11.

The scholarly data edition: publishing big data in the twenty-first century

Gábor Mihály Tóth

In the last two decades big textual data sets in the humanities have become increasingly more available (Schiuma and Carlucci 2018). As a result of large-scale digitisation projects by libraries and archives, we can expect that in the future even more truly big textual data sets will be released to the public. This trend raises a key question that is highly relevant for the future of digital editions:

> How do we facilitate access to and exploration of big textual data in the form of *scholarly digital editions*?

As scholarship has often pointed out, the simple release of data in the form of plain text is not a scholarly edition (Sahle 2016). Similarly, websites and online archives that make millions of texts searchable cannot be considered scholarly editions. All this raises another question:

> How should we edit and publish big textual data in a *scholarly manner*?

Digital humanities scholarship has elaborated a set of editorial principles that distinguish straightforward text releases from scholarly digital editions (Robinson 2002). However, as I will point out in the first section of this short essay, the application of these principles with truly big data is challenging; the traditional genre of scholarly digital edition can be applied to publish relatively small data sets (such as diaries, letters and poems of a single author or of a small group of authors) but it is hardly applicable with truly big data. We therefore need a new genre that I name *scholarly data edition*. In the second section of my essay, I will tentatively elaborate on this new genre through some of the editorial procedures the publication of big textual data involves. At the same time, I will attempt to establish a continuity between the scholarly digital edition and the scholarly data edition by re-using and redefining the editorial principles traditionally associated with scholarly digital editions. In conclusion, I will further discuss why the publication of scholarly data editions is crucial in the twenty-first century and how this new type of edition can further knowledge and scholarship.

Ideas and principles outlined throughout this essay are based on my own practical experience of editing and publishing an unprecedentedly large corpus (circa 60 million words) of nearly 3,000 Holocaust testimonies from three major US collections (Tóth 2021).

*

The hallmark of big data is its sheer size. We can measure the size of a textual data set in terms of the number of single documents it incorporates or the number of words (or, technically speaking, tokens) it includes. A truly big textual data set can easily contain tens of millions of tokens. It is obvious that to edit and publish this amount of data one needs recourse to the power of computers; yet, any human intervention to curate a big textual data set can be only very limited. For instance, traditional scholarly digital editions are often based on the manual transcription of documents by human experts. By contrast, the data underlying a data edition can be obtained by means of computational tools such as optical character recognition (OCR) tools. Another example of human intervention in the creation of digital editions is annotation (Barbera et al. 2013). As part of this process, human experts ascribe topics, keywords, names and places to different structural units of texts such as chapters and paragraphs. Sometimes human intervention in the process of digital editing aims to organise and structure texts and textual collections. In the case of big data, manual annotation is not feasible; neither can humans structure and organise millions of single documents. Instead, editorial teams need to rely on machine learning and data mining algorithms to annotate and organise documents in semi-automatic or completely automatic ways. Nonetheless, the application of computing tools to cope with the sheer size of a very big document collection goes against two key editorial principles traditionally associated with scholarly digital editions.

First, the traditional principle of critical accuracy is not attainable when editing and publishing big textual data. Jonas Carlquist offers a good description of what critical accuracy involves: 'the transcribed text must attain the usual levels of critical accuracy, meaning that the edition needs to follow diplomatic standards and be the product of expert work' (Carlquist 2004, 115, cited by Franzini et al. 2016).- Critical accuracy, for example, means that canonical names of places and persons ascribed to texts of an edition as part of annotation must be absolutely correct and the result of experts' research. Critical accuracy also means that a digital edition is a reliable and authoritative digital representation of a given source material. However, the use of machine learning and data mining algorithms is at odds with critical accuracy; these algorithms always feature a certain degree of inaccuracy. For example, sometimes a named entity recogniser, a specialised algorithm used to extract or mark names, places and entities in texts, will correctly identify a person; sometimes it will fail and treat a place as a person. In short, the degree of critical accuracy expected from humans cannot be expected from machines. As a result, it is questionable how we can apply the traditional principle of critical accuracy when editing and publishing big data.

Second, the principle of critical examination of texts is equally unattainable when editing big textual data. Patrich Sahle offers a succinct description of how critical examination of texts and digital editing are related:

> Reproduction of documents without critical examination is not scholarly editing. A facsimile is not a scholarly edition. A scholarly edition is marked by the critical approach towards the documents and the texts they contain (Sahle 2020).

Critical examination means the contextualisation of texts, which includes the study of their origins, meanings, purposes and so on. With millions of documents in a big textual data set, the contextualisation of each single document is not feasible. Contextualisation also involves the development of critical apparatus (that is, footnotes, comments, explanations and so on) that editors attach to a single text with the purpose of explaining its cultural, social or historical background. Again, with millions of single documents, this type of contextualisation is not a realistic undertaking. Generally speaking, the principle of critical examination addressing the micro level of textuality can hardly be applied with big textual data sets.

