Graph Models for Textual Data: Between Text and Information

Tara L. Andrews[†]
tara.andrews@univie.ac.at

Stefan Jänicke[*]
stjaenicke@vizcovery.org

Stefan Dumont[‡]
dumont@bbaw.de

Andreas Kuczera[◆]
andreas.kuczera@adwmainz.de

Thomas Efer[+]
efer@informatik.uni-leipzig.de

Joris J. van Zundert[◇]
joris.van.zundert@huygens.knaw.nl

[†]Digital Humanities, Universität Wien. Vienna, Austria.
[‡]Berlin-brandenburgische Akademie Der Wissenschaften. Berlin, Germany.
[+]Abteilung Automatische Sprachverarbeitung, Universität Leipzig. Leipzig, Germany.
[*]Abteilung Bild- und Signalverarbeitung, Universität Leipzig. Leipzig, Germany.
[◆]Akademie der Wissenschaften und der Literatur. Mainz, Germany.
[◇]Huygens Instituut voor Nederlandse Geschiedenis, Koninklijke Nederlandse Akademie van Wetenschappen. Amsterdam, The Netherlands.

Panel Topic and Organization

Historians and textual scholars alike often struggle with the task of properly describing their data so that it becomes tractable, while remaining meaningful, in a computational environment. Thaller (2018) argues that much of the struggle stems from a conflated understanding of data and information. While computer scientists are usually concerned with an accurate and lossless capture and processing of digital data, humanists tend to be interested in the processing of semantics that are situated by default, may carry several layers of contextualization that have accrued via transmission, and may carry interpretation added by a scholar or expert. This, Thaller argues, is information, not data, and the computational constructs and operations that are the building blocks of informatics (bits, integers, strings, Boolean values) are grossly inadequate to handle information as understood by a historian or textual scholar. Thaller points to interesting approaches to solve this discrepancy in a more fundamental informatics-oriented way—e.g. Devlin (1991), who suggests a computational approach to representing atomic bits of information rather than bits of data. Thaller himself (1993) suggested a graph-based approach to the modeling of historical information.

In this respect it is interesting that the increase in sheer computational power and speed in the last decade has made some approaches to information modeling possible that previously had to be relegated to the unfeasible. So we have seen, with the recent rise of graph models in digital textual scholarship (e.g. Schmidt and Colomb, 2009; Efer, 2017; Kuczera, 2017; Haentjens Dekker, Van Hulle, Middell, Neyt, and Van Zundert, 2015). The aim of this panel is to investigate in detail the different affordances and modalities that graph models offer to information modeling when it comes to text research. To this end we have gathered a number of highly visible scholars that are active in this area of development.

Our intent is to explore the multifarious potential of graph models; it is explicitly not to frame graph-based approaches as a solution to certain narrow, albeit real, problems of hierarchy-based markup. We think there is potential for graph models well beyond markup and annotation: that is, to develop into a generalized humanities information modeling technique, provided we can restrain the urge towards premature unification, standardization, and over-specification of the application domain (e.g. annotation and markup). We will therefore seek to discuss a fairly wide range of affordances that text-as-graph models give us. In order to facilitate our aim, the panel will begin with five 10-minute papers highlighting various topics concerning graph models for text. The presentations will be followed by a 10-minute general Q&A phase, after which the panel will then move into an open discussion round of about thirty minutes. Audience participation will be maximally fostered by the moderator, who will take questions from the audience in the room as well as selected questions posted to a Twitter hashtag, to be advertised for the discussion.

Panel Member Abstracts

Tara Andrews

Dynamic modelling of textual argument

It has been acknowledged for some time that a graph data structure is a very good way of representing and exploring variation in a textual tradition (Schmidt and Colomb, 2009; Andrews and Macé, 2013; Haentjens Dekker, Van Hulle, Middell, Neyt, and Van Zundert, 2015). My own recent work takes the variant graph as the starting point to model a digital edition—that is, the variants, the relationships between those variants, the correspondence between variants and the manuscripts or documents in which they appear, and the way in which all of this information is used to produce a new version of the text, which is to say, the edition.

