ADVANCES IN DIGITAL SCHOLARLY EDITING

ADVANCES IN DIGITAL SCHOLARLY EDITING

Sidestone Press

ADVANCES IN DIGITAL SCHOLARLY EDITING

PAPERS PRESENTED AT THE DIXIT CONFERENCES IN THE HAGUE, COLOGNE, AND ANTWERP

edited by
PETER BOOT
ANNA CAPPELLOTTO
WOUT DILLEN
FRANZ FISCHER
AODHÁN KELLY
ANDREAS MERTGENS
ANNA-MARIA SICHANI
ELENA SPADINI
DIRK VAN HULLE

Published by Sidestone Press, Leiden
www.sidestone.com

Imprint: Sidestone Press

Lay-out & cover design: Sidestone Press
Cover illustration: Tessa Gengnagel

ISBN 978-90-8890-483-7 (softcover)
ISBN 978-90-8890-484-4 (hardcover)
ISBN 978-90-8890-485-1 (PDF e-book)

Contents

Data driven editing: materials, product and analysis

Linda Spinazzè,[1] *Richard Hadden*[2]

& Misha Broughton[3]

Paper presented at 'Technology, Software, Standards for the Digital Scholarly Edition' DiXiT Convention, The Hague, September 14-18, 2015.

The remediation of cultural heritage documents into a digital environment – particularly through the disparate but related practices of mass digitization and digital scholarly editing – has a keen focus on textual and multi-media content. However, this focus sometimes occludes the fact that, working within a digital workflow, our core material is, in fact, data. This panel seeks to explore the possibilities of a more data-driven editing practice, one that sees not only our material (digital proxies, collections information, transcriptions, and metadata) but also our resulting products (corpora, editions) and all of our intermediary stages not as text or images or content, but as data per se. In the following sections, we will seek to reconcile the ambiguity inherent to humanities inquiry with the exactitude required of digital data, asking how we can 'read' this data, and what – if anything – is our responsibility as editors to provide access not merely to the final argument of our editions, but to the data that informs it.

1 linda.spinazze@gmail.com.
2 richard.hadden@nuim.ie.
3 wbrought@uni-koeln.de.

Source material as data – Linda Spinazzè

The case study which follows is concerned with crowdsourcing digital editing. In fact an overview of *Letters of 1916* project provides an occasion to explore the particular way in which a collection of texts can be edited digitally according to Web 2.0 philosophy of sharing and collaboration.

The *Letters of 1916* is a work in progress to create an online fully searchable collection of correspondence. The aim of the project is to gather and edit letters in the period leading up and just after the 1916 Rising by engaging the 'crowd'. The collection includes private letters, business missives, official documents, postcards, greeting cards and telegrams written around the time of the Easter Rising of 1916. On 24 April 1916, Easter Monday, in Dublin a small group of Irish nationalists decided to rebel against British rule. The General Post Office (GPO) served as the headquarters, where seven members of the Council who planned the Rising declared the proclamation of Irish Republic in front of this building. Within a week, the British army quickly had suppressed the rebellion and, on 3 May it started to execute the leaders of the Rising. Even though Ireland did not gain independence until 1922, it is a common opinion that the Easter Insurrection is the moment when everything changed, it is considered by historians as a sort of 'point of no return'[4]. The *Letters of 1916* project aims to help in understanding this 'change' better creating the new collection consisting of pieces of correspondence written between the 1st of November 1915 and the 31st of October 1916. Assuming that the words present in the letters[5] are the witnesses of different aspects of the society in that particular historical period, we are aware that such a collection can open new perspectives on the events and daily life at that time.

In contrast to a more 'traditional digital collection' which tends to be linked to a physical archive stored in a library, or for example, to a specific author already studied and edited, the *Letters of 1916* project brings together images of the correspondence from many different institutions, about 20 and also from private collections[6]. So, not only is the team or experts responsible for the upload, but often members of the public[7] undertake the process of uploading their family letters from scratch thanks to the platform created by the *Letters of 1916* team. In terms of crowdsourcing, the native platform of the project utilizes the Omeka software[8] alongside some plugins which carry out specific functionality.

---

4 The bibliography is huge; for a general reference about the subject we can just refer to one of the most recent (McGarry 2010) (see also the new 'Centenary edition', published in 2016).

5 It is worth to point out to this fact: «In 1914-1915, the last fiscal year during which records of letters posted were kept, approximately 192 million letters were mailed within Ireland, which works out at roughly forty-four letters per person», in Novick 1999: 350.

