Journal of the Text Encoding Initiative

Issue 8 | December 2014 - December 2015
Selected Papers from the 2013 TEI Conference

Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive

Trevor Muñoz and Raffaele Viglianti

Electronic version
URL: http://journals.openedition.org/jtei/1270
DOI: 10.4000/jtei.1270
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Trevor Muñoz and Raffaele Viglianti, « Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive », *Journal of the Text Encoding Initiative* [Online], Issue 8 | December 2014 - December 2015, Online since 23 September 2015, connection on 21 April 2019. URL : http://journals.openedition.org/jtei/1270 ; DOI : 10.4000/jtei.1270

Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive

Trevor Muñoz and Raffaele Viglianti

AUTHOR'S NOTE

This paper was originally presented at the 2013 TEI Members' Meeting, October 2–5, 2013, in Rome, Italy. We would like to thank the audience members from that original presentation for their engaging and probing questions. We would also like to thank Neil Fraistat, who collaborated on the development of the original presentation. Thanks also to Jennifer Guiliano, who commented on drafts of this piece. Finally, the ideas developed here owe everything to the team of people who have participated in the Shelley-Godwin Archive project. The Shelley-Godwin Archive is the ever-growing, collaborative product of many hands across a variety of institutions and disciplines. Please see http://shelleygodwinarchive.org/about for a full list of our collaborators. We thank them profusely.

1. Background of the Shelley-Godwin Archive

1 The Shelley-Godwin Archive is a project involving the Maryland Institute for Technology in the Humanities (MITH) and the Bodleian, British, Huntington, Houghton, and New York Public Libraries that will eventually contain the works and all known manuscripts of Mary Wollstonecraft, William Godwin, Percy Bysshe Shelley, and Mary Wollstonecraft Shelley. The S-GA project began in 2011 and completed its first phase of work in 2015. In October 2013, the project released a beta version of its online reading environment containing high-resolution images and accompanying TEI-encoded transcriptions of the three surviving manuscript notebooks containing Mary Shelley's drafts of *Frankenstein, or, The Modern Prometheus*.

2 The development of the S-GA from its original conception to current plans for second and future phases of work reflects wider shifts in emphasis within the digital humanities: from construction and online presentation of expert-curated digital collections (Palmer 2004) to experiments with more participatory forms of scholarship (Burdick et al. 2012). The central claims of the original grant proposal for S-GA (submitted in 2010)[1] reach back to rhetoric from earlier digital humanities and digital library development that centered on the possibilities for technology to enable virtual reunification of dispersed collections. Such motivations were more common to the framing of digital projects ten years older than S-GA—dating from the early- to mid-2000s (Deegan and Tanner 2002; for a fuller discussion see Punzalan 2013). Driven by the original vision of virtual reunification (largely without editorial intervention), the proposal for the project's first phase centered on imaging of manuscript materials from the partner libraries: "With the digitization of these items, particularly P. B. Shelley's notebooks, the Archive will make an invaluable contribution to Romantic scholarship—bringing together an entire group of widely scattered rare sources" (New York Public Library 2010).

3 Yet, the original plan for the S-GA also showed hallmarks of a field in transition—from one in which simply getting materials online was a central goal to one increasingly focused on allowing users to interact with online materials in additional ways, a trend loosely captured under the banner of "Web 2.0" (O'Reilly 2005). The catalogue of functionalities proposed for S-GA (all based on a prior collaborative project between several of the partners, *The Shakespeare Quartos Archive*) speaks to a nebulous and evolving conception of how those outside the project team might interact with this kind of digital scholarship. Users would have, the proposal declared, "capacity to collate texts, the

ability to overlay images of the original printed editions, compare images side by side, search full-text, and tag text with user annotations" [*sic*] (New York Public Library 2010). By the time work commenced in late 2011, the encoding of materials from the archive in TEI had been promoted to a more significant component of the project alongside digitization, while other potential features were deferred.

