Workshop “Advancing digital editing”: how to work with graph models, multiple perspectives, and flexible workflows

Introduction

Participants of this full-day workshop learn to work with TAG, an alternative (graph-based) data model for text, which allows them to explore a more flexible workflow of digital textual editing than is available with just the single hierarchy afforded by XML. The goal of the workshop is not (only) to learn to work with a new text editing tool, but primarily to acquire a way of thinking about digital text modeling that is not constrained a priori by external models or software assumptions built into software tools. During the workshop, participants will be introduced to the principles of a distributed architecture for digital scholarly editions as they learn to work with the implementation *Alexandria*. This implementation is built on the TAG data model and allows editors to import, edit, and export complex texts in a straightforward manner. It supports various approaches to scholarly editing and accommodates the encoding of multiple, coexisting views on text.

Background

Over the past decades, the field of digital text editing has witnessed some major developments, both practical (the development of advanced tools and technologies for text editing) and conceptual (an increased awareness of the modeling process). There are, however, at least three important areas of digital text editing in which we can identify room for further development:

1) *Models of text*. In general, we can assume that the practice of text modeling improves significantly when the data model is consistent with the conceptual model because the data model then supports the scholar’s process instead of restricting it. However, prevailing data models for text not always correspond to a scholar’s idea or understanding of text. An oft-cited example is the discrepancy between the understanding of text as an *Ordered Hierarchy of Content Objects*—the *OHCO* model that underlies XML—and the many textual phenomena that do not fit naturally within this model, such as (self-)overlapping, discontinuous and non-linear texts (Haentjens Dekker and Birnbaum 2017; Haentjens Dekker *et al*. 2018).
2) *Perspectives on text*. Another contrast can be seen between, on the one hand, the digital scholarly edition as a closed-off environment that offers one exclusive view of the text and, on the other hand, the edition as a *digital workstation* that provides for multiple coexisting perspectives on a

text. Because a perspective implies a certain hierarchical structuring of the textual data, expressing multiple coexisting perspectives in XML often leads to overlapping hierarchies.

3) *Workflow*. Over the years, many integrated editing environments with specifically designed graphical user interfaces (GUI) have been developed, yet these editing environments are rarely adopted by other editors outside the intended users. In fact, the creation of a unified model of the editorial workflow has proven to be unrealistic (Van Zundert 2018). One of the main reasons for this is that interfaces as well as workflows are highly project-specific and personal. A comprehensive GUI will inevitably constrain some editorial choices even as it facilitates others. Furthermore, it significantly hinders the sharing and reuse of scholarly tools and applications, and has resulted in multiple reinventions of the same wheel.

From the situation outlined above emerges the need for a technology that supports rich and various perspectives on text, and an editorial workflow that can be freely configured. Such a technology would hinge on three requirements: First, the data model is not limited to one hierarchical structure. Secondly, creating a transcription is no longer a matter of adhering to a single perspective on text: the technology supports adding an unrestricted number of layers of information. A third and final requirement is that the design of an editorial workflow is a decision of the editorial team in question. In short: the technology should enable a versatile approach to editing, instead of forming a mold in which every project has to fit. TAG is a new and flexible technology that addresses these requirements. The TAG data model, a hypergraph structure that can successfully address the various challenges posed by the modeling of complex texts, has been described as both “simple” and “brilliant” (Sperberg-McQueen 2017). Furthermore, the TAG data model doesn’t compromise sustainability or interoperability of textual data: texts encoded in TAGML can be converted to XML, although the down-conversion to a hierarchical format entails decision-making on the side of the editor.

Workshop program

The workshop would run for a single day, divided into morning and afternoon sessions of approximately 180 minutes each. The morning session concentrates on the theoretical background of data models for text, including a discussion of several textual phenomena and other features that are difficult to express in the tree structure of XML. We then introduce the richer graph model of TAG and the related markup language TAGML. The session concludes with hands-on work: participants will make a brief transcription of a text fragment in TAGML, with special attention to those textual features that pose challenges to XML as a

model. For pedagogical reasons we provide a data set during the workshop, but participants will perform their own document analysis and will markup the text from their own perspective(s).

In the afternoon session we dive more deeply into practice in the context of the entire editorial workflow. Participants work with *Alexandria* to process the marked-up files created during the morning session. Because *Alexandria* is operated via shell commands, we will allocate some time to ensure all participants will be able to perform the necessary operations on the command line. It is however not imperative to have prior command line experience as *Alexandria* is designed to be intuitive and straightforward to use. The last part of the afternoon is devoted to publishing: we convert the markup files to XML (which facilitates their accessibility outside our toolkit) and then publish them as HTML.

To sum up: the TAG workshop will cover the entire editing process, from transcription to publication, paying attention to each step along the way. This "pipeline" approach provides participants with a deeper awareness of the many conceptual and practical transformations that textual data undergoes. It enables both the production of an actual edition and a final high-level, abstract reflection on the importance of choosing an appropriate data model to express, process and analyze textual information. Over the course of the workshop, participants work with *Alexandria* and thus experience first-hand TAG's hypergraph model for text, the properties of which have significant benefits for the editorial process and for the way we conceptualize our texts. They will learn that expressing information about text, creating a digital edition, and shaping editorial practices doesn't have to be bound to a specific tool or technology.