6 Here the list of institutions which have allowed us to include images of letters and photographs in the *Letters of 1916* project: http://letters1916.maynoothuniversity.ie/learn/index.php/collaborate/institutions. Accessed 6.10.2017.

7 For a critical perspective on the gap between crowdsourcing and mission and values of cultural heritage organisations see Ridge 2014.

8 The *Letters of 1916* uses the *Omeka* 1. 5. 3 (http://omeka.org/); the transcription interface is based on the *Scripto* plugin http://scripto.org/; see forward for other add-ons. Sites accessed 6.10.2017.

Because of the crowdsourcing nature of the project is particular interested in the large participation of *amateurs*[9] and in the creation of a 'corpus that never was'[10], in such a digital collection the contact with the 'original text-bearing' objects is particularly fleeting. Precisely because the workflow quickly moves away from the material objects in favour of the digital data, the conversion from the 'material' to the 'digital' has to be particularly accurate. After taking the high resolution images the user has to upload the digital item via a form which helps to simultaneously create some basic metadata (such as title, creator, place[11]). In filling in this form, it should be clear, especially to the non-specialist that in this first phase they are contributing in creating a basic digital storage, that it is not a plain silo of photographs, but an actual database of items – of actual structured digital items. The high resolution digital images are surrogates of the original letters, and the archival of this digital material guarantees its curation and preservation. This is especially true in the case of certain private collections which often are stored inappropriately in their physical form (see Figure 1). After the uploading and structuring of the metadata, the new items are quickly revised by a member of the team who makes them accessible in the transcription area of the site, just ready for the next phase of the *Letters of 1916* workflow of editing.

Figure 1: Sometimes the private collections are not stored in the appropriate way, other times the letters are hidden away for decades. Here is an example of a metal biscuit tin filled with old letters and found in an attic by one of the contributors of the Letters of 1916 collection. Photo: ©Letters of 1916 (Kildare Launch, Maynooth University, May 2014).

9 For a definition of *amateur* inside the crowdsourcing philosophy, Owens 2014.
10 Paraphrasing the 'text that never was'; see Greetham 1999, or more recently 2014.
11 See form at: http://letters1916. maynoothuniversity.ie/images/HowToUploadALetter.pdf. Accessed 6.10.2017.

In order to also provide the community of users from the public audience with an introduction to TEI mark-up and more generally to a digital scholarly editing workflow, at this point the *Letters of 1916* project does not require a simple transcription but rather a 'structured' one, which incorporates basic encoding too. To combine the requirement of a TEI structured transcribed text the plain text-field by Omeka/Scripto is equipped with an adapted version of the 'Bentham toolbar'[12]. This plugin serves as a method for encoding some main feature contained in the letters (the features which provide information about the material aspect or 'semantic' details[13]).

In order to ensure that a digital scholarly edition is created from all these transcriptions, the team editors have to handle encoded information with formal errors, misunderstandings, omissions. In fact, inside the definition itself of 'edition', the accuracy assessment is one of the basic requirements. The question of how to dynamically proof the accuracy of the tagging remains. When the error is not about the 'well-formedness' of the mark-up, but is a real misinterpretation of the tag or a completely wrong reading, it is almost unpredictable. Is there a dynamical solution?

At the moment, we are concerned with figuring out a solution for proofing this kind of collection 'driven' by digital data on its own is unrealistic. So, considering that the human checking is necessary, the question is: how can we combine automated and manual editing effectively? And more importantly, can we just consider an edition a plain transcription, even if it is well structured and well formed?

The edition as data – Misha Broughton

If the resources of digital text editing are data, it is important to also note that its output is equally data. While this may seem self-evident in theory, it is a point easily forgotten in practice, where the aim of editing is so commonly the production of an edition. However, while this goal is certainly natural, our concept of what an edition is, or can be, is still limited to a print concept, or as Patrick Sahle would have it, 'the print paradigm.'[14]

What is the nature of a critical edition, print or otherwise? The MLA *Guidelines for Editors of Scholarly Editions* states that its 'basic task is to present a reliable text,'[15] with – I argue – a silent emphasis on the singular 'a.' Editions are composed, however, from multiple document witnesses and, often, a contentious transmission history. If this is the case, then editing – or, at least, editing to the edition – is a process of ablation, of whittling away at textual extraneities that do not support the privileged reading of that particular edition. And yet the text, the holographic, syncretic whole that we aim to represent through our endeavors, is surely bigger

