

Qu'est-ce qu'un texte numérique? — A new rationale for the digital representation of text

Joris J. van Zundert

Royal Netherlands Academy of Arts and Sciences, The Netherlands

Tara L. Andrews

University of Vienna, Austria

Abstract

In this article we aim to provide a minimally sufficient theoretical framework to argue that it is time for a re-conception of the notion of text in the field of digital textual scholarship. This should allow us to reconsider the ontological status of digital text, and that will ground future work discussing the specific analytical affordances offered by digital texts understood as digital texts. Following from the argument of Suzanne Briet regarding documentation, referring to Eco's understanding of 'infinite semiosis', and accounting for the reciprocal effects between carrier technology and meaning observed by McLuhan, we argue that the functions of document and text are realized primarily by their fluid nature and by the dynamic character of their interpretation. To define the purpose of textual scholarship as a 'stabilisation' of text is therefore fallacious. The delusive focus on 'stability' and discrete 'philological fact' gives rise to a widespread belief in textual scholarship that digital texts can be treated simply as representations of print or manuscript texts. On the contrary—digital texts are texts in and of themselves in numerous digital models and data structures which may include, but is not limited to, text meant for graphical display on a screen. We conclude with the observation that philological treatment of these texts demands an adequate digital and/or computational literacy.

Correspondence:

Joris J. van Zundert, Huygens
Institute for the History of the
Netherlands, Royal
Netherlands Academy of Arts
and Sciences, Amsterdam,
The Netherlands.

E-mail:

joris.van.zundert@huygens.
knaw.nl

In 1951 Suzanne Briet asked the question 'Qu'est-ce que la documentation?', ruminating on what it is that documentation does, what constitutes a document, and what does not. She departed from a linguist–philosophical definition of 'document'—'Tout indice concret ou symbolique, conservé, ou enregistré, aux fins de représenter, de reconstituer ou de prouver un phénomène ou physique ou intellectuel.' (Briet, 1951)—and ultimately proposed a new understanding of the concept of document, one much more fluid than the writings on paper that we usually associate with the term.

Un étoile est-elle un document? Un galet roulé par un torrent est-il un document? Un animal vivant est-il un document? Non. Mais sont des documents les photographies et les catalogues d'étoiles, les pierres d'un musée de minéralogie, les animaux catalogués et exposés dans un Zoo. (Briet, 1951)

A rock on the ground, for example, may not have information communication significance. The same stone as a part of a museum's geological collection, on the other hand, may document the type of rock

found in a certain geological layer or area. Her most famous example, that of an antelope, becomes documentation of its species as soon as it is captured by an explorer and housed in a zoo; even its corpse can be preserved and thus maintained as a document even after its death. Briet in this way expanded our notion of the ontological and epistemological status of the concept of 'document'.

For Briet, documentation 'was a scientific activity of the greatest importance' (Hearns Bishop, 2003, p. 12)—a positive and unifying force, a foundational work for science, and an inscription technology that allowed knowledge to be codified, connected, and spread (Bede, 2007). Most salient to our argument is that documentation is situated; it is an act of interpretation made from a particular cultural and historic context (Hearns Bishop, 2003, pp. 12–13). Thus, when Briet's antelope becomes a document and becomes a part of the document we call taxonomy, it is also a constituent part of a specific culturally induced worldview. The antelope document inscribes the meaning of an antelope according to a certain specific culture. Briet evokes the sheer prolific power of documentation to establish these meanings: 'In our age [...] the least event, scientific or political, once it has been brought into public knowledge immediately becomes weighted down under a "vestment of documents"': a new sub-species of antelope inspires a newspaper item, is described in various scientific articles, a specimen gets loaned to an exhibition, a taxonomic description is made, etc. (Briet, 1951, 2006; Hearns Bishop, 2003, p. 12).

Briet thus advocated a shift in understanding of what a document is, urging her readers to focus on its function rather than the form in which one normally expects to find it. If we take such a purely functional view, then a document is anything that, on the material level, is used by humans to communicate information to other humans.

Briet's understanding is thus that the concept of document is fluid. Documentation is not merely prolific, it is also transformative. Each act of documentation that sprouts yet more documents—that we understand now can be of any kind—transforms the inscription of knowledge that the source document contained. Bishop in this respect remarking on Briet notes that Briet's perspective is still important today

'because her theories allow us to view a wide variety of information objects in terms of their relationships'. These relationships are transformative: documentation 'is a surrogate artifact, it is an interpretation of the artifact. In effect, a new 'artifact' in the form of documentation is created to serve as a surrogate for the artifact' (Hearns Bishop, 2003, p. 15).