4 The decision to invest in TEI encoding of the S-GA materials along with a reading environment designed to take advantage of features of these textual data has allowed the project to remain current with developments in the digital humanities and scholarly editing not envisioned in the original proposal. Through work on the text encoding schema and reading environment, the S-GA project has refined its conception of how this type of scholarship can create new knowledge. Rather than a grab-bag of "Web 2.0" functionalities, engagement with text encoding, an interpretive method created by and specific to the digital humanities, forms the backbone of the S-GA project's work with these important materials.

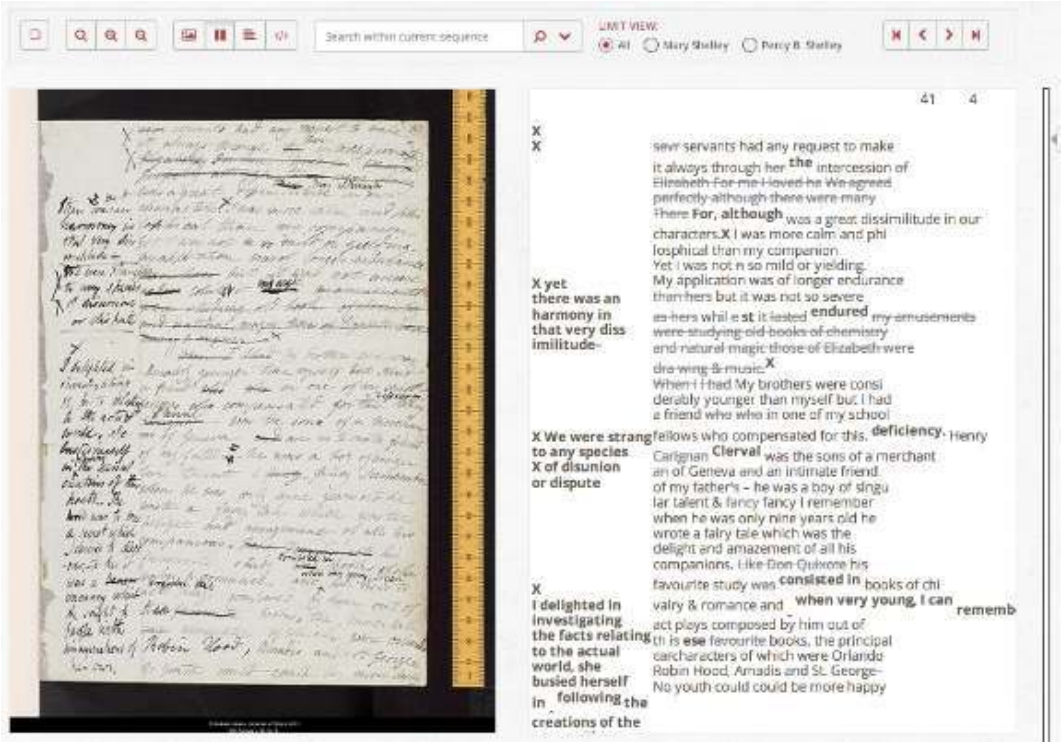
2. Motivations for the S-GA Encoding Scheme

5 The start of text encoding work on the S-GA coincided with the addition of the new "document-focused" elements to the TEI in the release of P5 version 2.0.1. These additions were the product of recent efforts by a subgroup of the TEI Manuscript Special Interest Group and have resulted in a considerable expansion of "Representation of Primary Sources" (chapter 11 of the TEI Guidelines). In the ontology that the new elements are intended to help express, digital text is encoded by describing the relationship of written traces to their physical carriers (TEI Consortium 2011). This encoding approach switches focus from text as communicative act or linguistic content to text as sign on some physical support; for the purposes of this paper this approach will be referred to as "document-focused" encoding. That is, encoders may formalize their understanding of how the text takes form on a surface; they will, for example, identify zones that group text topographically and the lines of text within such zones. Encoders may moreover track an author's actions on the page, identify textual revisions, movements, and deletions, and assert a temporal order for such actions. Describing how a text has been inscribed on a surface is an essential preoccupation of scholars who practice genetic textual criticism. The document-focused encoding strategy is closely but not exclusively identified with the interpretive goals of genetic criticism

throughout this discussion. Divergences from a strict genetic editing approach will be described below. The document-focused approach contrasts with an ontology of text in which encoders describe, through combinations of elements and attributes, what a certain portion of text "is." By surrounding characters with a <p> element, for example, an encoder formally asserts that those characters form a *paragraph*. This approach can quickly become more and more complex, depending on both the text and the encoder's research agenda. Herein this approach will be referred to as "text-focused" encoding (cf. Rehbein and Gabler 2013).