12 TEIToolbar from Transcribe Bentham project: http://www.ucl.ac.uk/transcribe-bentham/. Accessed 6.10.2017.

13 See explanation at: http://letters1916.maynoothuniversity.ie/images/ProofingXMLGuidelines.pdf. Accessed 6.10.2017.

14 Sahle, Patrick. 'about'. *A catalog of Digital Scholarly editions, v 3.0, snapshot 2008*. Accessed 6.10.2017.

15 'Guidelines for Editors of Scholarly Editions'. Modern Language Association. Accessed 16.1.2016

than any single reading, just as it is bigger than any single document witness that attests it. And yet, in effect, this is what the edition conceived under such these terms can not help but be: another document witness in the text's transmission history, albeit one authorized by an expert scholarly editor.

In a previous technological environment, the edition *could* be nothing but. Limited by the same constraints of the page space in which previous document witnesses were compiled, the print scholarly edition had but few methods (e.g. the critical apparatus, paratext, footnotes, marginalia) to do anything besides document the textual history largely as it was received. Though the advent of digital media technologies has brought many new affordances to the display, publication, and discoverability of scholarly editions, it brought little – if any – reconsideration of what the edition is. In our practical commonplaces, like the MLA Guideline cited above, we have re-inscribed the familiar shape of the print-document edition in the digital: a single, reliable text with a scattering of apparati to record the more important variation. Perhaps more importantly, though, this understanding of the edition as a print-like document also has influenced the logical model which informs our most prominent data model, TEI/XML. The Ordered Hierarchy of Content Object model of text,[16] which informs the XML markup language, was proposed as a method of organizing text data *specifically* for its similarity to print documents. For as far as we have come, technologically, we have arrived at little more than print documents migrated whole-cloth from pages to screens.

It is important to remember, though, that while the document-like (or text-like, if you will) methods of organizing data are a very venerable and mature technology, they are still only one possible method, and one far from perfect for all applications. While the form is familiar to editors for its similarity to the witnesses it collates, it is this very similarity that limits it dimensionally, making compositing of various textual features difficult, at best. And while it is certainly needful for the presentation of the 'reliable (reading) text' aforementioned, considerations of *presentation* and of *encoding* need not (and should not) be confused.[17]

These concerns would be entirely academic, of course, if not for the fact that the practice already is running afoul of all-too practical consequences of this document-like approach to encoding. The problems of hierarchy overlap (Renear *et al.* 1996) and limited data interoperability (Schmidt 2014) are, I argue, not only related but both stem from the same dimensional limitations imposed on digital textual encoding by a print-centric conceptual model and encoding scheme. It is all but impossible to fully represent a topographically complex three dimensional object in a two dimensional plane. How much more difficult must it be, then, to represent the layered complexities of multiple document witnesses – each at least a two-dimensional page space and some with their own collections of dimensionally extending commentaries, emendations, and apparati – in a conceptual space utilizing only the same set of dimensions and functions? The TEI community

16 See De Rose *et al.* 1997.

17 At present, I will leave the definition of 'reading' as the rather conservative one of reading linearly page-by-page (or its screen equivalent), though discussion certainly is warranted of the relation of the DSE to Moretti's 'Distant Reading,' Bloome's 'unruly reading,' and the large corpus of early work on hypertext reading practices.

has done wonders adapting to the shortcomings of the model, sometimes at the expense of the underlying logic of its assumptions.

For all of that, though, I predict that these problems – and more like them that we have not yet considered – will multiply far beyond our ability to make allowances for them under existing practice. Our understanding of text in its material form has been expanded by our years of work transmediating its content to the digital and it is this expanded understanding of (often printed) texts which leads to the desire to encode features or sets of features which challenge the underlying assumptions of our practice. What is needed is not more allowances in the current technologies to 'make it do what we want, ' but back trenching, a reconsideration of the logical model by which we encode that allows native expression of the dimensional complexity we have come to understand in text. In short, we must *display* our editions, but *encode* our data. Such an encoding would fulfill the requirements of what Elena Pierazzo has called the 'paradigmatic edition, ' an encoding that provides 'many alternative options for the same string of text in a nonlinear way,' (Pierazzo 2014) though perhaps going a step further to allow even different strings of text, different readings, different *editions*, in the same encoding.

