Scholarship in interaction: case studies at the intersection of codework and textual scholarship

Zundert, J.J. van

Chapter 3

The Case of the Bold Button: Social Shaping of Technology and the Digital Scholarly Edition[1]

> "First, let us observe two things missing from almost all electronic scholarly editions made to this point. The first missing aspect is that up to now, almost without exception, no scholarly electronic edition has presented material which could not have been presented in book form, nor indeed presented this material in a manner significantly different from that which could have been managed in print."

These are words by Peter Robinson, who spoke and wrote them in 2004 (Robinson 2004). I think little has changed in over a decade since and the observation still more or less holds. At the time, Robinson argued vehemently for digital scholarly editions that would move decisively beyond the realm of the possibilities of print publication. He was – and is – by no means the only one that has been advocating for such a shift. In fact, many have wondered how the digital medium, or the virtual environment, would change the nature and appearance of the scholarly edition. For that matter, grand

[1] This chapter appeared before as Van Zundert, Joris J. 2016. "The Case of the Bold Button: Social Shaping of Technology and the Digital Scholarly Edition." *Digital Scholarship in the Humanities* 31 (4): 898–910. https://doi.org/10.1093/llc/fqw012. It was updated to include a more expanded definition and explanation of the concepts of "graph" and "knowledge graph". A few sentences were added to clarify my use of "heuristics", and the meaning of "minimal" and "maximal edition".

perspectives on paradigmatic change due to medium change are not unique to textual scholarship. The introduction of a new medium or technology has always inspired great debate between advocates and antagonists of the next big thing. Self-proclaimed supporters of digital media usually advocate revolutionary changes. In the case of textual scholarship, for example, one may hear it proclaimed that the book is dead; good riddance, the advocates for "The Next Big Thing" (Bod 2013b:8) judge, for it was a clumsy, static, institutionally bounded, difficult to use, and outdated interface. Give way to open access, process orientation, dynamic interfaces, intuitive interaction, fluid text, social editing, etc. (cf., for instance, Siemens et al. 2012). With similar and undaunted zeal, Luddites lament the waning of solid scholarly practice: concentration span, close reading, philological interpretation, editorial practice, and convention (Fish 2011) – all sacrificed to the "Bitch goddess, QUANTIFICATION" (sic) as Bridenbaugh once put it (Bridenbaugh 1963). Or, for a more recent example in the Dutch literary and linguistics theatre, consider professor Rens Bod proclaiming the end of Humanities 1.0 (Bod 2013b), and Ph.D. student Marieke Winkler sincerely questioning that (Winkler 2013).

The screaming and kicking of Luddites aside the proponents of change do not seem really to get what they want. After many years of development of digital technology, the book is as alive as it ever was. We scarcely find digital editions, scholarly or otherwise, resembling the advanced models of dynamic, fluid, collaborative, and social texts such as those proposed by McGann (2010), Drucker (Lunenfeld et al. 2012:36), Shillingsburg (Thiruvathukal, Jones, and Shillingsburg 2010), Robinson (2004), Van Hulle (2010), Siemens (Siemens et al. 2012), and myself (Boot and van Zundert 2011). E-books are certainly impacting the market (AAP 2010; Cain Miller and Bosman 2011), but e-books are pure digital metaphors of the print book. Digital scholarly editions hardly have any impact (Porter 2013). What is more interesting is that digital scholarly editions are a far cry from what many expected them to be. We could suppose that this state of affairs is due to a lack of knowledge, skills, and technology support as has been indeed suggested before (cf. Courant et al. 2006). And it is probably true there are severe problems of teaching and training in our field, given that master and Ph.D. programs truly oriented on the digital humanities are only lately coming into existence. Yet, I think there might be more to the matter.

Maybe we need to answer to Borgman's call (2009): "Why is no one following digital humanities scholars around to understand their practices, in the way that scientists have been studied for the last several decades?" What do we see if we step back for a while from our work as textual scholars and digital humanities researchers and look at what is happening from the social sciences, in particular of Science and Technology Studies? Science and Technology Studies suggest to study technology development in its social context. In the past few years, I have studied the creation and development of the digital scholarly edition within the laboratory-like setting in the Huygens Institute for the History of the Netherlands. Here we find a relatively large – for humanities contexts in any case – IT Research and Development (R&D) group of on average sixteen persons working together with about sixty historians, textual scholars, and digital archivists. The research context consists of a dozen senior researchers, a similar amount of non-senior and associate researchers, a similar amount of Ph.D. candidates with various contracts ranging from predominantly full-time added staff to volunteer workers, and of course non-IT R&D supporting staff. As in many other contexts (cf. Nowviskie 2012) the relationship between the IT R&D group and scientific staff is some matter of internal debate in the institute. In part, the role of the IT R&D staff is in support, in part it is collaborative at the research level.

