International Journal of Digital Humanities (2019) 1:179–193
https://doi.org/10.1007/s42803-019-00016-6

RESEARCH ARTICLE

Opening the book: data models and distractions in digital scholarly editing

James Cummings[1]

Published online: 16 May 2019

Abstract
This article argues that editors of scholarly digital editions should not be distracted by underlying technological concerns except when these concerns affect the editorial tasks at hand. It surveys issues in the creation of scholarly digital editions and the open licensing of resources and addresses concerns about underlying data models and vocabularies, such as the Guidelines of the Text Encoding Initiative. It calls for solutions which promote the collaborative creation, annotation, and publication of scholarly digital editions. The article draws a line between issues with which editors of scholarly digital editions should concern themselves and issues which may only prove to be distractions.

Keywords Scholarly digital editions · Digital infrastructure · Textual editing · TEI XML · Markup and data models

1 Scholarly digital editions

The creation of scholarly digital editions is a complex endeavour which exposes and is dependent on our understanding of the theories of text, works, and documents that underlie our relationships with texts. My approach in textual editing projects tends towards the pragmatic, but there is a clear distinction between an objects to which we often refer as a 'work', i.e. an abstraction as understood by readers (including authors and editors), and a 'document', which is a particular instance of a physical manifestation of this text. Not all documents are faithful copies of the work, nor do they represent all possible ways of understanding the text in question.

> If we hold that each document is the work because without the document we would not have the work, we would have to see each different document as a different work. If one does not want to say that every copy of a work is a different

✉ James Cummings
James.Cummings@newcastle.ac.uk

[1] Newcastle University, Newcastle upon Tyne, UK

> work, then one must not say that the document and the work form an inseparable unity. If, on the other hand, one says a single work is represented differently by the variant texts in different documents, it seems necessary to also hold that one cannot apprehend the work as a whole without somehow holding its variant iterations in mind. Textual complexities resist simplification. How one conceives of the relationship between documents and works influences one's practice when editing; it is important to have a sense of the complexity of that relationship (Shillingsburg 2017, p. 188).

In editing a text then, the editor must attempt to communicate the understanding he or she has of both the document (and any additional related documents considered in scope) and the work as a whole. It is in 'holding its variant iterations in mind' that I would argue the true editorial objects are formed, and the representation of this in editions is inherently a lossy translation from the mental construct, whether or not the editions are print or digital, scholarly or otherwise. There is a long history of representing these mental constructs, i.e. what I consider a set of conceptual edited objects, in print editions, and the systems of encoding editorial understanding have a rich and complex history. As a side note, it should be recognised that in many discussions of editing, the assumption that it is solely concerning the relationship between multiple documents and their related works is often problematic for editors of works for which there is only a single witness. Single witness editions are no less editions. However, in the discussion of scholarly digital editions, we tend to focus on scholarly editing within the Lachmannian paradigm, on texts for which multiple witnesses exist, or at least complex textual apparatuses of one form or another, precisely because we seek edge cases on the basis of which to test, problematize, and construct our view of the nature of editorial activity. My view as a digital editorial pragmatist is that these edge cases are interesting, but while they must be dealt with, they need not distract us in the course of projects which focus on the creation of scholarly digital editions which function within the limits of existing solutions. I would argue that at times the academic investigation of scholarly digital editing focuses on the problems rather than the solutions, and as much as possible editors of scholarly digital editions should not be distracted from editorial tasks by technological concerns if these technological concerns do not affect their edition. There are cases, perhaps exacerbated by academic funding models, in which 'the perfect is the enemy of the good enough'.

In the creation of scholarly digital editions, the primary responsibility is the production of a scholarly edition that is no less rigorous than its print equivalent, but there is also the secondary responsibility to be truly digital in nature. That is not to say that a digitised edition (an edition which represents nothing significantly more or less than the possibilities of a print edition) is not a useful object. One could easily argue that the world would be a better place if we had digitised the full texts of existing print editions as a starting point. But in itself, a digitised edition barely exploits the fundamental shift in medium. Print editing, or the equivalent in digital form (such as static PDF editions), is restricted in the methods with which it presents the edited text, most commonly to a single perspective on the work with accompanying editorial information encoding additional witnesses or documentary information using standard formats, which the reader decodes.

> The edited text does not get closer to the documents, there is still no visual evidence, no making explicit of textual structures or semantic information, limited potential for multiple views on the text. This is why a digitised edition is not a digital edition (Sahle 2016, p.33).