However, if our encoding is not organized along the familiar and readily legible modes of the text document, how should it be organized? Even the most basic database or markup language provides a wealth of methods to interconnect related data, with features allowing for the relational or associative linking of content. My own proposal, currently under development in my doctoral dissertation at the

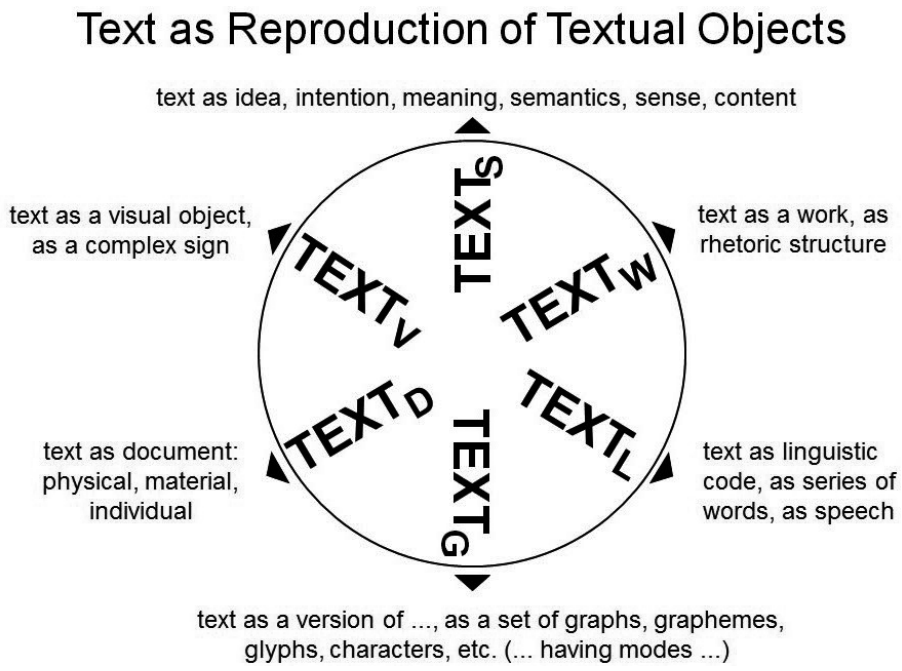

Figure 2: Patrick Sahle's Wheel of the Texts. Figure used by permission.

University of Cologne, is based on Patrick Sahle's 'Wheel of the Text' (Figure 2, seeing 'the text' as a locus of interpretation of discrete text-bearing objects, with features and values for observation dependent on the perspective of the observer. However, while Sahle's wheel indicates an equality of these perspectives, none alone sufficient to fully account for the features of other perspectives, my own approach sees a chain of *necessity* from the most document-centric perspective to the more text-centric perspectives. For instance, if we can not say that observations of the document-centric perspective are entirely sufficient to justify our observation of a linguistic code in the same text or the linguistic code of the work, we must say that the presence of a document is necessary to claim that a linguistic code is being employed and that the use of a linguistic code is necessary to the presence of a textual work. While our object of inquiry, then, is the abstract text, the 'communicative act' (Robinson) that is embodied merely in documents, we must acknowledge the presence of instantiated, embodying documents to make any claim that such an abstract text exists. Counterintuitively, perhaps, I propose that the best way to free this abstract text from the confines of a document-centric organizing mode is precisely by encoding data directly observed from documents and linking successive layers of sinterpretative perspective atop it.

The advantage of this system is three-fold: first, by separating layers of interpretative perspective, it provides a measure of vertical independence, separating the various observational perspectives from Sahle's wheel and thus avoiding hierarchy overlap common when trying to encode such features together in an in-line transcription. Second, it provides a measure of horizontal independence, allowing for the encoding of disparate editorial perspectives or features clusters in distinct groupings without reference to other perspectives that reference the same base (see Feature A/B/C in Figure 3). Third, though not represented in Figure 3, this approach allows for the encoding of *depth*, allowing even for the encoding of contradictory or mutually exclusive interpretations from the same editorial perspective (e.g. disagreeing transcriptions of the same region, differing tagging of prosody of the same transcription, etc.).