The adoption and application of technology is as much a social as it is a technical process. These processes are inevitably intertwined: technology does not determine but operates within and is operated upon in a complex social field (Bijker et al. 1987). The manifestation of such intertwined processes is directly visible in the field of digital humanities and in the development of the digital scholarly edition. Of course, the digital scholarly edition is a digital artifact brought to life in a context of heavy interaction between a highly digital technological community (computer scientist, software engineers and digital humanists) and a far less digital technology based community (textual scholars). This intricate and intensive interaction is a daily practice at the Huygens Institute for the History of the Netherlands. One of my tasks has been to guide the interaction between IT R&D, documentary editors, textual scholars, and researchers of literature and history, and to facilitate the ongoing methodological discussion between these cultures. I have had the privilege to study these processes from many angles: methodology, technology, model, role, audience, development, and so on.

As has happened in many similar research contexts, a transcription tool was developed at the institute to support the basic work of turning non-OCR-able texts from early printed works and medieval and modern manuscripts into digital machine-processable counterparts. The development of this tool, eLaborate[2], was based on a strategy of encapsulating and hiding XML markup – to be transformed to TEI encoding behind the scenes – with a graphical interface. In this way, the tool was meant to present minimal barriers to transcribers who came in a variety of levels of expertise on encoding. This indeed resulted in successful participation of significant numbers of volunteers unskilled in XML over a large set of projects. Also the GUI encapsulation of XML technicalities facilitated greatly the focus on community and project management (Beaulieu, Van Dalen-Oskam, and Van Zundert 2012).

Here I am not so much interested in the features or particulars of eLaborate. Instead I want to focus on one particular researcher-developer interaction I witnessed that, I think, stands as an example of a general and strong tendency in the scholarly community at large. The usability principle behind eLaborate is that any encoding or markup is treated as an annotation on arbitrary regions within the text. To this end, when a user has selected a certain region in the text with the mouse, a pop-up dialog appears allowing the user to enter annotative tags, comments, etc. The interface thus closely mimics a concept – using a highlighter and pencil to create annotations – that is known and tangible to anyone who has a basic experience in working with scholarly texts. The clear downside of this principle – if dogmatically applied – is that a user is faced with an enormous number of click-and-point-and-type annotation operations. Especially in cases of seemingly insignificant but frequent markup, such as with the indication of bold face print, this approach can strike the user as tediously and needlessly pedantic. It should be noted, however, that regarding the feelings of tediousness a distinction is likely to

[2]Cf. https://www.elaborate.huygens.knaw.nl

be made between senior scholars as transcribers and non-academic volunteer "crowd sourcees". Evidence from the projects, if mainly anecdotal, suggests that volunteer transcribers in fact may attach hundreds of tiny and similar annotations without complaint, but that the senior researcher will feel put at odds with her experience and practices when invited to do so. In any event, the upshot of this usability agony was a recurring and strong push in the user community to have a button labeled "bold" – in fact to have several such buttons for italics, underline, and other common very frequently appearing properties of text – lowering the volume of tedious annotation. To this day I remain convinced that we should not have implemented that button as we did.

The foundation for my conviction is that these buttons violate the rationale for XML over HTML, namely the strict and intentional separation of representational and semantic information. The most common interpretation of boldface type is that it is a material manifestation of the concept of emphasis. Even this is not universal – many other concepts may also be expressed by the use of boldface type. Thus, the provision of a button to record that some text is in boldface type introduces inevitable ambiguity in a descriptive system. At a later point in time there is no unambiguous way to tell what the function of the bold print was: it arbitrarily covers many intentions without delineating which of several possible textual concepts might apply.

However, more important for my argument here is that the implementation of this simple button reveals how technology is indeed shaped through its social context. The intent of eLaborate's approach was paradigmatic: its purpose was to allow editors of text to change from a representational paradigm to a semantic paradigm. We could have done this by forcing our users to become competent XML authors. Our users judged XML tedious and complicated, however, and complexity is a well-known "fail factor" working against the adoption of any new technology (Rogers 1983). Thus, to guide our users gently into a new paradigm, we had to create an interface that offered a clear and substantial advantage over existing technology and at the same time did not seem overly complex. The annotation "highlighter" pop-up seemed a good solution that tried to balance innovation with ease of use and some compatibility with a known paradigm. However, the annotation pop-up led to a tedious routine that in reality severely curtailed the ease of use. When ease of use is compromised to such an extent new possibilities inherent in a technology do not lead to a change of routine to accommodate the technology, and thus an adoption of the new paradigm does not occur. Instead, the perceived constraints lead to a change in the technology (Leonardi 2011). This is exactly what happened in the interaction between developers, users, researchers, and technology in the case of eLaborate. A bold button was introduced to remedy usability constraints: social shaping of technology at work.

