Editing for man and machine

Anne Baillot, Anna Busch

„Editing for man and machine“

Anne Baillot & Anna Busch
Conference of the European Society for Textual Scholarship
Leicester, Nov. 2015

The questions that we will present today (as well as some, but not that many, answers to them) emerged during the development of the digital edition 'Letters and texts. Intellectual Berlin around 1800'. This edition was realized in the context of a specific funding program and it was clear from the very beginning that the funding phase would not last any longer than five years – five years that are coming to an end in a couple of weeks.

First I will talk about the different features of our digital edition. I will show how these were deducted from the research question or, the other way around, how we tried to structure our edition so that it is able to answer our research questions.

But if our aim had been to collect information for the sole purpose of answering our research question, we would not have made the effort to develop an interface and to offer open access to our resources. This edition is both a research environment for our research group and a scholarly edition putting up resources for the community at large. This requires more thought about the definition of such a “community at large”: Who are the readers of a digital scholarly edition, what are their expectations, to what extent does it make sense to make efforts in order to meet these expectations? These questions will be presented in some more detail in the second part of the paper, in which Anne Baillot will try to define the audience for which an edition such as ours can be conceived and realized.

1. Presentation of the digital edition

The main objective of the research group “Berlin intellectuals 1800-1830” at the department of Modern German Literature at the Humboldt-University of Berlin is the analysis of intellectual relationships in Berlin around 1800. The Prussian capital was chosen as an eminent place of cultural transfer and knowledge exchange. Form and meaning of the participation of writers, publishers and scholars in public life play an important role, and we pay particular attention to the communication strategies they developed and how the close connections between several circles such as universities, academies, literary clubs and salons, were shaped.

Letters constitute the main part of the corpora of our digital edition, but other sorts of texts are published as well, for example work manuscripts of literary texts or lecture notes. The importance of letters in general as a source of information and also as texts with an inherent literary nature that adopt complex functions was one of the starting points for our project. The letters are previously unpublished or published only in an abridged form without marked changes and omissions mostly in editions that do not meet with today’s standards of text editing.

On the basis of handwritten manuscripts of letters and other texts which we transcribed and edited, we analyzed the conditions and developments crucial for intellectual communication in Berlin between 1800 and 1830

To do this, four thematic key aspects need to be considered:

To what extent did the establishment of the Berlin university in 1810 contribute to the intellectual self-conception of scholars and academics working as lecturers?

To what extent did the presence of the French in Berlin shape and define the political awareness of the intellectuals?

What communication strategies did male and female writers use to establish themselves?

How can a political statement be extracted from a literary or scholarly corpus?

We reviewed several other digital editions of letters as examples for our edition. You may know the two exemplary editions „Vincent van Gogh. The Letters“ and „Carl Maria von Weber - Collected Works (WeGA)“ which were in many aspects helpful to develop and implement our edition.

Our edition is based on using and adhering to the standards recommended by the German Research Foundation (DFG) for scholarly, web-based text representation which are amongst others: XML/TEI, Open Access, and the use of authority files.

The one-sided focus on the author, which is the basis for most printed editions, was substituted in favor of several possibilities to approach the manuscripts. These deciding conceptual considerations enable the user and reader to access the edition in different ways of which the one via the author stands not above, but on the same level next to the others. The edited letters and texts can also be accessed via their genre (letter, drama, novella, reports, lecture notes etc.) or by topics that are structured according to the already mentioned four basic research questions.

And that is what an edited letter looks like in our edition. Opposite the scan of the manuscript, we provide two versions of the transcription of its textual content: a diplomatic transcription and a reading version of the text.

The diplomatic transcription provides a transcription of the manuscript’s text to the letter with as little editorial interference as possible. All corrections, deletions, and additions found in the manuscript are reproduced. Characteristics such as line

breaks, horizontal alignment of the paragraphs or abbreviations are rendered unchanged. Missing parts of the text are not reconstructed. Therefore, the diplomatic transcription is suited for scientific text analysis, interpretation, and research questions for example concerning the text genesis.