An important aspect of this model of the edition is that it encompasses more than the structure as saved in a graph database; the model also extends into the custom Java code into which the database is embedded. The code thus serves a function of validation of new data that is similar to what an XML schema might provide, but has the additional capacity to work out the logical consequences of an editorial decision and add these consequences explicitly to the data structure. For example, if an editor claims that reading A and reading B are spelling variants of each other, and then claims that reading B and reading C are different grammatical forms of the

same root word, then it must follow that reading A and reading C have the same grammatical relationship to each other. In this way, a consistency of logic is preserved in the model that would be difficult to achieve without custom code, and the model of the edition must necessarily include that code. My contribution to the panel will be a discussion of how graph data structures, and the infrastructure that has been built around them, contribute to the creation of these sorts of complex and dynamic models.

Thomas Efer

Flexible building blocks

All models are wrong - on purpose. Through varying degrees of simplification, idealization, abstraction and formalization, they can produce just the right amount of complexity for the task at hand. Any such "good" model can become a valuable proxy for the "thing itself". An interesting variant is the (logical) data model: Oblivious of a "thing itself" it only provides building blocks with which to build larger digital models under certain well-exposed paradigms (see e.g. Flanders and Jannidis, 2017).

The main question of data model choice is neither if the collected data “naturally” fits into such a paradigm nor whether the thing itself really consist of such building blocks (Efer, 2017) Important is how the building blocks can help to handle the important data about the thing - on many levels: First, the digitization, discretization, formalization, categorization, binarization, … and storing of data entries. Second, the modes of accessing those entries together with their contexts; the options of querying for patterns and the creation of analyses and reports. Lastly, the annotation, augmentation and exchange of records (or the whole data collection), accompanied by auxiliary tasks such as version control and provenience tracking.

Property Graphs greatly assist with these aspects by providing flexible building blocks for modelling texts in the digital humanities (Efer, 2016). In this special application domain it is oftentimes not clear in advance, what specific storage, access and augmentation steps are required to answer future research questions. When there is a slight chance that a need for multi-aspectual hierarchization, flexible variant representation, complex and interlinked annotations or such could arise, a decision for a graph-based data model should be considered. In addition, when it is necessary to communicate the internal structure of the collection the and mechanisms of algorithms to humanities scholars, the graph paradigm with its simplicity can be a valuable means of communication.

Stefan Jänicke

Conveying visual meaning

Graph visualizations are the means of choice to communicate relationships among textual entities (Jänicke, Franzini, Cheema, and Scheuermann, 2015). Nodes of a graph may be representatives for texts of a collection, and links might illustrate similarities concerning metadata or textual contents (e.g., Eder, 2014). In social network visualizations, nodes usually

stand for individuals or characters, and links for acquaintance (e.g., Klein, 2012). Taking text-as-a-graph models, nodes represent elements of the text hierarchy (e.g., words, sentences, or paragraphs), and links can be used to express the text flow or intertextual relations. Unfortunately, most digital humanities applications use out-of-the-box tools for visualizing graphs, so that a meaningful conjunction between textual and visual features often remains undone. In addition, as illustrated in Figure 2, basic graphical elements of the graph like circles and arrows lead to a visual overload of the screen that impedes the close reading of the laid out text. In my work on *<placeholder-tool>* (<placeholder-ref>), I use text-as-a-graph to model relationships among text editions, which are visualized as stream graphs to outline patterns of line similarity (see Figure 1a), and a variant graph visualization (Jänicke, Geßner, Franzini, Terras, Mahony, and Scheuermann, 2015) is adopted to illustrate variances on word level among similar lines (see Figure 1b). In contrast to traditional graphing tools, this use case shows that close readings can be intuitively extended by displaying relational information.

