‘Over-tagging’ with XML in Digital Scholarly Editions

Elise Hanrahan, hanrahan@bbaw.de

NOTE: Obwohl ich mein Abstrakt auf Englisch geschrieben habe, kann ich den Vortrag gern auf Deutsch halten.

This talk looks at the phenomenon of over-tagging (term created here) in XML, which consists of exaggerated and unfocused tagging that concentrates on diplomatic characteristics. These tags are used for the display of the digital edition text on the computer screen--an especially questionable result when digital facsimiles exist.

To what degree computer technology affects practices and theories in scholarly editing is an open question. But whether or not we are in the midst of a revolution or simply experiencing a change of tools, there are certain influences that can already be observed. One is the availability of digital facsimiles combined with the use of XML.

How could digital facsimiles change online scholarly editions? Digital facsimiles challenge one of the claimed purposes of a scholarly edition—the recreation of the original manuscript.[1] This aim, often strived for by means of a diplomatic transcription, is particularly prevalent in recent editions. Elements of diplomatic transcribing can for instance be found in most German critical editions from the last twenty-five years.[2]

Yet the question ‘How *could* digital facsimiles change online scholarly editions?’ was posed because astonishingly the diplomatic method continues to be dominant. It is thus the argument of this talk that online editions do not reflect this significant alteration in the relationship between researcher and original source. Instead the same editorial method is being used, and the primary new development is the use of XML instead of Microsoft Word. Indeed not only do digital editions continue to be diplomatic, but the tendency has even increased.[3]

---

[1] It should be noted that this very strong focus on an ‘authentic’ recreation of the original handwriting is fairly new in Editionswissenschaft, starting in the 1970s and becoming very dominant in the 1990s with editors like Hans Zeller.

[2] For a very concise summary of the take-over of the material-paradigma, see: Rasch, Wolfgang, Wolfgang Lukas, and Jörg Ritter. "Gutzkows Korrespondenz – Probleme Und Profile Eines Editionsprojekts." *Brief-Edition Im Digitalen Zeitalter (Beihefte Zu Editio)* 34 (2013): 99.

[3] Elena Pierazzo addresses the popularity of digital documentary editions in: Pierazzo, Elena. "Digital Documentary Editions and the Others." *Scholarly Editing: The Annual of the Association for Documentary Editing* 35 (2014). Accessed November 10, 2014. http://www.scholarlyediting.org/2014/essays/essay.pierazzo.html.

Structure of the Talk

I. Arguments for why diplomatic transcriptions are no longer necessary when digital facsimiles are available
II. An examination of why diplomatic aspects in digital editions have surprisingly increased instead of decreased
III. Arguments and counter-arguments for continuing to transcribe diplomatically in digital editions
IV. Suggestions for alternative editorial priorities

I. Digital facsimiles are a game-changer

There are two main arguments for recreating the original manuscript in the edited text (thus using a diplomatic transcription), both of which online facsimiles challenge. The first is to bridge the gap between the original manuscript and the researcher. Before digital facsimiles it was quite possible that the researcher never set eyes on the original manuscript, which was carefully stored away in a library or archive. The edition thus strove to offer the researcher an objective depiction of the handwriting. Now, however, the researcher can look at the image of the handwriting online, thus the edition must no longer bridge this gap.

The second argument for a diplomatic transcription is to preserve the manuscript. If anything happened to the original source, the text of the edition could be used as a replacement. Additionally the edition protects the manuscript from being over-handled, because it functions as an authoritative substitute. There is however no longer a need to create a substitute for the manuscript, because a high-quality image exists.

Despite these arguments, in praxis one finds digital editions still ruled by the diplomatic trend. Why is this?

II. The diplomatic-tradition and XML

There are two reasons for the prevalence of diplomatically-influenced digital transcriptions. The first is that digital editions emerged at the same time diplomatic editing was the dominant method for scholarly editions.[4] It is therefore not surprising that the current method for print editions was transferred to the newly emerging digital editions.

The second reason is due to the very nature of XML. In XML, unlike in Microsoft Word, specifications for the visual presentation of the edited text are completely separated from the documentation of the original source. Hence in XML the editor is no longer limited by the

[4] That is, the end of the 1990s/start of the 2000s

space on the page for recording textual phenomena and can enter as many XML tags as desired, allowing a theoretically endless documentation of the characteristics of the manuscript. Combine this opportunity with an already diplomatic trend, and the result is a lot of diplomatic XML-tagging, sometimes to an incredibly minute degree. This very real phenomenon shall be called 'over-tagging'.