It is precisely the potential of a digital edition to be near-infinitely refactorable and dynamically to provide different views depending on external interactions that is one of its greatest strengths. However, far too much discussion of digital editions focuses on the presentation views of the edition. The real digital edition, that which best represents the set of conceptual editorial objects (whether textual, musical, image, or other forms of editorial object), is not represented by any one view of the edition.

> Therefore in digital editions the encoded texts themselves are the most important long-term outcome of the project, while their initial presentation within a particular application should be considered only a single perspective on the data. Any given view will be far from unique or canonical, as different usage scenarios call for different presentations—ranging from "reading text" to "interactive version" with popup content, to chart, graph, or map representations and beyond. Furthermore, all initial presentations are also ephemeral, bound to be either modified over time as technologies and forms of digital publishing change, or languish in obsolescence on a forgotten server (Turska et al. 2016, para 4).

One way of looking at the encoded edition, for example an edition created in TEI XML, is to consider it the true edition rather than any particular output. However, and perhaps surprisingly given my long history helping maintain the TEI Guidelines, I do not view the encoded edition in TEI XML as the best form. Rather, it is, in my assessment, the best *serialization format* for the underlying conceptual data of scholarly digital editions.

> The edition is in the encoding; this implies that encoded data is, in a certain sense, already a scholarly-mediated presentation of other data that exist in the original manuscript (Barabucci et al. 2017, p. 44).

While the encoded data is a good representation of the scholarly edition and one I care about deeply, a truly conceptual editorial object is malleable and recombinable, and an encoded edition, by itself, is not. The encoded edition is sufficient, for those literate in the method of encoding, to present an edition by itself, but this would substantially limit the audience of the edition. Editors should be able to understand the granularity and categorisation of an encoded edition, and they should recognise where they have abrogated any philological responsibility, but they should not necessarily be distracted by the underlying data format. However, by forcing us to formalise some of our assumptions, the structures and vocabularies of an encoded edition do help us foreground the theories of text we use when creating scholarly digital editions, and thus it is important that editors be at least familiar with the format of the encoded edition and any limitations placed by it on their activity (Cummings 2008).

If part of the point of editing a work is to make it more accessible, in all senses of that word, then usually some presentation view of the edition is required. With TEI

XML-based digital editions this usually involves transforming the data to a web-based serialization format (such as HTML or JSON being fed to an HTML container).

> This distinction [between TEI data and HTML presentation] leads to an important question: what constitutes the core of an edition? Its data or its presentation? It is possible to think of a critical edition as a collection of pieces of pure data? Or is a representation layer fundamental to the concept of 'edition'? (Barabucci et al. 2017, p. 37)

In order to maintain the malleability of a conceptual editorial object, it is not the presentation layer that is a requirement, as the presentation layer is merely one or more additional views on the data. Rather, it is the ability to reshape, query, transform, and reconceive of the data in the same way as (and in the case of a digital edition in more ways than) a reader might do when translating a printed critical apparatus into a mental construct representing the various document instances and their relationship to the work. People developing complex IT systems for the publication of (usually specific) scholarly digital editions might believe that their systems provide the necessary infrastructure. However, these systems tend to focus on the rendering of one or more presentational views on their datasets rather than providing a more direct interface to the set of conceptual editorial objects encoded in the underlying data. Given current technologies, I would argue that the true form of a scholarly digital edition would be better expressed as a well-documented API for the manipulation and description of editorial objects following an open international standard for the representation of digital text. This would not necessarily provide the presentational view on the data that most readers would require, but views of scholarly digital editions could be constructed on top of it. This would enable all forms of examination, querying, subsetting, and recombination of all editorial objects of all types. This infrastructure would, in no way, stop a scholarly digital edition from being a publication of knowledge and commentary on an individual work, but it requires that its underlying framework meet at least basic criteria to enable the edition's involvement in inter-edition commentary and research in the future. There is a clear difference between the knowledge in a scholarly digital edition and the knowledge which can be created across a collection of interoperable editions, but the creation of one should not preclude the eventual development of the other. Clearly this is unlikely to be created fully formed as a complete solution, but efforts for API-based access to serialized editorial objects (for instance with open annotations or URI-referenced encoding as first order objects) are a step in the correct direction which should be encouraged and amalgamated through a coherent infrastructure. Ideally, open data repositories for digital editions would build such APIs into their serving of the underlying data of digital editions. (The TAPAS archive seems to be moving in this direction, but even it has long way to go.)