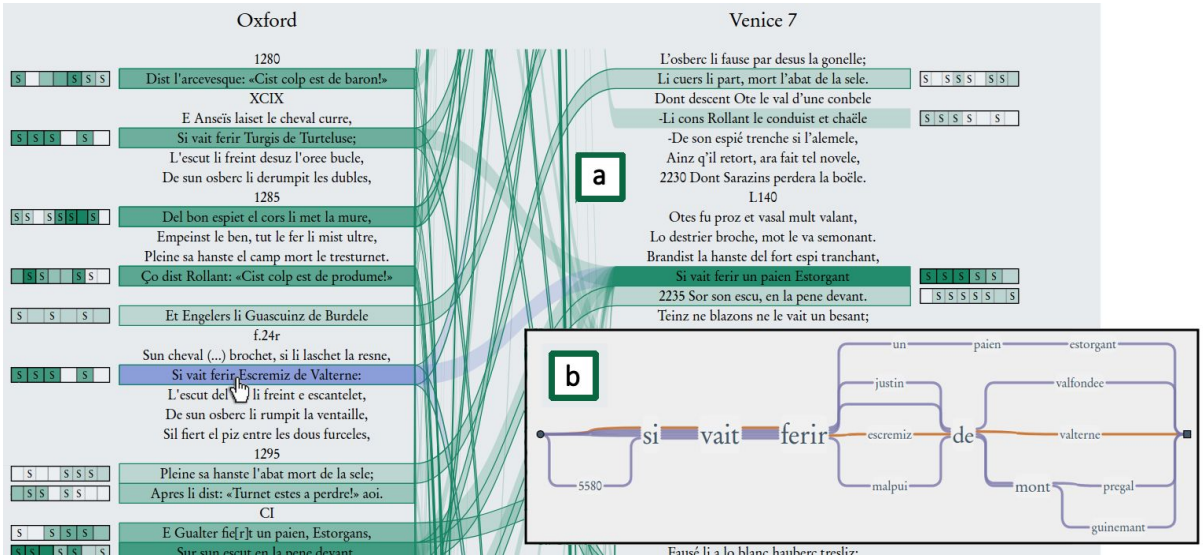

Figure 1: Textual variation on different text hierarchy levels

Regarded from another viewpoint, text-as-a-graph models are straightforwardly accessible to feed distant readings. While tag cloud visualizations are capable of illustrating unstructured textual information of text-as-a-graph collections, annotated metadata such as geo-references or time stamps can be collectively visualized in maps and timelines. Beyond pure text contents and metadata, I will discuss the adoption of layered graph drawings (Sugiyama et al., 1981) that are capable of conveying the complexity of text-as-a-graph models.

Andreas Kuczera

Complex annotation

When the Regesta Imperii (Zimmermann, 2000) were digitized, the project members had no hesitation about using XML for the Regesta-Texts (Weller, 2014; Schulz, 2017; http://www.regesta-imperii.de) but it was used in a very flat way, in order to make it easy to translate the XML directly to HTML. XML in connection with TEI has been a "standard" for digital scholarly edition projects for years now. But over the years data has become more and more connected.

For our most recent project proposal, to create the digital scholarly edition of the Windeck chronicle based on 5 different manuscripts and a print from the late 19th century, we thought about an xml-based approach but opted instead for text-as-a-graph and proposed a graph-based digital-edition.