6 Given that the majority of materials in the Shelley-Godwin Archive consist of autograph manuscripts, the editorial team quickly adopted several of the new elements proposed by the manuscript working group such as <sourceDoc>, <zone>, and <line> (with their greatly restricted content models) into its TEI customization. This document-focused approach has served the project well. It yields an encoding scheme that targets features of greatest interest to the scholarly editors who make up the initial user community (the principal investigator and collaborators). Also, focus on documents permits rigorous description of often complicated sets of additions, deletions, and emendations. In the case of the manuscript notebooks of *Frankenstein*, the encoding scheme for S-GA identifies each page as a <surface> containing one or more <zone>s.[2] Most pages have a main body of writing as well as a wide left margin, where Shelley's husband Percy wrote annotations and revisions to the developing text. These separate writing areas are encoded as <zone> elements containing writing organized into <line> elements. On top of this basic model, information about authorial hands (who wrote what) is encoded as well as revisions—including deletions, additions, substitutions, and transposed and retraced text. The reading environment created from this encoding is used to publish a semi-diplomatic transcription of the text alongside a facsimile image.

Figure 1. A screenshot of the primary reading interface of the Shelley-Godwin Archive.



7 As the discussion above suggests, the S-GA encoding scheme borrows concepts and terminology from genetic editing where the motivations align; namely in the representation of the *mise-en-page* of writing on the manuscripts. However, the SG-A does not seek to produce genetic editions of the works represented in the archive. Digital genetic editions usually emphasize the temporal sequence of authorial revision and make a stronger distinction between what is on the page and what is interpreted from the page. For example, the *Digitale Faustedition* project—which directly contributed to the expansion of the chapter of the TEI Guidelines on representation of primary sources—made this distinction by using two different encoding models for the same documents. The models correspond to the concepts of "record" and "interpretation" (*Befund* and *Deutung*) first introduced in 1971 by Hans Zeller (cited in Brüning, Henzel, and Pravida 2013). The "record" (*Befund*) consists of information about a primary source, of which the editor can make a detailed diplomatic transcription as the record of what is "found" on the source.[3] For example, some text that has been struck through may only be encoded as being struck through (i.e., not deleted). The "interpretation" (*Deutung*) records an editor's understanding of a writing act on the

page; for example, some struck-through text is interpreted as a deletion. The *Digitale Faustedition* project sees this interpretation as belonging to a linguistic domain; therefore, it conflates the marking of "deletion" with a more traditional text-focused encoding that includes linguistic and literary structures such as paragraphs and verses.

8 The S-GA, not being a genetic edition, conflates in one encoding model the transcriptional and editorial work more closely related to the document (roughly corresponding to *Befund* and *Deutung*). This conflation is not an accidental failure to conform to one or another ontology of text. The task of developing an encoding scheme to match the goals of the S-GA project pushed the editorial team to consciously borrow tools from both interpretive communities. That is, rather than seeking to produce an established edition or publication, the S-GA project has chosen to embrace the ambiguous nature of this type of digital humanities work (Price 2009) in pursuit of the goal of constructing the S-GA as a work site wherein the encoded text, along with its tailor-made reading environment, "operationalizes" certain aspects of editorial theory. The encoding scheme of the S-GA, drawing on a document-focused approach, is a formal model for operationalizing literary knowledge about how texts are constructed from written documents (Moretti 2013). Franco Moretti defines "operationalizing" as "the process whereby concepts are transformed into a series of operations … . Operationalizing means building a bridge from concepts to measurement, and then to the world" (103–4). In Moretti's case measurement involves quantification of features of literary texts, but measurement need not imply only quantification. In the case of S-GA, the encoding scheme is a formal model by which information about the *mise-en-page* of the various writing traces helps develop greater knowledge about the literary work *Frankenstein*.