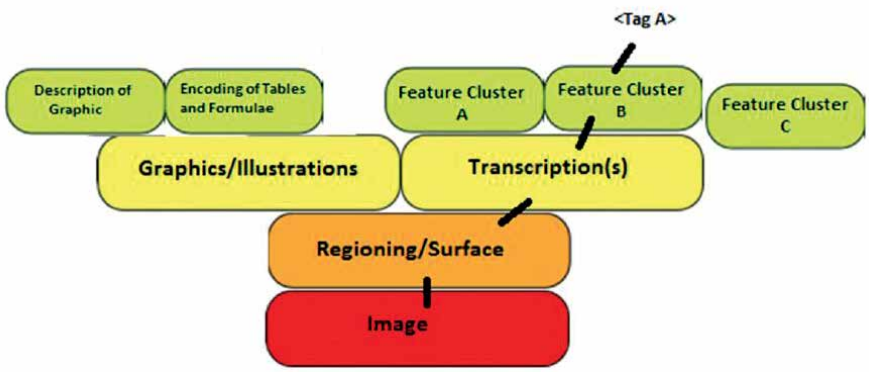

Figure 3: A layered approach to text modelling.

Editing for data – Richard Hadden

Digital Scholarly Editions have tended to follow a particular paradigm from their printed days, albeit a largely invisible one. This is the tendency to conflate the results of the actions of editing and the form of the edition. Such a view is natural enough: the text of the edition in print is bound very tightly to the material form in print, with specific adaptations designed to represent particular kinds of texts. The critical apparatus, for instance, may be seen as a way of representing the text of multiple witnesses, layered upon a base text. In this case, the presentation is coupled very strongly to the form of the edition.

It is possible – even arguably necessary – to consider a digital edition in the same light. If we view text as anything other than a pure abstraction, it is clear that the representation of text on screen is as vital to a reader's understanding as it is in print. Following Patrick Sahle's theory of a 'pluralistic' understanding of text (modelled as a 'wheel of text'), one can argue that the text of any edition – which is to say, that viewed by the reader – is the 'totality' of this plurality. If we ask, therefore, what is the edition in a digital, it is the text (as encoded) combined with its presentation.

Such a perspective, while valid, ignores the fundamental difference between a digital and a print edition: notably, that the text of the edition is stored as an abstraction, and only rendered in some form of interface on-demand by a reader's computer (or other device). As a result, there is inherently a disconnect between the edited text as an abstraction – data – and the edited text as rendered. What I will argue is that, though theoretically it is impossible to fully comprehend an edition and the text represented therein through only one aspect of its plurality, pragmatically at least we, as editors, should concentrate more forcefully on editing data, as that it what fundamentally drives an edition. To do otherwise is to ignore a fundamental reality of text in digital form, and, indeed, to deprive ourselves of a major benefit of the medium.

This is not to suggest that such a view is not already partially applicable. Since the bad old days of <font> tags and inline styling in HTML came to an end, web development already enforces a degree of separation of style (described using CSS) and data (encoded using HTML). With the advent of HTML5 and a greater range of semantically-meaningful, rather than presentationally-oriented, elements, such a divide is even greater. This is one abstraction of the text, albeit one only applicable to a web browser. Using TEI-XML to encode texts, a standard practice, increases this divide. We are able to describe text in much less generic ways (compared to HTML5). Further processing, using XSLT, for instance, to transform XML into HTML, to which CSS then can be added, which then can be rendered by a browser.

I would argue, however, that despite these separations of concerns, there is still too great a tendency to consider TEI encoding as merely the first step towards building an edition. Even though the actual building of an edition website may be the responsibility of someone with greater expertise – i.e. a project may employ editors to encode text and a web designer to build the site – too great a focus is placed, at the stage of encoding, on the end product. That is to say, we are too ready to abandon a greater level of expression (TEI) in order to produce a website

with some text on it. Altering the focus towards editing data – as an end in and of itself – rather than editing towards a final product, seems a way to avoid what, ultimately, appears to be work for no end.

In a recent paper Peter Shillingsburg argues that producing digital editions is too complicated, compared to the days when he could edit text and typeset the final edition (using LaTeX) all by himself. Now a greater range of expertise (web design, data processing, not to mention arcane procedures of server configuration) is required (Shillingsburg 2015). This is true as far as it goes, but to build, as he suggests, a system that would take care of everything is to lose sight of the benefits of the separation of data and presentation. There is no reason per se that he could not encode his edition directly into HTML – after all, if one can learn LaTeX, learning another relatively simple markup language and vocabulary cannot be too difficult; the two are broadly analogous. Such an approach, however, would involve throwing away a degree of abstraction, and ultimately constraini the use (or re-use, or elaboration) of the edited text.