One of the limitations of publications of scholarly digital editions is that only so many views on the underlying encoded edition data can be realised. Even when access is given to an API foregrounding manipulation of all the editorial objects, the nature of the access will be limited to the methods of interrogation conceived of at the time of its construction. One approach is to define an API that "sees digital documents as stacks of abstraction levels, each storing content according to a certain model." (Barabucci and Fischer 2017, p. 51) However, these digital documents will suffer the same restrictions

based on the limitations of how and when they are created. Even when more sophisticated layers of abstraction are provided, there must also be methods for as many low level operations on the encoded editorial data as possible. The editor of a scholarly digital edition, I would argue, should understand the separation of these levels of abstraction and the nature of the models they store on a conceptual level, but does not need to be distracted by the actual implementation of this system.

2 Opening the edition

In order for scholarly digital editions to reach their full potential as contributions to a wider academic environment of digital resources, it is not enough that they merely be accessible but (as thankfully is becoming a requirement of many funding bodies) they also need to be openly accessible. This may seem a minor difference, and it is often confused with them being 'freely' accessible, that is free at the point of use by researchers. In a world in which digital resources must become sustainable, there are cogent arguments against making them freely available to all, and while this is a regretfully retrograde step which mirrors the publication of print editions, online hosting of scholarly digital editions must still be resourced. However, ideally editions would be freely available to anyone who wants to use them, and institutions which resource such sustainability should be lauded for their attempts to do so. Nonetheless, 'openly' available here is meant to convey legal availability rather than financial accessibility, i.e. that a digital edition is openly pre-licensed for reuse with terms as open as possible, e.g. licensing with a Creative Commons Attribution license where feasible, rather than one that includes Non-commercial or Non-derivative conditions, since these conditions significantly limit the potential reuse of the edition. Assuming that data is openly licensed and the licenses follow open international standards and the open repositories of digital edition data will exist for an extended period of time, then the most interesting repurposing of any digital edition will likely not be done by the original creators. In other words, assuming the survival of a well-documented edition's data into the distant future, the edition (or the data) is much more likely to be repurposed as technology develops and exploited in ways which we could not have predicted. And yet, current reuse of digital edition data by others is very rare, and even with large text collections such as EEBO-TCP, the reuse often consists of making improvements to them in order to ensure they reach the minimum criteria for a scholarly digital edition. Editors of scholarly digital editions do not need to be distracted by the detailed legal implications of openly licensing resources, but they should understand the general categories of restrictions and how openly licensing their own project outputs benefits future research. True reuse of scholarly digital edition data is a laudable aim, and open licensing of data following well-documented open international standards is the necessary foundation of this (potentially overly-optimistic) open data utopia of the future.

While the following characterization may represent an idealistic view, one of the benefits of a scholarly digital edition is its infinite potential to be revisited, reformed, and updated in what is a perpetual beta state which subverts the publishing hegemony under which scholarly editing produces editions and a significant revision of an edition (perhaps in adding newly uncovered witnesses) would form a new 'second edition'.

However, this infinitely changing edition creates a new barrier to use through its very nature as an unstable object, unlike print editions, which are more stable but less flexible. While the inherent anxiety about the possibility of ever-changing digital objects is addressed by technological solutions (such as pointing at any particular stage in its version history), these solutions do not fully solve the problem. And yet, the nature of scholarly digital editions is such that we now talk about 'versions' or, in reference to the source data itself, 'revisions' of the editions. The editorial structures in the data which underlies any scholarly digital edition become the actual resource itself, only temporarily translated into a variety of presentational structures.

> We create information resources that are guided by abstract models and abstract descriptions of the objects at hand. The dogma of our current markup strategies is the separation or rather translation from form to content. Thus, we do not just transform our textual witnesses from one (material) media and form into another (digital) media and form. Rather, we try to encode structures and meaning of documents and texts beyond their mediality. And from this data we may or we may not create, and from time to time recreate, arbitrary forms of presentation in one media or another (Sahle 2016, p. 32).

That not all scholarly digital editing is intended to produce an 'edition' rendered in a presentation view is an important reminder that such editions are not merely publications, but are intended to be resources with which to venture answers to the research questions which prompted their initial funding. While this often leads to new and different ways of reading the edition, this is more a product of the context of digital editions.

> The digital edition allows readers to break away from mono-directional reading (as has also been vigorously discussed in relation to hypertext) (Vine and Verweij 2012, p. 134).

However one reads the edition, the underlying data may in fact have been created not for a scholarly digital edition as a publication, but as a resource to be interrogated, analysed, or queried, rather than published. The publication of a scholarly digital edition can, and perhaps more often should, be a mere byproduct of the real research undertaken. That such information resources, in this case datasets of editorial objects, become corpora for research analysis also enables us to work with reproducible methodologies where all aspects of the data, methodology, and results are transparent. Striving for reproducible research also enables us to publish in more transparent ways, where the data behind the graph which supports any research claims and even the tools used to undertake the analysis are provided. This enables others to check the conclusions in a way often perceived by society (falsely, I must note) as more 'scientific'.