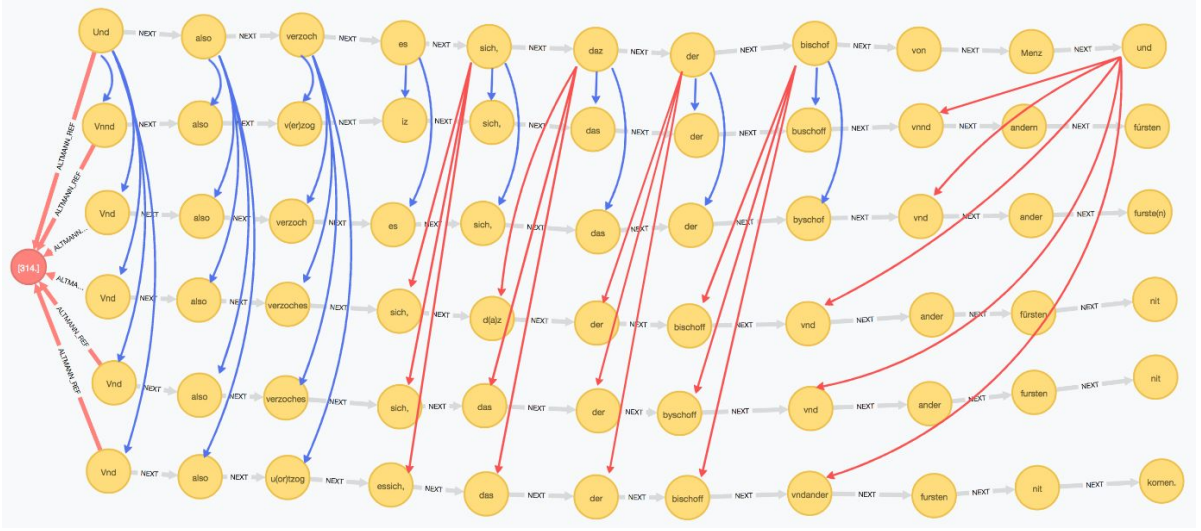

Figure 2: Text-as-a-graph model for six editions of the windeck chronicle

The picture shows the 6 texts as separate chains of word nodes, but in between there are edges with information about the relations between the word nodes. As we use a property graph, every word node can have multiple properties with further information about the word (e.g. Lemma, Word index number, Link to a Lexicon etc.)

This use of the property graph gives us at least these advantages:

1. No problems with overlapping hierarchies.
2. The possibility to use multiple annotation layers.
3. The possibility to model different users needs in one place.
4. With the flexibility of the graph existing annotation models can be adapted and even rearranged.

The example shows the flexibility in the granularity of the graph model. The question I will post for the panel is, can we find a lightweight model for text-as-a-graph that can handle text phenomena over a wide range of projects from different disciplines (Kuczera, 2017) ?

Joris van Zundert

Text Models as Simulacra of the Textual Condition

Baudrillard (1988), writing from a cultural critical perspective, concluded that "it is always a false problem to want to restore the truth beneath the simulacrum". Digital scholarly editions as digital facsimiles create a simulacrum too. Obviously, contrary to Baudrillard's pessimistic views, this simulacrum—a virtual world of digitally remediated texts—creates useful affordances for text analytical research. The current generation of digital scholarly editions amounts to iconolatry in a Baudrillardian sense: they reify the book rather than the text. But if an aim of textual scholarship is indeed to create an archive of philological fact (McGann, 2013), then philology should indeed push for the most ideal skeuomorphic designs that can be realized in the digital medium, for instance by embracing virtual reality facsimiles, even if such virtuality creates its own problems of representation (see for instance Burns, 2014:157).

What tends to be overlooked, however, is that digital computing is a full simulacrum in all its abstraction layers, from the bytecode up to the GUIs. There is no inherent truth in any of these computational layers and they do not represent any connection to textual reality except insofar as we interpret them as useful descriptions thereof. Therefore we commit to an epistemological fallacy if we enact screen essentialism (Kirschenbaum, 2008:27) as the sole actualization of digital textual scholarship. My argument is that if we solely travel the plane of GUIs and mostly reify digital simulacra of books, we do not exhaust the modalities of reading and affordances for interpretation that the computational simulacrum offers us. Beneath the opaque plane of GUIs, that depict rather shallow surfaces of texts, there is the as yet little traversed plane of the text model. There is little denying that the TEI represents the currently by far most widespread model in use in digital textual scholarship. But TEI is only one model in what is only one category of text models, which is to say, markup models. There are also, at the very least, the categories of relational database models and the category of graph models.

Graph models eradicate a number of the constraints that the database and markup models put on text modeling (Efer, 2016). More importantly to me, however, is that the advent of graph models for text signifies our ability to create many very different simulacra for the textual condition (McGann, 1991). Traveling down a level of abstraction to a plane where, instead of one model, many models become possible allows us to create very different hyperrealities (Baudrillard, 1988), each potentially highlighting excitingly different aspects of text and the textual condition.

