Chapter 17: Encoding and Analysis, and Encoding as Analysis, in Textual Editing

17 Encoding and Analysis, and Encoding as Analysis, in Textual Editing

Christopher Ohge and Charlotte Tupman

17.1 Introduction: Textual scholarship, encoding, and new modes of reading

Regardless of whether one views editorial work as merely mechanical or as heroic scholarly work, it is undeniable that it is a useful means of preserving and organising information (and of course it *is* heroic).[1] The word 'editor' comes from *editio*, which in its original sense could also denote a representation or exhibition, not merely the publishing of a work. Engaging in editing means entering into a long-standing tradition of textual exhibition in which an edition is a product of analysis, arguments and decisions as to the significance of the material. Yet even in book-making we are still producing something with dynamic processes, and now with the help of digital methods, the edition can take on new dimensions of exhibition.

Textual editing for digital publication encompasses not only the consensus of practitioners in philology, bibliography and textual scholarship, but also the creation of a computational pipeline that both curates the text and provides a structure for text analysis. The aspect of consensus bespeaks a praxis of digital scholarly editing that should, alongside the necessary theoretical debates about the nature of texts or the proper ways to choose and emend texts, be grounded in a spirit of pragmatism and thus a consideration of what actually works well in practice for the material under study. Elena Pierazzo rightly pointed out that in the strong and spirited debates between editorial theorists, 'the question has very rarely been which editorial framework was best for the type of document under consideration'.[2] The tendency is to posit what the author may have wanted, or what readers want now. We explore another relevant question: what do the documents require, and for what purpose? Textual editing is about making something—namely, assembling and building a text out of raw materials—and pragmatism provides a framework for making and knowing that allows editors to embrace the messy differences of the ideas within the web of cultural heritage.[3]

Editing also encompasses information management—namely, the guiding of information in a way that is useful to others. The philosopher C. S. Peirce often focused his attention on the dynamic processes whereby research moves toward reducing doubt and fixing belief in the abstract. His contemporary colleague William James invoked the physicist James Clerk Maxwell by asking for the 'particular go' of such research—'true ideas are those that we can assimilate, validate, corroborate and verify'.[4] This concept of the 'particular go' of ideas can be extended to the random detritus of textual production and cultural memory, but also to the curiosity, questions, disagreements and doubts about those things. Under which conditions are we warranted to assert something about a text? What *makes* a digital edition reliable, and what makes it useful to users? Instead of being bogged down in a facts-first mentality about texts qua texts or texts qua works, we examine the surviving documentary evidence, raise our research questions and examine the best applicable methods, and then describe the facts in terms of a workable interface. This is to emphasise an enquiry-first approach, rather than a theoretical one; inductive, rather than prescriptive. A digital edition is not simply for presentational purposes (reading) or for thought experiments; it is also a tool for some use that needs a plan based on the terrain (the textual-documentary landscape) as well as maintenance and rethinking. So in a crucial sense the very activity of textual editing involves a lot of analysis—critical judgments, arguments and principles about how best to build editions that exhibit texts. But how does digital editing take encoding and analysis further, into the realms of computational pattern-recognition and reading that the human brain cannot achieve?

Encoding and analysis, and encoding as analysis: many researchers in digital humanities have practiced encoding and text analysis as separate tasks.[5] This split is unfortunate. Some textual scholars have dismissed text analysis as merely counting, and some text analysis experts disregard the importance of authoritative texts and versions structured in XML. What exactly is text analysis and why include it in an editing project? Text analysis is a computer-assisted calculation of word counts and other statistics (e.g. word and sentence lengths, lexical uniqueness, unique word frequencies, average word use, sentiments, and topics) in a textual corpus. As one preeminent text analyst John F. Burrows suggests, 'the real value of studying the common words rests on the fact that they constitute the underlying fabric of a text, a barely visible web that gives shape to whatever is being said'.[6] The divide between encoding and analysis is evident in the current MLA guidelines for scholarly editions, in which the words 'data mining' or 'text analysis' do not appear as essential criteria.[7] Tara Andrews has made up for that lack by including 'analysis' in her four-part list of desiderata for digital editions.[8] Even so, automation and systemisation to facilitate analysis can reduce the customisation and flexibility that is essential to TEI encoders. We would advocate a notion of analysis that aims to be more expansive, flexible and geared to new ways of reading editions. We encourage text encoding and analysis as complementary activities because of the ways that the analysis can reveal the 'barely visible' aspects of the edition. We see analysis as not just the *how* of encoding, but the *why*, as well as the *what for*—ultimately, the use of the edition's data. This involves both the decisions for robust semantic markup that will facilitate text analysis and data mining, and the text analysis tools themselves that will provide further insight into the edition. This might have particular resonance for certain genres (particularly fiction and poetry) that would be enhanced by analytical and interpretive encoding.[9] An edition that adopts these principles would, in an ideal scenario, allow a user essentially to 'check out' a digital edition and immediately receive statistics on its content that would facilitate reading and exploration.

How an editor does or does not succeed will depend not just on the strategies of encoding for an interface, but also on the extent to which the underlying data can be analysed with computers rather than merely read on the screen. This is not necessarily about the 'front end' of the edition (the digital edition, or web site), but rather the data files that make up the edition: XML files, tables (in .csv or .tsv), dataframes, images, and scripts in XSLT, JSON, JavaScript, Python, R and other languages. Textual editing now includes the building of a computational infrastructure for digital publication as well as facilitating new ways of reading editions that are not possible in analog formats.

The traditions of philology and textual scholarship shape computational methodology, but digital research has also re-oriented traditional editorial workflows. It is not our purpose here to survey the history of philology, bibliography and scholarly editing; foundational guides already exist.[10] Rather, we encourage text encoders to read the history of scholarly editing—the debates about texts, accuracy, the decisions that editors make, the ways in which they present their texts and their scholarship. Is an editor an arbiter or an archivist of the texts? Should editors keep versions of texts intact as they were presented to the public or saved in repositories, or should they create a new text that is more readable, accurate or faithful to the author's wishes? Does a reader enter the edition through a single base text, with a record of variants, or multiple versions of texts, or multiple interfaces? The answers will dictate how one encodes and analyses the edition, and those decisions are best situated within the continuing conversations of editorial practice.

G. Thomas Tanselle says in his 'Varieties of Scholarly Editing' that editing takes part in a tradition of historical research, and therefore editorial decisions stem from the scholar's attention to accuracy and facts.[11] However, Tanselle never adequately defines what he means by history—judging by his writings one could assume that he takes it for granted that history is merely the study of the past. But history is, to echo Yuval Noah Harari, the study of change—and the stories that make change. The textual scholar John Bryant has also argued that all texts are ‘fluid’, and culture is in a constant state of adapting its cultural heritage, echoing Ralph Waldo Emerson’s aphorism in ‘History’ that ‘there is properly no history; only biography’ (as well as, ‘The poets made all the words, and therefore language is the archives of history’, in ‘The Poet’).[12] How best to represent that change is one of the chief tasks of the editor. And, when most people talk about textual scholarship, they usually mean the series of reforms of its predecessor—the field of philology represented at opposite ends by Karl Lachmann and Joseph Bédier—in the late-nineteenth and early-twentieth centuries, in which analytical bibliography and so-called ‘copy-text’ editing made scholarly editing seem more objective. These reforms, often grouped under the term New Bibliography, which were undertaken in a positivist intellectual era by luminaries in the field of early modern literature such as R. B. McKerrow, A. W. Pollard, W. W. Greg and Fredson Bowers, culminated in eclectic text editions, hybrid products of multiple versions put together by experts in the subject.[13]

As David Greetham has suggested, the critical editors of New Bibliography illustrated another iteration of a debate about the meanings of terms such as ‘authoritative’ as well as its constituents ‘form’, ‘intention’, ‘material’ and so on. ‘The history of textual scholarship can therefore be seen as a series of arguments—often resulting in intellectual and scholarly and personal conflicts, even feuds—over the meaning and significance of its most important terms, from the classical period to the electronic environment of the twenty-first century’, Greetham concludes.[14] Moreover, he shows that transcription and publication technologies have been useful to editors since classical times. Recent textual criticism has responded to the Greg-Bowers-Tanselle school of thought with ideas of textual authority that were indebted to mid- to late-twentieth-century literary criticism—particularly the French critique génétique—that gainsaid the seemingly obstinate fixity of the New Bibliographer’s ideas of ‘intention’ and ‘work’. Various ripostes might be classified as document-oriented (in the case of genetic text editing, fluid text editing) as well as reader-oriented (social text editing). A particularly useful feature of recent textual scholarship came from D. F. McKenzie and Jerome McGann’s work examining the various social processes and ‘bibliographic codes’ before and after the publication of works, processes which undermine the assumptions of textual authority and intention that are central to New Bibliography.[15] Many of these editorial practices have developed very neatly in a digital context. For example, the social text theory elucidated by Ray Siemens suggests that editing involves ‘communities of practice’, which can be better realised in a digital environment.[16] It has not been sufficiently shown, however, how workable some of these new practices might actually be, particularly in an era when so many communities of practice are still siloed in academia and often working in precarious roles.

At the other end of the spectrum is a worry about how the field of digital textual scholarship has been dominated by a commitment to ideologies of post-modern reader-response and deconstructionist theories that deny the truthfulness of claims to accuracy, textual stability and authorial intention. In ‘The Death of the Editor', J. Stephen Murphy suggests that recent editorial theorising has adopted the logic of Roland Barthes's 'Death of the Author' (i.e., that the unity of textual meaning lies not in the text itself but in its destination with the reader). Also, despite the reinvigoration of textual studies by many digital humanists, Murphy points out that their 'rhetoric may have been self-defeating' because it 'represents editors as antiquated, logocentric bullies opposed to readerly freedom and textual play'.[17] On one side is a theoretical attachment to 'the work' as an abstract reaching after the perfectly constructed eclectic text, and on the other is a relativism with regard to textual matters. Another way lies in between, particularly in the pragmatic ethos of digital humanities, to echo Christopher Ricks, to use 'hard thinking [that] is resolutely unelaborated beyond the exposition and application of principles'.[18] The important point is to take the debates seriously, and to make a principled judgment as to the most appropriate way to create a workable edition from the materials at hand.