If, thanks to Briet, the concept of 'document' is a fluid one, then so must be the concept of 'text'; while text is a feature commonly found in documents—indeed documents, in the broad understanding that Briet defined them, are the carriers of text—text is no more bound by the document than documentation is bound by the text it carries. However, that does not mean that text and document do not influence each other. The relationship between medium and message and between technology and meaning is reciprocal (McLuhan, 2003 (1964)). As Alan Kay (1993) put it: 'I had a very McLuhanish feeling about media and environments: that once we've shaped tools, in his words, they turn around and "reshape us."' The means of documentation in part shapes (or contributes to the shaping of) the interpretation that a document inscribes. A taxonomy for instance inclines toward enforcement of its categories: the decision to put a specimen in a certain category depends not only on the judgment of, e.g., a biologist but also on the trade-off to be made between the convenience of the existence of a category that more or less fits and the effort it takes to create one that might perhaps be more fitting, as well as the ramifications that this new category might have for the overall structure of the taxonomy. The medium of any text tends likewise to influence the shaping of the text—inscription in stone demands austerity, and a digital text editor invites prolixity.

The medium shapes the text; in the same vein then the medium shapes to an extent the meaning of the text. The prolific nature of documentation Briet pointed to is also the prolific nature of information: cultures use media of any kind to proliferate information. The medial shaping of text is causally connected to the remediation that happens when we push information onto a new medium to communicate, store, transform, or analyze it (Bolter and Grusin, 2000). But text is volatile and stable in

equal measure: even while the medium shapes a text, so that text can be conveyed across multiple media, it can still be recognized and meaningful as the same text. Text thus adapts with great ease to any new medium—from oral to pictorial to inscribed to time- or digitally based storage media—and will permeate nearly any medium within a short amount of time. Text is constantly moved, copied, translated, paraphrased, re-written, re-contextualized, and re-mixed. Each occurrence of any of these acts produces a new text, the result of the act, distinct from the text or texts that went into the act, and yet recognizable nevertheless as ‘the same text’. This is the textual condition, as it was defined by Jerome McGann (1991). We can understand text as a stable entity, but the harder we try to stabilize it, the more stubbornly it refuses to be bolted down.

Although philologists perceive the unwillingness of texts to be pinned down as ‘philological fact’ as a problem (McGann, 2013, van Zundert, 2016, pp. 375–60), for Umberto Eco this is their most salient feature, as it allows for ‘infinite semiosis’ (Eco, 1981). Eco argues that the sign has its roots in omens: natural phenomena that could be interpreted as predictors of natural events. Dark clouds forebode rain, and smoke above the forest signals a fire. This dynamic aspect of inference and interpretation is pivotal for the sign function. Only well after the invention of writing does an identity function begin to be attributed to signs as words; this process results from a unification of a theory of signs and one of language that, according to Eco, finds its culmination in the work of Augustine. In written language denotation, i.e. the assertion that a word (or sign) uniquely refers to some real-world antecedent, becomes strongly foregrounded. According to Eco: ‘problems derive from the fact that contemporary theories of sign have been dominated by a linguistic model, and a wrong one at that [...] where signs are conceived of as being intentionally emitted and conventionally coded, linked by a bi-conditional bond to their definition, subject to analysis in terms of lesser articulatory components, and syntagmatically disposed according to a linear sequence’ (Eco, 1981, p. 39). He refers to C. S. Peirce to argue that a sign’s primary function is not to signify some identity with

a real-world phenomenon, that is a word as a sign is not strongly linked to one single meaning, instead a sign works through inference, or interpretation. Reading, interpretation, and understanding do not operate by scientific induction or deduction to make predictions about the meaning of a sign. Rather, these are abductive processes: given a word, a reader instantly hypothesizes about possible meanings of that word and of its relation to words in its context, based on her own tacit knowledge. Reading a text is therefore not the decoding of a sequence of identity relations from text to real-world objects and events but the construction of meaning through a process, executed by the reader, of structuring hypotheses.

It is this phenomenon of abduction-based reasoning about the meaning of signs in a text that drives the infinitive process of interpretation and reinterpretation to which readers subject a text. This same phenomenon can be said to underlie Barthes’ (1975) concept of ‘writerly’ text: as a reader reads, she is constructing a new ‘cognitive’ text from a sequence of words. This text is ‘writerly’ because the reader is in a sense re-writing the text—she is both re-constructing its meaning and therefore also adorning and changing its meaning because the hypothetical or abductive way of decoding signs allows for—or rather, is the direct cause of—variation in interpretation. This variation would not be possible if words (and signs) really only existed as a unique referential relationship of identity.

Thus we cannot read but by interpretation. The text-as-signs is the inscription on the page, on the screen, or on the disk, but that is in a sense the least enticing aspect of text. It could be argued that the serialization of a text as words on paper or as bits in electronic storage exists merely as an affordance for the reader to hypothesize about the meaning of the ‘actual’ text that is being cognitively constructed as she reads. It is clear that this new ‘writerly’ and mental representation of the text is an ephemeral one, existing only in the mind of the reader. It is moreover a text that is ontologically different from what might be recorded in signs on the page. In order for its meaning to be fully realized, a text must thus undergo an ontological shift, from

existence as signs on paper to existence as a cognitive representation in a reader's mind. How could we say that what is in the mind of the reader is 'not' a text?