Schedule

Morning: Theory (180 minutes)

TIME	**TOPICS**
8h30 - 9h00	Start-up, installation (optional)
9h00 - 09h45	Exploration of challenges that scholars face when creating a digital scholarly edition. Overview of textual phenomena that we want to encode, explaining why they are hard to do in XML. Explaining that workarounds have consequences

	for the analysis and publication of the material. Introduction of the TAG hypergraph data model for text.
09h45 - 10h30	Explanation of the markup language TAGML; introduction of the concept of "layers of information"
10h30 - 11h00	Coffee break
11h00 - 11h45	Creating a transcription using the Sublime editor
11h45 - 12h30	Plenary discussion, taking stock, reflection

Afternoon: Practice (180 minutes)

time	**topics**
13h30 - 14h00	Introduction to the editorial workflow of *Alexandria* that follows from the new functionalities of the technology
14h00 - 14h45	Committing to the *Alexandria* repository; checking out a view from the repository
14h45 - 15h30	Conversion from TAGML to XML and subsequently to HTML for publication
15h30 - 16h00	Tea and coffee break
16h00 - 16h30	Explaining the principles behind a distributed architecture by means of an example that locally merges two transcriptions of the same text with different perspectives into one file (thus limiting the hands-on part to editing, diffing, and merging).

16h30 - 17h00	Plenary: reflection, discussion, and outlook

Target audience

The target audience of this workshop is textual scholars, scholarly editors, digital humanists and scholars interested in text modeling with *elementary prior experience* with TEI (or other XML), transcription, and textual editing. Experience with working on the command line, HTML and XSLT is a pro but not required. The tutorial has room for a maximum of 20 participants.

Technical requirements / technical support

Corpus: During the workshop, we will work with a small corpus of texts and an XSLT template to generate HTML. The corpus is prepared in advanced by the instructors in order to streamline the tutorial exercises.

Hardware: Participants must bring laptops with Windows/Mac OS X/Linux (No iOS / ChromeOS / Surface devices)

Software: Participants must install

- Java (>=1.8, e.g., https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html)
- Sublime Text 3 (https://www.sublimetext.com/3)
- GraphViz (https://www.graphviz.org/download/)
- *Alexandria* (https://github.com/HuygensING/alexandria)

Bios of instructors

Astrid Kulsdom

Astrid Kulsdom is a project manager and researcher in the Research and Development team at the Humanities Cluster, part of the Royal Netherlands Academy of Arts and Sciences. After completing her research Master's in Literary Studies at Radboud University in 2012, she has worked as a project manager for several government institutions. As project manager of the Research and Development team, she

combines her philological knowledge with her project management skills in order to effectively manage all strands of research within the team.

Bram Buitendijk

Bram Buitendijk is a software developer in the Research and Development team at the Humanities Cluster, part of the Royal Netherlands Academy of Arts and Sciences. He has worked on transcription and annotation software, collation software, and repository software. His recent work focuses on the development of *Alexandria*, a text repository designed for storing and editing documents, which is the reference implementation of the Text-as-Graph data model.

Elli Bleeker

Elli Bleeker is a postdoctoral researcher in the Research and Development Team at the Humanities Cluster, part of the Royal Netherlands Academy of Arts and Sciences. She specializes in digital scholarly editing and computational philology, with a focus on modern manuscripts and genetic criticism. Elli completed her PhD at the Centre for Manuscript Genetics (2017) on the role of the scholarly editor in the digital environment. As a Research Fellow in the Marie Sklodowska-Curie funded network DiXiT (2013–2017), she received advanced training in manuscript studies, text modeling, and XML technologies.

Ronald Haentjens Dekker

Ronald Haentjens Dekker is a software architect and lead engineer of the Research and Development Team at the Humanities Cluster, part of the Royal Netherlands Academy of Arts and Sciences. As a software architect, he is responsible for translating research questions into technology or algorithms and explaining to researchers and management how specific technologies will influence their research. He has worked on transcription and annotation software, collation software, and repository software, and he is the lead developer of the CollateX collation tool. He also conducts workshops to teach researchers how to use scripting languages in combination with digital editions to enhance their research.

References

- Haentjens Dekker, Ronald, and David J. Birnbaum. 2017. “It’s more than just overlap: Text As Graph.” Presented at Balisage: The Markup Conference 2017, Washington, DC, August 1 - 4, 2017. In *Proceedings of Balisage: The Markup Conference 2017.* Balisage Series on Markup Technologies, vol. 19 (2017). https://doi.org/10.4242/BalisageVol19.Dekker01.

- Haentjens Dekker, Ronald, Elli Bleeker, Bram Buitendijk, Astrid Kulsdom and David J. Birnbaum. "TAGML: A markup language of many dimensions." Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018. In *Proceedings of Balisage: The Markup Conference 2018*. Balisage Series on Markup Technologies, vol. 21 (2018). https://doi.org/10.4242/BalisageVol21.HaentjensDekker01
- McCarty, Willard. 2004. "Modeling: a study in words and meanings". Chapter 19 of *A companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion/
- McCarty, Willard. 2005. *Humanities computing.* Basingstoke and New York: Palgrave Macmillan.
- Sperberg-McQueen, C. M. 2017. "Text. You keep using that word …." Presented at Balisage: The Markup Conference 2017, Washington, DC, August 1–4, 2017. In *Proceedings of Balisage: The Markup Conference 2017.* Balisage Series on Markup Technologies, vol. 19 (2017). https://doi.org/10.4242/BalisageVol19.Sperberg-McQueen02.
- Van Zundert, Joris. 2018. "On not writing a review about Mirador: Mirador, IIIF, and the epistemological gains of distributed digital scholarly resources." In *Digital medievalist, 11* (1), p.5. DOI: http://doi.org/10.16995/dm.78