As an unintended consequence – as Robert Merton would have it – of this social shaping of eLaborate the paradigmatic intent of the innovation was now black boxed. This is not meant in the sense of Latour's definition that defines a black box according to general acceptance of the correctness of the inner mechanism (Latour 1987), but in the sense that the innovative aspect of the new paradigm was now completely unobservable and thus effectively unknowable to its intended audience. The unobservability of such a black-box model is also a known "fail factor" for innovation (Marinova and Phillimore 2003; Rogers 1983). I my experience I found interfaces to often have this unintended and usually unrecognized effect, and it is a problem that particularly affects graphical interfaces. A graphical user interface suggests a transparency of model and paradigm that is not truly there – in fact the graphical interface is as much an opaque barrier to the internal paradigm of a system as it is a means of engaging with that very system. Analogous to Robinson (2013b) and others, I would argue that software interfaces, such as the interfaces to digital text editions, are an intellectual argument about the internal model of a system rather than a neutral communication of that model to any user. Vice versa, when the interface undergoes social shaping as a result of the interaction between developer and user/researcher, that is also an expression by that user of an intellectual argument about the model.

In the case of the bold button, the user has not merely molded convenience into the interface. What also happened was that the intended paradigm – that of semantically oriented XML – was expressed in a paradigm which was more familiar to most users, namely that of representationally oriented HTML. But this effectively prevented the user from engaging with and getting to know the new paradigm, or at least a part of it. The bold button hid a class of semantically expressive potential behind a single representational "wrapper". As an extension of the Meno paradox (Nickles 2003), not only were the users unable to negotiate new knowledge, they had shaped the technology in a way that made it now impossible to engage at all with the new paradigm. User-centered design had led to the users shaping new technology so that it was congruent with the paradigm they were familiar with. The new was expressed in the ways of the old, but it also turned into something inaccessible and irrelevant. This unintended effect of an intended paradigm being encapsulated and effectively hidden by a more familiar paradigm is caused by what I will call paradigmatic regression: the social shaping of a technological interface such that it can no longer express essential properties of an intended paradigm. The pivotal error that was made with the introduction of the "bold button" was that the button did not express the digital paradigm. Instead, we did exactly the opposite: we facilitated the scholarly users' regression toward the paradigm of the book metaphor known to them. Thereby we confirmed that nothing had changed, that print convention was still the paradigm to use. As proponents of digital scholarship, we may tend to think we are free from this sort of paradigmatic regression. But we are not. Most if not all digital scholarly editions are still solidly rooted in book metaphors and print conventions, and I think it is exactly because of this silent regression. A brief and selective history of humanities computing may be telltale.

The beginnings of humanities computing and the development of the digital scholarly edition are usually dated to 1949 with the seminal work of Father Busa (Hockey 2004). Roberto Busa demonstrated the first practical applications of computational text processing by automating the tasks of indexing and context retrieval. However, the result was presented in a form already well known to scholarly editing: a fifty-six-volume print publication concordance. The computational aspect was used simply to automate and scale a tedious and error-prone editorial task. The utility and sense of this work goes without question. What interests me here, however, is that the automation was geared toward reiterating on a larger scale a scholarly task that was in essence well known and rehearsed; computational power was harnessed to produce an instrument well within the confines of the existing paradigm of print text and its scholarly applications.

The advent of the database and later the relational database prompted the curation and publication of several catalogs and indices of textual metadata, as well as the first repositories for digital text. This was of course a major enhancement of the capacity for discovery of texts and related metadata. Databases allowed for efficient and convenient discovery of text through the use of matching selection queries. Scholars such as Jerome McGann, Peter Robinson, Dino Buzzetti, Manfred Thaller, and others began to envision different forms of engagement with text made possible due to the availability of full-text repositories and metadata. Despite all this, the database did not change the essential way scholars engaged with the actual texts. Even if, for instance, Buzzetti and Thaller argued that a digital edition's "liability to processing" is the essential feature that sets it apart from conventional editions (Buzzetti and Rehbein 1998), texts were still perceived predominantly as intentionally ordered strings of words for human interpretation. Thus, notwithstanding ideas on how to engage with text in new ways separate from the reading, commentary, and interpretation that has traditionally been handled by humans, the digital scholarly editions produced in the last part of the twentieth century have again presented text to us essentially as a digitized book.

