7.

Predicting the future of digital scholarly editions in the context of FAIR data principles

Bartłomiej Szleszyński, Agnieszka Szulińska and Marta Błaszczyńska

Digital Scholarly Editions (DSEs) in Polish literary research – TEI PANORAMA (TEI.NPLP.PL)

Digital editing, even with the narrowing adjective 'scholarly' is a field that covers an extremely broad spectrum of activities and belongs to many traditional disciplines of the humanities. Let's start, then, by defining what 'digital editions' mean to us and how we apply the term in this chapter. Two of the authors are the creators of the TEI Panorama platform (TEI.NPLP.PL), literary scholars and digital editors (interested primarily in the practical side of editing), while the third is a head of an Open Science Unit at the Digital Humanities Centre. All of us work at the Institute of Literary Research of the Polish Academy of Sciences, which significantly determines both how we perceive DSE and how we approach the issues of data and its FAIRification. Hence if we are seeking an answer to the question of the future of digital editing, it is one we are practically engaged in on several levels. Our work and the solutions and priorities we have selected can be summarised in six points:

1) Our approach is shaped by our discipline and the categories of texts we edit. As members of the staff of the Institute of Literary Research,

we are engaged in literary studies in its broadest sense (our field of activity could thus be called digital literary studies), largely focused on writings from the more or less distant past. As an object of editing, we are primarily interested in literary works (prose, poetry, drama) and documents of literary life (such as the correspondence of writers). This entails a certain formal conservatism of the digital solutions we adopt – for the overwhelming majority, texts are paper-born and originally planned for publication in book form. However, the digital environment allows for showing them in infinitely more interesting ways, enabling the creation of editions that would be virtually impossible in paper form and providing text researchers with versatile tools for their interpretative work. To take a specific example: unpublished manuscripts with a very complicated arrangement of annotations and deletions can be shown without the editor's interpretation simplifying a complex manuscript into a single 'clean' version.

2) Digital literary research and editing are, in our perception, directly linked to the achievements of traditional literary studies. Thus, we focus on the evolutionary development of digital editing methods in direct cooperation with prominent 'traditional' editors in dialogue with the conceptions and history of scholarly editing. We are open to their needs and observations with regard to digital publishing as they will be also our future users.

3) On the technical side, let's start with the obvious: we use Text Encoding Initiative (TEI) standard in our editions. But how we do it is the result of a string of carefully considered decisions. The first was whether we would create our own software or use existing solutions – we determined that we would be better served by creating our own software, dedicated to the specific needs of the literary editions we would be handling. The second was about the structure of this software. We decided to create both a custom back-end TEI editor, allowing for the most intuitive possible input of tags, and a front-end software for presenting the tagged texts. The third was about how we use TEI – we made our practical decisions about the tagging system bearing in mind the specific format of the edition and its scholarly purpose.

4) All our editions are placed on a single platform, any expansion of the software (front end or back end) applies to work on all the corpora we develop. This makes it easier for us to ensure the sustainability and updating of the software and operate, bit by bit, as a national infrastructure for scholarly digital editions in the field of literary studies.

5) As academics, we are primarily concerned with the scholarly use of our editions, not necessarily going beyond academia – they are mostly created by professionals for professionals. At the same time we realise that digital editions, for many reasons, allow us to show what the process of scholarly text editing is much better than traditional editions.

6) We try to apply the principles of open science as widely as possible.

Each of these decisions has far-reaching implications in terms of what our work on scholarly editions looks like, as will be elaborated on later in the text.

In the following section, we will try to talk about the editions from the outlined area in the context of the research data (with a particular emphasis on FAIR principles).

TEI, FAIR, infrastructures – how can 'data' be described in DSEs?

In order to answer the question of opportunities and challenges in transferring the FAIR principles into DSEs, one needs first to focus on what 'data' means in the context of humanities, literary studies and – more specifically – scholarly editing. The FAIR principles relate to data's findability, accessibility, interoperability, and re-usability – and are assumed to be applicable to all research data. We will begin by discussing the specificity of approaching data in the humanities and recognising our own position in the disciplinary and national contexts.

