I Want it All, I Want it Now. Literature researcher meets programmer

Vanessa Hannesschläger[1[0000-0003-0938-0890]] and Peter Andorfer[1[0000-0002-9575-9372]]

[1] Austrian Academy of Sciences, Sonnenfelsgasse 19, 1010 Vienna, Austria
vanessa.hannesschlaeger@oeaw.ac.at
peter.andorfer@oeaw.ac.at

Abstract. This paper describes a collaborative project carried out in 2017. The initial motivation to do the project was a call for participation in a conference on text genetic editing in digital editions. A literature researcher (Vanessa) asked a programmer (Peter) to work with her on a little publication platform which would display an edition focusing on the text genesis of a specific play written by her main subject of study, the Austrian writer Peter Handke, to present at the aforementioned conference - he agreed. In turn, Peter asked Vanessa to come up with a research question about the text and a list of features she would need the platform to have. Her answer was simple, really - and it became the title of this paper, which speaks of what Vanessa wanted, what Peter wanted, how they did it and how that worked out. We will describe the data modelling, the automated and the manual processing of the data, the tools used, the technical implementation, the resulting *handke-app* and the challenges and benefits of two very different research perspectives in all these steps.

Keywords: Peter Handke, TEI, Digital Editing, Text Genesis, Automated Collation

1 Preface

1.1 A Subsection Sample

This paper describes a collaborative project carried out in 2017. The initial motivation to do the project was a call for participation in a conference on text genetic editing in digital editions. A literature researcher (Vanessa) asked a programmer (Peter) to work with her on a little publication platform which would display an edition focussing on the text genesis of a specific play written by her main subject of study, the Austrian writer Peter Handke, to present at the aforementioned conference - he agreed. In turn, Peter asked Vanessa to come up with a research question about the text and a list of features she would need the platform to have. Her answer was simple, really - and it became the title of this paper. As its reader, you first need to understand that to a Handke researcher, everything matters: every text witness, every correction, every date, every person involved, every pen used, every coffee stain, every archive holding a Handke collection, every place in which a part of the writing process took place, every location of the play, every book read by the author, every language used in the text, every biography of every person that might have been inspiration for one of the characters of the

play, every production of the play, every person involved in every production of the play, every published version of the text, every translation of the text, and of course all other texts that Handke wrote and the cross references with this one text - to just name a few of the relevant aspects. As Vanessa had written her diploma thesis about the play in question, she already had all these data, though not in a machine readable form. When Peter asked Vanessa to focus and pick the most important aspects so that he could start developing data models and technical solutions for them, she repeated her answer. In this paper, we describe what things we managed to put into practice and where the realities of time management stopped us from constructing the Swiss army knife Vanessa had originally envisioned (and Peter had at no point in time intended to build).

2 Introduction and Scope

Peter Handke's play *Immer noch Sturm* (2010) (*Storm still*, 2014)[1] is a perfect source text for a study on modelling the genesis of a literary text: Five text stages (with several sub-stages) can be identified even before page proofs.[2] The numerous smaller and bigger adjustments Handke made from stage to stage, but also the large quantity of accompanying material (such as the author's preparatory reading notes) make this corpus a good starting point for a study on the possibilities and boundaries of sustainable encoding of text genetic processes.

In our study, we transcribed the respective first page of each text witness and added TEI P5 markup. The result of our work was published under the title *handke-app*.[3] The encoding focused on issues such as the distinction between immediate and later manual corrections or the representation of the proven integration of preparatory reading notes and thus served as a list of requirements for functionalities of the web application which would represent these encoded texts. The goal of this application is to provide the best possible support for the analysis and research process.

3 Material

Storm still is a formally complex text for the stage about the Carinthian Slovene resistance against the Nazi occupation during the Second World War. The *dramatis personae* include a nameless "I", "who is not reasonably distinguishable from the author's persona"[4], and his Carinthian Slovene maternal relatives. Its formal structure is remarkable and shows the author's continuous development of his own approach to "epic theatre", in which he combines elements of the antique tragedy with epic narrative from a first person perspective.[5]

The text genetic material is as remarkable as the text itself. The first version was written from December 15th, 2008 to February 22nd, 2009 (in pencil). Before the page proofs were created in July 2010, four further text stages with several sub-stages emerged due to continuous adaptations and changes to the text made by the author. In addition to this extensive (and fully preserved) material, additional notes, books, and other materials (all kept by the Salzburg literary archives) give further insights into the process of creation.[6]

The situation described above poses a number of questions and problems. These lead to the question that inspired our project: What does a digital edition have to be able to do in order to serve a literature researcher's needs? The answer: It has to represent all existing knowledge and research about the edited text, thereby generating new knowledge.