In contrast, the reading text focuses on readability. It provides - just as the diplomatic transcription - a reliable textual basis but it aims to present an easy-to-read view of the transcription that offers a quick orientation, which is especially helpful when there are a lot of deletions and additions in the text. The focus is solely on the 'basic text' of the author's hand, all other hands are omitted. Corrections, deletions and additions are carried out, original line breaks are ignored, abbreviations are written out. Parts of the text that are missing but can be reconstructed with a high certainty are supplied.

The annotation is also carried out according to this double-track method: On the one hand, it reflects the subtleties of the manuscript text, and on the other hand, it offers details concerning the context. As editors, we can provide this annotation for both the diplomatic transcription and the reading text, or just for one view.

Both of these views are generated from the same XML/TEI file. The TEI files contain the transcription of the text of one letter or manuscript each, the markup of the text, the annotation and the metadata. Next to these, the TEI files are also the basis for the view of the metadata the listing of entities and index entries, the view of the XML/TEI encoding and the generating of the PDF file. These several aspects affect the encoding that has to incorporate encoding decisions for the diplomatic transcription as well as the reading text.

The different formats of text representation - next to the diplomatic transcription and the reading text also the XML/TEI encoding and the PDF file with the index entries - are freely available for the reader under a CC-BY-license. In addition, the reader has access to all relevant metadata and the index entries: with information on the sender and addressee, the origin of the manuscript and its context, the repository and the history of acquisition, the editor, etc. And as you have already seen in the screenshots, the general presentation of the website itself and all metadata are available in three languages: German, English and French. (some text bits are still missing, but we are working on it!)

Overall, our edition is not radically different from the major current digital scholarly editions. Its charm lies mostly in its specific research question and in the combination of a series of features (be it structural or on the level of the interface).

2. The editor and his reader

General considerations

Scholarly editions are generally conceived by scholars for scholars, especially in the German context in which we work. Format and price of scholarly editions make it almost impossible to reach a wider audience. Knowing that one's edition will be primarily (and most of the time, exclusively) used by scholars, editors tend to define their editorial standards according to their own needs and habits.

Digital scholarly editions are different from most scholarly print editions in terms of accessibility. While some editions are password-protected hidden behind a paywall, many of them are available for free in open access. Potentially everyone can access and use a digital scholarly edition that is made available in this way. It does not mean that these editions have to be conceived for a wider audience, but it certainly changes their general conditions of existence in that it is possible to address and actually reach other readers.

It is much more difficult to define the expectations of a non-scholarly audience than those of a scholarly audience. Trying to reach a wider audience is a business for which scholars are not properly trained. Not only are we not trained for it: it is also unproductive in terms of career benefits – again, I am speaking for the German Academic context in particular. On the contrary, a higher complexity is usually worth more in terms of academic reputation and hence capital, than a greater accessibility of the research results. The aspiration to address a different reader than one is oneself hence remains widely unsupported by the Academic system in terms of evaluation and in terms of training.

The decision to offer a reading version on the same level as the other ones is a philosophical and a political one. What especially mattered to the whole team during the development work of our digital edition was to make all 6 versions of the text information equally readily available and to present the transcription on the same level as the digital image of the manuscript. It came with the decision not to transcribe on the level of the sign, but to give the reader the freedom to establish the link between the digital image and the transcription by him or herself. I call this philosophical because it implies that each reader brings his or her own reading habits and is still able to make his or her sense out of the edition: it is a freedom based on the assumption of education. And I say that it is political in the sense that we do not consider the document that the archives allow us to display as less important or less relevant than the transcription we editors (who are scholars) are offering to the reader. Putting archival performance and scholarly performance on the same level is a political statement in the German Academic context. Also, it turned out to have a direct influence on our choice of corpora, since not all archives would allow us to publish free digital images of their manuscripts.

These philosophical and political choices represent an implementation of our conception of text editions as a tool aiming at enabling readings of all sorts, a structured space that empowers the reader with access to text and information for him/her to structure. Of course, we direct the reader (we all know that it would be delusional to think that one could set up a “neutral” edition that would allow all possible readings). But we try to hand out digitally to our readers the critical elements that are needed for an autonomous reading. The attempt to offer such an “enabling” edition (enabling the reader to act as the designer of his/her own reading) is not new, but it requires to take into account specific aspects in the case of digital editions.