The burgeoning digital methods evident in this Companion open up more avenues for debate in textual scholarship, but to evoke Ricks again, a fair amount of hard thought should be applied to editorial *and* computational methods; this means thinking about why the edition is being built, for what research purposes and how to make that research data accessible. Editing, according to A. E. Housman's concise definition, is the 'science of discovering errors in texts, and the art of removing them'.[19] Besides the apt reminder always to exercise doubt, Housman's lasting contribution was to emphasise the role of individual judgment in editing. Couple this, too, with Samuel Johnson's elegant maxim that in editing, a scholar is 'to correct what is corrupt, and to explain what is obscure'—but also to keep in mind, 'The editor, though he may less delight his own vanity, will probably please his reader more, by supposing him equally able with himself to judge of beauties and faults, which require no previous acquisition of remote knowledge'.[20] It is a constant negotiation for scholarly editors to decide the extent to which they should make themselves known by intervening in the construction and explanation of the contents of the edition, and it is now possible for that negotiation to be openly documented.

How has digital research re-oriented traditional editorial workflows? Text editing in print and digital involves an act of data modeling, which is based on the epistemological grounding of the editors. Pierazzo has posed an important question about whether the methodologies of editing can 'be pursued digitally or does the digital medium necessarily provide a new theoretical framework?'[21] This sets up a distinction between implementing old methods (and outputs) of editions versus creating a brand new methodology. While she leaves the question open, we argue here that old methods are informing digital editions at the same time as computational methods are enhancing what might be called the 'old' methods. Daniels and Thistlethwaite have asserted that 'Digital technologies have radically altered the traditional structure of habits in the scholarly workflow.'[22] All editors are output-driven, but some spend more time on analytical features. Others are more interested in presentation (reading texts). Also, even though 'user-driven' digital editions are a laudable goal, the purpose of the edition is still to employ the expertise of a textual scholar for the benefit of users. Even if we acknowledge that the 'user' is an abstraction that can never be reconciled with the myriad desires of scholars, students and general readers who use editions, it is still a pragmatic concept that grounds traditional approaches to the design of any digital object or software. Pierazzo is right to suggest that the workflow should be a primary concern in digital editing, along with producing the right type of edition as dictated by the material. The principles of the editors are also crucial (and are not always stated as often or as clearly as they might be). What is required in this rapidly advancing ecosystem of digital scholarship is a methodology grounded in pragmatist epistemology, one that seeks to use computers to make texts work better for research questions, and one that uses success in action to balance textual scholarship and the creation of a computational pipeline.[23]

The building of a computational pipeline should include decisions about the accessibility of data: a digital editor should consider whether their XML data is amenable to the analysis of researchers from outside the project. Tim Berners-Lee proposed five levels of open data specifications:[24]

- Available on the web (whatever format) *but with an open licence, to be Open Data*
- Available as machine-readable structured data (e.g. Excel instead of image scan of a table)
- As (2), plus non-proprietary format (e.g. CSV instead of Excel)
- All the above, plus: Use open standards from W3C (RDF and Sparql) to identify things, so that people can point at your stuff
- All the above, plus: Link your data to other people's to provide context.

But data is not quite 'open' if a digital researcher needs a complicated specialist's manual to figure out how to analyse the layers that make up the edition. In other words, 'structured' data does not necessarily entail easily analysable data. What is missing in Berners-Lee's list of accessibility is analysability.

Berners-Lee's ideals can be enhanced with the desiderata set forth by Julia Flanders and Neil Fraistat,[25] that digital editions should be:

1. **interoperable** with each other and with other texts using professional standards such as the TEI;
2. **layered and modular,** so that the edition is separate from an interface which allows for redesigning interfaces;
3. **multimodal**, providing analysis of the text but also of other paratextual materials;
4. **dynamic**, encouraging user interaction;

5. **scalable**, allowing for microscopic and macroscopic inquiry;
6. **everted and interconnected**, so that edition data can be used by others;
7. **sustainable**, so the community can access the material.

As much as most scholarly editors would like to see these desires fulfilled, what Flanders and Fraistat suggest is not always practicable for scholars faced with time and resource constraints. It is hard to think of any current digital editions that successfully hit all seven targets, unless they are one of the rare, large-scale digital humanities projects that have attracted seven figures of funding. It is also arguable whether all editions should be dynamic, or whether interface needs to be a concern at all for those who simply want to preserve a small-scale text in an XML repository, which in itself should be considered a genuine service to scholarship. A minimalist—or minimal computing—approach to achieving all seven might be a laudable goal, but we would argue that the third point, that of multimodality (possibly combined with scalability and usability and interconnectedness), should be a primary concern to all editors, for it is an aspect of digital editing that is truly innovative in a way that printed books cannot be.

Constraint is an important factor that must be addressed in encoding, analysis, and workflow.[26] If you are a beginner to digital editions, it would be a mistake to think of the encoding as an encyclopedic markup enterprise. It can be intimidating to see the 23 Modules of the TEI Guidelines, and to wonder how one could understand all the possibilities that various elements, attributes and values can offer. Yet it is not the purpose of the edition to populate it with as many tags as possible. Rather, one should start by asking questions such as:

- How do the materials cohere with the methods of textual scholarship?
- What is the purpose of this edition?
- What are the arguments about the best way to organize and present the texts?
- What are the arguments for the significance of the text, and the priorities of its audience, and how do these affect our encoding choices?
- What analytical tools could enhance our understanding and demonstrate the significance of the research?

From there one can then construct an analytical model of encoding decisions to guide best practice. Sometimes, in order to accomplish the research aims based on the above analytical questions, a researcher will need to customise TEI or create new elements and attributes. That is entirely legitimate: there is a reason that the TEI Guidelines are not called the TEI Laws. They can be customised depending on the needs of researchers. The analytical dimension will also include some general principles about the goals of text analysis: what new forms of reading can brought to light as a result of the encoding?

A recent debate in textual scholarship between historical (or documentary) and critical ('copy-text') editors illustrates the importance of thinking in terms of encoding as analysis. By historical-documentary editions, we mean the transcription of a source text as exactly as possible; and by critical editions, we mean those that choose the most authoritative base text (often called the copy text) for the edition, the reading text of which will be emended if other authoritative readings exist in other versions (otherwise called 'witnesses'). One of the complaints about digital documentary editions—and their materialist cousins, genetic and versioning editions—is that they only work for a small audience of specialist scholars. Genetic editions aim to show the creative process in surviving documents, whereas versioning texts usually show documentary texts of multiple existing versions, usually in parallel. All of these editions are potentially unreadable or tedious, at worst; most people want to read one clean, accurate text.[27] The problem with that argument is that it is only forceful in the context of normal codex-based human reading processes. A complicated digital versioning text, or a transcription of a heavily revised and uncompleted manuscript, can be 'read' in a novel way with the tools of text analysis and querying. The text analysis can be an adjunct to the normal reading process (the edition should be machine-readable, after all). Moreover, as Bryant and other editors at the Melville Electronic Library have shown, the digital interface can make the reading process of difficult manuscripts smoother than that of their print predecessors by directly engaging the reader with the material context rather than relying on complex genetic symbols (especially when multiple reading interface options are involved).[28] The criticism leveled by Robinson against the limited page-by-page transcription of the digital documentary edition *Jane Austen's Fiction Manuscripts* does not take into account the intellectual value that text analysis could bring to that edition's XML data. His problem that we might 'distance ourselves and our editions from the readers' is actually more of an interface issue than a worry about the usefulness of the edition's data.[29] And let's not forget: humans are not the only readers now; the machines are too, and they can help humans find new information in complicated texts.

How, then, to begin? One simple place to start is to do what Syd Bauman, James Cummings, and Julia Flanders have suggested: create a spreadsheet of the kinds of elements that would be most useful to the project.[30] Researchers should also include a list of research hypotheses and intended analysis outcomes in this spreadsheet. In addition to the research questions that will guide the editing and encoding, one should also consider whether the edition data will be subject to analysis with Voyant Tools, AntConc, Python NLP, or R (or all of these in combination).[31] What kinds of information resulting from text analysis would benefit the research? The spreadsheet should identify at least two things: the set of elements, attributes and values, and the set of analytical aims of the encoding. A more advanced approach would be to implement the information from the spreadsheet into a project ODD file. ODD—which stands for 'one document does it all'—sets the constraints of encoding, along with the set of rules of the proper vocabularies and hierarchies. The Roma JS application, which is under development but available in beta, makes this process even easier to accomplish than it is with the TEI's existing Roma tool for generating customisation files.[32] Most TEI XML projects will only use around 25 or so elements, so there is actually no need to implement a TEI-all document template (which allows for around 500 elements), or to get overwhelmed with identifying a large group of elements for more sophistication.[33] The inclusion of an analytically-driven ODD file constrains the encoding in addition to bolstering the functionality of the finished edition.

17.2 Using EpiDoc to edit classical texts

For those less familiar with TEI and its scope, it is worth noting that there are subsets of the TEI designed specifically to constrain the available elements, attributes and values in order to encode particular types of texts: to take an example, the EpiDoc initiative produces a schema and guidelines for encoding scholarly editions of texts such as inscriptions and papyri (not limited to ancient materials).[34] EpiDoc caters not only for transcription and editorial interventions, but also for describing the object on which a text is written or inscribed, as well as its history. Rather than using the full set of options available within the TEI, the authors of EpiDoc have considered which elements an epigrapher or papyrologist will need to use to produce an edition. They have selected specific attributes and values for these elements in order to make the TEI as useful and relevant as possible to specialists in these areas, and the features that are expressed through EpiDoc are documented in a set of explanatory guidelines designed to help the user through the process of encoding.[35] In addition, there is a supportive community that runs training workshops and a mailing list for further discussion.[36] This approach has proved successful, in that EpiDoc is now considered the standard method for encoding inscriptions and papyri for digital publication and interchange, without being prescriptive about the precise workflow that a project should follow.