Where the act of documenting creates a surrogate of a phenomenon or artifact, the act of reading produces a cognitive surrogate of a text that is itself already a surrogate: the signs on paper that represent a cognitive text that once existed in an author's mind. To exist as communicated meaning, the text has undergone not one but two shifts in its ontological status, and arguably has become three texts along the way—and yet it is said that both author and reader have acted upon the same text.

Documentation codifies and inscribes how a certain culture understands its world. As Briet showed, the volatility of the concept of document and the prolific nature of documentation are fundamental to this function. Without a rapid proliferation of copies, surrogates, derivatives, and remediations, documentation fails in one of its major purposes. Text inscribes information that codifies some understanding about the world. Its carriage by these rapidly proliferating documents makes text volatile with respect to form, medium, and—following McLuhan, 2003 (1964)—meaning. The abductive process of interpreting and understanding them, that according to Eco we cannot escape, makes their meaning yet more volatile. Moreover, to achieve its purpose of communicating some meaning, the volatility of texts—and quite possibly that of documents too—must also 'fundamentally' encompass a negotiation of boundaries between modes of being, that is an ontological shift from being as signs-in-a-medium toward being as cognitive representation.

Thus, the functions of document and text are realized through processes of dynamics. The idea then that the purpose of textual scholarship is to 'stabilize' a text, is an audacious one. Even as the philologist works to stamp an authority on a particular version of a text, the text itself replicates cognitively with every exemplar considered, and its meaning shifts with its medium. The end result, of course, is a new text which can claim to be a faithful representation of the cognitive text of the editor, informed by the texts of the exemplars. This new

text may stake a claim to supplant prior editions, but it is very difficult to argue that it supplants any of the exemplars, and even its claim to authority over prior editions can be questioned. Essentially, a new set of signs has sprung into existence that will produce yet more texts, each of which may be just as prolific as its siblings and ancestors. Thus in the quest for authority and stabilization of a text, the philologist cannot help but have a multiplicative effect.

In the time before the rise of digital scholarly editions, the sheer audacity of this multiplication in service to authoritative stability was not so clear. As long as the system of print production and mass distribution of books endures within textual scholarship, the authority of the particular inscribed version of a text that the philologist seeks to impose has been amplified by the inherent authority of that version gaining access to the academically and commercially controlled channels of distribution and replacing older versions on the shelves of libraries and bookstores.

The scholar who, on the other hand, goes on to make a digital edition drives the proliferation of the ontological status of text even farther, and farther perhaps than he or she realizes. Peter Shillingsburg has argued that even a simple digital transcription cannot but be an imprecise and often erroneous representation of a written text. The questions he raises defy simple answers:

But especially transcription—even 'text only' transcriptions—involves interpretation (Is it an *i* or *e*? Is it underlined or crossed out? Is the obscured letter a *k* or a *t*? Should that upright have been crossed as a *t* or is it an *l*?—were the bushes lopped or topped?). And these questions about the text can be multiplied if one asks what is the meaning of underlining or italics (Is it for emphasis or to indicate a foreign word, title of a book, or name of a ship?). And so, it is asked, can the surrogate be unmediated, representing exactly the original, such that the user need not see the physical document? And if transcription is always interpretive, is all the interpretive analysis by transcribers of a piece? Is it futile to distinguish levels of intervention so that the

decision about the e/i or t/l, the decision not to include crossed out words, the decision about the emphasis/ship's name, and the decision to add links to related documents just a continuum of editorial intervention from minimal to unlimited? (Shillingsburg, 2014, p. 164)

Just as Eco argues that contemporary theories of sign have been dominated by a linguistic model foregrounding the identity between sign and referent, leaving little space for the more fundamental operation of inference for formation of meaning, we argue that contemporary theories in textual scholarship have been dominated by a foregrounding of the document and the sign as purely discrete (insofar as they are able to be typeset) attributes of text. Shillingsburg's statements show how even at the most intimate level of the glyph philological editing is seen as a process of abstraction and creation of discrete surrogates of what is in reality a multi-attributed material representation of text.

Moreover, his discussion of markup code, and his claim that it 'tends to interfere with repurposing' of a text, indicates that textual scholars are working in a medium whose properties and qualities they often do not yet fully grasp. It is clear, as Shillingsburg argues, that digital transcriptions are not simply 'the text' itself. Even less so are digital editions—that is to say, digital versions of texts that go beyond simple transcription, or digital presentation of print-based editions. These are texts whose medium lends them qualities that defy translation to the physical or print medium, as observed by Sahle (2008). Notwithstanding the claim to representation of a handwritten or print text that these editions generally make, these digital texts are not simple surrogates or stand-in artifacts for originals (Hearnshaw, 2003, p. 15) nor are they merely

philological evidence (cf. for instance Greetham, 1994, pp. 2, 42, 43). To the degree that we can summarize Briet's argument as 'infinite proliferation of the document' and Eco's argument as 'infinite proliferation of meaning', we should regard these texts fully as texts in their own right.