One of the challenges in tackling data in the humanities lies in marrying the perspective of scholars and the newly developed professional personnel focused on research data management, such as open data officers, data stewards or librarians with specific data-related interests. Sometimes one can notice a tension between the first group, focused on conducting the research and often perceiving the data activities such as the creation of a data management plan at the beginning of the project as more of a task to be performed by support staff (*data reflection as an administrative task*) and the latter who aim at increasing the awareness of the significance of data in the scholarly context (*data reflection as part of the scholarly workflow*, where data management becomes a 'reflective process that exposes and tweaks existing behaviours, rather than one that introduces specific tools' – Edmond and Tóth-Czifra 2018, 1). The argument that many data stewards put forward is that, while they can help and support the data-related activities at each step of the project with their specific knowledge and expertise, it is the researcher him- or herself who understands the project best and is able to provide the greatest insights into the data to be created, collected, processed, analysed, published and/or re-used. The pressure is high when we consider how consuming the data management activities are. Such pursuits also often remain poorly rewarded within the existing evaluation systems, discouraging individuals from deeper engagement. Therefore, it seems to make sense for the researcher and the data specialist to work together so that they can use their complementary competencies (which, in a way, we realised, having written this article together). When thinking about the future of DSEs, we should also seriously consider the real possibilities of re-standardising existing TEI standards (Maryl et al. 2021, 164).

While the acceptance of the notion of data in the humanities has been growing over the past few years, in reality it has been adopted by specific groups of researchers rather than become part of the mainstream. There may also be some methods and communities that encourage data reflection within humanities more than others – for example, Erzsébet Tóth-Czifra discussed previous studies revealing confusion around the notion of 'data', pointing out that it would be

interesting to investigate 'whether there is any correlation between data awareness and the level of integration of computational methods into the respective research workflows' (Tóth-Czifra 2020, 251).

The FAIR principles present us with some general ideas on how scholarly data ought to be managed. However, knowledge gathering, methods and approaches are most often domain-based in the humanities. It is often within disciplinary communities that most common standards are discussed, established and solidified or rejected. As a result, what a historian may understand as 'data' may be quite different from a cultural studies scholar or a linguist. This will also affect the way in which they perceive FAIR principles. In this paper, as mentioned in the introduction, we focus on the approach of literary scholars – and more specifically, scholarly editors with a literary studies background (this seems to be the relevant place to point out that editors who identify as philosophers or historians might have a different understanding and areas of focus). What also needs to be taken into account is the fact that all the authors are based in Poland – in the case of humanities, local contexts and national languages also form part of the important community in which scholarly cultures develop.

However, the advantages of FAIRifying humanities data – such as data in scholarly editions that we discuss in this chapter – are often similar to those of natural sciences because, for members of the research community 'the value of making data FAIR, and accessing FAIR data, is unprecedented access to research assets and analytical tools to interrogate those assets' (Harrower et al. 2020, 6). At the same time, we will keep in mind that there are several dangers associated with overstandardisation. While work towards minimal norms and principles in data curation is to be encouraged, setting up the bar initially too high will isolate big portions of data, possibly eventually leading to data loss, the opposite of our aim.