> When producing a scholarly edition, an article, or an introduction to an edition in a reproducible way, we publish not only the text in its final format including the prose with possible figures and tables, but also the data (in our case typically annotated transcriptions) as well as the computer code use in the analytic work.

> This enables other users– including our future selves – to redo, build upon and adjust the work without the need to start over (Speed Kjeldsen 2017, p. 135).

It is not just the reproducibility of the research that is important, but the underlying approach. Such a transparent approach to scholarly editing is not a neo-liberal quantification of computational literary studies as only containing objective data to be analysed (there is no such thing as neutral editorial encoding), but merely a foregrounding of our assumptions, methodologies, data, and results, whether we use Digital Humanities methodologies or 'Experimental Humanities'.

> [W]e present a different approach to the application of digital techniques to humanities research, a branch of experimental humanities in which digital experiments bring insight and engagement with historical scenarios and in turn influence our understanding and our thinking today (De Roure and Willcox 2017, p. 194).

Editors of a scholarly digital edition should not find the exposing of their research methods distracting; their editorial tasks produce a dataset upon which experiments which are core to a humanities research approach can be based. Moreover, as the humanities inevitably becomes an increasingly collaborative undertaking, any approach that assists us in making all aspects of scholarly digital editing more transparent from the outset can only be seen as useful.

3 Data models and the TEI

The de-facto standard for a data model to be used in creating scholarly digital editions is the Guidelines of the Text Encoding Initiative (TEI, http://www.tei-c.org/). This is a community-developed open international standard which provides a set of recommendations for the encoding of digital texts. Yet it is inaccurate to say that the TEI is a data model itself. Used properly, it is more of a framework for constructing and documenting data models for particular editorial projects. In many cases, the TEI defines objects for encoding texts, but it does so in a way which has been called ontologically agnostic. That is, it defines a particular markup object for encoding a specific textual phenomenon, but it does not always prescribe how to determine the nature of that phenomenon. For example, the TEI's <title> element is defined as '[containing] a title for any kind of work', but TEI does not specify how to determine whether or not something is in fact a title of a work. This extends to all sorts of editorial interventions and encoding, where the editor is still left to determine whether a string of characters is indeed the textual phenomenon in question. In reality, this is a pragmatic level of indirection which enables the standard to be used by vastly different editorial communities. Moreover, TEI customisation can provide equivalences to existing ontologies if the project is intended to relate an understanding of TEI encoding to particular real-world concepts. The individual encoding of textual phenomena represents the editor's interpretation of the objects which exist in the real world, and while these signs may be encoded according to different methods, the editor's choice for how to encode any particular instance of a textual phenomenon has at its root a materialistic cause which we should not confuse with its conceptual categorisation.

> Note that here I am not negating the whole pluralist view of textuality: I am only denying the unlicensed (and undesirable, in my view) consequence that texts are not really existent objects. The fact that we can describe reality at different levels does not imply that the objects we describe do not exist *in se*: this fallacy is a direct consequence of the confusion between ontology and epistemology, a confusion that I want to get rid of (Ciotti 2017, p. 87).

At the heart of creating a TEI data model is the process of customisation that the TEI framework uses to document, in a literate programming vocabulary, the relationship of the vocabulary of the TEI to the application that is being undertaken in any particular project. The TEI provides a processable form of customisation using the TEI ODD format, which enables both the constraining of the overall scheme and its extension into new areas.

At time of writing, the TEI P5 Guidelines version 3.5.0 have 573 elements, but no particular scholarly digital edition would be expected to make use of all of them.[1] Though, to be clear, it is not just the inclusion/exclusion of elements that might form part of a customisation. All aspects of the TEI framework (elements, attributes, classes, modules, prose, examples, content models, intended processing, and much more) can be modified for any particular project. Indeed, in proper use of the TEI, customization is not only recommended, but almost required for:

> These Guidelines provide an encoding scheme suitable for encoding a very wide range of texts, and capable of supporting a wide variety of applications. For this reason, the TEI scheme supports a variety of different approaches to solving similar problems, and also defines a much richer set of elements than is likely to be necessary in any given project. Furthermore, the TEI scheme may be extended in well-defined and documented ways for texts that cannot be conveniently or appropriately encoded using what is provided. For these reasons, it is almost impossible to use the TEI scheme without customizing it in some way (TEI Guidelines, Chapter 23: 'Using the TEI', Section 23.3 'Customization' http://www.tei-c.org/release/doc/tei-p5-doc/en/html/USE.html#MD).