Over-tagging refers to an exaggerated amount of XML tags that do not pursue a specific research question, but are in praxis only used for the display of the edited text on the screen. Over-tagging could be for example using a character in the line below an indentation to mark the length of an indentation, tagging the exact location and angle of marginalia, tagging orthographical elements like the long *s* in German, or tagging line breaks that are not semantically meaningful.

While there is nothing inherently wrong with tagging these kinds of textual phenomena, it is important to ask, what is the purpose of these tags? Such tagging is especially problematic when it comprises a large part of the XML schema. And one must add, no matter how detailed a transcription is, it can't recreate the image of the handwriting and the vast amount of data that the image carries. And not only is the usefulness of the results debatable, over-tagging takes a great deal of time and energy. Other fundamental editorial tasks fall by the wayside--tasks such as editorial commentary, to name only one of many. There are many other gaps to be bridged between the reader and original source besides the material gap. An essential benefit to be gained from more reflective tagging practices is the time to focus on these other editorial tasks.

III. Counter-arguments: machine searchable and a reader-aid

One argument for over-tagging is machine readability. This means that tagged textual information can be found in automated searches. Yet does anyone truly search for aspects such as indentation size? An editor might argue: ‘Perhaps not now, but someone could in the future. And what’s more, someone could discover that this seemingly insignificant textual characteristic actually carries a semantic meaning’. Such an answer reveals the editorial tradition that still strongly underlies digital editions today and is inseparable from diplomatic transcribing. This is the philosophy that literally everything on the page could be significant.[5]

There are two responses to this. First of all search masks are currently not being made to search for diplomatic characteristics. In praxis these tags are used for the display of the text only. It is true that, if desired, search masks could be made for this purpose. It is also true that

[5] This of course leads back again to the authenticity/materiality movement from Hurlebusch, Zeller and others. For an example of such a perspective that does not directly lead to the solution of a diplomatic transcription (but instead to digital facsimiles), see: Richter, Elke. "Goethes Briefhandschriften digital – Chancen und Probleme elektronischer Faksimilierung." *Brief-Edition Im Digitalen Zeitalter (Beihefte Zu Editio)* 34 (2013): 53-75.

it is impossible to prove that a certain aspect of a page is not meaningful. However, the ‘everything could be significant’ position is not a feasible editorial method. It is definitely not the basis on which to build a well-functioning XML schema. In reality, at the end of the project there exists a very large amount of information that is solely used for the display of the text on the computer screen.

Another argument for over-tagging is that the resulting text is a kind of facsimile-reading-aid.[6] This is however a problematic stance. Firstly, critical editions are not made primarily to be reading-aids for facsimiles, although they can be helpful for this purpose. Critical editions are editorial arguments and offer a readable edited text according to that argument. A facsimile-reading tool is potentially very useful, but is something different than a critical edition. Secondly, a diplomatic transcription was not conceived for this aim and is probably not the best method. There are certainly much better ways to help guide a researcher through a facsimile than simply mirroring the facsimile in the edited text, especially in consideration of technological possibilities.

IV. Different priorities for digital editions

Relinquishing over-tagging means more time for editors to concentrate on other aspects of editing, such as commentary and semantic tagging (including not just person or place names, but also more abstract themes, such as concepts found in the texts). There would also be more time to tag the creative process of the author (such as capturing the layers of the text's development by tagging crossed out/added words). There would be more time to enter meta data, to link through standard IDs like VIAFs for persons and geonames for places, and to simply to think about how to use the digital space to the researcher's best advantage.

In many ways, instead of reflecting on what doors technology opens for critical editions and thus shaping technology to this end, editors have let technology define them, losing sight of priorities in today’s digital world. For instance, an essential current challenge for digital editions is to avoid the 'island' problem—single editions floating in the internet without a real connection to one another. Minute diplomatic tagging does not address this problem (standard IDs and meta data to some degree does). Yet it isn’t a question of what is important and what is not--all editorial tasks are important--it is a questioning of appreciating what an edition has to offer and carefully considering the energy invested and the benefits gained. ‘Over-tagging’ is perhaps a very small piece of the debate on digital editions, but it could point to a general direction and is therefore worthwhile to consider in this context.

[6] This idea is touched on in Pierazzo(2014), 4.