However, "all" existing knowledge is a relative term in this context. We could have focussed on the work context - the play and its relations to Handke's other works -, the biographical context - the characters and their relations to Handke's real life family members -, the historical context - the representation of historical events and the text's relations to sources about these events[7] -, or the text genetic context. The latter is a good choice for various reasons, one of them being the availability of information and data about the text genesis via the platform Handke*online*.

4 Encoding

Even though the platform Handke*online* is a data treasure for Handke research, it has some clear disadvantages from a technical perspective. It does not provide the possibility to extract any structured data, let alone process it. Therefore, quite some manual work was necessary.

The following data were relatively easily transformable into structured data:

- Dates: Handke has been documenting the writing dates of his first text versions for many years, and this is also true for *Storm still*. Every writing date is noted in the manuscript next to the text written on the respective day. Usually (and also in this case), he also notes the dates on which he worked on subsequent versions of a text.
- Persons and institutions: Information on people and institutions formally involved in the production of a text stage (e.g. transcriber, editor, owner of the manuscript) had already been collected by Handke*online* and could therefore be transformed into structured data easily.
- Places: Thanks to meticulous documentation of writing places in the manuscripts, the identification and subsequent geo referencing of places relevant to the genesis of the text was unproblematic.

On the other hand, the translation of the following information into machine-readable form posed certain challenges:

- Preparatory reading: In preparation for writing *Strom still*, Handke read extensively about the history of the Carinthian Slovene resistance against the Nazi occupation of Carinthia during World War II. His reading focused on partisan memoirs. The most important book read in this context was Karel Prušnik-Gašper's *Gemsen auf der Lawine* [Chamoix on the avalanche].[8] Handke's triple reading of this book (each several years apart) can be dated exactly as the book, in which he made notes and annotations in different colors during each reading and also noted beginning and end dates, is available in the archives.[9] For other books, the data is more fuzzy: He took notes and collected quotes, which are partly dated and have been preserved, however the books themselves have not. In addition, further readings of books to which no notes have been preserved can be proven by identification of

direct quotes - however, the time of his reading can only be estimated. This is one of several examples of substantial information about the text genesis which has been investigated and confirmed by research, but can still not be pinned down and transformed into precise, machine-readable data.

- Source indication: The research articles about Handke's *Storm still* quoted above show that the reconstruction of the text's genesis was a task of many years. Vanessa, one of the authors of this paper, has worked on this text since it was published and dedicated her diploma thesis as well as several papers and a lot of her work done for Handke*online* to the investigation of its becoming, its meaning, and its interconnections to other texts. She was also the data provider for the app we developed. Thus, we did not source our data from other research, but rather deduced it from the data provider's previously accumulated knowledge about the topic. For this reason, we did not add a file including bibliographic information on sources used for the app. Another reason was the additional time it would have required to connect such a file to the individual data points. While we are confident that this was a legitimate decision in terms of research ethics, we still see the problem that the app does not indicate if a given data point was sourced from Handke*online*, Vanessa's thesis or a paper.
- Provenance history: Even though all text genetic material for *Storm still* is kept by the Salzburg literary archives today,[10] it was previously owned by several individuals and therefore arrived at the archives not as one collection, but in parts and with time. In addition to this, only parts of the material belong to the archives, other parts are privately owned and only kept by the archives as a permanent loan. The history of this material and its paths is complex, but it is known - in principle. This is informal knowledge among Handke researchers and fans. Therefore, reliable and exact data suitable for structured analysis cannot be deduced from this information which some might refer to as gossip - it might be that from a present perspective, but looking, say 200 years to the future, this might be valuable information for researchers who could be interested in the author's network or the market prices for manuscripts at the beginning of the millennium. In the long run, not preserving this information might therefore mean a loss.