Let us now turn directly to the issue of data in the area thus charted. The most obvious data that is produced during scholarly digital editions is, of course, the TEI encoded texts. It is good practice to

share the code of already completed digital editions as we do on our platform and as many other sites with DSEs do. The idea is that such code should be, first – understandable to other editors, regardless of the language of the text being edited (they can read its structure and encoded properties), and, second – compatible with other TEI encoded texts and suitable for automatic processing. That is, if you put it in the terms just mentioned – fully FAIR. While the first assumption is basically fulfilled, the second works to a limited extent. This is due to the fact that it is difficult to unnotice (although some try with all their might) the grown elephant in the room of TEI editors. Well, the flexibility for which the TEI standard is often praised (and which seems to be one of the reasons for its popularity) but leads to the fact that every digital editing project uses TEI in a more or less different way, creates limitations for interoperability and re-usability of such data, like TEI code from a specific project. Therefore, it is good practice for any digital edition to present in as much detail the specific ways in which TEI is applied. Thus, one could somewhat provocatively ask: 'is TEI then a standard?' (or: 'how much TEI is a standard?') or even declare that: 'TEI is not FAIR'. However, this issue should be seen in a broader perspective, for the above recognitions do not make TEI useless. Rather, they should prompt reflection on optimal practices for using it in editorial projects. We should consider how to make the TEI code as interoperable and re-usable as possible to allow the most extensive exchange of data between projects.

One way is demonstrated by the DraCor platform (Fischer et al. 2019), which collects drama corpora in various languages, tagged in the TEI subset dedicated to dramatic works – TEI Drama. It uses fairly basic markers of dramatic structure to visualise it in different ways for each drama. On the one hand, it proves in practice that 'I' (interoperability) and 'R' (re-usability) from FAIR principles are in fact possible to implement in projects using TEI. On the other hand, it should be borne in mind that if one tried to visualise those elements that are not presented in detail in the TEI Guidelines (for example, types of didascalies), unification would be practically impossible – in the absence of detailed guidelines, each project is forced to develop them in its own way. Therefore, one solution is to collect corpora

labelled in a specific subset of TEI and use basic structural labels, without going too deep into their details.

Another way to fix it – and one we would like to devote a little more space to – is to move in the direction of building a path towards an infrastructure. In her article (Pierazzo 2019), using a metaphor taken from the world of fashion, Elena Pierazzo showed the alternative to be found between DSEs tailored for one particular edition ('haute couture') and those created, as it were, on a conveyor belt basis ('prêt-à-porter'). This catchy metaphor, while proposing a certain (very important) order for reflecting on DSEs, at the same time somewhat simplifies the issue. Indeed, the serialisation and repetitiveness of DSE productions on individual platforms is sometimes gradual – in addition to platforms dedicated to only one work – such as the *Faust* edition (Bohnenkamp-Renken, Henke and Jannidis 2018) or those collecting very many editions/corps (such as the aforementioned DraCor), there are also many intermediate solutions such as the well-known Melville edition (Bryant and et al. 2017) collecting a number of quite diverse editions by the same author (Ohge 2021, 41–53). We would prefer to propose the metaphor of 'factories' for infrastructures with an approach that leans towards automation or 'manufactures' when most of the editing work (such as text marking) is done manually.

If we were to answer the question about (tentatively at this point) the future of digital editing and about the possibility of standardising digital editions and their FAIRification in particular, we propose to build ever larger infrastructures – at the national and European level. The example of such an editing platform is the TEI PANORAMA (TEI.NPLP.PL), that can be called a 'manufacture' for editions from the field of Polish literary research. In addition to many other advantages, infrastructural approach is also extremely useful for standardisation – all corpora tagged on the same infrastructure are fully compatible in terms of how TEI tags are used (and, consequently, how the same phenomena are being visualised). If we combine this with the openness of the software tool code (in our case, we make the code available at the request of our partners, but we plan to

make it fully open), in this way we popularise a particular way of using TEI, thus reducing the dispersion among scholarly digital editing platforms.

To conclude this part of the reflection, one more aspect of scholarly digital editions and re-use of data should be mentioned – they are also a tool for researchers to visualise, collate, search and display various kinds of statistical information. Viewed from this perspective, DSE data is as re-usable (and useful) as the tools that process TEI-tagged texts make it possible.

In the following section, we will show a more detailed landscape of scholarly digital editing using TEI in order to make an attempt at presenting the possible future in scholarly editing (suggesting that similar solutions can be applied at other national – and indeed, at the European – levels) and its practices.