According to Hockey (2004), in the early to mid-1990s a great deal of interest and discussion arose in the scholarly community concerning what an electronic edition might look like. However, with the "notable exception of work carried out by Peter Robinson", few of these publications were realized in an actual implementation. Once "theory had to be put into practice and projects were faced with the laborious work of entering and marking up text and developing software, attention began to turn elsewhere". As with the bold button example, we find that a new technology turned out to provide too little practical facility to lead to successful innovation. Yet there is more to the matter.

The "Next Big Thing" of the last decade of the twentieth century was the World Wide Web, founded on the technologies of the Internet and hypertext. As Landow has pointed out, "computer hypertext – text composed of blocks of words (or images) linked electronically by multiple paths, chains, or trails in an open-ended, perpetually unfinished textuality described by the terms link, node, network, web, and path" precisely matches Roland Barthes' ideal textuality (Landow 2006). If we need to point to a single moment and opportunity in history when the very fabric of a new technology was made suitable to a scholarly community for the expression of relations and structures, not just within single texts but especially between texts, it was the moment of the invention of hypertext. That the opportunity arose cannot have been surprising, as the essential mechanism of hypertext – the hyperlink – was the technological implementation of a long-standing idea that knowledge and information are interlinked. Already pioneers such as Paul Otlet in the early twentieth century could contemplate information systems that would link knowledge in the form of formalized multidimensional relations between documents (Rayward 1994). What is actually rather surprising is that such long-standing epistemological knowledge about the relation of different chunks of information within documents and congruent ideas from post-structuralist literary criticism such as Kristevas intertextual references (Mitra 1999) found so little expression in digital scholarly editions. The expressive power of that single pivotal element of the original HTML 1.0 specification, the A element with its invaluable HREF property, implemented by Tim Berners-Lee and itself an echo of Theodor Nelson's ideas of transclusion (Nelson 1995), should have reverberated within the scholarly community. The Hypertext Markup Language (HTML) enables anyone to link one HTML document to another by inserting the anchor element and pointing its HREF attribute to the web address of the other document. So when you see a link on a webpage (e.g.the blue headings on search results in Google) then in the source code of that page the link is encoded similar to this example:

```
Further work has been pursued by researchers of the
  <a href="https://www.huygens.knaw.nl/">
    Huygens Institute
  </a>
.
```

On a webpage that <a> element ensures the words "Huygens Institute" will be a clickable link, and clicking the link will forward you to the URL (i.e. web address) encoded in the href attribute of that same element.

This simple mechanism should have created ample opportunity for editors to give expression to the linked and intertwined natures of cultures of text, literary criticism, and (digital) textual materiality that go to the heart of the field (Van Mierlo 2006). The hyperlink created a native digital expression for the act of referencing, an expression of knowledge very much at the core of textual description, interpretation, and criticism. Thus, here was a unique opportunity to change from a paradigm of print publication to a paradigm of interconnected texts expressing knowledge.

The scholarly editing community, however, adopted the "markup" rather than the "hyper" part of the hypertext markup language, by developing Goldfarb's SGML eventually into the TEI-XML descriptive standard (Goldfarb 1996; Renear 2004). At the time, these dialects of markup technology were used primarily to mark up texts as they are represented in books – the fact that I do not think anyone has but flippantly suggested marking up Web pages in TEI-XML may stand to prove the point. The scholarly community predominantly turned hypertext markup into a descriptive model of the book, and we have produced digital book metaphors as digital scholarly editions ever since. As with the bold button, a new technology was not explored but rather encapsulated by a known paradigm. The hyperlink was meant not to be a descriptive tool, but to link information in different documents. Yet its foremost use in scholarly editing has been to link contents, chapter headings, and indices to pages in self-contained digital editions. Roberto Busa had "a vision and imagination that reach beyond the horizons of many of the current generation of practitioners who have been brought up with the Internet". He imagined scholarly editions on the Internet combined with analysis tools (Hockey 2004), a horizon that has been reiterated by many (cf., for instance, Buzzetti 2009). However, digital editions developed in a completely different direction. The processing involved is mostly aimed at rendering the text for consumption by human readers. To defy the intent of the hyperlink has been in my view among the most remarkable feats of paradigmatic regression in the textual scholarship community. One can wonder though whether this is a bad thing. If we accept the bilateral dynamic between audience and innovation, then why would we care when some innovations do not succeed? If the book metaphor paradigm suffices for our needs, does this not indeed suffice?