If we edit towards data rather than an edition, we run into at least some conceptual problems, not least: what exactly are we making? I would argue that a TEI document is, in itself, an edition, with at least equal status to a beautifully-rendered and functioning website. After all, it is (for me at least) as easy to 'read' a TEI-XML encoding of a text as to understand the arcane symbols employed in, say, a typical printed critical edition. At the same time, it must have a degree of primacy: a website built by transforming XML into a web-based interface is clearly derivative. As a result, it is not possible to completely disregard the end product when editing the data. Encoded data is not a neutral, 'pure abstraction' – as can be said of any form of editing – and neither is it total. If we wish, therefore, to produce a certain kind of edition, it is necessary that enough detail is encoded to make this possible. But we should aim for a form of neutrality – or, better put, a degree of agnosticism with regards to the final product. This is, after all, what the TEI does, by inviting us to describe the text of a work or document rather than its endpoint.

The great benefit of this is both in the re-use and further elaboration of data. Re-use is, of course, one of the fundamental points of the TEI: by providing a set vocabulary, it should be possible for the data created by one project to be reused in another (many digital scholarly editions make their TEI data available for this purpose). However, this potential is seldom realised, I would suggest chiefly because even TEI encoding is geared in too great a degree towards its end transformation into a HTML. As projects necessarily are limited in scope, this is hardly surprising: editors encode as much information as they need, and in the way that they need it, to produce the kind of edition they aim to make.

An approach to circumventing these obvious restrictions is to treat editing as a form of 'progressive enhancement' (to borrow a web design term) of data: editing is treated as a modular and incremental workflow, where the objective is to elaborate upon the data as it exists, so far as this might allow new ends to be achieved. Such tasks may be carried out by those working on the initial project, or (re)users of the data further down the line. Moreover, elements of data already encoded may be used algorithmically by automated processes and scripts.

Application to the *Letters of 1916* project

Linda Spinazzè already has described some of the workflow of the *Letters of 1916* project. I aim here to outline how the principles of this data-centric approach to editing have been, are being, and (I hope) will be employed.

The first aspect to note is the very clear delineation of phases in the project workflow, in terms of activity and personnel involved (this is one distinction from Shillingsburg's desire for end-to-end production of his own editions). Part of this is, of course, necessary as a result of the crowdsourcing nature of the project. The first stage is the capturing of digital images of letters en masse, by project team members visiting archives (I should say 'principally by', as some, though a small minority of images are uploaded directly by contributors). The letter images are uploaded to be transcribed by the 'crowd', who also add a limited number of TEI-XML tags (not necessarily accurately). At this stage, we have data that arguably can be distributed as an edition – albeit not a very good one.

The transcribed data then is extracted from the crowd-transcription environment (Omeka) and enhanced using a range of automated scripts written in Python. The text, which is stored as individual pages in Omeka, is joined into a single TEI document for each letter; metadata added to the letter in Omeka is used to construct TEI elements such as <correspDesc> and <revisionDesc>; and further semantic information is added automatically based on the limited encoding already completed.

The 'compiled' TEI documents then are sent to be proofed for text and markup by project team members using a purpose-built, web-based editing tool, which tracks edits to the documents using automatic commits to a git repository. At this stage, further data is added, such as normalising names of senders and receivers from a canonical list. The workflow thus far is strongly data-focused, with effort geared towards producing accurate and valid TEI encoding; also, each stage can be viewed as a progressive elaboration of semantic information over the 'base' transcription.

At this stage, it is necessary to consider the plans for the forthcoming edition. This has been designed by another project member, and is designed as a full-text searchable edition, with a provision of the full letter-text and side-by-side page and image views. This new site has been designed to store text as pre-rendered HTML. However, it uses only some of the encoded TEI elements. As such, it can be seen as consumer of the edition data, while the focus of editing remains on the data itself. The new edition's importing process is adapted to the data, rather than the other way round. By using TEI, much of this adaptation can be foreseen; though where this is not the case (for instance, the use of specific elements to indicate document structure), it is for the importing scripts to adapt. This being the case, and following the argument made thus far, it can be seen as one consumer among many potential consumers: it satisfies one potential use of the data, but by no means all of them.