The transcriptions of the various text stages are previously nonexistent data which were newly created for the *handke-app*. As our goal was not to provide a full edition of *Storm still*,[11] but rather a technology test, we only transcribed the respective first pages of each text witness and encoded it with TEI-P5 markup[12].

5 General Set-up

The described project was a pilot study. Work on the research content of the project, i.e. development of the research question, transcription, and annotation (edition) of the text witnesses, was carried out by Vanessa. Peter was responsible for all technical aspects as well as development of the data management workflow. Both contributors are employees of the *Austrian Centre for Digital Humanities* of the *Austrian Academy of Sciences* (ACDH-OeAW) which provided the necessary (server) infrastructure.

6 Document Centered Approach

The outcome of the project was shaped by the team's decision to follow a document centred approach for the transcription. Each text witness was transcribed in an individual XML file. The <teiHeader> elements of each file contain the metadata specific to the text witness (archive holding the manuscript, physical traits, history of its genesis) structured in a TEI conformant way (as far as that was possible). The next step was to transcribe the respective first page of the text witness and encode specific text genetic phenomena within the respective text witness and model the formal structure of the text (using <pb>, <lb>, and <p>).

As the pilot study focused on text genesis, we refrained from encoding text genre specific phenomena according to the TEI module *Performance Texts*[13], e.g. indication of a <speaker>, or a more in-depth literature analytic markup e.g. providing information on time, place, or characters, as we found this to be insignificant due to the incompleteness of the transcription.

We encoded the work's genesis, i.e. the systematically documented changes, deviations, and variants, in the next step, which was collation. While methods of collation strongly vary between disciplines, research projects, and even individual researchers in "analogue" text research[14], it is a strongly formalized approach in the digital humanities. This is necessary as collation is in this case primarily carried out by machines. According to the *Gothenburg model*[15] developed in 2009, the following steps have to be taken:

1. The respective witnesses have to be divided into comparable chunks of text; this process is called *tokenization*. This is generally done on the word level (which the machine defines by identifying strings of symbols separated by spaces).
2. In a second step, the tokens of the respective witnesses are compared to each other. For the (likely) case of differing amounts of tokens, so-called *gap tokens* have to be inserted.
3. Based on this comparison, analysis can be carried out. However, the authors of the *Gothenburg model* have pointed out that this task might be beyond the machines' limits, especially when it comes to the task of identifying how deviations in various witnesses are related to each other and if differing sequences of tokens are additions, deletions, or transpositions of text parts: "While alignment results can still be judged in terms of their quality to some extent, transposition detection can only be done heuristically as one can easily think of cases, where it is impossible for a computer 'to get it right'."[16]
4. The final step is the synthesis if the results of collation. Just as in "traditional" text studies, the result can be a critical apparatus that documents deviations from a base text version in other witnesses. Depending on the technical implementation of the *Gothenburg model*, it can also be a graph and/or a tabular representation.

The project team was aware of two concrete implementations of this model, namely *CollateX*[17] and *Juxta Commons*[18]. The decision for *Juxta Commons* was made due to more user friendliness (which helped Vanessa in doing her part), i.e. *Juxta Commons*

is a web service with a graphic user interface while *CollateX* requires a local installation and some familiarity with use of the command line.

With the help of *Juxta Commons*, Vanessa was able to collate the individual text witnesses and thus to encode (or let encode) the text genesis in a quick, systematic, and machine-readable form. The genesis of the text was annotated by *Juxta Commons* according to the *Parallel Segmentation Method*[19], which is characterized by the notation of the various readings next to each other (parallel), which facilitates the comparison of variants. A short example:

```
<app>
  <rdg wit="#wit-24801 #wit-24800">eine Sitz-</rdg>
  <rdg wit="#wit-24794">Eine Sitzbank</rdg>
  <rdg wit="#wit-24795">Eine<lb/> Sitzbank</rdg>
  <rdg wit="#wit-24793">Eine Sitzbank</rdg>
  <rdg wit="#wit-24798">Eine Sitzbank</rdg>
  <rdg wit="#wit-24799">Nichts</rdg>
  <rdg wit="#wit-24796">Eine Sitzbank</rdg>
  <rdg wit="#wit-24797">Eine Sitzbank</rdg>
</app>
```