Despite their infeasibility, we cannot entirely give up on these two principles. Sahle, Robinson and other theoreticians of traditional digital editions are right, claiming that a simple digital reproduction of texts does not meet the standard of scholarly editing. What do critical accuracy and critical examination mean in the context of a data edition? To answer this question, I will outline some of the editorial practices that the preparation of a data edition involves.

Preprocessing

As base data, data editions rely on digitally born or already digitised materials such as OCR-ed texts. The base data often comes in raw formats such as XML, JSON, CSV or plain text. The very first editorial step in the process of creating a data edition is the preprocessing of raw data. This might include a number of substeps.

First, raw data is often unstructured; that is, paragraphs, chapters, titles and other structural elements of texts are not identified and separated. Hence, as part of the preprocessing stage, editors of data editions need to apply computing tools to distinguish structural units of texts. Second, if raw data was computationally generated with tools such as OCR or automatic voice recognition, it might contain a high number of misspellings and other types of errors. In case of erroneous base data, the editor has to accomplish the task of data correction. If the base data was generated by different projects relying on different computational tools, the editor has to normalise it. Third, the base data sometimes includes not only texts in raw format but also metadata, that is, information (date and place of compilation, name of the author and so on) about the content of texts. Metadata often requires normalisation and harmonisation, especially if it was provided by different institutions. Fourth, the preprocessing of raw data might involve the computer-assisted annotation of texts. On the one hand, this can take place in the form of linguistic annotation. As part of this process, computational tools are used to separate texts into words as distinct units (tokenisation); furthermore, computational tools are applied to identify dictionary forms of words (lemmatisation) and to ascribe grammatical categories to each word (part-of-speech tagging). On the other hand, the computer-assisted annotation of texts often aims to recognise and mark specific types of words such as names and places.

Both the principle of critical examination and the principle of critical accuracy can be meaningfully applied throughout the process of preprocessing. In order to cope with the challenges (widespread presence of errors, lack of normalisation and so on) that a raw data set poses, the editor has to examine the data and survey the possible errors and variations. In this context, critical examination means the systematic and comprehensive survey of a data set with the purpose of discovering its shortcomings and deficiencies. The principle of critical accuracy in turn means the informed selection of suitable computational tools that can efficiently address these deficiencies and improve the quality of the base data. As part of the informed selection, the editor of a data edition is expected to run tests and check the performance of the selected computational tools. Finally, preprocessing is an editorial practice accomplished with critical accuracy if it incorporates two further principles: transparency and reproducibility. The editor has to be transparent about the computational procedures he or she applied and the entire process of preprocessing has to be reproducible.

Transparency can be achieved by means of various measures. The first, and perhaps the most important, measure is the thorough documentation of the code used to process a raw data set; this must include the code itself, as well as the publication of the code in open-access archives and repositories such as Github and Bitbucket. The second measure that can assure transparency is the plain explanation of how the code used to process a given data set works. This must be comprehensible to nonprofessionals; hence it is different from the documentation. Third, an editor of a data edition is transparent if he or she documents and explains the blind spots of the algorithmic solutions applied throughout the data processing. For instance, suppose there is a large textual collection containing a large number of historical place names. The capacity of a named entity recogniser to identify historical place names is limited, which is therefore an inevitable blind spot. In brief, transparency means unlocking the black box of algorithmic procedures and making these procedures accessible to a lay audience.

Discovering and presenting hidden layers of textuality

Big data is featured not only by its sheer size but also by the impossibility to explore it as a whole. One can read a novel or a collection of poems from the beginning to the end and study it as a whole. For example, one can explore connections between different paragraphs and follow how a common theme such as love is developing through a novel. But one cannot read millions of documents and explore connections. Generally speaking, connections between texts in a big textual collection are invisible and resist human exploration; they are thus part of hidden layers of textuality. I contend that the goal of a data edition is to facilitate the holistic exploration of a large textual universe, including the hidden layers of textuality; the task of the editor is to discover these hidden layers and make them accessible to the readers. We can further explore the difference between hidden and visible layers of textuality through the following illustrative examples.