To answer this question we must ask: to whose needs do digital scholarly editions actually cater? Given the designation, they should cater to scholars and researchers, but do they? The latest developments in digital scholarly editing are linked to the possibilities created for Computer-Supported Cooperative Work (CSCW) – a term that was coined by the IBM research group headed by Greif (1988) – by networked computing, the Internet, and the rise in computer literacy. Essentially CSCW is a label that can be put on any collaborative activity that is supported by Web or Web 2.0 means. Crowdsourcing as a means of dividing large work-loads has been around for a while and has been a specific implementation of CSCW ever since Web 1.0 technologies turned into Web 2.0 technologies. Many have proclaimed crowdsourcing to be the advent of the social edition – most prominently Ray Siemens (Siemens et al. 2012) – which redefines the editor's role to be that of a team leader concerned with proper workflow, quality control, and overseeing managerial and funding aspects (Sahle 2013), whereas concrete editorial tasks are delegated to social communities formed around specific texts. Questions have been raised about the actual effectiveness of crowd sourcing (Causer, Tonra, and Wallace 2012). But more importantly, recent studies show that the old rule of thumb of the collaborative Internet – that 10% of the workforce provides 90% of the labor (cf. Brumfield 2012; Brumfield, Klevan, and Vershbow 2012) – still holds for any open collaborative project, implying that many crowdsourced editions are not in fact truly social. Moreover, when Peter Robinson said "All readers may become editors too", he was not simply referring to a cheap labor force for source transcription, to be conveniently discarded the moment a transcription phase is done (Robinson 2004). Instead, like Ray Siemens proposed, he envisioned a "social edition" that embodies the ideas of open notebook science (cf. Shaw, Buckland, and Golden 2013) and renders all aspects of the editorial process – e.g. annotation, commentary, and interpretation – open to public engagement (Siemens et al. 2012). But we in the scholarly community are not at all at ease with letting go of our presumption that scholarly editing is a highly skilled practice that does not provide for easy delegation of tasks. It is challenging to truly consider the extent to which we can open up the scholarly process of creating a digital edition to leave the tedious tasks typically associated with high quality scholarly inference to the wisdom of the crowds – in the case of literary analysis, this often includes, for instance, the painstaking tracing of names, annotation of plot, and clarification of meaning. In current practice, however, the digital scholarly editorial tasks beyond the transcription phase remain reserved either for the single authoritative author or for a small group of qualified editors. In this way, most scholarly digital editions adhere to an authoritative publication paradigm. We use big all-encompassing words like "social", "open", and "community", but in fact we are again regressing to authoritative processes that remain well within the paradigm of the print edition. Although on the verge of being harsh, it is nevertheless fair to state that digital scholarly editions cater to the needs of the scholarly editors, not to users and researchers as knowledge producers.

Along another tangent: Edward Vanhoutte (2011) pointed out the possibilities of targeting different audiences with different visualizations of the same edited digital text resources. So-called "minimal editions" – essentially filtering down all resources to provide a polished and uncomplicated reading text – could cater to a broader audience while "maximal editions" would cater to the use of scholarly researchers, providing all scholarly explanations, variations, annotations, and so forth. Several digital scholarly editions do show signs of this sort of differentiation. We can point to the Van Gogh Letters (Jansen, Luijten, and Bakker 2009) as something of a midpoint between the minimal and maximal edition. The Samuel Beckett Digital Manuscript Project (Van Hulle and Nixon 2011) and the pre-production version of the Digital Faust Edition (Brüning, Henzel, and Pravida 2013), that I was allowed to peruse while it was in development, certainly should qualify as maximal editions. However, these digital scholarly editions again reiterate in their GUIs the metaphor of the 'read-only' book.

Only very few digital scholarly editions do provide what I think is paramount for true interaction with editions or scholarly text resources: the capacity to negotiate the edition and its text as data over Web serviced Application Programming Interfaces (APIs). APIs allow for computer-to-computer negotiation of texts, opening them up to algorithmic processing and reuse. My primary reason for arguing that we need digital scholarly editions as API accessible texts is not, as some may expect, to enable quantified computational approaches such as those that Matthew Jockers and Franco Moretti have presented (Jockers 2013; Moretti 2007), or the stylometric analysis desired by many others (Van Dalen-Oskam and Van

Zundert 2007; Kestemont 2012). It is highly useful and convenient to have the text of scholarly editions available as an open Web service, so that my computational colleagues and I can do our principal component analyses, bootstrap consensus trees, clustering analyses, and any other analysis that can possibly be envisioned.