9 To achieve this goal, the editorial team needed to be able to produce two distinct representations of the S-GA materials so as to provide rigorous, semi-diplomatic transcriptions of the fragile manuscripts for those with an interest in the compositional practices of a significant group of British Romantic authors, and also to make available clear "reading texts" for those who are primarily interested in the final state of each manuscript. Thus, what was needed was not only the powerful new formalization of document-focused encoding but also a mechanism to enable movement back and forth between document-focused and text-focused models of the materials in the Archive. The development of the document-focused encoding scheme has been described above. The work of automating the production of usable "reading texts" encoded in text-

focused TEI markup from data that is modeled according to a document-focused approach proved much more challenging. Conversion and interchange between these two models poses challenges for encoding practice and workflow, for data provenance and maintainability, and for reading environment and presentation.

3. Encoding Workflow Challenges

10 The conflict between representing multiple hierarchies of content objects and the affordances of XML is well known, and the TEI Guidelines as well as the professional literature of the text encoding community discuss several possible solutions (TEI Consortium 2011; Renear, Mylonas, and Durand 1993; Roland 2003; Piez 2013). One of these solutions is to designate a primary hierarchy and to represent additional hierarchies with empty milestone elements that can be used by some processing software to construct an alternate representation of the textual object. The approach taken by the S-GA team to produce both document-focused and text-focused TEI data is a version of the milestone-based approach. The document-focused elements form the principal hierarchy while milestone elements are supplied to support automatic conversion to text-focused markup (which will contain elements such as `<div>`, `<p>`, `<lineGrp>`, etc.).

11 This solution places increased burden on document encoders to maintain "correctness," thus potentially lowering data consistency and quality. For instance, empty element milestones representing the beginning and ending of textual features have no formal linkages as part of the document-focused document tree. Encoders must supply identifiers and pointers to indicate these linkages. Ensuring that these identifiers and pointers pair correctly must be accomplished with some mechanism other than the RELAX NG validation that checks conformance to the rules specified in the TEI schema. In S-GA this is partly addressed by a number of Schematron rules added to our TEI schema customization. These further checks, however, add an additional step within the processing workflow that must be balanced against the need for a simpler and efficient encoding workflow. As noted above, managing multiple hierarchies through the use of milestones is not new. The experience of the S-GA team suggests that the new possibilities available through the increased expressiveness of the TEI Guidelines, which include the additional document-focused

elements, also increase the scope for projects to produce data that reflect two divergent ontologies, and thus to encounter the difficulties involved in the "workarounds" for multiple hierarchies more frequently.

4. Maintainability and Provenance Challenges

12 In addition to posing challenges for maintaining workflows with good quality control while producing data, use of the milestone strategy for multiple hierarchies (ontologies) decreases the reusability of the textual data produced. The project relies on an automatic process to convert the document-focused encoding into a text-focused one. This process consists of a set of XSLT transformations authored by Wendell Piez, who served as a consultant to the S-GA project in 2013. These transformations are structured as a pipeline—progressively remodeling the document-focused TEI data to a more familiar text-focused TEI.[4] Some of the stages involved in this process include, for example, identifying chapter boundaries that span across multiple `<surface>`s (which for convenience are maintained in separate files), and then combining the content of these surfaces into a single `<div type="chapter">`. While some transformations can be handled heuristically, others require a "hint" for the processor. To support this automated conversion, the S-GA team needed to go beyond purpose-built milestone elements like `<delSpan>` and `<addSpan>` and, in effect, semantically overload the general purpose `<milestone>` element using attributes. The value of an attribute on `<milestone>` indicates which text-focused element is intended to appear in a particular location:[5]

```
<zone>
 <!-- previous lines -->
 <line>
  <del rend="strikethrough">also</del>
  <metamark> [</metamark>
  <mod>
   <del rend="overwritten">w</del>

   <milestone unit="tei:p" spanTo="#mp1"/>

   <add place="intralinear">W</add>
  </mod>
  hen I was about
  <mod>
   <del rend="strikethrough">twelve</del>
   <add place="superlinear">fourteen</add>
  </mod> years
 </line>

 <line>old we were at our house near</line>
 <!-- more lines -->

 <anchor xml:id="mp1"/>

</zone>
```