It would be worth exploring how best to guide and inform those who plan to undertake such a project. The Women Writers project,[37] led by Julia Flanders at Northeastern University, has produced an extremely useful guide to strategy and workflow for encoding projects.[38] Although it is based on the encoding of early printed books, it serves a much wider set of users in offering a step-by-step guide to concepts, strategies, project management and design, including document analysis, markup, error-checking, post-processing and documentation. It is a valuable guide for anyone thinking of undertaking an encoding project, and the recommendations we make here build upon its foundation. How might appropriate workflows best be designed, from examining source materials to publishing and analysing data that serve a project's scholarly aims? In attempting to define a workflow, we should first establish what aims are being served. Those of the researchers on the project? Those of known groups likely to use the resource? Those of future or potential users with different/wider research questions? Financial practicalities, skill sets and varying availability of personnel throughout the project will also have their own influences on workflow. We make all the suggestions below with these caveats in mind.

To illustrate how an editorial workflow might be designed that caters to text encoding and analysis as complementary activities, we take the example of the work that an epigrapher typically undertakes when editing an inscription, or corpus of inscriptions, for digital publication, and how this might be enhanced to include plans for text analysis.[39] It seems timely to consider this, not least because a number of tools that could enhance such work are now available but are not as widely used as they could be, as noted recently by Bodard and Stoyanova when they remarked that 'We have yet to fully integrate any of this activity into the workflow of the epigrapher or papyrologist [...] and further training in this area would doubtless result in better integration with EpiDoc guidance'.[40]

The traditional editor of an epigraphic text will begin with the reading and transcription of the inscription from a combination of autopsy, photography and perhaps also the making of a 'squeeze' (an impression of the inscription made using squeeze paper, water and a specially made brush with tightly packed bristles designed to ensure an accurate rendering of the inscribed letters). In addition to traditional methods, digital techniques including RTI (Reflectance Transformation Imaging) can be included in the epigrapher's toolkit to enhance the reading of difficult or damaged letters.[41] If the text has been published, previous readings of the inscription will be considered, particularly where these were made when the stone was in better condition and more of the inscription was visible (for instance, before an inscription was damaged or exposed to weathering). Typically the epigrapher might work with a notebook and pencil before transferring their reading of the inscription to digital form.

Those who have been trained in EpiDoc encoding might create a 'born digital' text, entered immediately into a <div type="edition"> with minimal structural markup as a first step. Most epigraphers using traditional methods, however, will initially write up their reading of the text in their preferred text editor. They will record the diplomatic version of the text, i.e. what can be seen on the stone, and then as a separate task will produce an edited version that includes expansions of abbreviations, supplied letters or words, indications of unclear letters and so forth. At this stage the editor will include the 'Leiden conventions', sigla that indicate editorial interventions in the text, and will write an apparatus criticus discussing previous readings or unclear or otherwise notable sections of the text.[42] They will also consider details of the object on which the inscription appears, such as material, measurements and decorative features; previously known locations; dating criteria; letter heights; and bibliography. Finally, to explain the historical context of the inscription and its significance, the editor will write a commentary.

So far, this workflow would be familiar to any epigrapher working today, even if they might approach it in a slightly different order. Anyone considering a digital publication would familiarise themselves with the EpiDoc Guidelines[43] and take advice from the community mailing list[44] or one of the training workshops[45] as to the appropriate markup to structure the text and metadata, and to represent the editorial interventions in the text, but this aspect of the digital workflow is not markedly different from that of the print editorial workflow: indeed, the traditional workflow directly informs and influences the digital. The types of information recorded here are drawn from centuries of consensus about what epigraphic scholarship entails, epitomised by the gargantuan works of the *Corpus Inscriptionum Latinarum* and *Inscriptiones Graecae*.[46]

Where an encoded version of an epigraphic text begins to depart from the work of the traditional epigrapher is in the level of detail that can be achieved in the semantic markup of the edition. This goes beyond solely diplomatic transcription, encompassing more explicit details of editorial interventions as well as the encoding of information about specific entities within the text, and this stage is reflected in the digital workflow: a substantial amount of time is likely to be devoted to the encoding of entities that are identified as being of research interest to the project and/or its users. For instance, this might include encoding information about people or places that appear in the text, and providing links to internal or external authority lists containing further details about those entities, such as biographical details for people and coordinates for places, and possibly also some information about related entities. Whether or not the important entities have been established at the beginning of the project will depend on whether the content of the inscriptions was known before the project began, and often markup requirements will have to be adapted as new data come to light during the course of the work. However, in an ideal situation these entities will have been identified and decided upon before the project begins, so that effort is expended in the appropriate areas of the encoding process. It is always necessary to select the markup carefully: in a world where we could encode almost any aspect of a text and its physical support, the time, funds and available expertise will always be limiting factors (just as they are in non-digital projects). The digital workflow, then, might at this stage include the encoding of features such as places, events, dates, individuals, names, commemorative relationships, age, sex, social status, and occupation, depending on the focus of the project or that of its expected end users.

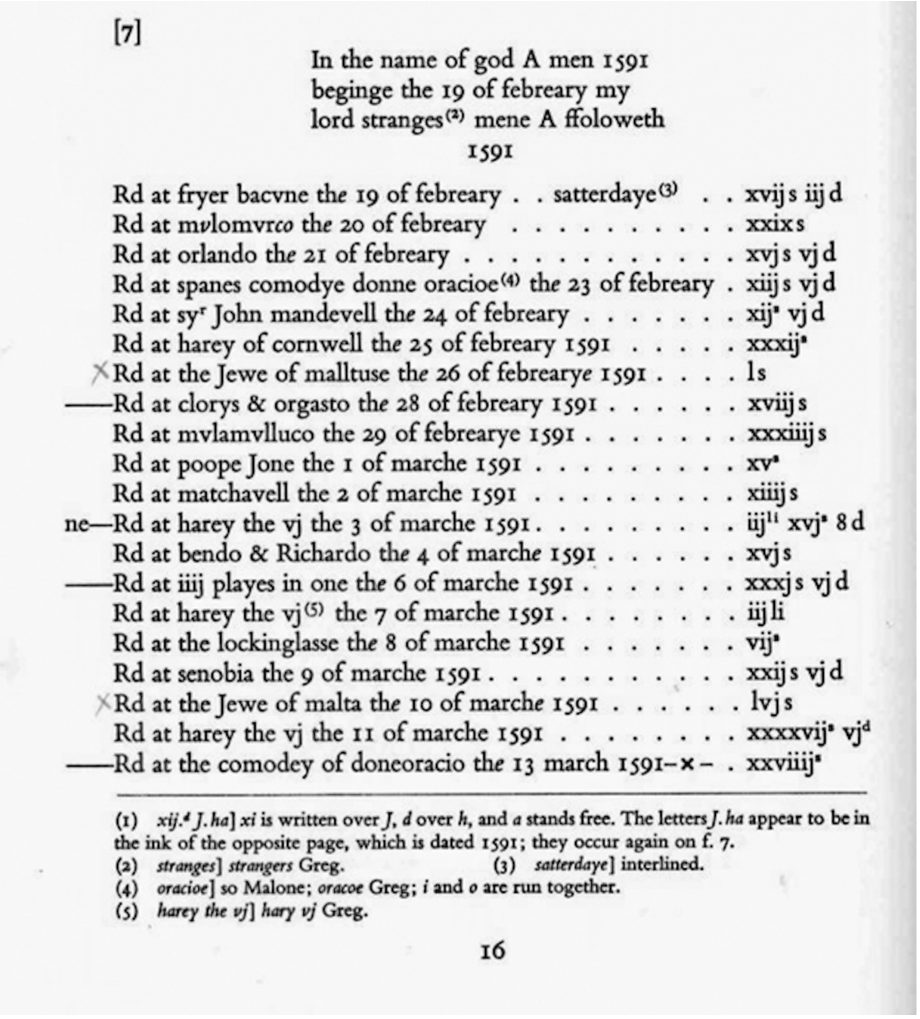
In a digital project, decision-making about the desired indices, tables of content, and other facets is ideally done at the earlier stages, to allow not only for planning the encoding that needs to be done in order to produce these features of the edition, but also to establish what *else* could be done beyond the markup itself. As Bodard and Stoyanova observe, 'the rigorous intellectual effort of indexing in a tradition[al] project is changed in the digital process, but not replaced by an automated process.' Referring to the order in which these skills are taught in EpiDoc workshops, they note that 'this structure follows the workflow of an epigraphic project, where the indices, tables of contents, lists of lemmata etc. are produced at the end of the project from the encoded XML files.'[47] While the generation of indices might still be done at the end of a project, the *thinking* about what indices are needed is best done as early as possible. The same is true for the creation and structuring of internal authority lists (and/or identification of the relevant external authority lists), although inevitably these will be populated as the project progresses.

The presentation of the texts (i.e. the design of the user interface) is likely to involve an iterative process of testing an initial design, receiving feedback and developing the interface further in several stages, whether the project uses the Kiln-based EFES (EpiDoc Front End Services),[48] an eXist-based approach with the EpiDoc stylesheets,[49] or a custom framework. Is the project intending to present users solely with the editors' readings and commentaries of the texts, or seek user annotations or submissions of new readings? If the latter, how much encoding knowledge is assumed on the part of the user? Should the interface allow for a relatively simple means of annotating a text, to encourage a greater level of participation amongst those who have not learnt to encode?