The nature of the TEI framework in providing methods for extensible meta-schemas (from which TEI users generate schemas to validate their document) can result in vastly different views of the TEI. These views may be so varied as to be almost mutually incompatible, and yet having the common framework at their basis is always going to be more beneficial than a multitude of different schemes. The documentation of the fragmentation found in a TEI ODD customisation file actually enables easier interoperability and interchange between digital editions than if no such documentation existed.

> Such documentation of variance of practice and encoding methods as a TEI ODD meta-schema preserves then helps to enable real, though necessarily mediated, interchange between complicated textual resources (Cummings 2014).

[1] The number of elements the TEI Guidelines currently include is available on the element reference page from version 3.2.0 onwards. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html

The potential for different projects to define their own meta-schemas creates a fragmentation of overall consistency among projects. However, because they all define their variance from the same source in a machine processable form, the divergences are not as great as one might expect. (Though to say 'same source' ignores the rolling releases of the TEI framework, i.e. the objects available to one customisation may have been greatly modified by the time another customisation is created.) The TEI ODD customisation methods provide a mechanism by which the meta-schema can document the version of the TEI on which the customisation is based. And yet, even the phrase 'version of the TEI' is inaccurate. In most uses of the TEI this is a sufficient description, but it is possible in a single TEI customisation to use multiple sources, which might be different versions of the TEI framework or indeed other standards entirely.

The TEI customisation-literate programming mechanisms can also be used to document entirely non-TEI schemas. To further complicate matters, in some sophisticated uses a TEI ODD will 'chain' customisations in order to provide a variation on an existing TEI customisation. For example, a project might decide that its needs are very similar to the EpiDoc schema (a pure TEI P5 subset) but that it needs additional elements or different attribute values or wants to customise the examples to its materials. The project would indicate that the source is not the TEI directly, but the compiled EpiDoc customisation, which itself points to the TEI. In processing these chained customisations to generate schemas, each of the customisations is flattened in turn against its source. Any form of TEI customisation is potentially quite complicated because of the generalistic nature of the framework of which it is a part. And yet this also provides quite rigorous methods of documenting the variance between schemas in a way that can be processed on a more general level. An editor of a scholarly digital edition should understand, at least on a conceptual level, the customisation of any formal vocabulary they are using and the relationship of this vocabulary to the categories of textual phenomena and the editorial activity they are undertaking. However, an editor whose well-resourced project team has included assistance in this area need not be distracted by the methods by which this customisation is implemented.

A recent change in the TEI further extends the TEI ODD customisation language vocabulary with the ability to document intended processing models. This is a significant departure for the TEI, which has more usually held that the processing of the encoded edition is a completely separate activity from that which it defines, the encoding of the textual phenomena according to an agreed framework. The TEI vocabulary for customisation now provides users a mechanism with which they can indicate, in an entirely implementation-agnostic method, how they intend a particular element (or other TEI object) to be processed for a variety of outputs. This mechanism does not specify precisely how to handle elements, but it gives a general behaviour recommendation and might indicate some details of formatting. For example, this processing model documentation might indicate that for abbreviations and expansions embedded inside a TEI <choice> element that something processing it should implement a behaviour of 'alternate' which would somehow provide both the abbreviation and expansion to the user. Furthermore, the processing model documentation in the TEI customisation can indicates which of the abbreviations or expansions should be used as the 'default' content and which should provide the 'alternate'. One can imagine using this instruction in web output to provide a tooltip with expansion and the text showing the abbreviation. In a print publication the same 'alternate' behaviour might generate a

footnote to provide a similar effect. One of the reasons for doing this in a hands-off implementation agnostic manner is to predetermine not the nature of the processing, but instead the nature of the intended output. Another benefit is the shrinking of the code-base necessary to maintain publication solutions. Instead of writing code to deal with every occurrence of TEI elements, one could create a system which examines this documentation and reacts to the models it contains. Indeed, early experiments show that this can be beneficial in simplifying the maintenance of such code (Turska 2017, p. 364). One of the intentions behind the documentation of the processing model, however, is to benefit not just software developers but also editors of scholarly digital editions. The format is designed to be simple enough that an editor could easily change whether the abbreviation or expansion is shown or whether it is highlighted in bold or italics. While editors of digital scholarly editions need not be distracted by how the processing is implemented, if this processing model documentation is being exploited by their publication processing they will enjoy significant benefits if they understand the base format well enough to have control over the presentation.