But there is another reason, in my view more important yet overlooked, to consider anchoring digital scholarly editions on a data model that is not oriented around a book metaphor. This motivation derives from the growing and increasingly unsettling gap I find between the close reading of scholars using conventional hermeneutic approaches and the "big data" driven distant reading supported by probabilistic approaches – a discrepancy which is also signaled by others (e.g. Capurro 2010). On the one hand, we see a conventional scholarly approach in which texts are mindfully and meticulously produced, detailed, and interpreted. On the other hand, we find a deterministic and probabilistic approach that focuses on large-scale data analysis and which is, through its statistical aspect, reductive in nature. To the hermeneutic scholar, distant reading approaches are therefore "lossy", prone to discarding some of the substance, and quite incapable of capturing essential hermeneutic knowledge (cf. Ramsay 2011c). It is often the statistical outliers and not just patterns of similarities that are telltale to textual scholars and historian in their hermeneutic explorations. At present there is no model connecting these worlds of close and distant reading. Rather, the distance between them is growing, which threatens not only to set the scholarly community of textual and literary studies against itself, but also to waste the opportunity for a true and meaningful advance in our capabilities for computational-based humanities research.

If we are to close this gap, we need a model for digital text that allows for both hermeneutic and statistical approaches so that these approaches can truly inform each other. To this end we need to revisit and reconsider how we anchor digital editions on the hypertext model. The slavish adherence to the book metaphor, even in XML form, will not take us into a realm where texts and editions are published as online APIs for processing by computational means. Yet, also models of quantification fall short as they are narrowly defined for statistical methodology. Because such models are not data models, they do nothing as to expressing description, encoding, or annotation. We are in need of a model that actually provides for all of the above. That is, a model that provides for the capturing, encoding, and annotating of a text and also for processing the edited or raw resource to enable analyses by both conventional hermeneutics and quantified approaches. Lastly, this model must be recursive: it must be able to capture all resulting information from an analysis and add that information into the model itself. Only then new knowledge gained from the model can be used 'natively' for a next cycle of both qualitative and quantitative analysis. Such a model captures all editorial and research aspects and outputs of scholarly activity in an encompassing lifecycle. But even more important: only such a model provides for a way to bridge the widening gap that is coming into existence between the hermeneutic tradition and new quantified means. Computational method can do far more than just counting, averaging, and comparing histograms. But currently computational approaches ignore many of the properties of text and textual materiality that are important to hermeneutic engagement. Current quantified approaches lack therefore the ability to model and computationally process the close reading aspects of text engagement.

Thus what we lack is something we could call tongue in cheek near distant or near close reading. More formally and in line with current debate, I think we should qualify what we lack as an enabler of computational heuristics for capta (Drucker 2011). There is no computer language that specifically supports the type of abductive reasoning that textual scholars and literary researchers do. Their heuristic is a scholarly adequate but not very formally qualified set of rules and activities underpinned by knowledge and experience from reading, studying, and discourse. These heuristics are applied to combine both evidence and plausible assumption to infer interpretations of by definition situated – that is, context dependent – cultural artifacts. It is thus a method to construct interpretation, the result of which is what Drucker calls "capta". Arguably either 'near close reading' or 'near distant reading' both capture in their own ambiguity exactly the properties of textual scholarly data, knowledge and method that quantified approaches tend to overlook: extremity of sparseness, inconsistency, vagueness, ambiguity, multi-interpretability, and uncertainty. There is no readily available means for such qualitative computing. Qualitative modeling and computing are still highly explorative fields (cf. Forbus 2008), and yet, abilities to compute and reason over qualitative data are coming into existence. As the creators and providers of the raw materials that such qualitative computational approaches should operate on, editors of digital scholarly editions should consider how text as data is to be provided.

Knowledge graphs are, I think, extremely well suited for this. Knowledge graphs represent our objects, concepts, and properties of interest and their relations in a network. They consist of nodes that usually represent objects or concepts. These nodes are connected by lines, called edges, that usually represent relations. As a matter of fact, the constructor of a graph is completely free to determine the meaning of edges and nodes. If he or she chooses so the edges might be the concepts and the nodes types of relations. One can imagine that all tokens in a text might be individual nodes and that the edges determine the linear order in which they appear in the text. Another graph might represent characters in a novel and the edges might model the relations ("friend", "sister", etc.) between them. In essence knowledge graphs are a formalization to express predicate semantics, two nodes represent object and subject, the edge indicates the predicate (figure 3.1).

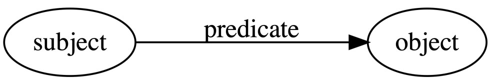

Figure 3.1: The predicate principle of graphs.

Graphs are a most generalized class of construct of which ontologies and thesauri are subclasses with more specific constraints. In their most generalized form graphs are stupefyingly easy to express in computer readable form using, for instance, the DOT format or language (Hayes-Sheen 2017):

```
graph {
  mary -- rose[label="friends"];
}
```

This example results in the graph of figure 3.2, when processed with a suitable program such as Graphviz[3]. The salient point is not that this results in nice visualizations, but that this simple formalization serves as a facile interface between human expression of knowledge and information that can be processed by the computer and be reasoned with by algorithms. A very baseline example of this would be to infer who is a friend of a friend in a more complex network.