We are in agreement with Bodard and Stoyanova that were EpiDoc encoding skills to be taught alongside Linked Open Data (LOD) and Named Entity Recognition (NER) skills in Python or other languages, we would see the immediate benefits of encoding and text analysis as complementary and connected tasks.[50] In addition we would recommend the inclusion of text analysis tools such as Voyant[51] and AntConc,[52] at least in respect of making participants aware of their potential (even if teaching specific skills in using them is beyond the bounds of that particular workshop). The drive of projects such as Pelagios[53] to produce user-friendly interfaces for the creation of LOD[54] (and, of course, Papyri.info[55] for the creation of EpiDoc encoding 'underneath the hood') means that confidence in at least some of these areas can be developed within the bounds of a relatively short learning time, although this is not a substitute for bringing in the appropriate expertise for a specific project, or for learning the fundamentals oneself, which will always bring a deeper understanding not only of what one is doing but of what might be possible. It is not within the scope of this chapter to debate the relative merits of user-friendly interfaces versus more in-depth training in the fundamentals of encoding and analysis, but what the former provides is the means for analytical tools to be explored by the epigrapher at the planning stages, and decisions made about how the encoding could be designed to facilitate further analysis. In doing so, we should plan for an iterative process, not least because experimentation is an important aspect of text analysis: while as researchers we have our own particular questions and priorities, we will inevitably need to modify or generate new questions as a result of undertaking this work.

17.3 Encoding and analysis in modern English texts

These considerations in epigraphy remain relevant to other time periods. Consider Philip Henslowe's diaries, written between 1592–1609, which detail his financial transactions, as well as fine-grained information about the daily operations of his playhouse. The manuscript is the best surviving source of information of English Renaissance theatre. This vital record has been published in print editions, but it is difficult to read and clearly aimed at specialists who already understand the linguistic and numerical data.[56] The recent addition of digital facsimiles of the original manuscripts has made it possible to create a digital edition, which is currently underway under the direction of Dr. Yuanbo (Edgar) Mao.[57] Any digital edition of a document like Henslowe's diary should be guided by analysis; the diary is a dataset, consisting as it does of rows of data relating to play titles, performance dates, loans and ticket sales. Examining the Foakes and Rickert edition of the following receipts from February 1592 shows the difficulty of following a print edition for this kind of text:



[7]

In the name of god A men 1591
beginge the 19 of febreary my
lord stranges[2] mene A ffoloweth
1591

	Rd at fryer bacvne the 19 of febreary . . satterdaye[3] . .	xvij s iij d
	Rd at mvlomvrco the 20 of febreary	xxix s
	Rd at orlando the 21 of febreary	xvj s vj d
	Rd at spanes comodye donne oracioe[4] the 23 of febreary .	xiij s vj d
	Rd at sy^{r} John mandevell the 24 of febreary	xij^{s} vj d
	Rd at harey of cornwell the 25 of febreary 1591	xxxij^{s}
×	Rd at the Jewe of malltuse the 26 of febrearye 1591	l s
——	Rd at clorys & orgasto the 28 of febreary 1591	xviij s
	Rd at mvlamvlluco the 29 of febrearye 1591	xxxiiij s
	Rd at poope Jone the 1 of marche 1591	xv^{s}
	Rd at matchavell the 2 of marche 1591	xiiij s
ne—	Rd at harey the vj the 3 of marche 1591.	iij^{li} xvj^{s} 8 d
	Rd at bendo & Richardo the 4 of marche 1591	xvj s
——	Rd at iiij playes in one the 6 of marche 1591	xxxj s vj d
	Rd at harey the vj[5] the 7 of marche 1591.	iij li
	Rd at the lockinglasse the 8 of marche 1591	vij^{s}
	Rd at senobia the 9 of marche 1591.	xxij s vj d
×	Rd at the Jewe of malta the 10 of marche 1591	lvj s
	Rd at harey the vj the 11 of marche 1591	xxxxvij^{s} vj^{d}
——	Rd at the comodey of doneoracio the 13 march 1591–×– .	xxviiij^{s}

[1] *xij.^{d} J. ha*] *xi* is written over *J*, *d* over *h*, and *a* stands free. The letters *J. ha* appear to be in the ink of the opposite page, which is dated 1591; they occur again on f. 7.

[2] *stranges*] *strangers* Greg.

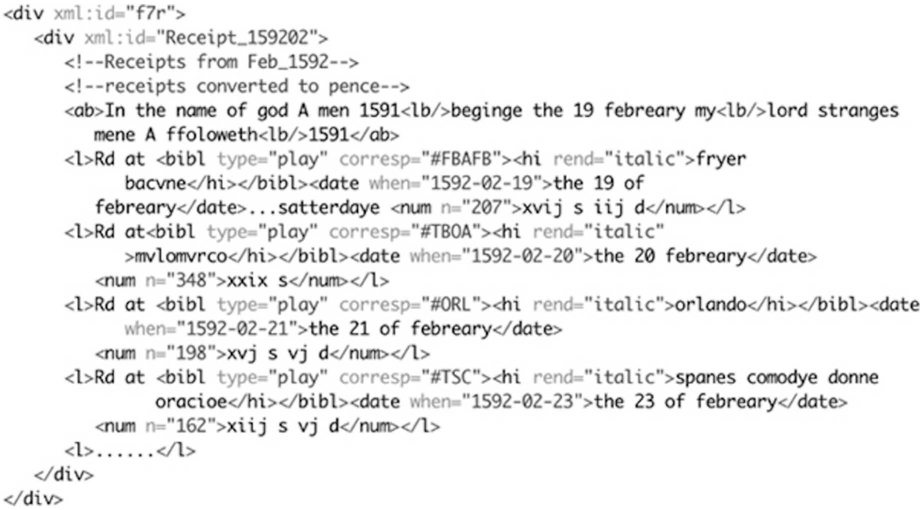
[3] *satterdaye*] interlined.

[4] *oracioe*] so Malone; *oracoe* Greg; *i* and *o* are run together.

[5] *harey the vj*] *hary vj* Greg.

*Figure 17.1*A scan of the February 1592 receipts from the Foakes and Rickert print edition of *Henslowe's Diary*

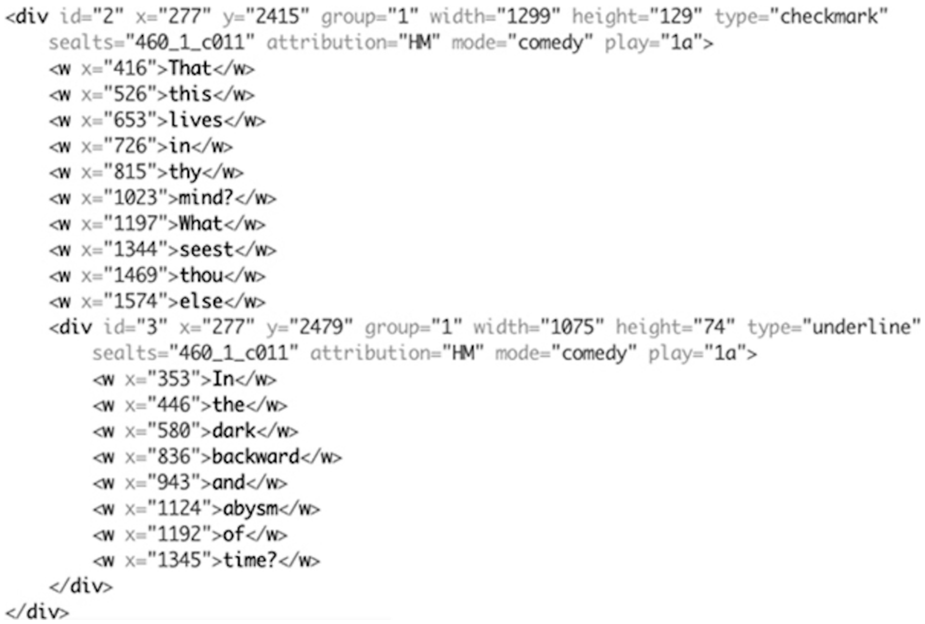
Here is a draft TEI XML snippet of the same receipts from February 1592:



```
<div xml:id="f7r">
   <div xml:id="Receipt_159202">
      <!--Receipts from Feb_1592-->
      <!--receipts converted to pence-->
      <ab>In the name of god A men 1591<lb/>beginge the 19 febreary my<lb/>lord stranges
         mene A ffoloweth<lb/>1591</ab>
      <l>Rd at <bibl type="play" corresp="#FBAFB"><hi rend="italic">fryer
            bacvne</hi></bibl><date when="1592-02-19">the 19 of
         febreary</date>...satterdaye <num n="207">xvij s iij d</num></l>
      <l>Rd at<bibl type="play" corresp="#TBOA"><hi rend="italic"
            >mvlomvrco</hi></bibl><date when="1592-02-20">the 20 febreary</date>
         <num n="348">xxix s</num></l>
      <l>Rd at <bibl type="play" corresp="#ORL"><hi rend="italic">orlando</hi></bibl><date
            when="1592-02-21">the 21 of febreary</date>
         <num n="198">xvj s vj d</num></l>
      <l>Rd at <bibl type="play" corresp="#TSC"><hi rend="italic">spanes comodye donne
               oracioe</hi></bibl><date when="1592-02-23">the 23 of febreary</date>
         <num n="162">xiij s vj d</num></l>
      <l>......</l>
   </div>
</div>
```

Encoding this project with TEI best practice in mind, one can not only structure the entries in the diary with descriptive markup, but also regularise the data for statistical analysis. The @when attributes in <date> elements regularise the data for the purposes of analysis, such that one could now identify all elements from February 1592 even if they are written in a different way. Moreover, the <num> elements and @n attributes exemplify good practice, but they also encapsulate encoding as analysis: while regularising the numerical values, a researcher should also aim to think about how to produce statistical calculations on Henslowe's recordings. That is, how could a digital edition improve the reading experience of the text? What text analysis tools would best suit this project? One could aim to include Python or R scripts for processing mathematical and subject calculations. Analysis of editions in this fashion can perform whole-text as well as node-level data mining.[58]

In bibliography and genetic criticism, marginalia studies stand out as another revealing example of encoding as analysis, particularly at the node level. *Melville's Marginalia Online* (MMO) is a highly functional virtual archive, digital bibliography, and searchable edition of Herman Melville's library that is still very much in progress. The encoding decisions in the initial phase of the project could have followed the TEI, but the aim of the project was to create a searchable database of Melville's markings and annotations that matched the word-level results with their corresponding digital facsimiles. The resulting coordinate-capture XML encoding does just that:

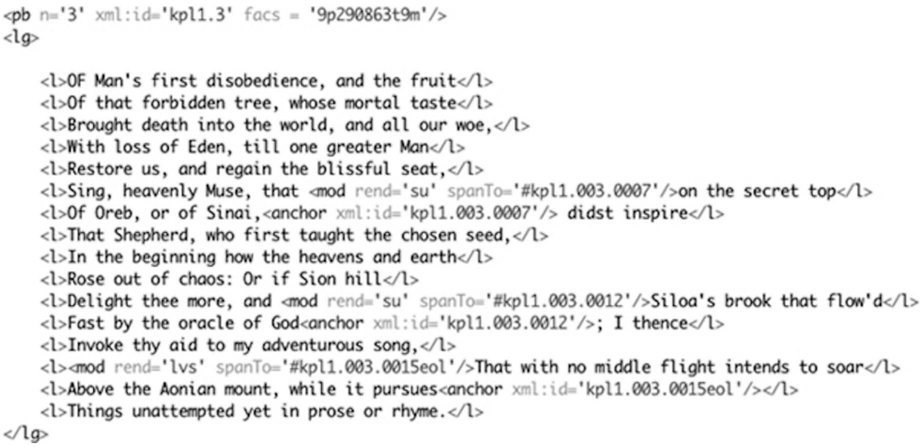


```
<div id="2" x="277" y="2415" group="1" width="1299" height="129" type="checkmark"
    sealts="460_1_c011" attribution="HM" mode="comedy" play="1a">
    <w x="416">That</w>
    <w x="526">this</w>
    <w x="653">lives</w>
    <w x="726">in</w>
    <w x="815">thy</w>
    <w x="1023">mind?</w>
    <w x="1197">What</w>
    <w x="1344">seest</w>
    <w x="1469">thou</w>
    <w x="1574">else</w>
    <div id="3" x="277" y="2479" group="1" width="1075" height="74" type="underline"
        sealts="460_1_c011" attribution="HM" mode="comedy" play="1a">
        <w x="353">In</w>
        <w x="446">the</w>
        <w x="580">dark</w>
        <w x="836">backward</w>
        <w x="943">and</w>
        <w x="1124">abysm</w>
        <w x="1192">of</w>
        <w x="1345">time?</w>
    </div>
</div>
```

This is Melville's first marking (with an embedded additional marking) in *The Tempest*, from his seven-volume set of Shakespeare's plays. Each instance of marginalia is contained within a <div>, which includes several attributes identifying various bibliographic and holographic information. Clearly this is not TEI-compliant, but it is functional as to its purpose, which is to enable word searches of marginalia with corresponding highlighting of search results in the digital facsimile of the page from Melville's book. Of course TEI encoding would make it easier to refine analysis (of, say, marginalia differences between poetry and prose structures), and the project has plans to incorporate stand-off TEI to complement the existing coordinate-capture markup. Yet the fact that each instance of marginalia is encoded with a <div>, and that each <div> has additional attributes (such as the marking @type, the play's @mode, the play's @title, and the @sealts attribute, which identifies bibliographic information as well as the page number in a single value) means that the data is already amenable to text analysis. Also, each word encoded within a <w> allows for fine-grained markings-level and word-level analysis.

Complemented with the plan to encode Melville's marked texts were a series of XSLT and R scripts for performing text analysis on Melville's reading data.[59] With the services of Performant Software, MMO has also started to implement a complementary analysis interface, based on Voyant tools, which shows general statistics of reading data. The fragmentary nature of marginalia makes text analysis even more important as a tool for understanding. XSLT scripts created HTML tables of all the markings that could be sorted by word count as well as bar graphs of total words marked per play and play mode (comedy, history, tragedy). R code adapted from Jockers produce linguistic calculations on the lexical uniqueness of the markings. Other R code adapted from Silge and Robinson create sentiment analyses of Melville's marked content. And Voyant generates word clouds and graphs of most frequent word data. The illustrations and figures resulting from the text analyses illustrate Melville's varying forms of engagement in his readings, bringing into particular relief hitherto unanalyzed and under-appreciated aspects of his marginalia. Word frequencies point the way toward ideas and themes that interested him; lexical uniqueness and word-sentiment values of marked passages shed light on the rhetoric and perspectives to which he gravitated. The visualizations of reading evidence bolster conceptions of the writers that influenced him, including Melville's attending to Shakespeare's profundity in concise, philosophically bleak themes and perspectives. The node-level text analyses showed the value of using text analysis techniques to complement close reading, as we were able to show how his marginalia in Shakespeare inspired passages in *Moby-Dick*. Text analysis, therefore, does not always have to be concerned with large swaths of data; it can also enhance the reading of smaller data sets.

Another digital marginalia project, the Keats Library,[60] has encoded Keats's heavily marked copy of Milton's *Paradise Lost* in TEI, but the encoding encounters overlapping hierarchy problems, therefore requiring a 'Trojan Horse' markup scheme of using empty elements with @spanTo pointers to their corresponding @xml:id.[61] Here is how Dr Dan Johnson (Notre Dame) encoded Keats's marking at the beginning of *Paradise Lost*:



```
<pb n='3' xml:id='kpl1.3' facs = '9p290863t9m'/>
<lg>

    <l>OF Man's first disobedience, and the fruit</l>
    <l>Of that forbidden tree, whose mortal taste</l>
    <l>Brought death into the world, and all our woe,</l>
    <l>With loss of Eden, till one greater Man</l>
    <l>Restore us, and regain the blissful seat,</l>
    <l>Sing, heavenly Muse, that <mod rend='su' spanTo='#kpl1.003.0007'/>on the secret top</l>
    <l>Of Oreb, or of Sinai,<anchor xml:id='kpl1.003.0007'/> didst inspire</l>
    <l>That Shepherd, who first taught the chosen seed,</l>
    <l>In the beginning how the heavens and earth</l>
    <l>Rose out of chaos: Or if Sion hill</l>
    <l>Delight thee more, and <mod rend='su' spanTo='#kpl1.003.0012'/>Siloa's brook that flow'd</l>
    <l>Fast by the oracle of God<anchor xml:id='kpl1.003.0012'/>; I thence</l>
    <l>Invoke thy aid to my adventurous song,</l>
    <l><mod rend='lvs' spanTo='#kpl1.003.0015eol'/>That with no middle flight intends to soar</l>
    <l>Above the Aonian mount, while it pursues<anchor xml:id='kpl1.003.0015eol'/></l>
    <l>Things unattempted yet in prose or rhyme.</l>
</lg>
```

This is a workable solution for encoding marginalia in TEI, but unlike the Melville example, the transcription does not match up with the digital facsimile in the interface.[62] How does the descriptive markup facilitate analysis, and what kinds of text analysis could be accomplished? The marginalia encoding above is an impressive way to deal with the shortcomings of overlapping hierarchical markup. It is also not primarily intended for analysis; like many digital projects, it is designed for front-end viewing (users interacting with the finished interface) than it is for the ability to generate analyses within the project, or to produce alternative analyses of the data. Sometimes, as with the Melville example, editors must accept the gains and losses of TEI functionality as against the analysis.

As Berry and Fagerjord have observed, “The encoding system […] need[s] to be carefully planned not only to enable effective data retrieval, but also in order to get data in…”[63] Here they touch on one of the key questions for encoding projects: in designing the markup, are we considering encoding primarily as a means of (a) recording and storing information about our texts; (b) disseminating our research findings to others via a project website; (c) enabling others to produce new analyses based on our data; or (d) some combination of all the above? The way we see the purpose of the encoding will inevitably shape our decisions about what to include and prioritise, and this should ideally be discussed at the initial design stages of a project, as it will influence the underlying framework as much as the web interface.

A similar back-end versus front-end crux shows in the Shelley-Godwin Archive digital edition of the *Frankenstein* Notebooks. An impressive work of documentary editing, the edition was also designed not for text analysis but for viewing on the Web. Again, this makes sense because we would expect most people to browse through a document-based edition on the Web, but there is currently no functionality for gaining access to a single concatenated XML file of all the Notebook transcriptions. If one were available, various text analyses would be made possible, from the more rudimentary sort of element searches (say, identifying all of Percy Shelley’s edits to the manuscript and applying linguistic analyses) to stylometric analyses that could tease out the differences between Percy’s language and Mary’s. The current site, however, does not provide this: one can download all of the project’s files (for which they should be commended), but due to the particular aims of that project their file system is almost inaccessible to the average outside researcher. The notebook XML is an index to the individual diplomatic transcriptions with <xi:include> elements that point to other XML files based on each transcription of the manuscript page.[64] Even with a combined XML file, one would still need to validate the XML against their schema in order to transform (or run analyses of) the XML, which puts up another barrier for the digital researcher outside of the project. This suggests that a workflow designed at once for valid TEI encoding and for text analysis should (regardless of the filing system and site architecture) at least make available a downloadable single XML file (or directory of XML files) amenable to analysis by others. A workflow designed with both encoding and analysis in mind could also make text analysis tools available on the front-end, but that would be more expensive than simply making available pared-down XML files. These questions ultimately come down to workflow design, and the efficiency of a researcher’s workflow within a team.

At the risk of stating the obvious, the ethos of collaboration and teamwork in digital humanities should be applicable to digital editing projects. No one researcher will be able to do all the transcriptions, create a computational pipeline, and design a database and website—a selected team of professionals with complementary skills will bring all those pieces together. Yet it is crucial that digital editors create a workflow that includes encoding—ideally following TEI standards—while also keeping in mind how other digital researchers might want to access the project data for data mining and analysis.