13 This solution is explained in the project's documentation, and the convention used would be (one hopes) evident after cursory examination of the data. Nonetheless, the desire to make available two models of the text forced the S-GA team to add markup to the project's canonical document-focused data. This makes the encoding more unique to the S-GA project and less easily consumable by future users with different goals.

14 To avoid the conceptual and technical challenges involved in automating the transformation between text-focused and document-focused representations, the two sets of data could each have been created by hand (rather than automatically generated) and maintained separately. Indeed, this is the approach followed by the *Digitale Faustedition* project, where a distinction between what the project calls "documentary" and "textual" transcription was considered necessary not only

as a reaction to encoding problems, but also as a practical application of theoretical distinctions between documentary record and editorial interpretation (Brüning, Henzel, and Pravida 2013). The *Faustedition* project team, however, still encountered technical challenges when trying to correlate and align these two transcriptions automatically. Use of collation and natural language processing tools helped with this problem, but eventually more manual intervention was needed.

15 The S-GA team felt that maintaining two data sets representing different aspects of the textual objects would have led to serious data consistency, provenance, and curation problems. As the example of the *Faustedition* project shows, separate representations must be kept in sync with project-specific workflows developed for this purpose. In the case of S-GA, documentary transcription is the main focus; the greatly increased cost and time involved in also maintaining a textual transcription would have reduced the size of the corpus that could be encoded and thus the amount of materials from the archive that could be made fully available under the initial phase of the project. These exigencies prompted the project's attempts to automate the generation of text-focused TEI data from the core document-focused data that the project editors were creating.

5. Presentation Challenges

16 The display and presentation of document-focused encoding is another technical challenge introduced by the new TEI elements. Rendering a diplomatic transcription is more easily achievable in a coordinate-based system; the S-GA project, therefore, adopted SharedCanvas, a data model developed by Stanford University and a coalition of partners, which allows editors (and potentially future users) to construct views out of linked data annotations. Such annotations, expressed in the Open Annotation vocabulary, relate images, text, and other resources to an abstract "canvas." S-GA is developing and deploying a viewer for SharedCanvas that uses HTML5 technologies to display document-focused TEI elements that are mapped as annotations to a SharedCanvas manifest, a Linked Open data graph that ties all the SharedCanvas components together. The encoding scheme does not record position coordinates for every zone and line, but the positions of main zones to be painted on a canvas are automatically inferred, and base HTML display rules govern the rendering of text within these zones.

17 SharedCanvas not only provides S-GA with a framework to publish TEI transcriptions, but also enables the Archive to move further toward a sustainable participatory infrastructure. The SharedCanvas model allows for further layers of annotations to be added dynamically to a manifest; the S-GA already makes use of this for appending search result highlights to the linked data graph and for displaying these to the user. Eventually, the project aims to use the same mechanism to enable user comments and annotations. The engagement of students and other scholars will be driven by the possibility of creating annotations in the Open Annotation format, so that any SharedCanvas viewer will be able to render them. It remains a matter for the future development of the project to understand whether annotations can be added dynamically to the source TEI—especially those pertaining to transcription and editorial statements—or whether these secondary annotations, created after the main encoding of documents is complete, should always be managed separately from the source TEI data.

6. The Need for Interchange Between Document-Focused and Text-Focused Models

18 There is intellectual power and utility in both document-focused and text-focused approaches to creating digital texts; the scholarly community has gained by the increased expressiveness of maintaining two ontologies within the standard governed by the TEI community. There are also intellectual and practical reasons why it is undesirable to maintain data reflecting these two models as separate or severable representations. The experience of the S-GA editorial team lends support to Peter Robinson's claim that "document, text and work exist in a continuum, and [that] the questions of intention, agency, authority, and meaning exert pressure at every level of reading" (2013, 114). The ability to move along this continuum is an affordance that a digital text should support because it is in the alternation between these two ontological models that editors enact the construction of their particular form of humanistic knowledge.