One popular area for discussion in explorations of scholarly digital editing is the handling of critical apparatuses of multi-witness texts. These are works that are represented in the edition by multiple documents (extant or theorised) in order to produce a coherent editorial view of the text. This is one area where people sometimes argue against XML as a serialization format. As this is the current format in which TEI is expressed, those making this argument often find themselves arguing against this open, international, community-developed set of recommendations. Instead, other formats are suggested by proponents of one solution or another, mostly based on hiding the serialization format from the user. And yet, once an interface is placed between the editor and the underlying code which represents their decisions, then in many ways it is only the granularity of information and its relationships which matter, not the serialization format. In such a system, though editors are assisted in their work, 'the tools themselves and their heuristics are not questioned, as long as they do what they are "told"' (Pierazzo 2015, p. 109).

One of the facile arguments that is often made by those opposed to an XML-based solution is that XML (and thus TEI) is unable to handle overlapping hierarchies. This is, of course, a falsehood. People who perpetuate this myth, however, are usually doing so innocently with a naive understanding of XML as a format assuming all XML representations are created as embedded in-line markup. (Cummings 2018) The creation of markup structures that rely on an embedded hierarchical use of XML is an increasingly dated notion of how XML is used in complex resources such as scholarly digital editions. Increasingly, scholarly digital editions are based on distributed and multi-faceted sets of resources. The idea that all the markup of an edition is embedded within a single hierarchy of XML and encoded inline, while often the tempting when one is creating early digital editions, is now a strawman used for to propose a conflict between overlapping hierarchies. While much ink has been spilt on the merits and shortcomings of the various solutions to the problem, this discussion is pursued primarily by people seeking more elegant solutions for the markup languages of the future. For many projects, the in-built solutions, such as the use of milestones (such as <pb/> to mark page breaks) where one hierarchy (usually the intellectual) is preferred over another (usually the physical), are sufficient. Swapping between hierarchies displayed as milestones is no longer the complex processing activity that was once imagined.

There are numerous other methods for overcoming the supposed limitations of XML without departing from its specification, including out-of-line or standoff markup. It is perfectly reasonable in XML to employ a simple technique of remotely pointing into basic structures (with URI-based pointing or other standoff mechanisms) to provide encoding and annotations which might be at risk of overlapping. This is not a non-XML solution, as it is entirely possible for out-of-line markup to exist as pure XML. For example, the TEI Guidelines provide recommendations for how an <app> element recording an editorial apparatus entry may be stored completely separately from the base text to which it refers. In addition, the apparatus readings may now surround larger structures (such as whole divisions or paragraphs), and not merely phrase-level content. With regard to the use of out-of-line markup, it does not matter if the objects being stored out-of-line are variant readings, physical vs intellectual structures of the document, or something else. If the text is encoded to a sufficient degree of granularity, then all of these supposedly conflicting hierarchies can be expressed in separate out-of-line markup that points to the site of overlap. My own stance as a pragmatic digital editor is to encode at an orthographic word level of granularity (whose markup can be added by simple scripts). While this might mean some redundancy when recording sub-word changes, this is balanced by the ease of processing at this level. While out-of-line markup is a very simple and powerful mechanism that can be used to cut across an infinite number of hierarchies, it does so at the cost of human-readability. The underlying problem, which explains why solutions such as this, which employ out-of-line or standoff markup, are not popularly used by all digital editions, is that of support from tools, not only in the creation of editorial objects and annotations of data, but also in its processing. There are limitations in the creation of markup for scholarly digital editions that may cross the boundaries of common embedded markup structures, but these limitations are the result of a lack of tools with which to create the markup in standoff or out-of-line forms, rather than any particular serialization format. Other proposed formats, such as JSON (a very useful serialization for frontend manipulation) or RDF (a useful graph technology for conceptual annotation), have as many well-understood problems as formats like XML, and in the creation of scholarly digital editions in TEI, they are more accurately understood as generated outputs from the TEI source. Nonetheless, solutions to these problems are not beyond the scope of current technology, but when projects create solutions, the solutions are usually for very specific use-cases rather than generalised applications. When a scholarly digital edition project creates a significantly detailed frontend to hide the encoding structure, it becomes unnecessary to start proposing entirely new data formats and to eschew the vocabularies of existing open international standards. Significant user-friendly technology in this area would benefit the creation of scholarly digital editions no end, especially if these solutions built on the improvements to the recommendations of the TEI, such as the processing model documentation. The TEI framework is a mature, rich, and complex method of documenting our relationships with text (in its many forms). While editors of scholarly digital editions should not be overly distracted by the implementation of underlying technology for the creation and publication of their editions, they should not be dissuaded from using de facto standards, such as the TEI, merely because they do not wish to understand any of the technological background to their editions. Developers who would throw away frameworks like the TEI because they dislike the current serialization format (XML), because of their own technology choices, or want to reinvent the wheel (and do not realise that they can do so within the framework) are short-sighted.