One of the present difficulties is that projects rarely elucidate the decisions they made about workflow, and whether they saw text or data analysis as a crucial part of their documentation. Dunn has observed that workflows in arts and humanities 'are highly individual, often informal, and cannot be easily shared or reproduced.'[65] We would recommend that however individual a project's workflow might be, it is worth sharing the decision-making process as part of that project's outputs. The scarcity of such documentation is undoubtedly a barrier to creating improved processes through an understanding of the issues projects have encountered, how they have resolved them, what their priorities have been, and, if possible, what effects their decisions have had within and even beyond the project. Commitment to describing workflow might best be made at the point of applying for funding, as otherwise there can be a tendency to view workflow documentation as an output that fits firmly under the 'ideal world' umbrella, rather than being considered an important part of the project's outcomes.

Conclusion: Encoding as analysis

With the above in mind, we set out here a suggested workflow for an encoding project that will promote analysis and knowledge production from the text. As noted, Flanders and her team at the Women Writers project published a guide that we consider highly effective and to which we are suggesting some additions and reordering rather than replacements.[66] Flanders and her collaborators set out seven main steps:

- **Planning your project**: representation of text, details of transcription and encoding, editorial method, and additional information such as glosses;
- **Project analysis**: duration of project, reasoning for encoding, editorial philosophy, considerations of audience, team, and how users will access text;
- **Document analysis**: similarities and differences amongst documents, genre, chronology, language, physical support, and legibility;
- **Transcription and markup**: digitising the text, encoding methods including automated markup, creation of template, stylesheet development;
- **Error checking**: proofreading for typographical errors and encoding errors, using tools and/or by hand;
- **Post-processing**: automated encoding, discovery and correction of er-

rors, transformations to other formats for publication and archiving;

- **Documentation:** schema, encoding practices, editorial practices, tools and procedures.

While this is not quite a step-by-step guide, as several areas overlap and it is not always possible to divide the tasks quite so neatly, the main areas of the work of an encoding project are represented here. To this we would add an evaluation of text analysis between 'Document analysis' and 'Transcription and markup': in other words, after you have analysed your documents but before actually beginning your markup, it would be worthwhile to consider the types of text analysis that would be relevant, and how you could enable such analysis in the way you encode your texts. This could even include testing some of the available open source tools for text analysis.

We would also recommend that documentation be considered not as a separate step, but as an integral part of each stage. Many of us have encountered projects on which the documentation is left until the end and time runs out. To avoid this, it would be worth documenting the stages of the project, and the decisions made, as they happen. Our suggested workflow would, then, look more like this:

- **Planning your project + documentation**
- **Project analysis + documentation**
- **Document analysis + documentation**
- **Text analysis evaluation + documentation**
- **Transcription and markup + documentation**
- **Error checking + documentation**
- **Post-processing + documentation**

This that the project wants to optimise its encoded materials for text analysis by others, rather than undertaking the text analysis; if the latter, discussion should be included in the 'Planning your project phase' and the task itself following the markup phase.

The same principles that call for a standardised language of documentary editing in Classics should easily transfer to documents in modern literature. Yet, as Tanselle showed in his seminal article 'The Editing of Historical Documents,' it is far from true that that has consistently happened:[67] for various reasons faithfulness and diplomacy can give way to modifications of original documents in the name of readability. Yet another problem with documentary editions is how to optimise the information in documents that are difficult to read in print.

A palpable challenge in the age of computers is the temptation to 'make it new'. This modernist doctrine did not advocate throwing out the old and coming up with something entirely new. The 'new' should be informed by, and in response to, tradition, as T. S. Eliot memorably put it in 'Tradition and the Individual

Talent'. Textual editing is simply doing more with the aid of computers to guide research and reading practices. One way in which digital methodology and text analysis is in a sense 'new' is that it goes beyond texts into other data of culture. With the advent of digital curation, editions of material culture can be encoded and analysed for the benefit of literary culture and vice versa.

Any project will have its share of false starts and ill-judged decisions that lead to re-doing some of the work, and these 'failures' are of course helpful to document in themselves. But as the examples above suggest, old notions of print-based workflow stand in the way of an appropriate computational pipeline that would make analytically-informed, machine-readable documents complement text analysis tools that provide new modes of reading evidence. Part of the continuation of the print-based model is due to a belief that somehow print books are more lasting, that they do not exist in some immaterial form or ethereal cloud. But even digital projects rely on physical things, and all things decay. A print-out of an XML file will be more useful to future historians than printed books. Why? XML files include more information about a text—or group of texts—that make up an edition. As Greg Crane has pointed out, we are still in the incunabula phase of digital editions.[68] If that is true, it would be smarter to focus less on layered web applications and more on curating the underlying XML data—data which can be shared and used for digital text analysis to create new modes of reading and critical interpretation.

Notes

- 1 So said Greg Crane in his 2010 article 'Give us Editors! Re-Inventing the Edition and Re-thinking the Humanities', in J. McGann (ed.), *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia, March 26-28, 2010* (pp. 81–97). Retrieved from http://cnx.org/content/col11199/1.1/ Editing is also 'messy, destabilizing, and above all, dynamic', Crane adds (p. 83).

- 2 Pierazzo, E. (2015). *Digital Scholarly Editing: Theories, Models, Methods.* Farnham, Surrey: Ashgate: 77.

- 3 Pierazzo 2015, pp. 88–90. Pierazzo surveys the 'epistemic virtues' that are part of textual scholarship ('truth-to-nature', 'objectivity' and 'trained judgment', as supplied by Daston and Galison), and what we are suggesting here is an alternative epistemology of pragmatism, or a success-in-action model that privileges coherence warranted assertibility in textual claims.

- 4 Quoted in Blackburn 2017, p. 63.

- 5 This was the subject of a panel at the 2012 Digital Humanities conference by Syd Bauman, David Hoover, Karina van Dalen-Oskam and Wendell Piez, Text Analysis Meets Text Encoding: ‘Recent DH conferences have comprised, in addition to other activities, two distinct sub-conferences – one focusing on text encoding in general and TEI in particular, and the other on text analysis, authorship attribution, and stylistics. The separation between the two is so extreme that their participants often meet only at breaks and social events.’ Given the fact that many practitioners of text analysis tend to be interested in txt files and very large data sets, it makes some sense that they would be less interested in the ‘small data’ of editions. Retrieved from http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/text-analysis-meets-text-encoding.1.html

- 6 Burrows, J.F. (2004). Textual Analysis. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), *A Companion to Digital Humanities.* Oxford: Blackwell. Retrieved from http://www.digitalhumanities.org/companion/. Burrows's 1987 study of Jane Austen, *Computation into Criticism.* Oxford: Oxford University Press, is still one of the finest examples of text analysis. Another recent one is Matthew Jockers's *Macroanalysis* (Urbana-Champaign and Chicago: University of Illinois Press 2013).

- 7 The Guidelines can be found at https:/ /www.mla.org/Resources/Research/ Surveys-Reports-and-Other-Documents/Publishing-and-Scholarship/ Reports-from-the-MLA-Committee-on-Scholarly-Editions/ Guidelines-for-Editors-of-Scholarly-Editions#editor. Of interest too is the MLA's recent white paper on electronic editions, ‘MLA Statement on the Scholarly Edition in the Digital Age’, https://www.mla.org/content/download/52050/1810116/rptCSE16.pdf. The words ‘text analysis and data mining’ do appear at the very end of the white paper, but only in the context of making print editions available online for such tasks, rather than being an essential aspect of the encoding process.

- 8 Andrews, T. (2013). The Third Way: Philology and Critical Edition in the Digital Age. *Variants,* 10, 61–76.

- 9 For more on analytic and interpretive encoding, see Chapter 17 of the *TEI Guidelines.* Retrieved from https://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html.

- 10 The best are Philip Gaskell's *A New Introduction to Bibliography* (Oak Knoll, 1978), David Greetham's *Textual Scholarship: An Introduction* (Garland, 1994), and William Proctor Williams and Craig S. Abbott's *An Introduction to Bibliographical and Textual Studies* (1999).

- 11 Tanselle, G. Thomas. (1995). Varieties of Scholarly Editing. In D. Greetham (Ed.), *Scholarly Editing: A Guide to Research.* New York: Modern Language Association. 9–32: p. 9.

- 12 Bryant, J. (2002). *The Fluid Text.* Ann Arbor: University of Michigan Press.

- 13 See Tanselle's 'Varieties', pp. 18–22; also Suarez and Woudhuysen.
- 14 Greetham, D. (2013). A History of Textual Scholarship. In N. Fraistat & J. Flanders (Eds.), *Cambridge Companion to Textual Scholarship.* Cambridge: Cambridge University Press (pp. 16-41): p. 19.
- 15 McKenzie, D.F. (1999). *Bibliography and the Sociology of Texts.* Cambridge: Cambridge University Press; McGann, J. (1992). *A Critique of Textual Criticism* (Charlottesville: University of Virginia Press).
- 16 Siemens, R. (2016). Communities of practice, the methodological commons, and digital self-determination in the Humanities. *Digital Studies/le Champ Numérique.* doi: http://doi.org/10.16995/dscn.31
- 17 Murphy, J.Stephen (2008). Death of the Editor. *Essays in Criticism* 58.4, 289–310. p. 294
- 18 Ricks, C. (1981). In Theory. *London Review of Books* 3.7, 16 April 1981: 3–6.
- 19 Housman, A.E. (1921). The Application of Thought to Textual Criticism, originally published in *Proceedings of the Classical Association* 18, 68–69, and reprinted in Keleman, E. (2008). *Textual Editing and Criticism: An Introduction.* New York: Norton.
- 20 Johnson, S. Proposals for printing by subscription the dramatic works of William Shakespeare, corrected and illustrated by Samuel Johnson. In W. K. Wimsatt, Jr. (Ed.), (1960). *Samuel Johnson on Shakespeare.* New York: Hill and Wang, 19–20.
- 21 Pierazzo, E. (2015). *Digital Scholarly Editing: Theories, Models, Methods.* Farnham, Surrey: Ashgate: p. 15.
- 22 Daniels, J. & Thistlethwaite, P. (2016). *Being a Scholar in the Digital Era. Transforming Scholarly Practice for the Public Good.* Bristol: Policy Press. p. 9.
- 23 Pragmatism is not just a word that encompasses an attitude toward practice and practical success in action; it is also a logic that facilitates 'a method for the analysis of concepts' (Peirce, 'A Definition of Pragmatism,' p. 56). It bears reminding that later in the same piece (p. 57) he emphasises, 'Thinking is a kind of action, and reasoning is a kind of deliberate action.'
- 24 Tim Berners-Lee (2007). Linked Data. Retrieved from https://www.w3.org/DesignIssues/LinkedData
- 25 Fraistat, N. & Flanders, J. (2013). Introduction. In N. Fraistat & J. Flanders (Eds.), *Cambridge Companion to Textual Scholarship.* Cambridge: Cambridge University Press, pp. 13–14.
- 26 For more on this, see Bauman, S. (2008). Freedom to constrain: where does attribute constraint come from, mommy? Presented at Balisage:

The Markup Conference 2008, Montréal, Canada, August 12–15, 2008. In *Proceedings of Balisage: The Markup Conference 2008. Balisage Series on Markup Technologies*, vol. 1. doi:10.4242/BalisageVol1.Bauman01

- 27 This debate is nicely summarised in Pierazzo 2015, pp. 78–9.