Challenges for new users in TEI-oriented digital editing world (and how to overcome them)

TEI is undoubtedly a popular choice in a lot of digital humanities projects, including DSEs. Looking at the *Catalogue of Digital Editions: The Web Application.* (Fanzini et al. 2016), we can find 165 digital editions with filters 'scholarly: yes' (as this catalogue gathers also nonscholarly editions); 'digital: yes'; 'edition: yes' and 'XML-TEI transcription: XML-TEI is used'. Thus, with the total of 261 entries in the database, DSEs make up over half of them. Yet, there are no filters for disciplines, thus we cannot check how many of them are DSEs of literary texts or are within the range of literary research. Nonetheless, it is worth mentioning that the fact that one of the filters pertains to a particular standard is a sign of its significance in that field.

However, we do not imply that only numbers count. As a manifestation of TEI popularity in academic circles, we perceive a range of entities enhancing scholarly communication based on TEI. Here are some examples:

annual conferences like TEI Conference;

a wide range of tools and services designed to work with and enhance TEI standard, including the TEI Publisher, CETEIcean, Oxygen, LEAF-writer, FairCopy;

databases and corpora with requirement of data in TEI: DraCor, CorrespSearch;

coursers like *Text encoding and the Text Encoding Initiative* and *Digital Scholarly Editions: Manuscripts, Texts and TEI Encoding* on #dariahTeach;

communities such as E-editions and Special Working Groups at TEI Consortium, like *Correspondence*, *Manuscripts*, *Ontologies*;

The Journal of the Text Encoding Initiative on OpenEdition, edited by the Text Encoding Initiative Consortium;

and, of course, textual outputs, for example, articles about TEI: for instance, according to the GoTriple, a discovery portal for open SSH resources, 36 open documents with 'Text Encoding Initiative' were published in open access only in 2020.

What is also worth mentioning is that many of such entities follow TEI's values by being open and community-driven.

The pros of using TEI are also well acknowledged by DH communities: the fact that this standard was designed for humanities, being based on stable language XML, running on every browser, tags with familiar naming and functions like <witness> for 'contains either a description of a single witness referred to within the critical apparatus, or a list of witnesses which is to be referred to by a single sigil', grouping in modules with terminology that is relatively familiar to philologists like *critical apparatus*. And there are many entities and communities around it, as the list above proves.

However, it seems that TEI grew so large and powerful with so many projects and tools, that it is still challenging for a new user to start a project in a standardised way. For years there was no default open source and affordable tool to choose, when a scholar wished to annotate a literary text and create a digital edition.

TEI Publisher is growing to that status, yet it emerged fairly recently in comparison to years of TEI usage in humanities projects. As was already mentioned, we do have a vast range of DSE projects created in various environments, with the use of different workflows, data management plans (or even without them), so the main question for a default solution here is connected to the topic of re-usage of existing projects. The case of re-creation of Van Gogh Letters, one of the first DSEs of letters, with TEI Publisher, is promising in its demo state. Yet the original version (Jansen et al. 2009) is still believed to be a 'primal' digital edition with a full set of source data.

Another case considered as a challenge in FAIRification of our literary data is a history of a subset called TEI Simple, especially designed for modern texts that 'permit[s] modern web applications to easily present and analyze the encoded texts, mapping to other ontologies, and processes to describe the encoding status and richness of a TEI digital text' (TEI Simple Repository on Gthub). As for the TEI Panorama platform, it appeared as a perfect solution for our first (and, as it turned out, not the last) digital scholarly edition of correspondence between twentieth-century poets on emigration. Two obstacles were met during this case. One of them is that for the second DSE (and the third, fourth) we needed to expand this subset urgently as TEI Simple was really basic (which was indeed a core of this subset to be fair) and it does not cover enough 'base', for example, for modern drama literary texts.