Suppose that we have a collection of approximately 100,000 lyrics. We can hypothesise that there are leitmotifs (recurrent and common themes such as love, farewell, death and so on) connecting the songs in this collection. How can we explore these leitmotifs? A simple text search would not help much. Word search finds texts where a given word occurs; it thus uncovers the visible layer of textuality. But common themes connecting texts in a collection are often expressed metaphorically. Furthermore, songwriters use a great variety of vocabulary to describe a common theme. Because they cannot be retrieved with an explicit word search, leitmotifs belong to the hidden layers of textuality. To make leitmotifs explorable in a data edition, the editor can apply topic modelling (Lafferty and Blei 2009). This is a text and data mining algorithm that explores common themes in a collection; it is also a tool to assign topics to texts in a collection. The editor and his or her team can then build a specific critical apparatus that renders the result of topic modelling and make hidden connections between the lyrics explorable.

We can also view a large textual collection as a set of possibilities, which is another hidden layer of textuality to be made explorable in a data edition. Consider a hypothetical data edition of early modern printed news. One might want to explore the attributes that are ascribed to a given social group such as noblewomen. As we read the news, these attributes are in constant change. Sometimes attribute A is ascribed to noblewomen; sometimes attributes B and C are ascribed to noblewomen. The attributes used to feature noble women in the data set are not firm; they are just possibilities that are sometimes realised. Possible attributes are part of the hidden layers of textuality because just by reading a handful of texts in a big data set, we cannot explore them. To explore attributes as possibilities, we can apply collocation analysis, which is a standard method in corpus and computational linguistics (Cantos-Gómez and Almela-

Sánchez 2018). Collocation analysis shows the possible words surrounding a given word, including the likeliness these words follow the other word. Again, collocation analysis can be part of the critical apparatus supporting the exploration of the hidden layers of textuality.

As a whole, I think that the distinctive feature of a data edition should be its capacity to support the exploration of hidden layers of textuality. This feature gives rise to a number of novel editorial practices and responsibilities. First, a key editorial task is the selection of appropriate text and data mining algorithms that can uncover these hidden layers. Second, running the selected algorithms remains the responsibility of the editor. Finally, the development of a critical apparatus that presents the hidden layers of a given big textual data or makes them explorable is another pivotal editorial practice throughout the process of developing a data edition.

Just as with the preprocessing stage discussed above, the principle of critical accuracy, including transparency and reproducibility explained above, can be meaningfully applied throughout these tasks as well. Again, critical accuracy means the informed selection of algorithmic solutions complemented with the consideration of the scholarly communities' need. There is potentially an infinite number of hidden layers in a big textual data set; the editor's role is to target the ones that are important from a scholarly point of view. Generally, the exploration and the presentation of hidden layers of textuality is a scholarly activity if it is embedded in existing scholarship and if it furthers knowledge.

The contextualisation and the critical examination of texts underlying a data edition

As discussed above, both printed and digital editions are expected to include the critical examination and the contextualisation of the text or texts to be published. This can take place on the macro and the micro level of textuality. The macro level addresses the general historical, the social and the intellectual circumstances amid which a given work was born; it might also address the philological background of a text, that is, for example, the existence of manuscript variants, possibility of different readings and so on. The micro contextualisation takes place in the form of footnotes and comments attached to the single paragraphs and sentences of the running text. I contend that a data edition also requires critical examination and contextualisation; however, with a data edition this is possible only at the macro level.

Contextualisation in a data edition is similar to contextualisation in a traditional digital edition, though it needs to contain additional elements as well. As part of the contextualisation, the editor has to outline the historical and social context of the entire data set. Additionally, he or she needs to discuss how the data set was originally recorded and constructed. This discussion might address the limitations of a given data set. For instance, the editor might discuss the lacunas and losses in a data set; he or she might also discuss the errors due to shortcomings of the original data collection.

The critical examination of big data should also take place in the form of descriptive statistical analysis (Olson and Lauhoff 2019). This aims to summarise the basic characteristics of a data set by focusing on three areas: measures of central tendencies, measures of variability and distributions. These three areas have specific meanings in the context of textual data. Single texts in a large collection such as the hypothetical collection of lyrics were most probably authored in different years and in different geographical locations. By studying the distribution, the editor can discover and present how texts are spread out in space and time; he or she can show those years and places that are particularly well represented and those years and places that are underrepresented. Measures of central tendencies aims to uncover the averages: average length of single texts, average number of documents produced in a given year or in a given country or spatial location. This helps readers assess the extraordinary or the ordinary nature of a single document. For instance, Bob Marley's *Positive Vibration* consists of 214 words (lyrics downloaded from https://genius.com/Bob-marley-and-the-wailers-positive-vibration-lyrics and word count measured by https://wordcounter.net/, websites last accessed 10 November 2023). Is this a long or a short lyric? We can answer this question only if we know the average length of a reasonably large number of other lyrics. Another example is the song *Richest Man in Babylon* by the Thievery Corporation. This song contains 96 unique words or technically speaking types (lyrics downloaded from https://genius.com/Thievery-corporation-the-richest-man-in-babylon-lyrics and number of unique words counted by https://wordcounter.net/, last accessed 10 November 2023). To which extent is this extraordinary? Finally, measures of central tendencies need to be complemented with the study of variability. This shows the dispersion in the data set. For instance, it can show to what extent the reader can expect deviation from the central tendency. Again, this supports the assessment of a given document's extraordinary or ordinary nature. Descriptive statistical analysis might also include the presentation of outliers and prototypical examples of documents. In short, with a descriptive analysis the editor can offer a thorough overview of a large document collection and help readers foresee what they can expect when browsing thousands of documents; readers can in turn sharpen their reflective attitude towards the data presented in the edition.