- 28 The obvious utility of this can be seen in the Melville Electronic Library's 'fluid-text' edition of *Billy Budd, Sailor*, which is based on Melville's final and uncompleted manuscript of his novella. The edition replaces the genetic symbols used by its print predecessor with TEI encoding that not only gives all of the evidence of the manuscript's genesis but also matches directly to the facsimile image of the surviving manuscript. It also has three reading views: a diplomatic transcription, a 'base' text that renders the revised manuscript, and a lightly edited reading text. Using text analysis, one could identify patterns, topics and other important aspects of the genetic data from this document-based text of *Billy Budd*, and an additional interface called 'revision narratives' guides the reader through the variants. The developers on the project are aiming to make a single, concatenated XML file available for researchers who might want to perform these kinds of analyses. See also Ohge, 'Melville Incomplete,' which surveys the gains of the MEL *Billy Budd* over its print predecessors.

- 29 Robinson, P. (2013). Toward a Theory of Digital Editions. *The Journal of the European Society for Textual Scholarship* 10, 126–127. doi: https://doi.org/10.1163/9789401209021_009

- 30 For the module's coursepack, see http://www.wwp.northeastern.edu/outreach/seminars/uvic_advanc (particularly pages 6 and 11–12).

- 31 For more on how to implement text analysis strategies with Voyant Tools, see Geoffrey Rockwell and Stéfan Sinclair's *Hermeneutica* (MIT Press, 2016; companion at http://hermeneuti.ca/). Jockers's *Text Analysis in R for Students of Literature* (Springer, 2014) is currently the best introduction to R programming for literary projects. Another very useful book is Julia Silge and David Robinson's *Text Mining with R: A Tidy Approach* (available online at https://www.tidytextmining.com/). The *Programming Historian* (https://programminghistorian.org/) also features excellent tutorials on AntConc, Python and R, among many other computing topics.). The *Programming Historian* (https://programminghistorian.org/) also features excellent tutorials on AntConc, Python and R, among many other computing topics.

- 32 The current Roma tool is available at https://roma.tei-c.org/. Roma JS code can be accessed at https://github.com/TEIC/romajs.

- 33 See Cummings, J. (2014). The Compromises and Flexibility of TEI Customisation. In C. Mills, M. Pidd & E. Ward. *Proceedings of the Digital Humanities Congress 2012*. Studies in the Digital Humanities. Sheffield: The Digital Humanities Institute. Available online at: https://www.dhi.ac.uk/openbook/chapter/dhc2012-cummings.

- 34 https://sourceforge.net/p/epidoc/wiki/Home/ For a survey of EpiDoc and its aims, see H. Cayless et al., “Epigraphy in 2017”, *Digital Humanities Quarterly* 3.1 (2009). Available online at: http://digitalhumanities.org/dhq/vol/3/1/000030/000030.html

- 35 Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S. et al. (2007-2017). EpiDoc Guidelines: Ancient documents in TEI XML (Version 9). Available: http://www.stoa.org/epidoc/gl/latest/.

- 36 MARKUP list: https://lsv.uky.edu/scripts/wa.exe?A0=MARKUP

- 37 Women Writers Project (1996–2016). Northeastern University. Retrieved from http://www.wwp.northeastern.edu/

- 38 Women Writers Project Guide to Scholarly Text Encoding. (2007). Brown University Women Writers Project. Retrieved from http://wwp.neu.edu/research/publications/guide/index.html

- 39 For an introduction to epigraphy and the work of an epigrapher, see John Bodel’s chapter ’Epigraphy and the Ancient Historian’ in J. Bodel (Ed.), (2001). *Epigraphic Evidence. Ancient History from Inscriptions.* London and New York: Routledge (pp. 1–56).

- 40 Bodard, G. & Stoyanova, S. (2016). Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy. In G. Bodard and M. Romanello (Eds.), *Digital Classics Outside the Echo-Chamber.* London: Ubiquity Press (pp. 51–68): 59. doi: https://doi.org/10.5334/bat

- 41 Mytum, H. & Peterson, J.R. (2018). The Application of Reflectance Transformation Imaging (RTI) in Historical Archaeology. *Historical Archaeology* 52: 489–503.

- 42 Krummrey, H. & Panciera, S. (1980). Criteri di edizione e segni diacritici. *Tituli* 2, 205–215; Panciera, S. (2006). I segni diacritici: riflessioni e proposte. In S. Panceira, *Epigrafi, Epigrafia, Epigrafisti. Scritti vari editi e inediti (1956–2005) con note complementari e indici Vol. II.* Roma: Edizioni Quasar. (pp. 1711–1717). The EpiDoc Guidelines have Leiden equivalence as a minimum, and in fact can be used to encode much greater detail than is permitted by the Leiden conventions.

- 43 Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S. et al. (2007–2017). EpiDoc Guidelines: Ancient documents in TEI XML (Version 9). Available: http://www.stoa.org/epidoc/gl/latest/.

- 44 MARKUP list: https://lsv.uky.edu/scripts/wa.exe?A0=MARKUP

- 45 EpiDoc training workshops: https://wiki.digitalclassicist.org/EpiDoc_Summer_School

- 46 The first volume to be produced by the *Corpus Inscriptionum Latinarum* (*CIL*) project under Theodor Mommsen was published in 1863. To date *CIL* encompasses some 17 volumes in 70 parts. *Inscriptiones Graecae* (*IG*) is a continuation of the original *Corpus Inscriptionum Graecum* directed by August Böckh. To date it has published 49 fascicules.
- 47 Bodard, G. & Stoyanova, S. (2016): 55.
- 48 EpiDoc Front End Services. Retrieved from https://github.com/EpiDoc/EFES
- 49 Pietro Liuzzo's eXist-db test app: https://github.com/EpiDoc/OEDUc
- 50 Bodard, G. & Stoyanova, S. (2016): 55.
- 51 Voyant Tools. Retrieved from https://voyant-tools.org/
- 52 AntConc. Retrieved from https://www.laurenceanthony.net/software/antconc/
- 53 Pelagios Commons. Retrieved from http://commons.pelagios.org/
- 54 Pelagios' Recogito tool for semantic annotation, for instance: https://recogito.pelagios.org/
- 55 Papyri.info. Retrieved from http://papyri.info/
- 56 The most recent print edition, edited by R.A. Foakes and R.T. Rickert, is largely an updating of W. W. Greg's 1904 London edition (*Henslowe's Diary*, by Philip Henslowe, 2nd ed. Cambridge: Cambridge University Press, 2002). Our thanks go to Dr Mao for sharing his XML on his project-in-progress.
- 57 The facsimiles of the Henslowe diary are available at http://www.henslowe-alleyn.org.uk/essays/hensloweddiary.html.
- 58 By 'node-level' we mean either XML elements (tags) or specific attribute values within those tags. In R, for example, one can do this with the XML library, which has the necessary functions for parsing XML documents (https://cran.r-project.org/web/packages/XML/index.html). MMO made significant use of the XML library package in R.
- 59 See Ohge, C. & Olsen-Smith, S. Digital Text Analysis at *Melville's Marginalia Online*, and Ohge, C., Olsen-Smith, S. & Barney Smith, E. (2018). "At the Axis of Reality": Melville's Marginalia in *Dramatic Works of William Shakespeare. Leviathan: A Journal of Melville Studies* 20.2: 1–16, 37–67.
- 60 Keats Library. Retrieved from http://keatslibrary.org/paradise-lost/
- 61 Trojan Horse markup uses empty elements to indicate the start and end of regions that cannot be contained within XML content elements. See Sperberg-McQueen's recent demonstration of 'Trojan Horse' markup at the 2018 Balisage markup conference: Sperberg-McQueen, M. (2018). Representing concurrent document structures

using Trojan Horse markup. In *Proceedings of Balisage: The Markup conference 2018.* Balisage Series on Markup Technologies, vol. 21. doi: https://doi.org/10.4242/BalisageVol21.Sperberg-McQueen01

- 62 For a good overview of the problems of encoding marginalia in TEI, see Estill, L. (2016). Encoding the Edge: Manuscript Marginalia and the TEI. *Digital Literary Studies*, 1.1. doi:
- 63 Berry, D.M. & Fagerjord, A. (2017). *Digital Humanities. Knowledge and Critique in a Digital Age.* Cambridge: Wiley: 52.
- 64 For more on using <xi:include> for stand-off markup, see Chapter 16 of the *TEI Guidelines.* Retrieved from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html#SASO
- 65 Dunn, S. (2016). Dealing with the Complexity Deluge. VREs in the Arts and Humanities. *Library Hi Tech* 27, 205–216.
- 66 Women Writers Project Guide to Scholarly Text Encoding. (2007). Brown University Women Writers Project. Retrieved from http://wwp.neu.edu/research/publications/guide/index.html
- 67 Tanselle, G.T. (1978). The Editing of Historical Documents *Studies in Bibliography,* 31, 1–56.
- 68 Crane. G. (2010): 81. For a more sustained treatment of some of these issues, see Cummings, J. (2018). 'A world of difference: Myths and misconceptions about the TEI', *Digital Scholarship in the Humanities.* doi: https://doi.org/10.1093/llc/fqy071

References

1. Andrews, T. (2013). The Third Way: Philology and Critical Edition in the Digital Age. Variants, *10*, 61–76.
2. Bauman, S. (2008). Freedom to Constrain: where does attribute constraint come from, mommy? Presented at Balisage: The Markup Conference 2008, Montréal, Canada, August 12–15, 2008. In Proceedings of Balisage: The Markup Conference 2008. Balisage Series on Markup Technologies, vol. 1. doi:10.4242/BalisageVol1.Bauman01
3. Bauman, S., Hoover, D., van Dalen-Oskam, K. & Piez, W. (2012). Text Analysis Meets Text Encoding. DH2012 Book of Abstracts. Retrieved from http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/text-analysis-meets-text-encoding.1.html
4. Berners-Lee, T. (2007). Linked Data. Retrieved from https://www.w3.org/DesignIssues/LinkedData.htm
5. Berry, D. M. & Fagerjord, A. (2017). Digital Humanities. Knowledge and Critique in a Digital Age. Cambridge: Wiley.