Despite the risks associated with the use of a medium that is not fully understood, Shillingsburg advises us to 'just do it'—to go forth and create the editions, to create a text which may be intended to represent a physical text but is nevertheless new, and digital. The scholar who takes his advice is confronted with the peculiar fluidity, perhaps the agency¹ even, of the text that she has created. This is a feature of the digital medium in which we have chosen to work—one of its qualities that we, as yet, only dimly understand.

The digital medium can be transformative and discrete in its ambiguity. ASCII art is an excellent example of this (e.g. Fig. 1). What exactly is the text here? How should it be read, and how should it be represented if it is remediated? This is, however, not to say that text must be digital to be fluid, or that the potential for such double entendre resides peculiarly in the digital. The commingling of linguistic sign and visual image that is usually expressed as ASCII art today is in fact age old, as can be seen in Fig. 2. This kind of glyphic art stresses not only the transmedial or fluid nature of text; it also points to its continuous character even in this discrete medium. That continuity of interpretation, certainly present within artistic expression before the computational age, becomes even more pronounced in some Internet memes.

Or consider emojis—in essence iconography but now registered in the writing system called Unicode, pinned down in meaning by a standard but infinitely variable in interpretation depending on the style of implementation of that standard. This is



Fig. 1 ASCII Art, produced by the authors using the online Text to ASCII Art Generator (patorjk.com/software/taag/)



Fig. 2 Fragment of British Library, MS Harley 647, f.11r.
© British Library Board, reproduced with permission of the British Library Board

fluid transmediality at its finest—emojis are pictorial in nature and defined as glyphs, while ASCII art is composed of glyphs and used to produce something pictorial. The transmedial, fluid, and continuous nature of text and therefore of ‘textuality’ is now far more obvious than it ever was. The digital environment amplifies the continuity of these natures, just as the digital environment seems to amplify all problematic qualities of text (O’Donnell, 2015).

Our claim of course is not that text is fluid. This is well known (cf. Bryant, 2002, Levy, 1994). We also consider it well established by now that the very concept of text is fluid. What we claim here is that the digitally enabled humanities for the greater majority have fallen into a habit of considering digital texts as mere digitized surrogates of non-digitally inscribed texts, that is as documents. The height of sophistication of digital publishing among publishers still tends to be to offer a digital publication, meaning a PDF or an ePub version of a book—even while the idea of so-called ‘born digital’ texts (files produced by text editors, blog posts, tweets, and so forth) has found traction throughout the world. Or, to put another analytic frame on this, for the most part digital textual scholarship seems to be stuck in a paradigm of remediation. Bolter and Grusin contend that all mediation is remediation, that is all new media express themselves through the encapsulation of older media. ‘Each new medium is justified because it fills a lack or repairs a fault in its predecessor, because it fulfills the unkept promise of an older medium’ (Bolter and Grusin, 2000, p. 60). Indeed, the rhetoric surrounding the digital

scholarly edition is revolutionary, whereas closer inspection of the practice reveals little novelty; most digital editions seem to be dutiful remediations of print publications (Karlsson and Malm, 2004).

When scholars speak of a ‘digital text’ what they usually have in mind is the visible rendering of a digitally inscribed text, which usually takes a form visually very similar to a physical text, allowing the option of re-inscription on, e.g., paper. The on-screen display of the digital representation is technically an interface to the digitally inscribed text, but from the ontological perspective of the scholar, it is ‘really’ an interface to a real or potential physical text. This conforms to the assertion by Bolter and Grusin that ‘digital media can never reach [a] state of transcendence, but will instead function in a constant dialectic with earlier media’ (Bolter and Grusin, 2000, pp. 49–50). Yet perhaps the most significant remediation of text is not occurring at the level of the graphical interface. There may certainly be a mediation—in the original Marxist analytic sense of the process of negotiating a balance of power between social groups—between scholars producing conventional print editions and those creating digital editions. But the more important, yet less apparent, remediation is a similar renegotiation of what text is, between those scholars who understand digital text as the visualization in a graphical interface and those scholars and programmers who write, work with, and experience the digital code and models of text of which these visualizations are merely screen-oriented representations.