19 The motility inherent in this model of digital text projects creates significant pressure to address the problem of "interchange" between and among textual data modeled according to different schemes. Syd Bauman has provided a valuable operational definition of interchange in the context of text encoding. Following Bauman's argument, "interchangeable" data is that for which some human intervention (changing data to suit a new system or modifying a system to process new

data) is required but for which this intervention can be accomplished without direct human communication with the data originator—because the data is in some way standard or documented (Bauman 2011). Indeed what the S-GA team has pursued is a strategy for staying document-focused in terms of data creation while preserving the ability to produce and share text-focused encoded data by specifying the appropriate semantics within the more widely used text-focused ontology of digital text and developing data-transformation pipelines that use those semantics as a guide.

20 For Bauman it seems that interchange ("blind interchange" as he delineates it) represents a best compromise between interoperability (full equality of semantics across different text models) and the liberty or expressiveness that motivates scholarly text encoding in the first place. This form of interchange is made possible by "a lot of adherence to standards" and extensive documentation of deviation from those standards (see also Flanders 2009). This argument is deployed against skeptics of the value of markup or advocates of other approaches to curation of digital textual data. The case of the two ontologies of text now co-existing within the TEI Guidelines is somewhat different—both are part of the TEI standard. The need for interchange between data modeled according to these different approaches is real and urgent for a project such as the Shelley-Godwin Archive, which will increasingly depend on the ability to flip back and forth between different representations of the data most relevant to different communities seeking to use the Archive as a site for their own knowledge-making.

21 The debates around if and how TEI is a usable standard for interchanging scholarly information about texts are by now very old. The dilemma faced by the S-GA in attempting to create specialized document-focused data but also to generate and share text-focused data—all of it "TEI data"—suggests a complication of and a possible extension to Bauman's conclusions about interchange. Bauman's argument is still couched in terms of polarity even as it suggests a "common-sense" relaxation of intensity toward the whole question of TEI's suitability as a standard, a kind of lowering of expectations from interoperability to interchange. Geoffrey C. Bowker and Susan Leigh Star suggest that "standardization has been one of the common solutions" to the problem of "how objects can inhabit multiple contexts at once, and have both local and shared meaning" but that the vocabulary of standardization is insufficient "to characterize the heterogeneity and the processual

nature of information ecologies" (Bowker and Star 1999, 293). The more nuanced conception of the issues and interactions involved in the types of problems that Bowker and Star develop could be useful to the TEI community.

22 Following earlier scholars in the domain of social studies of science, Bowker and Star describe a process of balancing "local constraints, received standardized applications, and the re-representation of information" (1999, 292). This could easily be describing the process of developing a TEI project. When these arrangements become "ongoing stable relationship[s] between different social worlds and ... shared objects are built across community boundaries," Bowker and Star refer to the results as "boundary objects" (1999, 292). Thus Bauman's description of "interchange" around the standard of the TEI above, and Star's assertion that "boundary objects are a sort of arrangement that allow different groups to work together without consensus" (Star 2010, 602), seem well aligned. In this discussion, the TEI encoding standard is the boundary object: a "set of work arrangements that are at once material and processual ... resid[ing] between social worlds (or communities of practice) where [this object] is ill structured" (604). Star observes that multiple groups take advantage of the "interpretive flexibility" of boundary objects and customize them for local purposes. In this sense, the S-GA's use of the TEI's formal mechanisms to produce a custom schema and the workflow-driven introduction of project-specific markup and markup conventions (overloading milestones) discussed earlier are hallmarks of work with boundary objects.

