marking-up Historical Documents in the Humanities. *it - Information Technology*, 58(2):63–69, DOI: 10.1515/itit-2015-0030.
- Schulz, J. (2017). A review Of: Regesta Imperii Online, Ed. By Deutsche Kommission für die Bearbeitung der Regesta Imperii e.V., 2001-2017. <http://www.regesta-imperii.de/>. *RIDE*, 6, DOI: 10.18716/RIDE.A.6.5.
- van Acker, L. and Klaes-Hachmöller, M. (1993-2001). *Epistolarium. [Hildegard von Bingen]*. Brepols, Turnholti.
- van Zundert, J. (2016). Barely Beyond the Book? In Driscoll, M. J. and Pierazzo, E., editors, *Digital Scholarly Editing: Theories and Practices*, pages 83–106. Open Book Publishers, DOI: 10.11647/OBP.0095.05.
- van Zundert, J. (2019). Why the Compact Disc Was Not a Revolution And «Cityfish» Will Change Textual Scholarship, or What Is a Computational Edition? *Ecdotica*, 15:129–156, <https://pure.knaw.nl/portal/en/publications/why-the-compact-disc-was-not-a-revolution-and-cityfish-will-chang>.
- Witt, J. C. (2018). Digital Scholarly Editions and API Consuming Applications. In Bleier, R., Bürgermeister, M., Klug, H. W., Neuber, F., et al., editors, *Digital Scholarly Editions as Interfaces*, volume 12 of *Schriften des Instituts für Dokumentologie und Editorik*, pages 219–247. BoD, Norderstedt, <http://nbn-resolving.de/urn:nbn:de:hbz:38-91182>.