Representation for reading purposes only scratches the surface of what a digital text is. As soon as scholars set out to apply the first computational analyses to text and to create the first digital editions, text started to flow into the digital environment. As it did, it brought its textual condition to the digital environment, even as properties of the digital were imparted on the nature of text. The people who worked within the digital environment began to create a particular category of text that was digital in nature. Programming languages, for instance, applied algebraic and textual constructs, so that they could be more easily read and applied (anon, 1954). These executable texts were then

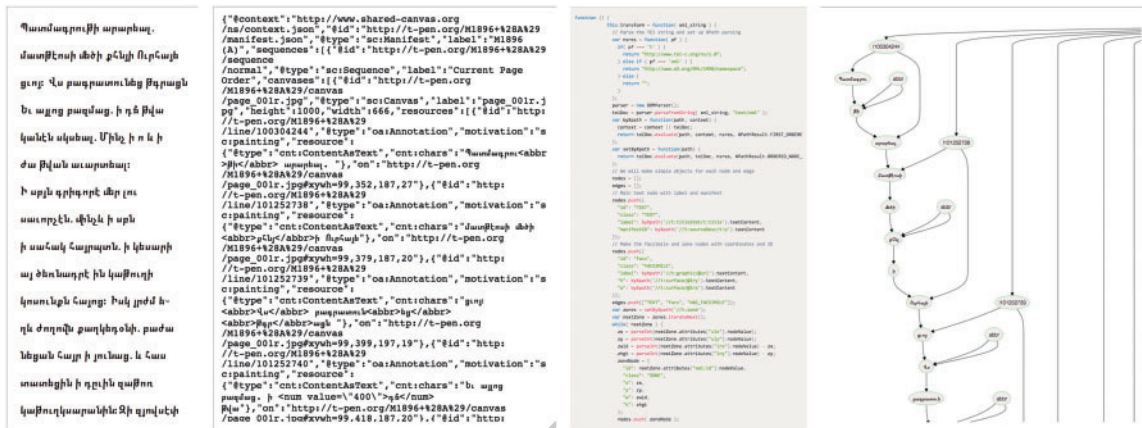


Fig. 3 Examples of four texts: a so-called plain text, a JSON encoding of a manuscript transcription, the text of a JavaScript source code document, and a graph depiction of textual markup

used to model databases, which in turn were used to model other texts into these databases (Jones, 2016). Computational linguists began compiling vast corpora of texts, using textual tags to annotate them, making them distinctly ‘different’ from the physical texts they were derived from. All these textual constructs imported into the digital environment became products of their idiosyncratic environment, defined foremost by their ‘digital’ properties.

Our claim is that these texts belong to a distinct ontological category. They are true digital objects with inalienable digital properties. Even a plain-text transcription is not a mere imitation of a real-world text, but should be considered as a text in and of itself. Until now we have not usually considered these texts as being texts in their true digital form. But what happens if we broaden our perspective to accept all of these as texts: databases, XML files in their XML form, source code in its legible form as well as in the form of the results of its execution, whatever visual form those results may take (cf. for instance Fig. 3). From this perspective, and with the help of hindsight, it becomes clear that the history of digital textual scholarship has been by and large one of ‘patching’ the perceived inadequacies of digital text to allow it to function more like ‘normal’ physical text—thereby inadvertently misunderstanding and disregarding the digital nature and existence of digital text.

The simplest form of digital text is arguably the string—a linear series of binary signals that encode characters according to some predefined table. Its origin is connected to the physical and technical requirements of telegraphs and earlier signal transfer technologies such as semaphores (Petzold, 2000). The ability to regard information as an unidimensional stream of discrete dichotomic bits was essential to the work of both Turing (1937) and Shannon (1948). They and we have nonetheless been aware all along that a linear series of characters can never capture the multidimensional properties of text. It cannot represent structure, semantics, relations, or perspectives internal or external to the text. Because of this, the computational string was ‘patched’ to become a data structure, initially with typesetting codes to instruct printing machines on how to produce typographically beautified texts (Goldfarb, 1996). Markup and hyperlinks were invented at a later stage, patching the string to allow for more multidimensional connections within and between digital texts. Markup in the form of HTML arguably became the most preferred of these patches, alongside XML in general and, in humanistic/scholarly contexts, TEI in particular. These ‘patching’ technologies were developed to allow digital text to behave more like analogue text. They helped to remediate what we knew about the properties of ‘real-world’ text in the digital environment.

Arguably the most advanced 'patch' we have come up with so far is the knowledge graph. The graph as an interface to a real-world text has been gaining currency (Andrews and Macé, 2013; Dekker *et al.*, 2015; Schmidt and Colomb, 2009; van Zundert and Andrews, 2016). As a model for the representation of the multidimensionality of text, the graph model takes us far beyond the limitations of the linearity of the string. It also takes us well beyond the limitations of the hierarchy of the string-segmented-into-a-tree that is markup. This has considerable advantages for the digital 'representation' of real-world text. Especially where that representation must encompass multidimensional aspects, such as ambiguity, narrative structure, variance, annotation, and so forth.

Nevertheless there is a further, more fundamental step beyond re-representation that should be taken. This is a step that was in fact already taken when text crossed into the digital medium and a new ontological category was created, but we as textual scholars have failed to acknowledge it. As long as we keep treating digital texts indeed as 'models' of text, digital models moreover whose only purpose of being is to depict themselves as digital re-representations of analogue texts, we deny these models their ontological status of actually 'being-a-text' in and of themselves. This is what we claim: the graph, the database, and the JSON-LD file that now are regularly created and maintained to function as data structures for the representation of text are in fact texts, and they should be considered as that: as texts.