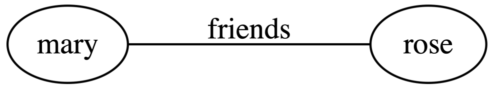

Figure 3.2: A most basic graph, showing two nodes and one edge.

Graphs are not new to us, nor to our field. The World Wide Web is a graph, a network of nodes and edges connecting information. In a sense, every digital scholarly edition put online has in fact been made part of a graph therefore. In recent years, graphs have found various more explicit applications also in the field of digital humanities, most notably as a data model for describing textual variation between different witnesses of the same text (Schmidt and Colomb 2009). The properties of the graph model, however, allow it to be a generic model capturing the information tied to a digital scholarly edition on all conceivable levels of granularity. Two examples may show this potential conceptually.

The first example was kindly explained to me by Moritz Wissenbach who at the time was a technical lead working on the Faust digital edition from Würzburg university. Imagine a knowledge graph as a network with nodes and edges. In this hypothetical graph, we designate three nodes to represent texts A, B, and C. An interface to the graph allows us to add edges and nodes to this network. What is essential here is that the underlying model is a graph, the graphical display may take many forms but need not necessarily be a visual network itself. Suppose now a textual scholar X states that text A was conceived before text C. This statement can be represented as a directed relational edge (or predicate if you like) 'precedes' between A and C as depicted in figure 3.3. Now assume another researcher Y at another point in time, and

[3]http://www.graphviz.org/

not necessarily even knowing anything about text A, independently of researcher X, concludes that text B was conceived after text C. This statement can be captured by putting an edge 'precedes' between C and B. The tiny graph as depicted in figure 3.4 now holds the accumulated knowledge. However, note that the combination of independent observations now adds up to more than just the sum of its parts, for "traversing", "walking", "reasoning over", or "computing over" the graph – all these terms essentially express the same operation of computationally inferring knowledge from the graph – gives us the added knowledge that A must have preceded B.

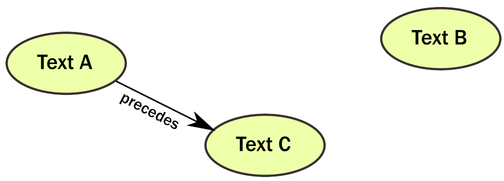

Figure 3.3: Nodes in a conceptual knowledge graph.

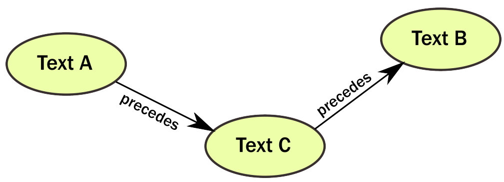

Figure 3.4: Edges multiply knowledge.

The second example is taken from CollateX, which is a tool to automatically collate variant texts (cf. http://collatex.net/). The result of such comparisons can be stored as graphs, e.g. figure 3.5. Such graphs cannot be said to be quantified, they express rather the qualitative word variance between texts. But the application of the graph stretches wider. As in the previous example, we can add statements (knowledge) about this text to the graph by adding nodes and edges. The example in figure 3.6 shows two statements made by superseding nodes on partly overlapping regions of the text. They express in a hypothetical fashion how these regions should look for a reader of an

EPUB publication of the text to be read on an eReader. Note how overlap, a well-discussed problem for hierarchical models (Sperberg-McQueen 2002), is not relevant to such a non-two-dimensional graph model.

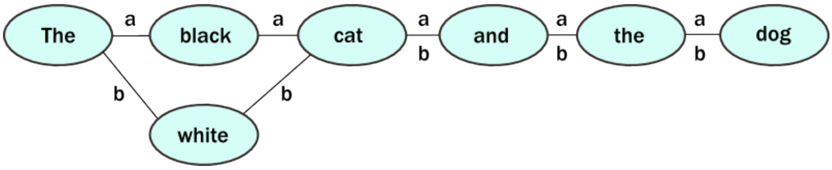

Figure 3.5: Conceptual knowledge graph representing textual variation in two texts a and b.

It should be carefully pointed out that knowledge graphs as a model are not to be equated with the currently popular ideas on semantic Web and RDF. RDF can necessarily only be a static representation of a certain state of such graphs that rather would be dynamic representations of changeable knowledge. To solve this initiatives such as the Open Annotation Collaboration are proposing extensions to the World Wide Web and Semantic Web models to support annotation of linked data including temporal "aware" annotations (Haslhofer et al. 2011). It is out of scope for this chapter to examine whether such models would provide for the needed reciprocality and dynamics for graph model-based digital scholarly editions. As the Web in its current form is not real-time read/write enabled, it is hard to imagine though how it would provide for such highly dynamic webs of knowledge interaction. The relation between RDF/Semantic Web and graph models is somewhat analogous to the relation between TEI and XML. A TEI conformant XML document is a singular instantiation of (a part of) the TEI model. The TEI model itself however is represented by the dynamic set of guidelines defined for the description of text and document structures.