4 Publishing scholarly digital editions

Even where one is not worried about multiple hierarchies or complex out-of-line markup and is creating an edition which is straightforward, the publishing of a scholarly digital edition is still a needlessly complicated affair. Given the technology that already exists and the solutions which have been reinvented time and time again, it is unconscionable that public research funds are used to produce bespoke publication engines unnecessarily again and again. Slowly generalised but customisable and detailed publication infrastructures (such as TEI Publisher http://www.teipublisher.com) are being developed, but they still have a long way to go. It is unusual for an individual to have all the skills necessary to edit a work properly, create an encoded edition, and develop a publication framework. Some of us who have some skills in multiple aspects of these areas are usually less developed in other areas.

> While there are scholars who have achieved such an impressive skillset, it also seems evident that they are setting the threshold very high and that it is not likely that this profile will become very common in the foreseeable future, if at all (Pierazzo 2015, p. 115).

The real answer, of course, is that the creation of scholarly editions, whether digital or not, has never been an individual enterprise. Just as an author in the age of incunabula had a sense of printing technology but did not fully understand the techniques printers used, editors should not be distracted by the publication infrastructure for their editions. Usually, the publishers and printers of print editions took on many of the activities that are now cognate to the frontend developers and web hosting for digital editions. However, as mentioned earlier, so far, no single generalised software for the publication of scholarly digital editions has had mass uptake by the community. And as the research for scholarly digital editions becomes more collaborative (though ignoring the potential of crowdsourcing and citizen science for digital editions), a solution that lowers the bar for the production and publication of digital editions would inherently need to be a collaborative platform. Instead of creating solutions that are individual to any specific project's needs, we need collaboratively to build small modular improvements on top of a generalised infrastructure for the creation, publication, and analysis of scholarly digital editions.

> All of these tools, however, act like small unconnected islands. They expect input and output data to match their own data format and data model, both narrowly tailored to their task and following their own idiosyncratic vocabulary (Barabucci and Fischer, p.48).

What is needed is a generalised infrastructure to which a larger community of scholarly editing projects contribute and which leverages existing technologies for handling scholarly digital editions. This infrastructure should require little or no specialised knowledge for its use by an editor of scholarly editions. Having the requisite skills for work as an academic researcher in a modern digital age should be sufficient to produce a digital edition. Even if the skills of encoding the edition in TEI XML are required by the editor (and my experience in teaching TEI is that this is a basic skill that

all modern editors are more than capable of learning if they honestly have the desire to do so), the additional annotation, text-image interactions, collaboration with colleagues, and publication and interrogation of this data should be done through a standard easy-to-use interface based on the most common open international standards.

> If digital editing should become the standard practice for preparing editions, digital tools, which are easy to handle and do not require much technical or even programming skills are needed. Moreover, we need useful standardization processes, which lead to an unhindered and unrestricted usage of digital tools (Speer 2017, p. 199)

That no single solution has been widely adopted by a majority of projects is an indication of the disparate nature of the desires of those producing scholarly digital editions, the strength of the 'not invented here syndrome', and the limitations of the existing software. But even when the software is available, it often does not meet the needs of those outside the specific project because it was created with very specific and often fragile approaches to the editorial endeavour.

> Let us state clearly that the described issues are not due to the fact that the implementations of the tools are incomplete. The root cause lies, instead, in the fragile theoretical foundations upon which these tools are built. (Barabucci and Fischer, p. 50)

Editors of scholarly digital editions should not be distracted by the lack of single cohesive solutions to the creation, annotation, and publication of digital editions. Instead, a de facto community-based solution should be created to meet their needs. Scholarly digital editions and the solutions that support them must learn from the history of the print edition and fully exploit the digital medium through which they are expressed.