References

Andrews, T. L., and Macé, C. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*, *28*(4): 504–21. 10.1093/llc/fqt032.

Baudrillard, J. (1988a). *Selected Writings*. (M. Poster, Ed.). Stanford: Stanford University Press.

Baudrillard, J. (1988b). Simulacra and Simulations. In *Selected Writings*. Stanford: Stanford University Press, pp. 166–184.

Burns, J. E. (2014). Digital Facsimiles and the Modern Viewer: Medieval Manuscripts and Archival Practice in the Age of New Media. *Art Documentation: Journal of the Art Libraries Society of North America*, *33*(2): 148–67. 10.1086/678515.

Devlin, K. (1995). *Logic and Information*. Cambridge University Press.

Eder, M. (2014). Stylometry, network analysis, and Latin literature. *Proceedings of the Digital Humanities 2014.*

Efer, T. (2016). *Graphdatenbanken für die textorientierten e-Humanities*. Dr.rer.nat., Universität Leipzig. http://www.qucosa.de/fileadmin/data/qucosa/documents/21912/Dissertation_Thomas_Efer.pdf (accessed 17 June 2018).

Efer, T. (2017). Introducing NoXML for the Digital Humanities. In Eibl, M. and Gaedke, M. (eds.), *INFORMATIK 2017*. Gesellschaft für Informatik, Bonn.

Flanders, J., and Jannidis, F. (2015). Data Modeling. In Schreibman, S., Siemens, R., and Unsworth, J. (eds.), *A New Companion to Digital Humanities*. Chichester: Wiley Blackwell, pp. 229–237.

Haentjens Dekker, R., Van Hulle, D., Middell, G., Neyt, V., and van Zundert, J. (2015). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Literary and Linguistic Computing*, *30*(3): 452–70. 10.1093/llc/fqu007.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. 10.2312/eurovisstar.20151113.

Jänicke, S., Geßner, A., Franzini, G., Terras, M., Mahony, S., and Scheuermann, G. (2015). TRAViz: A Visualization for Variant Graphs. *Digital Scholarship in the Humanities*, *30*(suppl_1): i83–99. 10.1093/llc/fqv049.

Kirschenbaum, M. G. (2008). *Mechanisms: New Media and the Forensic Imagination*. MIT Press.

Klein, L. (2012). Social Network Analysis and Visualization in 'The Papers of Thomas Jefferson'. *Proceedings of the Digital Humanities 2012.*

Kuczera, A. (2017). Graphentechnologien in den Digitalen Geisteswissenschaften. *ABI Technik*, *37*(3): 179–196. 10.1515/abitech-2017-0042.

McGann, J. (2013). Philology in a New Key. *Critical Inquiry*, *39*(2): 327–46. 10.1086/668528.

McGann, J. J. (1991). *The textual condition*. Princeton, N.J: Princeton University Press.

Schmidt, D., and Colomb, R. (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, *67*: 497–514.

Schulz, J. (2017). Regesta Imperii Online. *RIDE: A review journal for digital editions and resources*, 6. https://ride.i-d-e.de/issues/issue-6/regesta-imperii-online/ (accessed 17 June 2018).

Sugiyama, K., Tagawa, S., and Toda, M. (1981). Methods for visual understanding of hierarchical system structures. *IEEE Transactions on Systems, Man, and Cybernetics, 11*(2), pp. 109-125.

Thaller, M. (2017). Historical Information Science: Is there such a Thing? New Comments on an old Idea [1993]. *Historical Social Research / Historische Sozialforschung. Supplement*, (29): 260–86.

Thaller, M. (2018, April 24). On Information in Historical Sources. *A Digital Ivory Tower*. Billet. https://ivorytower.hypotheses.org/56 (accessed 17 June 2018).

Weller, T. (2014). Die Regesta Imperii online. *Rheinische Vierteljahrsblätter*, *78*: 234.

Zimmermann, H. (2000). *Die Regesta Imperii im Fortschreiten und Fortschritt* (Vol. 20). Köln [u.a.].