The second barrier comes from the the fact that this subset is no longer supported. Of course, TEI Simple was also a ground for development of the TEI processing toolbox (and the TEI Publisher), thus its role for the future standardisation processes is unquestionable. Yet, at some point it was no longer possible to strictly follow

the TEI Simple schema, which is considered to be problematic, when it comes to data FAIRification.

Although absolute unification of the DSE creation process with TEI as a standard is impossible, two other tendencies may help in order to navigate new users to this kind of digital work.

Workflows, defined as 'sequences of operation/steps performed on research data during their life cycle' are an innovative type of digital outputs and might be converted as data itself. Whereas a part of team workflows might be sometimes presented in the editorial note section of the DSE, creating this kind of document increases its re-usability and interoperability by linking to a specific tool. Comparing a vast number of various teams' workflows might help in identifying common needs and gaps for current and future creators of DSE using TEI. For instance, a workflow *Customizing TEI to Check Pointers* (Bauman 2022) is a great start for anyone who wants to add a Uniform Resource Identifier (URI) into his/her TEI schema. It would be advisable to gather those kinds of resources in one place, ideally a place designed for digital scholarly editing.

A proposed tailored adaptation of the TEI standard not only in the lingual, but also in the cultural context of a particular literary text may seem an idea that would lead to further scattering of data in DSE projects. Yet remembering the dangers of overstandardisation discussed above, it is a necessary step for teaching purposes, for instance in the context of the use of TEI by students at universities. As Allés-Torrent and del Rio Riande (2020, 32), who conducted a number of lessons about TEI for Spanish students, observe, 'even though there are a lot of open access materials on the web on DH training and DSE in TEI in English, it is not enough for the Spanish-speaking community to translate them, since it is necessary to re-create the problems and adapt existing materials to their own needs and examples.'

Promoting TEI in literary texts in the context of culture, language and historical momentum might also be a way to identify phenomena

which are not reflected in a dedicated set in TEI P5 Guidelines, but have a great impact on national cases, such as the political censorship on Polish literature in the 1950s and 1960s. Achieving a level of consistency on the country level in DSE projects still seems like a formidable challenge, yet definitely worth facing and working on.

Towards infrastructures and standardisation – on a possible (bright) future of DSEs

In conclusion, by taking TEI Panorama (TEI.NPLP.PL) as an example of a platform for DSEs expanding into a larger infrastructure, we can reflect on the direction of similar ventures and thus, on the future of this aspect of scholarly digital editions.

The TEI Panorama platform has reached a considerable critical mass at this point – scholarly editions of dramas, novels, works in verse and correspondence are being created on it. Its various functionalities allow, among other things, to show versions of a given work, manuscript properties, count statistics and create complex networks of links between tagged entities. At the same time, it remains the only such infrastructure in Poland, so it is gaining interest from many scholarly institutions that plan to make editions using it. We can try to describe two futures – the near future, almost at hand, which is already beginning to come to fruition, and the more distant one, less certain, but according to current trends quite a probability.

In the first one, TEI Panorama will eventually become the main Polish infrastructure for creating scholarly literary research editions. As a result, all these editions will use the same TEI standard – so they will be, at the national level, fully FAIR. This standardisation will be further enhanced by the fact that the software code will be fully open. So the nearer future may bring integration of infrastructures at the national and disciplinary level.

And what might happen in the more distant future? It seems that a positive and quite likely scenario will be that infrastructures will cross

national and disciplinary borders, providing the tools needed at each stage of the scholarly digital editing process. This will, of course, require a restandardisation of the ways in which TEI is used and extensive reflection on the differences in editions and infrastructures – and it seems that the result should be worth the effort. But that's a story for a slightly different occasion.

References

Allés-Torrent, S. and del Rio Riande, G. 2020. 'The Switchover: Teaching and Learning the Text Encoding Initiative in Spanish.' *Journal of the Text Encoding Initiative* 12. https://doi.org/10.4000/jtei.2994.