These results were manually cleaned in order to obtain better readability by wo/man and machine:

```
<app>
  <rdg wit="#wit-24801 #wit-24800">eine Sitz-</rdg>
  <rdg wit="#wit-24794 #wit-24793 #wit-24798 #wit-24796
    #wit-24797">Eine Sitzbank</rdg>
  <rdg wit="#wit-24795">Eine<lb/> Sitzbank</rdg>
  <rdg wit="#wit-24799">Nichts</rdg>
</app>
```

Here we encountered one of the challenges of communication and coordination between a “tekkie” and a “human”: This code optimization could easily have been done via a small script instead of manual cleaning, had Vanessa thought to ask for it. However in addition to this, Vanessa also had to manually add in manuscript corrections by the author which had been encoded previously, but which *Juxta Commons* failed to include in the collation. An example of this is the <subst> element in the following passage:

```
<app>
  <rdg wit="#wit-24799">I</rdg>
  <rdg wit="#wit-24794 #wit-24795 #wit-24793 #wit-24801
#wit-24800">EINS</rdg>
    <rdg wit="#wit-24798 #wit-24797">ERSTER AKT</rdg>
    <rdg wit="#wit-24796">
      <subst>
        <del>ERSTER AKT</del>
        <add hand="#handke" place="below">EINS</add>
      </subst>
    </rdg>
```

</app>

Looking back, it would have been more efficient to do without text critical markup in the first transcription and only adding it in after collation; but as the team had never worked with this tool before, we were not aware of this problem beforehand.

The reason for the choice of this text witness centred approach was ultimately of a technical / pragmatic nature: For encoding, we used the *oXygen*[20] XML editor for the very simple reason that Vanessa was already familiar with using this tool and the ACDH-OeAW owns licenses for it.

7 Implementation

The web application for the *handke-app* was implemented using *exist-db* for the following reasons: As the transcriptions and annotations were done in XML format, use of a native XML database stood to reason. Additionally, *exist-db* can easily be integrated in *oXygen*; and last, but not least Peter is experienced in working with *exist-db* and its functionalities that facilitate the development of data-driven web applications.[21]

The following features were successfully implemented in the *handke-app*:

- Text views 1[22]: Access to individual text witnesses via a traditional table of contents. Individual text views include extensive meta information (document title, archive holding the manuscript, original title, transcriber, license) as well as the text (including additions, deletions, etc.). A scan of the original manuscript page is also available. [23]
- Text views 2[24]: This entry point allows users to see text witnesses next to each other in order to compare them. For this view, the *EVT Viewer*[25] was used. From a technical perspective, it was very pleasant to see that the *EVT Viewer* was able to process the files created by Juxta Commons ad edited by Vanessa without further adaptations of the application's code.
- Text views 3[26]: The result of the *Juxta Commons* collation can also be exported as a "traditional" apparatus in a static HTML file.
- Indices 1[27]: Events. This page collects and lists meta information retrieved from the <teiHeader>s of all XML files. Thus, all events related to the text genesis (individual writing days, Handke's preparatory reading of books, corrections by Handke and others, etc.) are collected in a list here. In addition, they were visualized on a map with an included timeline. Due to the mentioned inaccuracies and problems of standardizations of certain informations (e.g. precise dates of preparatory reading), this visualisation's meaningfulness is limited.
- Indices 2[28]: Persons. This page lists all persons involved with the material, be it as owner or in the creation process (transcriber, editor, etc.). Persons are attached to a location (where their contact with the text took place). These locations are visualized on a map.
- Indices 3[29]: Places. The same map as on the person page and a list of the places.

- Indices 4[30]: Institutions. A list of all institutions involved. Indices 3 and 4 are not particularly useful due to the small amounts of data, but were so easily implementable from a technical point of view that we decided to include them anyway.
- Analyses 1: Deletions and additions. Two graphs show the amount of deletions and additions (i.e. the frequency of the TEI tags <del> and <add>) in all text versions.
- Analyses 2[31]: User requests. Users can choose a TEI tag and query for its frequency in all text versions. As only a small amount of tags was used in the project at hand, this feature is not particularly useful for this specific data set. It was included because Peter wanted to test the necessary efforts of implementing this feature.