As a result of such an approach, it is possible to envisage further uses for this data, both in terms of alternative editions, possibly using data-analysis techniques such as topic modelling, and, more importantly perhaps, further elaboration of the base data: thus far, encoding has steered clear of more graphical features of

the document (such as official stamps) which could be added later; the marked-up addresses can be used to add geolocation data. Moreover, data already marked up could be used to train classifiers to automate the markup of the next 'generation' of letters to pass through the workflow: work on this has been attempted already, using decision-tree classifiers to identify lines in the text with particular significance, such as addresses and dates. Such an approach also can be used to identify named entities within the text body, which currently are not marked up.

The obvious downside to such a data-driven approach is consistency. If the underlying data of an edition is constantly – and actively – changed, what are the implications of this for a scholarly edition, of which academic rigour demands stability? To allow versions to exist concurrently, the project uses two approaches. Firstly, the data is stored in a git repository, which tracks all changes to documents, and also allows the data to be cloned, edited and re-merged as necessary. Further to this, the TEI markup makes extensive use of the <revisionDesc> element: each change to each page made by transcribers is logged as a revision, with the text of each 'version' stored in an XML comment (this is necessary as the transcribed text is not necessarily valid XML) for future reference. Each scripting operation logs its effect in a revision as well.

As with the TEI-encoded text itself, this revision data is not oriented towards a particular use: instead, it is simply made available to potential consumers to make use of as required.

```
      <revisionDesc>

         <change when="2014-05-25T19:26:34" who="#Badzmiek"> Page 2637 modified:
            <!-- Holy Ghost Missionary College
Kimmage Manor,
Dublin 1 July 1916

Dear Monsignor Hogan,
I cannot but accede to your request to conduct the Retreat for the opening of the coming Scholast
Please let me have a copy of the Regulations usually followed, and I should, also, like to have a
With every best wish,
Faithfully yours,
John J. Murphy C.S.S.P.
The Right Rev. J. F. Hogan, D.D.,
President, &c. -->
         </change>

         <change when="2014-06-28T15:24:21" who="#Badzmiek"> Page 2637 modified:
            <!-- <address>Holy Ghost
Missionary College.
Kimmage Manor,
Dublin</address> <date> 1 July 1916</date>

<salute>Dear Monsignor Hogan,</salute>

<p>I cannot but accede to
your request to conduct the Retreat
for the opening of the coming Scholastic
Year. So you may count on me for that
purpose.</p>
<p>Please let me have a copy of
the Regulations usually followed, and
```

Figure 4: Illustration of the revisionDesc in the Letters of 1916 TEI files.

This final point illustrates the pitfall of this data-centric approach. With each phase divorced from the next, and with a greatly lessened possibility for revision of a previous process at a later stage, rigour at each point is essential. Each elaboration of data is built upon a pre-existing foundation, which must be secure. At the same time, the benefits for ongoing usefulness of editorial activity make such an approach worthwhile.

References

DeRose, Steven J., David G. Durand, Elli Mylonas and Allen Renear. 1997. 'What is Text, Really?' ACM SIGDOC Asterisk Journal of Computer Documentation.

Greetham, David C. 1999. *Theories of the Text*. London: Oxford University Press.

Greetham David C. 2014. "Retexting the Barthesian Text in Textual Studies." In *The Conversant. The Renaissance of Roland Barthes, a Special Issue*, edited by Alex Wermer-Colan. (http://theconversant.org/?p=7880).

McGarry, Fearghal. 2010. The Rising: Easter 1916. Oxford: Oxford University Press.

Novick, Ben. 1999. "Postal censorship in Ireland, 1914-1916." *Irish Historical Studies* 31(123): 343-356.

Renear, Allen, Elli Mylonas, and David Durand. 1996. 'Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies.' In Research in Humanities Computing. Oxford University Press.

Ridge, Mia. 2014. *Crowdsourcing our Cultural Heritage*. Farnham-Burlington: Ashgate.

Owens, Trevor. 2014. "Making Crowdsourcing Compatible with the Missions and Values of cultural Heritage Organisations." in *Crowdsourcing our Cultural Heritage*, edited by M. Ridge. Farnham-Burlington: Farnham-Burlington: Ashgate: 269-280.

Pierazzo, Elena. 2014. 'Digital Documentary Editions and the Others.' In Scholarly Editing: The Annual of the Association for Documentary Editing 35.

Robinson, Peter. 'The Concept of the Work in the Digital Age'. Pre-publication draft.

Schmidt, Desmond. 2014. 'Towards an Interoperable Digital Scholarly Edition'. Journal of the Text Encoding Initiative 7.