As editors, we anticipated reading scenarios and used them as a basis for our design development. We certainly had to limit our creativity due to the financial constraints. We reached a result that can be considered honorable with regard to the conditions in which the edition was produced, but which is in many ways not satisfactory, especially in terms of reader-friendliness. Wanting to accommodate smaller and bigger monitors lead to the fact that we had to fix some elements (especially frames) in such a way that the result lacks fluidity and some frame elements are too dominant. This will not be corrected in this project life, but it will certainly be one of the first things I would improve if I had money for it. The first approach, namely the homepage and the fluidity of the ramifications starting from the homepage, is essential if you want to build a reading relationship.

We don’t know much about the validity of the reading scenarios that we speculate on. In fact, after a few years, we came to the conclusion that such an edition as ours, with its columns and additional information etc. was very inconvenient in terms of readability, in the sense that it is uncomfortable to read the text itself. The idea to offer a pdf version additionally to the 6 different html displays and the query interface emerged from the diagnosis that this edition was not adequate to extensive reading. The pdf version can be downloaded either for each document or for a whole corpus; the registers are generated accordingly. The pdf generator was our way of making our edition more readable. I must say that I was at first not very convinced about it and would have preferred to invest more money in the html design. But it turns out to be an excellent choice regarding the monographs and theses which we have to hand in (from the Master thesis to the habilitation), since only printed material counts as actual work in the world of Modern German Literature in Germany and pdf is easier to print than html.

We are hoping to learn more about the reading habits of our audience through the analysis of the logfiles which is currently being realized in the context of the DiXit program by Anna-Maria Sichani. She contacted me a few weeks ago and indicated that this kind of study has only been realized so far for the Van Gogh edition. The study of the logfiles of the Van Gogh edition confirms well-known reading attitudes:

simple query is used much more often than advanced query, the links play a major role in the way the reader accesses information. In that sense, the editor is still the one giving the major orientation impulse in the access to information. In the context of her research, Anne-Marie Sichani will analyze the logfiles of 4 different editions, among which ours. We hope to learn more about the way our readers navigate in our edition and seize this opportunity to be able to correct our projected scenarios. For instance, the analysis of the Van Gogh edition showed that the first and the last letter presented on the webpage are the ones that are most often consulted. The search for famous paintings is also a major anchor point for the users of the Van Gogh edition. How will this turn out in an edition that is not centered on one author, one corpus, and decidedly aims at deconstructing canonical approaches to the history of literature? The results of this analysis will certainly be profitable not only because of what we can learn about the expectations of our readers in particular, but also because of what can be deducted from the analysis of four different editions. To what extent does the orientation in our edition differ from the others? Are there constants in the approaches by the readers or do they depend strongly on the way the edition is conceived and implemented? Can we distinguish reader types or reading patterns, what do they tell us about the interest the readers are giving to the edition? We would like to encourage you to play around with our edition right now since we are in the middle of the month in which the logfiles are being harvested for Anna-Maria to work on.

What may seem to deal solely with theoretically insignificant implementation issues reflects in fact a major question related to the relationship between edited text and reader: Which role does the editor of a digital scholarly edition play in the interaction between both instances? How much text and how much design is there to be put in the balance?

Reading vs. using

“Reading” covers only part of the way text is accessed by the audience of a digital edition. In this case, the term “using” is certainly more adequate to the multiplicity of approaches made possible by the digital media.

In a digital scholarly edition like ours, access to text is being enabled by:

- Different entry points (genres, authors, etc.)
- Queries and corresponding interfaces
- Links within the edition and outside of it

Each of these forms of access to text questions the concept of “reader”. In the case of letter editions, dedicated digital tools have been developed in order to facilitate

these different types of access to text and offer a new type of user-friendliness in the work on letter corpora.

Crosslinks as user guides

Let me get back to our edition in order to take you to the big picture in a second step.