*

As a conclusion, despite the fact that big data sets proliferate in the humanities and beyond, their explorations have remained challenging. The heart of the matter is that to explore big data one inevitably needs training in text and data mining; however, today most humanities scholars are not well equipped with skills in text and data mining. The lack of these skills is a significant barrier to the study of big data in the humanities. A new type of scholarly edition that I named data edition in this essay is therefore needed. As I have argued here, data editions are meant to accommodate big textual data sets; even more importantly, they are meant to make big data accessible to and explorable for the scholarly community.

Throughout my short essay I attempted to point out what makes a data edition a scholarly edition. In short, I believe that principles traditionally applied to create scholarly digital and analogue editions can be applied with big data as well, though they need to be redefined and include new elements such as transparency and reproducibility. On the one hand, the consistent application of these redefined principles is what makes the development of a data edition a scholarly work. On the other hand, a data edition, just like a traditional digital edition, is a scholarly work if it contributes to knowledge and scholarship. I argue that a data edition effectively furthers knowledge if it unlocks the hidden layers of textuality and helps the scholarly community explore and study them. The role of the editor in the process of uncovering the hidden layers of textuality and furthering knowledge can be understood through another analogy with traditional digital and printed editions. The editor of a traditional digital edition furthers knowledge by enlightening the content and the context of a given text; as a result of this enlightening process the text is becoming understandable and it can 'speak clearly' to the reader (Sahle 2016, 26). With a data edition, the editor furthers knowledge by *enabling* the data to speak for itself. Big data does not speak for itself; it is the data edition, and the editor behind it, that makes the data speak for itself. To conclude, a data edition is a scholarly work if it *facilitates* the process of 'speaking for itself'.

References

Barbera, M., Meschini, F., Morbidoni, C. and Tomasi, F. 2013. 'Annotating Digital Libraries and Electronic Editions in a Collaborative and Semantic Perspective.' In *Digital Libraries and Archives. IRCDL 2012. Communications in Computer and Information Science,* edited by Agosti, M., Esposito, F., Ferilli, S., Ferro, N. Springer. http://dx.doi.org/10.1007/978-3-642-35834-0_7.

Cantos-Gómez, P. and Almela-Sánchez, M. 2018. *Lexical Collocation Analysis: Advances and Applications*. Springer.

Franzini, G., Terras, M. and Mahony, S. 2016. 'A Catalogue of Digital Editions.' In *Digital Scholarly Editing: Theories and Practices*, edited by Driscoll, M.J. and Pierazzo, E. Open Book Publishers. http://dx.doi.org/10.11647/obp.0095.09.

Lafferty, J. and Blei, D. 2009. 'Topic Models.' In *Text Mining: Classification, Clustering, and Applications*, edited by Srivastava, A.N. and Sahami, M. Chapman and Hall/CRC. http://dx.doi.org/10.1201/9781420059458.ch4.

Olson, D. L. and Lauhoff, G. 2019. *Descriptive Data Mining*. Springer.

Robinson, P. 2002. 'What Is a Critical Digital Edition?' *Variants: The Journal of the European Society for Textual Scholarship* 1: 43–62.

Sahle, P. 2016. 'What Is a Scholarly Digital Edition?' In *Digital Scholarly Editing: Theories and Practices*, edited by Driscoll, M. J. and Pierazzo, E. Open Book Publishers. http://dx.doi.org/10.11647/obp.0095.02.

—. 2020. 'Editions-Browser.' A Catalog of Digital Scholarly Editions. https://www.digitale-edition.de/exist/apps/editions-browser/about.html.

Schiuma, G. and Carlucci, D. 2018. *Big Data in the Arts and Humanities: Theory and Practice*. CRC Press.

Tóth, G. M. 2021. *In Search of the Drowned*: *Testimonies and Testimonial Fragments of the Holocaust*. Yale Fortunoff Archive. https://lts.fortunoff.library.yale.edu/.