That most scholars do not regard the idiosyncratic aspects of databases, graphs, text files, and so forth, as idiosyncratic properties of a kind and category of texts in its own right, is an effect of the fact that digital text production is still rooted very firmly in a representational philosophy. Almost all digital text production is geared toward recreating, within a digital environment, in a familiar guise, the comfortable and familiar aspects of continuous and fluid texts-in-the-physical-world. Even while we have used digital text in this way, the properties of these digital 'versions', which is to say the digital properties of these texts in their own right, were unintentionally neglected. Graphs, markup, and

strings seen solely as representations of text-in-the-real-world will always strike us as inadequate on some level. As Shillingsburg argued, it is not possible to make a perfect translation or copy of an analogue text into the digital realm. Kirschenbaum (2008) has convincingly shown that digital texts are physical too and that we must acknowledge their materiality. Along the way he confirms that our digital models have a full claim to the status of texts, for they too are material texts—ones that require machine mediation to be read, and have therefore a different sort of materiality, but nevertheless still material and still 'texts'.

In 2013 Jerome McGann, apparently driven to despair about the perceived volatility and ephemerality of digital texts, argued that textual scholars should regroup toward the philological-physical fact of the glyph on paper (McGann, 2013). We argue here that scholarship should rather venture in the opposite direction, embracing digital texts for what they are: texts adorned with properties that are both inalienably textual and inalienably digital. David Berry (2014) argues for the need to critically examine digital objects such as digital information streams, now that these objects increasingly help to constitute contemporary society and culture (cf. also Jones, 2014). We would add to this the argument that code and digital data structures are included among the digital texts that increasingly constitute contemporary cultural artifacts and scholarship. These texts are thus worthy of our philological consideration. We call attention here to their ontological and epistemological status and import within textual scholarship.

We should indeed go even further: where Berry (and others) call for the consideration of the 'surface' or the 'interfaces' of these data structures as digital objects in themselves, we contend that the data structures and models are themselves the objects worthy of our scholarly scrutiny. These are, after all, texts in themselves. When a scholar has modeled the semantics, the structure, or for that matter any characteristics of a text in a database, and she has added some logic or style sheets to depict a visualization of those characteristics onto the 'canvas' of a computer screen, then that depiction may be the representation of the text of some

physical exemplar. In the process, however, new cognitive texts and new documentary evidence of these, in the form of those very database models and style sheets, were created. Not only the visualization but also the digital objects that produce the visualization have become documents and texts in themselves.

The fluidity of document and the infinite semiosis of text cause a proliferation of documents and texts that each have inalienable unique properties that may be bound to the specific materiality and medium of the document and the text. In a scholarly context it is negligent not to acknowledge these idiosyncratic properties, and to regard them as mere inconvenient and unsatisfying incongruities between the physical print text, the digital representation, and the digital model. These incongruities are what make digital texts texts in their own right, and they point toward the differing ontological status of digital and print text. These texts cannot be and, in fact, actively resist being identical. Purporting that they are, or can be, and that they are only representations of physical texts, and nothing more, is epistemologically shortsighted. None of the texts we produce can have an inherent scholarly primacy over the others, simply on the basis of its form—the print text says things that the TEI encoding does not, the TEI encoding says things that the JSON does not; the JSON says things that the graph does not; and saliently: vice versa. They are all texts, and their forms are intrinsically bound up in the expression of their essence.

One reason that scholars have paid less attention to digital data structures and information models as texts in their own right may be that digital texts require their own specific literacy to be read and written. Digital structures and objects are texts that contain programming code, or require programming code to be created, analyzed, visualized, etc. That is, these texts are made up in part of signs whose meanings scholars will recognize from other sorts of texts (e.g. characters, words, and syntactic and semantic structures), but they also consist of signs still rather alien to scholars without programming experience, such as string denotations, punctuation semantics, variables, loops, and subroutine statements. These signs, innate to the realm of code

and computation, require a different, additional literacy to be fully understood and interpreted. The ability to code and encode are necessary prerequisites, but computational literacy goes beyond learning the syntax and semantics of a particular programming language. As Annette Vee noted: ‘But, unfortunately, when “literacy” is connected to programming, it is often in unsophisticated ways: literacy as limited to reading and writing text; literacy divorced from social or historical context; literacy as an unmitigated form of progress’ (Vee, 2013, p. 43). Vee argues that literacy refers to a set of skills without which one is no longer able to navigate one’s world. Code and digital texts as technologies are not yet infrastructurally critical to textual scholarship. However, the textual scholar who does want to engage with digital texts as ‘digital’ texts requires a specific literacy. Epigraphical literacy, codicological literacy, and computational literacy are essential in the understanding, respectively, of a stone inscription, of a medieval manuscript, and of a digital text, each one in its specific mode of being.