Bauman, S. 2022. 'Customizing TEI to Check Pointers.' Women Writers Project. Northeastern University Women Writers Project. 2022. https://www.wwp.neu.edu/research/publications/documentation/other/checking_pointers_in_ODD.html.

Bohnenkamp-Renken, A., Henke, S. and Jannidis, F. 2018. 'Historisch-kritische Edition von Goethes Faust.' Digital edition. Faustedition. 2018. http://www.faustedition.net/.

Bryant, J. et al. 2017. 'Herman Melville Electronic Library.' Digital edition. Versions of Billy Budd: A Fluid-Text Edition. https://melville.electroniclibrary.org/versions-of-billy-budd.html.

Doran, M., Edmond J. and Nugent-Folan, G. 2022. 'Seeing Shapes in the Cloud: Perspectives from the Humanities on Interdisciplinary Data Integration.' http://www.tara.tcd.ie/handle/2262/98529.

Edmond, J. and Tóth-Czifra, E. 2018. 'Open Data for Humanists, A Pragmatic Guide.' https://halshs.archives-ouvertes.fr/halshs-02115443.

Fischer, F. et al. 2019. 'Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama.' In *Proceedings of DH2019: 'Complexities',* Utrecht University. https://doi.org/10.5281/zenodo.4284001.

Franzini, G., Andorfer, P. and Zaytseva, K. 2016. *Catalogue of Digital Editions: The Web Application*. https://dig-ed-cat.acdh.oeaw.ac.at/.

Galvini, G., Sessa, C., Wallace, D., Taylor-Wesselink, K., Ohlmeyer, J., Lyall, C., Fletcher, I. et al. 2021. 'Report of Workshops and Analysis of IDR/AHSS Integration Learning Cases.' SHAPE-ID: Shaping Interdisciplinary Practices in Europe. Zenodo. https://doi.org/10.5281/zenodo.4439665.

Harrower, N., Maryl, M., Biro, T., Immenhauser, B. and ALLEA Working Group E-Humanities. 2020. 'Sustainable and FAIR Data Sharing in the Humanities: Recommendations of the ALLEA Working Group E-Humanities.' Berlin: ALLEA - All European Academies. Digital Repository of Ireland. https://repository.dri.ie/catalog/tq582c863.

Jansen, L., Luijten, H. and Bakker, N., eds. 2009. Vincent van Gogh –The Letters. https://vangoghletters.org/vg/.

Maciej, M., Błaszczyńska, M., Szulińska, A., Buchner, A., Wciślik, P., Zlodi, I. M., Stojanovski, J. et al. 2021. 'OPERAS-P Deliverable D6.5: Report on the Future of Scholarly Writing in SSH.' Zenodo. https://doi.org/10.5281/zenodo.4922512.

Ohge, C. 2021. *Publishing Scholarly Editions: Archives, Computing, and Experience*. Cambridge University Press. https://www.cambridge.org/core/elements/publishing-scholarly-editions/D5A9FCEA4DECF1DE798B938BA48B2ED3.

Pierazzo, Ea. 2019. 'What Future for Digital Scholarly Editions? From Haute Couture to Prêt-à-Porter.' *International Journal of Digital Humanities* 1 (2): 209–20. https://doi.org/10.1007/s42803-019-00019-3.

Social Sciences & Humanities Open Marketplace. 2022. https://marketplace.sshopencloud.eu/.

TEI P5: Guidelines for Electronic Text Encoding and Interchange. 2022. https://guidelines.teipublisher.com/index.html.

TEI Simple Repository on Github. 2016. https://github.com/TEIC/TEI-Simple.

The GoTriple Platform. 2022. https://www.gotriple.eu/.

Tóth-Czifra, Et. 2020. 'The Risk of Losing the Thick Description: Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem.' In *Digital Technology and the Practices of Humanities Research*, edited by Edmond, J. Open Book Publishers. https://doi.org/10.11647/obp.0192.10.