23 Yet, according to Star, a less-studied dynamic in the use of boundary objects is the way "groups that are cooperating without consensus tack back-and-forth between [local and shared] forms of the object" (605). This is where the experience of S-GA becomes particularly relevant. By seeking to maintain at the level of the data model the kind of flexibility that Peter Robinson sought to achieve through processing and presentation (Robinson 2009), the encoding scheme that S-GA has developed for itself and the automatic conversion processes that operate on it enact the tacking-back-and-forth that Star describes. The implications for the wider TEI community reside in the linkage Star articulates between boundary objects and the development—and, it would seem to follow, maintenance—of infrastructures and standards.

24 According to Star and her collaborators, infrastructures and standards are a "scal[ing] up" from the back-and-forth use of boundary objects (2010, 605). In this sense, the introduction of a document-focused ontology within the TEI alongside the more common text-focused approach is an opportunity as well as a challenge. Increased commitment to one or the other ontological model of text increases the difficulty that other interpretive communities will face in adapting the digital text to their local meanings and practices. The process of developing the S-GA exposed challenges related to workflow and quality control, maintainability, and presentation. Yet, the concept of boundary objects and their extension into infrastructures and standards provides a framework for articulating the value of constructing a digital text object that spans current boundaries in editorial theory and practice. Digital texts that span the interpretive communities of different schools within textual editing and literary scholarship, by applying pressure to notions of interchange, promote circulation within the system of the standard that contributes to its greater health.

BIBLIOGRAPHY

Bauman, Syd. 2011. "Interchange vs. Interoperability" In *Balisage: The Markup Conference 2011 Proceedings*. Montréal, QC. doi:10.4242/BalisageVol7.Bauman01.

Bowker, Geoffrey C., and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=13186.

Brüning, Gerrit, Katrin Henzel, and Dietmar Pravida. 2013. "Multiple Encoding in Genetic Editions: The Case of *Faust*." *Journal of the Text Encoding Initiative* 4. http://jtei.revues.org/697. doi:10.4000/jtei.697.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner, and Jeffrey Schnapp. 2012. "The Social Life of the Digital Humanities." In *Digital_Humanities*, 73–98. Cambridge, MA: MIT Press.

Deegan, Marilyn, and Simon Tanner. 2002. *Digital Futures: Strategies for the Information Age*. New York: Neal-Schuman Publishers.

Flanders, Julia. 2009. "Dissent and Collaboration." Paper presented at Digital Humanities 2009, College Park, MD, June 22–25.

Moretti, Franco. 2013. "'Operationalizing': Or, the Function of Measurement in Literary Theory." *New Left Review* 84: 103–19.

New York Public Library. 2010. “Application to the National Endowment for the Humanities (NEH) Humanities Collections and Reference Resources: Shelley-Godwin Archive.” https://securegrants.neh.gov/publicquery/main.aspx?f=1&gn=PW-50939-11.

O’Reilly, Tim. 2005. “What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software.” *O’Reilly Media*. September 5. https://web.archive.org/web/20140204114855/http://oreilly.com/web2/archive/what-is-web-20.html.

Palmer, Carole L. 2004. “Thematic Research Collections.” In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, chap. 24. Oxford: Blackwell. http://www.digitalhumanities.org/companion/.

Pierazzo, Elena. 2011. “A Rationale of Digital Documentary Editions.” *Literary and Linguistic Computing* 26 (3): 463–77. doi:10.1093/llc/fqr033.

Piez, Wendell. 2015. “TEI in LMNL: Implications for Modeling.” *Journal of the Text Encoding Initiative* 8. https://jtei.revues.org/1337. doi:10.4000/jtei.1337.

Price, Kenneth M. 2009. “Edition, Project, Database, Archive, Thematic Research Collection: What’s in a Name?” *Digital Humanities Quarterly* 3 (3). http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html.

Punzalan, Ricardo L. 2013. “Virtual Reunification: Bits and Pieces Gathered Together to Represent the Whole.” PhD diss., University of Michigan. http://hdl.handle.net/2027.42/97878.

Rehbein, Malte, and Hans Walter Gabler. 2013. “On Reading Environments for Genetic Editions.” *Scholarly and Research Communication* 4 (3). http://src-online.ca/index.php/src/article/view/123.