Vee’s argument is the most recent in a discourse spanning at least four decades, which includes inter alia Stephen Ramsay (2011), John Unsworth (2002), Friedrich Kittler (1993), and Donald Knuth (1984)—a discourse that puts forth the argument that working with digital texts requires some proficiency in coding, and that this proficiency is easily recognizable as literacy: reading and writing, but of a different kind. Roots of the argument can be traced back to the work of Adele Goldberg and Alan Kay, who were involved with the creation of Smalltalk, which can be regarded as the mother of all object-oriented programming languages. Kay and Goldberg were specifically interested in how programming could be taught, an experience that profoundly influenced Goldberg’s thinking on literacy, convincing her ‘that literacy should involve computing-based technologies and the expectation that our knowledge and skills will continually change, rather than define literacy as being pencil/paper/book-based’ (Goldberg, 2010, p. 24). However, literacy (be it computational or written-language literacy) cannot be reduced to the skills of reading and writing. Kay’s sobering observation was that, after 30 years of experience, the success of

teaching computing literacy still depended on the “hacker phenomenon”, that, for any given pursuit, a particular 5% of the population will jump into it naturally, while the 80% or so who can learn it in time do not find it at all natural’ (Kay, 1993, p. 81). More salient however is another observation he makes: ‘The connection to literacy was painfully clear. It isn’t enough to just learn to read and write. There is also a literature that renders ideas. Language is used to read and write about them, but at some point the organization of ideas starts to dominate mere language abilities’. That is, literacy does not only consist of the basic skills of reading and writing a certain set of symbols. Following Eco, interpretation and understanding come from tacit knowledge-based inference. Reading, writing, and thus also coding are about fluency of words and of symbols, whereas the fluency we need is a fluency in ideas and concepts. In the case of coding literacy, this means an experienced understanding of basic algorithms, coding constructs, and programming patterns, and it is a literacy that requires a number of years in training and experience, rather than a few months.

It is hard for scholars who lack this literacy to conceive of code and data structures as just another semiotics, another meaningful way to express texts. It is clear that being non-literate in code and encoding makes it extremely hard to appreciate ‘digital’ texts as what they are essentially: texts. What is then left is the mere use of code and data structures as another tool for representational approaches, for the depiction of a print or manuscript text in a digitized guise mimicking the exemplar as closely as possible. We lose sight of the fact that there are ‘native’ digital ways of looking and working with digital texts, reading and writing them, when we remain within our representational philosophical confines. That limited understanding not only provokes us to concentrate almost exclusively on standards for representing texts it also prohibits us from investigating the textual nature of the digital text.

Just as Briet argued the epistemological status of the rock as document, we should grant the proper ontological and epistemological status to the digital objects that we have so far used merely for textual representation. Just as a rock can be a document, a serialization or a source code is certainly a text.

References

- Andrews, T.L. and Macé, C.** (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*, **28**(4), 504–21.
- anon** (1954). *Preliminary Report: Specifications for the IBM Mathematical FORMula TRANSLating System, FORTRAN*. New York, NY: International Business Machines Cooperation. <http://www.computerhistory.org/collections/catalog/102679231> (accessed 13 November 2016).
- Barthes, R.** (1975). *S/Z: An Essay*. New York, NY: Hill and Wang.
- Bede, M.W.** (2007). What is documentation? English translation of the classic French text, by Suzanne Briet (Lanham, MD: Scarecrow, 2006). *College and Research Libraries*, **68**, 199–200.
- Berry, D.M.** (2014). *Critical Theory and the Digital*. New York, NY; London; New Delhi etc.: Bloomsbury Academic.
- Bolter, J.D. and Grusin, R.** (2000). *Remediation: Understanding New Media*. Cambridge, MA: MIT Press.
- Briet, S.** (1951). *Qu'est-ce que la Documentation?* Paris: Édit.
- Briet, S.** (2006). *What is Documentation? English Translation of the Classic French Text*. Lanham, Md: Scarecrow Press. <http://ella.slis.indiana.edu/~roday/briet.htm> (accessed 19 October 2015).
- Bryant, J.** (2002). *The Fluid Text: A Theory of Revision and Editing for Book and Screen*. University of Michigan Press. <http://books.google.nl/books?id=1w4wpOdPbu4C>.
- Dekker, R., van Hulle, D., Middell, G., Neyt, V., van Zundert, J.** (2015). Computer supported collation of modern manuscripts: CollateX and the Beckett digital manuscript project. *Literary and Linguistic Computing*, **30**(3), 452–70.
- Eco, U.** (1981). The theory of signs and the role of the reader. *The Bulletin of the Midwest Modern Language Association*, **14**(1), 35–45.
- Goldberg, A.** (2010). Oral history of Adele Goldberg. <http://archive.computerhistory.org/resources/access/text/2013/05/102701984-05-01-acc.pdf> (accessed 8 November 2016).
- Goldfarb, C.F.** (1996). The roots of SGML—A personal recollection. <http://www.sgmlsource.com/history/roots.htm> (accessed 26 August 2014).
- Greetham, D.** (1994). *Textual Scholarship: An Introduction*. New York & London: Garland Publishing Inc.
- Hearns Bishop, M.** (2003). Briet’s antelope: Some thoughts on Suzanne Briet (1894-1989) and conservation documentation. *WAAC Newsletter*, **25**(1), 12–16.