6. Blackburn, S. (2017). Truth. London: Profile Books.

7. Bodard, G. & Stoyanova, S. (2016). Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy. In G. Bodard and M. Romanello (Eds.), Digital Classics Outside the Echo-Chamber. London: Ubiquity Press (pp. 51–68): 59. doi: https://doi.org/10.5334/bat

8. Bodel, J. (Ed.). (2001). Epigraphic Evidence. Ancient History from Inscriptions. London and New York: Routledge.

9. Bryant, J. (2002). The Fluid Text. Ann Arbor: University of Michigan Press.

10. Bryant, J., Kelley, W., & Ohge, C. (Eds.). (2019). Fluid-text edition of Billy Budd, Sailor. Melville Electronic Library. Retrieved from https://melville.electroniclibrary.org/versions-of-billy-budd.html

11. Burrows, J. F. (1987). Computation into Criticism. Oxford: Oxford University Press.

12. Burrows, J. F. (2004). Textual Analysis. In S. Schreibman, R. Siemens, J. Unsworth (Eds.), A Companion to Digital Humanities. Oxford: Blackwell. Retrieved from http://www.digitalhumanities.org/companion/

13. Cayless, H., Roueché, C., Elliott, T. & Bodard, G. (2009). Epigraphy in 2017. Digital Humanities Quarterly *3*(1). Retrieved from http://digitalhumanities.org/dhq/vol/3/1/000030/000030.html

14. Crane, G. (2010). Give us Editors! Re-Inventing the Edition and Rethinking the Humanities. In J. McGann (Ed.), *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia*, March 26–28, 2010 (pp. 81–97). Retrieved from http://cnx.org/content/col11199/1.1/

15. Cummings, J. (2014). The Compromises and Flexibility of TEI Customisation. In C. Mills, M. Pidd & E. Ward. Proceedings of the Digital Humanities Congress 2012. Studies in the Digital Humanities. Sheffield: The Digital Humanities Institute. Retrieved from https://www.dhi.ac.uk/openbook/chapter/dhc2012-cummings

16. Daniels, J. & Thistlethwaite, P. (2016). Being a Scholar in the Digital Era. Transforming Scholarly Practice for the Public Good. Bristol: Policy Press

17. Dunn, S. (2016). Dealing with the Complexity Deluge. VREs in the Arts and Humanities. Library Hi Tech, *27*, 205–216.

18. Elliott, T., Bodard, G., Mylonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S. et al. (2007–2017). EpiDoc Guidelines: Ancient documents in TEI XML (Version 9). Retrieved from http://www.stoa.org/epidoc/gl/latest/

19. EpiDoc Front End Services. Retrieved from https://github.com/EpiDoc/EFES

20. Estill, L. (2016). Encoding the Edge: Manuscript Marginalia and the TEI. Digital Literary Studies, *1*(1). doi: https://doi.org/10.18113/P8dls115971

21. Foakes, R. A. & Rickert, R. T. (Eds.). (2002). Henslowe's Diary. 2nd ed. Cambridge: Cambridge University Press.

22. Fraistat, N. & Flanders, J. (Eds.). (2013). Cambridge Companion to Textual Scholarship. Cambridge: Cambridge University Press.

23. Gaskell, P. (1978). A New Introduction to Bibliography. New Castle, Delaware and Winchester: Oak Knoll.

24. Greetham, D. (1994). Textual Scholarship: An Introduction. New York: Garland.

25. Greetham, D. (2013). A History of Textual Scholarship. In N. Fraistat & J. Flanders (Eds.), Cambridge Companion to Textual Scholarship. Cambridge: Cambridge University Press (pp. 16–41): p. 19.

26. Housman, A. E. (1921). The Application of Thought to Textual Criticism. Proceedings of the Classical Association, *18*, 68–69, reprinted in Keleman, E. (2008). *Textual Editing and Criticism: An Introduction.* New York: Norton.

27. James, W. (1907). Pragmatism. New York: Longmans, Green & Co.

28. Jockers, M. (2013). Macroanalysis. Urbana-Champaign and Chicago: University of Illinois Press.

29. Jockers, M. (2014). Text Analysis in R for Students of Literature. New York: Springer.

30. Keats Library. Retrieved from http://keatslibrary.org/paradise-lost/

31. Krummrey, H. & Panciera, S. (1980). Criteri di edizione e segni diacritici. Tituli *2*, 205–215.

32. McGann, J. (1992). A Critique of Textual Criticism. Charlottesville: University of Virginia Press.

33. McKenzie, D. F. (1999). Bibliography and the Sociology of Texts. Cambridge: Cambridge University Press.

34. Modern Language Association Committee on Scholarly Editions. (2016). MLA Statement on the Scholarly Edition in the Digital Age. Retrieved from https://www.mla.org/content/download/52050/1810116/rptCSE16.pdf

35. Murphy, J. S. (2008). Death of the Editor. Essays in Criticism *58*(4), 289–310.

36. Mytum, H. & Peterson, J.R. (2018). The Application of Reflectance Transformation Imaging (RTI) in Historical Archaeology. Historical Archaeology, *52*, 489–503.

37. Ohge, C. (2019). Melville Incomplete. American Literary History, *31*(1), 139–150.

38. Ohge, C. & Olsen-Smith, S. (2018). Computation and Digital Text Analysis at Melville's Marginalia Online. Leviathan: A Journal of Melville Studies, *20*(2), 1–16.

39. Ohge, C., Olsen-Smith, S. & Barney Smith, E. (2018). "At the Axis of Reality": Melville's Marginalia in *Dramatic Works of William Shakespeare*. Leviathan: A Journal of Melville Studies, *20*(2), 37–67.

40. Panciera, S. (2006). I segni diacritici: riflessioni e proposte. In S. Panceira, Epigrafi, Epigrafia, Epigrafisti. Scritti vari editi e inediti (1956-2005) con note complementari e indici *Vol. II*. Roma: Edizioni Quasar. (pp. 1711–1717).

41. Peirce, C. S. (1998). How to Make our Ideas Clear. In Chance, Love, and Logic. Lincoln: Bison Books.

42. Peirce, C. S. (1997). A Definition of Pragmatism. In Pragmatism: A Reader. Ed. Louis Menand. New York: Vintage.

43. Pierazzo, E. (2015). Digital Scholarly Editing: Theories, Models, Methods. Farnham, Surrey: Ashgate.

44. Proctor Williams, W. & Abbott, C. S. (1999). An Introduction to Bibliographical and Textual Studies. New York: Modern Language Association of America.

45. Programming Historian. Retrieved from https://programminghistorian.org/

46. Ricks, C. (1981, April). In Theory. London Review of Books, *3*(7), 3–6.

47. Robinson, P. (2013). Toward a Theory of Digital Editions. The Journal of the European Society for Textual Scholarship, *10*, 126–27. doi: https://doi.org/10.1163/9789401209021_009

48. Rockwell, G. & Sinclair, S. (2016). Hermeneutica. Cambridge: MIT Press.

49. Siemens, R. (2016). Communities of Practice, the Methodological Commons, and Digital Self-Determination in the Humanities. Digital Studies/le Champ Numérique. doi: http://doi.org/10.16995/dscn.31

50. Silge, J. & Robinson, D. (2020) Text Mining with R: A Tidy Approach. Sebastopol: O'Reilly Media. Also available online, retrieved from https://www.tidytextmining.com/

51. Sperberg-McQueen, M. (2018). Representing concurrent document structures using Trojan Horse markup. In *Proceedings of Balisage: The Markup conference 2018. Balisage Series on Markup Technologies*, *21*. doi: https://doi.org/10.4242/BalisageVol21.Sperberg-McQueen01

52. Suarez, S. J. & H. R. Woudhuysen. (2010). The Oxford Companion to the Book. Oxford: Oxford University Press. Retrieved from https://www.oxfordreference.com/view/10.1093/acref/9780198606536.001.0001/acref-9780198606536-e-3354

53. Tanselle, G. T. (1978). The Editing of Historical Documents. Studies in Bibliography, *31*, 1–56.

54. Tanselle, G. T. (1995). Varieties of Scholarly Editing. In D. Greetham (Ed.), Scholarly Editing: A Guide to Research. New York: Modern Language Association. 9–32.

55. The TEI Guidelines for Electronic Text Encoding and Interchange. Retrieved from https://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html

56. Wimsatt, W. K. Jr. (Ed.), (1960). Samuel Johnson on Shakespeare. New York: Hill and Wang.

57. Women Writers Project (1996–2016). Northeastern University. Retrieved from http://www.wwp.northeastern.edu/

58. Women Writers Project Guide to Scholarly Text Encoding. (2007). Women Writers Project, Northeastern University. Retrieved from http://wwp.neu.edu/research/publications/guide/index.html