CollateX, Juxta and the Challenge of Smart Text Repositories

Gregor Middell
Universität Würzburg
gregor.middell@uni-wuerzburg.de

For software which aims at meticulously comparing and correlating resources from possibly very different origins – that is: software for computer-supported collation – its successful application depends fundamentally on the interoperability of its input. If one understands interoperability technically as "the ability to reuse data [...] outside of the technological system in which they were originally instantiated"[1] then collation tools can only deliver accurate results to the degree that they are able to reuse the information available about a text. The same relationship between interoperability and accuracy, when looking at it from a hermeneutic angle, translates to the practical observation that the thicker[2] a formalized description of compared texts is when provided to a collation tool, the more reliable correlations (and vice versa: differences) between those texts can be computed.

The state of the art in terms of computer-supported collation has been thus far to compare texts on the level of their character content. Their markup (if available at all) would be ignored at best and removed in the process (if embedded) at worst.[3] One reason for this very basic limitation of what collation tools can achieve automatically when confronted with a marked-up digital text today, lies in the specific problem of interoperability as it is framed by the markup community. Whether it is the challenge of supporting a necessarily wider range of more loosely expressed semantic constructs in encoding standards like TEI-P5,[4] the long-standing problem of dealing with commonly encountered, yet poorly expressed overlapping structures in markup languages based on context-free grammars like XML,[5] or whether it is today's common imbalance in encoding practices, where a high emphasis is put on dense, holistic encoding and a far lesser emphasis on ensuring that such information can actually be decoded in other contexts later on;[6] the analysis of all of these rather special, yet interconnected problems of current markup techniques resulted in an insight during the development of CollateX and Juxta, which is at the center of this presentation: In order to make collation tools truly interoperable with a wide range of texts and at the same time pushing the boundaries of what can be made possible in terms of automating this tedious and error-prone task, one has to address some of the limitations caused by the way in which marked-up texts are encoded and processed.

Against this background firstly the implementation of an alternative document model will be presented, which…

---

[1] McDonough, James: "XML, Interoperability and the Social Construction of Markup Languages: The Library Example." In Digital Humanities Quarterly. 3/2009 (3). http://digitalhumanities.org/dhq/vol/3/3/000064/000064.html (seen 29.01.2012)
[2] Geertz, Clifford. "Thick Description: Toward an Interpretive Theory of Culture." In The Interpretation of Cultures: Selected Essays. New York: Basic Books, 1973. 3-30.
[3] Schmidt, Desmond. "Merging Multi-Version Texts: a Generic Solution to the Overlap Problem." Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11 - 14, 2009. In Proceedings of Balisage: The Markup Conference 2009. Balisage Series on Markup Technologies, vol. 3 (2009)
[4] Bauman, Syd. "Interchange vs. Interoperability." Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2 - 5, 2011. In Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7 (2011).
[5] Johnsen, Lars G., and Claus Huitfeldt. "TagAl: A tag algebra for document markup." Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2 - 5, 2011. In Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies, vol. 7 (2011).
[6] Mueller, Martin. To members of the TEI-C Board and Council. August 4, 2011. http://ariadne.northwestern.edu/mmueller/teiletter.pdf (seen January 29, 2012)

- is inspired by LMNL[7] as well as other experimental markup languages,
- supports markup via arbitrarily overlapping annotations on a text,
- can construct document instances from scratch as well as via schema-independent transformations of XML documents,
- is persisted to a relational database,
- features a predicate-based query language, and
- provides several export-/serialization-formats, e.g. in XML or JavaScript Object Notation (JSON).

The currently developed, web-based version of Juxta, CollateX and some use cases from an ongoing project creating a digital genetic edition of Goethe's "Faust" will serve as examples as to how this document model facilitates both, the consideration of character content as well as of markup during the collation process and how in so doing the application of computer-supported collation offers new ways of "thickening" a text's digitally expressed description, which ultimately means: its interpretation.[8]

Secondly the potential of such a document model will be explored when implanted into the framework of a web-based, distributed micro services architecture as it is envisioned and put to test by Interedition. Some of the document model's properties, notably its reliance on a text's immutability in terms of character content and its JavaScript-oriented annotation data model, makes it suitable as the foundation for a missing yet crucial building block of distributed architectures supporting textual scholarship: smart(er) text repositories. Looking at some major obstacles in designing a collaborative workflow around reusable tools in digital textual scholarship, e. g. identification and replication of textual data among institutions, layering of annotations, locality of data or locality of services in distributed text processing environments, possible solutions to these issues based on the given document model will be outlined and prototypes of such solutions will be presented.

[7] Piez, Wendell. LMNL in Miniature. An Introduction. Amsterdam Goddag Workshop. December 1-5, 2008. http://piez.org/wendell/LMNL/Amsterdam2008/presentation-slides.html (seen January 29, 2012)
[8] Buzzetti, Dino, and Jerome McGann. Electronic Textual Editing: Critical Editing in a Digital Horizon. In Electronic Textual Editing, ed. L. Burnard et al.New York: MLA. 2006. 51–71.