The accessibility of our data is ensured by implementing authority files and standards like the Integrated Authority File (GND) for persons, GeoHack for places, XML and TEI, ISO-Codes, and Open Access. Crucial is also the connection with other repositories and editions by implementing BEACON. BEACON is a simple file format, hosted by Wikipedia, which allows for linking to one another contents on the basis of the GND number, the German authority file for persons. This means, whenever a person in our edition appears (be it in the edited texts, the annotation, the metadata) that another BEACON-using project also records, there will automatically be a direct link to the respective page of this project. And the other way around, the other project will automatically get the link to our edition. The German National Library and many regional libraries, archives, biographical and bibliographical projects and many others use the BEACON format which is therefore an easy way to connect the contents of your project with a wide range of scholarly web services.

Currently, we do a test run of mutually interconnecting the contents of our digital edition with Kalliope, a database and National Information System for collections of personal papers and handwritten manuscripts in mostly German archives and institutions. To our encoding of the metadata as well as in the respective view of these meta data in our edition, we only recently added the link to the according data set of said manuscript in Kalliope. And the other way around, Kalliope points to our edition. The advantage for the user while browsing through the database Kalliope and looking for a certain letter is that he or she is directly guided from the metadata in Kalliope to the facsimile, the transcription and the annotation in our edition. Here, cataloguing and indexing by a library and scholarly editing and research close ranks.

One of the developers of our edition, Sabine Seifert, is an active member of the TEI Special Interest Group Correspondence. This group has been working on a comprehensive model for encoding correspondence and guidelines for best practice over the past years. They concentrate on correspondence-specific metadata within the TEI header and their encoding. Their work has been presented as “the TEI council’s greatest achievement of the year” at the TEI conference in Lyon a month ago.

For the encoding of the metadata, the SIG Correspondence developed the new element <correspDesc>, the correspondence description, that contains the most important correspondence-specific information on the letter or any other piece of correspondence, like telegram, e-mail, text.

The idea behind this was to rethink how correspondence editions can be linked - as there is a general and growing demand of correspondence projects for interchange and linked data facilities. To enable this, a reduced subset of <correspDesc> with more limited encoding choices was selected. This abbreviated form is called CMI, Correspondence Metadata Interchange format. It heavily relies on authority files and standard formats to meet with the naturally diverse encoding decisions of various editions and projects. This CMI format shall serve as the basis for linking and connecting metadata. A first result, set up by Stefan Dumont from the Berlin-Brandenburgische Akademie der Wissenschaften, is the web service 'correspSearch'. This web service makes the correspondence-specific metadata of different German-language correspondence editions searchable with one query and gives an idea of what is possible when correspondence projects can be linked. Anyone willing to join the platform can contact Stefan Dumont (or us). The consequent implementation of authority files, standards and the use of TEI with the new <correspDesc> element is one step in that direction.

What can we do with a digital edition of letters?

Making an edition machine-readable (like we do by using the CMI format) is, of course, much easier on the level of the metadata, which are quite standardized, than on the level of the text. Still, text and data mining should also address specific aspects of correspondence editions.

One characteristic of letters is that you generally are not the first one to read them when you discover them in an archive. Not only have they been addressed to a person or a group of persons in the first place (making us, unavoidably, secondary readers), many of the letters we at least are working on have already been edited in the last centuries. But not in extenso, no: they have been abridged, overwritten, corrected according to the expectation of the audience in the time that they were edited. So we are confronted with manuscripts which many editors have been working on, knowing only too well that we are only at the historical end of a long chain of approaches of the text. What do we do differently from our predecessors (apart from the fact that we do not write on the manuscripts, but time leaves its traces whether we want it or not and our digital images are also only the representation of a specific moment in the history of the manuscripts)?

At least, we can try to learn how our predecessors worked, how they abridged, re-wrote and hence drew a picture of their time's readers. In a Master thesis in Machine Learning and German Literature that I am co-directing, M.A. student David Lassner is basing his work on the deletions and additions in our edition (12 000 deletions, half as many additions). The aim is to develop an algorithm that can first define, then

predict the mechanisms of censorship at work in the writing/editing process. The size and the structure of the corpus are optimal.

When I started working on this digital edition, and especially its correspondence part (which is the bigger one), I thought it will all come down to being able to draw the intellectual networks in intellectual Berlin around 1800. Now I know better: it all comes down to the text.