- Jones, S.E.** (2014). *The Emergence of the Digital Humanities*. New York, NY; London: Routledge.
- Jones, S.E.** (2016). *Roberto Busa, S.J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. New York, NY; London: Routledge, Taylor & Francis Group.
- Karlsson, L. and Malm, L.** (2004). Revolution or remediation? A study of electronic scholarly editions on the web. *HUMAN IT*, 7(1), 1–46.
- Kay, A.C.** (1993). The early history of smalltalk. *ACM SIGPLAN Notices*, 28(3), 69–95.
- Kirschenbaum, M.** (2008). *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press.
- Kittler, F.** (1993). Es gibt keine software. In *Draculas Vermächtnis*. Leipzig: Reclam Verlag, pp. 225–42.
- Knuth, D.E.** (1984). Literate programming. *The Computer Journal*, 27(1), 97–111.
- Levy, D.M.** (1994). Fixed or fluid? Document stability and new media. In *ECHT 94 Proceedings of the 1994 ACM European Conference on Hypermedia Technology*. Edinburgh; New York, NY: ACM Press, pp. 24–31. <https://pdfs.semanticscholar.org/f340/b718af6a208e47-c143e4d25c1da9dcaf9f23.pdf> (accessed 15 March 2017).
- Manovich, L.** (2013). *Software Takes Command*, Vol. 5. New York, NY; London; New Delhi etc.: Bloomsbury Academic.
- McGann, J.** (1991). *The Textual Condition*. Princeton: Princeton University Press.
- McGann, J.** (2013). Philology in a new key. *Critical Inquiry*, 39(2), 327–46.
- McLuhan, M.** (2003). *Understanding Media: The Extensions of Man (Critical Edition)*. In Gordon, W. T. (ed.) (First published 1964). Berkeley: Gingko Press.
- O'Donnell, D.P.** (2015). A first law of humanities computing? *Blog*. <http://people.uleth.ca/~daniel.odonnell/Blog/the-first-law-of-humanities-computing> (accessed 17 June 2016).
- Petzold, C.** (2000). *Code: The Hidden Language of Computer Hardware and Software*. Redmond: Microsoft Press.
- Ramsay, S.** (2011). *Reading Machines: Toward an Algorithmic Criticism (Topics in the Digital Humanities)*. Chicago: University of Illinois Press.
- Sahle, P.** (2008). About “a catalog of: Digital scholarly editions”. <http://www.digitale-edition.de/vlet-about.html> (accessed 11 November 2016).
- Schmidt, D. and Colomb, R.** (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67(6), 497–514.
- Shannon, C.E.** (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423.
- Shillingsburg, P.** (2014). From physical to digital textuality: Loss and gain in literary projects. *CEA Critic*, 76(2), 158–68.
- Turing, A.M.** (1937). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(1), 230–65.
- Unsworth, J.** (2002). What is humanities computing and what is not? In Braungart, G., Gendolla, P. and Jannidis, F. (eds), *Jahrbuch für Computerphilologie*, 4. <http://computerphilologie.digital-humanities.de/jg02/unsworth.html> (accessed 8 July 2013).
- van Zundert, J.J.** (2016). Author, editor, engineer—Code & the rewriting of authorship in scholarly editing. *Interdisciplinary Science Reviews*, 40(4), 349–75.
- van Zundert, J.J. and Andrews, T.L.** (2016). Apparatus vs. graph: New models and interfaces for text. In Hadler, F. and Haupt, J. (eds.) *Interface Critique. Kaleidogramme*. Berlin: Kulturverlag Kadmos.
- Vee, A.** (2013). Understanding computer programming as a literacy. *LiCS*, 1(2), 42–64.
- Woolgar, S., and Cooper, G.** (1999). Do artefacts have ambivalence? Moses’ bridges, Winner’s bridges and other urban legends in S&TS. *Social Studies of Science*, 29(3), 433–49.

Note

1. Following the debate surrounding the potential politics and agency of artifacts (cf. Woolgar and Cooper, 1999), some form of agency for documents and (thus) texts can be assumed. Although we would not attribute direct agency to texts, artifacts (and thus documents and texts) may effectuate a ‘deferred’ agency of, for instance, an author. This notion of deferred agency relates to poetic notions such as catharsis and, e.g., Bertold Brecht’s ideas on theater as a political forum. Among others Manovich (2013), Berry (2014), and van Zundert (2016) argue that ideas on such deferred agency actually are very relevant with respect to the recent form of text that software code is.