Renear, Allen H., Elli Mylonas, and David Durand. 1996. “Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies.” Final version, January 6. In *Research in Humanities Computing: Selected Papers from the ALLC/ACH Conference*. Oxford; New York: Clarendon Press; Oxford University Press. http://hdl.handle.net/2142/9407.

Robinson, Peter. 2009. “What Text Really Is Not, and Why Editors Have to Learn to Swim.” *Literary and Linguistic Computing* 24 (1): 41–52. doi:10.1093/llc/fqn030.

———. 2013. “Towards a Theory of Digital Editions.” *Variants* 10: 105–27.

Roland, Perry. 2003. “Design Patterns in XML Music Representation.” In *ISMIR Conference Proceedings 2003*. http://jhir.library.jhu.edu/handle/1774.2/50.

Star, Susan Leigh. 2010. “This Is Not a Boundary Object: Reflections on the Origin of a Concept.” *Science, Technology & Human Values* 35 (5): 601–17. doi:10.1177/0162243910377624.

TEI Consortium. 2011. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.0.1. Last updated December 22. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/2.0.1/doc/tei-p5-doc/en/html/.

NOTES

1 The authors joined the S-GA Project in 2011 and 2013 respectively.

2 As the TEI Guidelines note, `<surface>` may represent an opening in a codex within a single element. However, since the original items in this case have been disbound for conservation treatment, `<surface>` always refers to a single recto or verso.

3 Some interpretation is always involved in determining what is important among the facts that can be "found" in the document. If anything, dealing with *Befund* in a digital context poses new complications, because digital diplomatic transcriptions are less bound by the limits of paper-based reproduction. Elena Pierazzo (2011, 463) has pointed out that "[f]or print the choice of which features to include in [a] transcription is limited largely by the limits of the publishing technology. In contrast, the digital medium has proved to be much more permissive and so editors need new scholarly guidelines to establish 'where to stop.'" The challenges in establishing clear expectations for a digital diplomatic transcription reveal how the *Befund* is very much subject to editorial interpretation: editors still need to choose what to record from among what is evident from the source.

4 The automated transformation workflow was originally managed using XProc to compose the various stylesheets but was converted to an Apache Cocoon block for ease of maintenance by project staff at MITH.

5 This solution is loosely analogous to some of the ways custom data attributes are intended to be used under the HTML5 Specification: http://www.w3.org/html/wg/drafts/html/CR/dom.html#embedding-custom-non-visible-data-with-the-data-*-attributes.

ABSTRACT

The introduction in 2011 of additional "document-focused" (as opposed to "text-focused") elements represents a significant additional commitment to modeling two distinct ontologies for textual data within the standard governed by the Text Encoding Initiative (TEI) Guidelines. A brief review of projects using the new elements suggests that scholars generally treat the "document-focused" and "text-focused" models as distinct and even severable—the tools of separate interpretive communities within literary

studies. This paper will describe challenges encountered by members of the development and editorial teams of the Shelley-Godwin Archive (S-GA) in attempting to produce TEI-encoded data (as well as an accompanying reading environment) that supports both document-focused and text-focused approaches through automated conversion. Based on the experience of the S-GA teams, the increase in expressiveness achieved through the addition of document-focused elements to the TEI standard also raises the stakes for "interchange" between and among data modeled according to these parallel approaches.

INDEX

Keywords: digital archive, diplomatic transcript, schema design, manuscript encoding, linked data, interchange, Romanticism

AUTHORS

TREVOR MUÑOZ

Trevor Muñoz is assistant dean for digital humanities research at the University of Maryland Libraries and an associate director of the Maryland Institute for Technology (MITH). He works on developing digital research projects and services at the intersection of digital humanities centers and libraries. Trevor's research interests include electronic publishing and the curation and preservation of digital humanities research data.

RAFFAELE VIGLIANTI

Raffaele Viglianti is a research programmer at the Maryland Institute for Technology in the Humanities (MITH). He is involved in front-end development and TEI-based digital editorial projects. Raffaele holds active roles in both the TEI and MEI communities and his research on digital editing spans both TEI [textual] and MEI [musical] practices.