8 Conclusion

Summing up, we can conclude that this pilot study was fruitful both for the "tekkie" and for the "human". While there were challenges in communication and cooperation at certain stages, we managed to broaden each other's view and understanding considerably and both benefited from the cooperation.

Our work showed that a complex work genesis including a number of text witnesses can be encoded efficiently by following a text witness centered approach and subsequently using machine supported collation. This is especially true when the result can be processed using existing software such as the *EVT Viewer*.

Another positive result is that we were able to show that digital methods can provide modes of analyses (i.e. quantitative queries about text phenomena) that a traditional apparatus would not be able to offer (even though the result was not meaningful in the case of our pilot study due to the limited amount of data).

We also learned that the *Text* Encoding Initiative's guidelines, while the unquestionably best approach for encoding text inherent phenomena, reach their limits when used for encoding "real world phenomena" related to text genesis such as places, persons, or events. Even though the TEI offers elements for encoding these phenomena,[32] the interconnection of these entities to each other and to the text witnesses is not well specified and largely varies from project to project. Therefore, connecting the data created in the pilot study to other projects will likely be difficult to impossible. Use of a more comprehensive model that does not only focus on encoding text, but also extra textual realities (e.g. *CIDOC CRM*[33]) might have been a better choice.

From the literature researcher's point of view, the pilot study was also very fruitful. Though the result is not the "Swiss army knife" originally imagined, it is a pretty nifty tool that can do some unanticipated things. While sometimes challenging, the methods and tools used were manageable even for a non-"tekkie" and worth every effort considering the results. The visualizations and features inspired a deeper understanding of the text and its becoming as well as completely new perspectives - even though Vanessa thought that she knew this text inside out even before she started doing this study, hav-

ing worked on it for numerous years. We will spare you the explanation why it is amazing to learn that Handke only transformed the "ninety nine apples" on the apple tree mentioned in the first stage direction of this play into "99 apples" in the very last text witness. But believe us: it truly is mind blowing, and we would never have found this out without the *EVT Viewer*.

From the developer's point of view, the pilot study was fruitful in regards of getting a deeper understanding of the "Swiss army knife" metaphor. On a first glimpse, a Swiss army knife is one single tool which can do many things. Though on a second look it could also be understood as a collection of many different kind of tools wrapped up between to red plastic halfs. Transposing this metaphor into the digital (humanities) world, we reach the idea of many (micro) services/tools, each one tailor made for a single task (i.e. the *oXygen* XML editor for encoding text, *eXist-db* for storing data, *Juxta Commons* for collating text) "glued together" by some basic website, providing links to all those services / tools / data. The main challenge in the interaction with the non-"tekkies" is the communication process needed to break down a huge research question like "I want it all" into its many components. This is doable. But needs time and patience.

Finally, we have to mention that since this pilot study, we have worked together on various projects and gotten better at understanding each other's perspectives and languages better and better over time. Our cooperation is ongoing, as is the work on the *handke-app*, as we noted on its start page: "Please be aware: This is work in progress. If you find any mistakes or have suggestions for further development, please create an issue in the project's code-repo on GitHub."[34]

References

1. Handke, P.: Immer noch Sturm. Suhrkamp, Berlin (2010), respectively Handke, P.: Storm Still. Trans. Chalmers, M. Seagull, London, New York, Calcutta (2014).
2. See the list of text genetic material for *Storm still* on Handke*online*, http://handkeonline.onb.ac.at/node/57/material, last accessed 2019-02-04.
3. handke-app, https://handke-app.acdh.oeaw.ac.at/, last accessed 2019-02-04, respectively acdh-oeaw/handke-app: First release (Version v1.0), http://doi.org/10.5281/zenodo.1195978, last accessed 2019-02-04.
4. Kastberger, K.: Lesen und Schreiben. In: Kastberger, K., Pektor, K. (eds.): Die Arbeit des Zuschauers. Peter Handke und das Theater. Jung und Jung, Vienna, Salzburg (2012), pp. 35–47, p. 44 [transl. VH].
5. See Hannesschläger, V.: Real Life Fiction, Historical Form: Peter Handke's "Storm Still". In: Boldrini, L., Novak, J. (eds.): Experiments in Life-Writing. Intersections of Auto/Biography and Fiction. Palgrave, London (2017) [Palgrave studies in Life-writing 1], pp. 145–165.
6. See Hannesschläger, V.: „Geschichte: der Teufel in uns, in mir, in dir, in uns allen." – Zur Rezeption von Familiengeschichte und Historie in Peter Handkes Immer noch Sturm. Vienna (2013) [Diploma thesis].
7. See Hannesschläger, V.: Peter Handkes „Immer noch Sturm" und Karel Prušnik-Gašpers „Gämsen auf der Lawine". In: Wieser, L. (ed.): Karel Prušnik-Gašper: Gämsen auf der Lawine. Materialien. Wieser, Celovec (2016), 13–18.