5 Conclusion

The use of open international standards for the creation of scholarly digital editions is necessary if the resources spent on them are not to be squandered. The TEI does a good job in being flexible and customisable to individual scholarly digital editing projects. Where feasible, it is better to use this at least as a storage and preservation format than to invent even more standards. The search for better serialization formats and the reinvention of encoding formats, while an important endeavour for markup theorists, is a distraction pragmatic digital editors should ignore. Similarly, the creation of openly available resources is the future for any truly collaborative international research, and editors should adopt common legal solutions, such as creative commons, so as not to be distracted by unnecessary legal intricacies. The publishing of scholarly digital editions and the distractions of concerns about a particular presentation view of the edition should be discarded in favour of the adoption of consistent digital editorial publication methods where feasible. However, more work needs to be undertaken on the production of generalised software for editing tasks that truly supports the flexibility of out-of-line and standoff markup technologies within existing standards like the TEI. However,

this work should not be undertaken by scholarly editorial projects, who are the customers in this enterprise. I would contend that large amounts of public funding should not be set aside merely for the open publication of digital editions, as there is no technological barrier to achieving this if a consortium of projects desires to do so. (I would want to see that funding used to create the generalised infrastructure proposed above that such projects would use.) Much as the TEI has become the de facto standard for the data of scholarly digital editions, it is time for software infrastructures to be adopted for a more consistent environment that benefits all. As editors of scholarly digital editions, we need to have some understanding of the mechanisms of the production of our editions without being distracted by the underlying technological issues, unless we are interested this distraction.

References

Barabucci, G., & Fischer, F. (2017). The formalization of textual criticism: Bridging the gap between automated collation and edited critical texts. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing: Papers presented at the Dixit conferences in the Hague, Cologne, and Antwerp* (pp. 47–54). Leiden: Sidestone Press.

Barabucci, G., Spadini, E., & Turska, M. (2017). Data vs. presentation: What is the core of a scholarly digital edition? In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing: Papers presented at the Dixit conferences in The Hague, Cologne, and Antwerp* (pp. 37–46). Leiden: Sidestone Press.

Ciotti, F. (2017). Towards a new realism for digital textuality. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing: Papers presented at the Dixit conferences in the Hague, Cologne, and Antwerp* (pp. 85–90). Leiden: Sidestone Press.

Cummings, J. (2008). The text encoding initiative and the study of literature. In S. Schreibman & R. Siemens (Eds.), *A companion to digital literary studies* (pp. 451–476). Oxford: Blackwell.

Cummings, J. (2014). The compromises and flexibility of TEI customisation. In C. Mills, M. Pidd, & E. Ward (Eds.), *Proceedings of the Digital Humanities Congress 2012. Studies in the digital humanities*. Sheffield: HRI Online Publications. https://www.dhi.ac.uk/openbook/chapter/dhc2012-cummings. Accessed 13 May 2019.

Cummings, J. (2018) A world of difference: Myths and misconceptions about the TEI. Digital Scholarship in the Humanities.

De Roure, D., & Willcox, P. (2017). Experimental humanities: An adventure with Lovelace and Babbage. In *2017 IEEE 13th international conference on eScience 978-1-5386-2686-3/17* (pp. 194–201). https://doi.org/10.1109/eScience.2017.32.

Pierazzo, E. (2015). *Digital scholarly editing: Theories, models and methods*. Farnham: Ashgate.

Sahle, P. (2016). What is a scholarly digital edition? In M. J. Driscoll & E. Pierazzo (Eds.), *Digital scholarly editing: Theories and practices* (pp.19-40). Cambridge: Open Book Publishers. https://www.openbookpublishers.com/htmlreader/978-1-78374-238-7/ch2.xhtml. Accessed 13 May 2019.

Shillingsburg, P. (2017). Enduring distinctions in textual studies. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing: Papers presented at the Dixit conferences in the Hague, Cologne, and Antwerp* (pp. 187–190). Leiden: Sidestone Press.

Speed Kjeldsen, A. (2017). Reproducible editions. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing:*

Papers presented at the Dixit conferences in the Hague, Cologne, and Antwerp (pp. 135–140). Leiden: Sidestone Press, 2017.

Speer, A. (2017). Blind spots of digital editions. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing: Papers presented at the Dixit conferences in the Hague, Cologne, and Antwerp* (pp. 191–200). Leiden: Sidestone Press.

Turska, M. (2017). TEI simple processing model: An abstraction layer for XML processing. In P. Boot, A. Cappellotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E. Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing: Papers presented at the Dixit conferences in the Hague, Cologne, and Antwerp* (pp. 361–364). Leiden: Sidestone Press.

Turska, M., Cummings, J., & Rahtz, S.P.Q. (2016). Challenging the myth of presentation in digital editions, *jTEI*, 2016. http://journals.openedition.org/jtei/1453. Accessed 13 May 2019.

Vine, A., & Verweij, S. (2012). Digitizing non-linear texts in TEI P5: The case of the early modern reversed manuscript. In B. Nelson & M. Terras (Eds.), *Digitizing medieval and early modern material culture* (pp. 113-136). , *New Technologies in Medieval Renaissance Studies* (Vol. 3). Toronto: Iter.