Knowledge graphs can grow dauntingly complex very quickly, as may be inferred from figure 3.5. Because such complexity also poses a problem for querying and performance on the computer science side of things, we have not seen wide application of graphs until now – let alone as a model for humanities data. However, meanwhile knowledge graphs in the same fashion as shown in these tiny examples back the social network applications of,

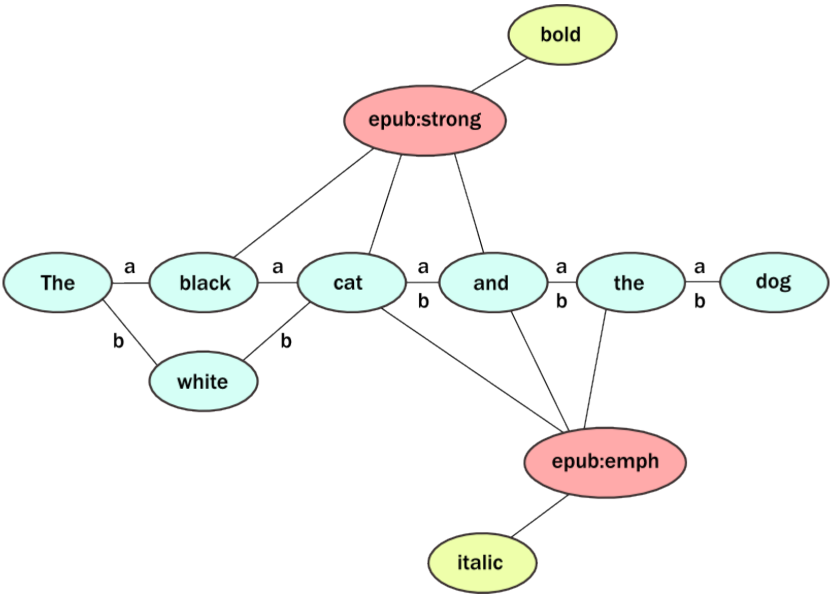

Figure 3.6: Overlapping semantic and representational knowledge added to the graph of figure 3.3.

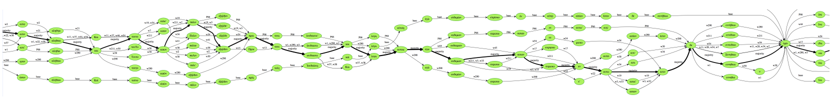

Figure 3.7: A graph representing a bible verse in various redactions.

for instance, companies like Facebook and Google. Graph databases[4] like Neo4j, ArangoDB, and AllegroGraph have made application-level models feasible. This paves the way toward exploring the potential of graphs for expressing the information and knowledge represented in digital scholarly editions. In reality when putting text and editions on a graph, as users we may not experience them as graphs, but rather as any visualization or data representation we want to derive from the graphs. By footing such representations and visualizations on a graph model, we provide an underlying truly generic and interoperable means for representing, editing, annotating, and visualizing text, its relations, its multi-perspectivity, and its materiality in digital scholarly editions. At the same time and through the same data model we provide a means for qualitative and quantitative computing over the information contained in the graphs representing our editions. Thus, with a graph model, we provide a more expressive data model for digital scholarly editions, allowing for the modeling and computation of both statistical and hermeneutic approaches.

Providing a digital scholarly edition with the backbone of a network graph would mean anchoring text on a fundamentally different model than that of the current prevalent digital book metaphor. All digital book metaphors are until now essentially closed off inconvenient mixtures of multiple page and string oriented hierarchical models. What we cannot achieve through the book paradigm is walking the various alternatives of the graph that expresses interpretations and knowledge about the document in consideration. That is, we cannot algorithmically get at and process the text with all its annotations, comments, and additional information on authorship, materiality, interpretation, etc. The reason for this is that the book paradigm keeps us locked in and focused on a finite representational state of the text: it is oriented toward closing down the text. In contrast, graph models provide an elegant open way to connect information to the text in an infinite extensible fashion. Whether machine negotiated or by human interpretation, new information can be attached to any particular item in the graph in the same way, thus becoming information that can be processed by both scholar and algorithm. Thus, the essential difference is that the same model can cater to capturing hermeneutic inference and computational analysis results. But

[4]https://en.wikipedia.org/wiki/Graph_database

we will only successfully explore that potential if we quit the social habit of shaping back new models into old paradigms.