8. Prušnik-Gašper, K.: Gemsen auf der Lawine. Der Kärntner Partisanenkampf. Drava, Celovec (1980). Handke read and annotated this edition himself (see Handke*online*, http://handkeonline.onb.ac.at/node/1566, last accessed 2019-02-04).
9. Handke*online*, http://handkeonline.onb.ac.at/node/1566, last accessed 2019-02-04.
10. Salzburg literary archives, collection: Handke, Peter (LAS) and collection: Handke, Peter (Leihgabe Widrich) (PH-PAW).
11. This was infeasible due to copyright reasons on the one hand, and limited (time and financial) resources on the other.
12. TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2017), http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf, last accessed 2019-02-04, respectively http://www.tei-c.org/release/doc/tei-p5-doc/en/html/, last accessed 2019-02-04.
13. See TEI Consortium: TEI P5: Performance Texts, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DR.html, last accessed 2019-02-04.
14. See e.g. Sahle, P.: Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 1: Das typografische Erbe. Norderstedt (2013) [Schriften des Instituts für Dokumentologie und Editorik – Band 7].
15. TEI SIG Manuscripts: The "Gothenburg model": A modular architecture for computer-aided collation (2011), https://wiki.tei-c.org/index.php/Textual_Variance, last accessed 2019-02-04.
16. Ibid.
17. CollateX, https://collatex.net, last accessed 2019-02-04.
18. Juxta Commons, http://juxtacommons.org, last accessed 2019-02-04.
19. TEI Consortium: TEI P5: Parallel Segmentation Method, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html#TCAPPS, last accessed 2019-02-04.
20. SyncroSoft: oXygen XML editor, https://www.oxygenxml.com/, last accessed 2019-02-04.
21. Andorfer, P.: dsebaseapp (2016ff), https://github.com/KONDE-AT/dsebaseapp, last accessed 2019-02-04.
22. https://handke-app.acdh.oeaw.ac.at/pages/toc.html, last accessed 2019-02-04.
23. The scans were kindly provided by the Salzburg literary archives.
24. https://evt.acdh.oeaw.ac.at/#/critical?d=doc_1&e=critical, last accessed 2019-02-04.
25. Edition Visualisation Technology – EVT Viewer, http://evt.labcd.unipi.it/, last accessed 2019-02-04.
26. https://handke-app.acdh.oeaw.ac.at/pages/juxta-play.html, last accessed 2019-02-04.
27. https://handke-app.acdh.oeaw.ac.at/pages/events.html, last accessed 2019-02-04.
28. https://handke-app.acdh.oeaw.ac.at/pages/persons.html, last accessed 2019-02-04.
29. https://handke-app.acdh.oeaw.ac.at/pages/places.html, last accessed 2019-02-04.
30. https://handke-app.acdh.oeaw.ac.at/pages/organisations.html, last accessed 2019-02-04.
31. https://handke-app.acdh.oeaw.ac.at/pages/stats-dynamic.html, last accessed 2019-02-04.
32. See TEI Consortium: TEI P5: Names, Dates, People, and Places, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html, last accessed 2019-02-04.
33. CIDOC Conceptual Reference Model (CRM), http://www.cidoc-crm.org/, last accessed 2019-02-04.
34. https://handke-app.acdh.oeaw.ac.at/pages/index.html, last accessed 2019-02-04.