Journal of the Text Encoding Initiative

Issue 8 | December 2014 - December 2015
Selected Papers from the 2013 TEI Conference

From Entity Description to Semantic Analysis: The Case of Theodor Fontane's Notebooks

Martin de la Iglesia and Mathias Göbel

Electronic version
URL: http://journals.openedition.org/jtei/1253
DOI: 10.4000/jtei.1253
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Martin de la Iglesia and Mathias Göbel, « From Entity Description to Semantic Analysis: The Case of Theodor Fontane's Notebooks », *Journal of the Text Encoding Initiative* [Online], Issue 8 | December 2014 - December 2015, Online since 09 June 2015, connection on 07 May 2019. URL : http://journals.openedition.org/jtei/1253 ; DOI : 10.4000/jtei.1253

From Entity Description to Semantic Analysis: The Case of Theodor Fontane's Notebooks

Martin de la Iglesia and Mathias Göbel

1. Project Overview

1 The German writer Theodor Fontane (1819–1898) produced many novels, poems, essays, and other types of texts, including the notebooks which we are currently editing (Radecke 2010, Radecke 2013). Fontane's notebooks constitute one of the last bodies of his manuscript materials that is still not entirely published; only a few excerpts have been published, and these do not comply with standards of textual criticism and editing (Radecke 2010). From 1859 until the end of the 1880s, Fontane filled notebooks with a mass of miscellaneous materials, such as diary entries, letter drafts, plans for poetic writing, prose sketches and drafts, lecture notes, drafts for theater and arts reviews, book excerpts, and also notes and sketches accumulated on his journeys. A total of 67 of these notebooks are known to exist, most of them measuring approximately 10 × 17 cm and containing between 64 and 180 leaves. Some pages have been left blank, so we are dealing with

less than 10,000 pages with writing on them. Fontane's handwriting is often hard to decipher, which means that we need to put a lot of effort into transcribing it. Once published, though, this transcription will be a considerable benefit to the reader.

2 We plan to release our edition in 2017 when our project is scheduled to end. The project is titled *Genetic-Critical and Annotated Hybrid-Edition of Theodor Fontane's Notebooks Based on a Virtual Research Environment*.[1] The term *hybrid edition* indicates that it is going to be published both online, in open access form, and as a printed book by Walter de Gruyter. The term *Virtual Research Environment* in the title of the project refers to TextGrid,[2] which we use in combination with the oXygen XML Editor[3] to produce our TEI-encoded transcription as well as to store both the digital images and the XML files permanently. The project is funded by the German Research Foundation (DFG) and carried out by the Theodor Fontane Research Centre at Göttingen University, in cooperation with Göttingen State and University Library. The Berlin State Library is the owner of the notebooks and an associated partner of the project. Our core project team consists of three philologists at the University, including the project manager and official editor of the edition, Gabriele Radecke, and, at the university library, an IT and a metadata specialist. The workflow has been set up as follows: the philologists survey the material, determine the editorial concept and principles, transcribe the notebooks, and encode them in TEI according to a TEI P5 subset scheme based on the elements for genetic editions (e.g., `<sourceDoc>` and `<surface>`) specified by the metadata specialist. This TEI code is then XSL-transformed into HTML and integrated into the project's website by the IT specialist. This transformation can be invoked from within the TextGrid environment at any time. Thus, the HTML file not only is used for publication on the website, but also supports the editors in transcribing and encoding the text of the notebooks by providing an on-the-fly visualization of their work which is often easier to check for errors than the XML text. (A detailed description of this workflow can be found in Radecke, Göbel, and Söring 2013.)

3 All of the examples in this paper use data from one of the 67 notebooks (shelf mark C07). C07 contains some of the notes taken during or shortly after Fontane's second journey to the German regions of Thuringia and Franconia in the summer of 1873. Fontane visited several different places there, including the battlefields of the Napoleonic Wars of the early nineteenth century, and sites related to the sixteenth-century Protestant reformer Martin Luther. Some notes in C07 refer to a book about Thuringia that Fontane planned to write, but apparently never did (Wüsten 1973).

2. Encoding References to Entities in Theodor Fontane's Notebooks

4 Obviously, a notebook about Thuringia and its history is bound to contain many references to entities, primarily places and persons, and also dates. In order for these references to be analyzed, they first had to be verified by the philologists and then tagged as references to entities in our TEI code. Despite the progress in automatic Named Entity Recognition technology (Wettlaufer and Thotempudi 2013), it turned out to be less useful for our purposes as the dataset we are dealing with is relatively small and we aim for high precision in identifying references to entities, including indirect references such as pronouns. Therefore we are identifying and tagging all references manually. For notebook C07 we have already completed this task, using the TEI element `<rs>` (referencing string) to point out words in the notebooks which refer to entities.

Example 1. Example `<rs>`.

```
  <line>5. <seg><rs type="direct" ref="#Luther">Luther</rs></seg> tritt als
Mönch</line>
  <line>in das <seg><rs type="direct" ref="#Kloster_EF">Auguſtinerkloſter</rs></
seg></line>
```

5 Reference attributes (`@ref`) point to nodes located elsewhere in the TEI dataset. It should be noted that the organization of the TEI dataset and the location of the entity notes therein is of no importance to the reference linking mechanism described here. At the current stage of our project, the TEI code for each of the 67 notebooks is stored in its own TEI document, with the entity indexes stored within each TEI header. However, combining the 67 TEI documents into one large document with the entity information stored in a single header, or placing all entity data in a separate TEI document altogether, would be just as feasible. In the entity data nodes, additional information is provided on the entities, such as hyperlinks to authority file records, or a classification into *person* or *place*.

Example 2. Example of `<person>` and `<place>`.

```
<sourceDesc>
 <listPerson>
  <!-- ... -->
  <person xml:id="Luther">
   <idno type="GND">118575449</idno>
  </person>
  <!-- ... -->
 </listPerson>
 <listPlace>
  <!-- ... -->
  <place xml:id="Erfurt">
   <idno type="GeoNames">2929670</idno>
   <place xml:id="Kloster_EF">
    <idno type="OpenStreetMap">164587492</idno>
   </place>
  </place>
  <!-- ... -->
 </listPlace>
 <!-- ... -->
</sourceDesc>
```

6 We will show further entity types in a later example. In the case of the authority files, the pointers from the TEI code can be either simple URIs (or just authority file identifiers stored in `<idno>` elements from which URIs can be easily built), or qualified hyperlinks (e.g., in `<relation>` elements) using a vocabulary such as CIDOC CRM (Le Boeuf et al. 2013) in order to express relations which may support a Linked Data structure. The processing of these links to authority records as described in this article works in either case, as long as valid URIs are provided. The philologists validate the entries, and if there is an error within the hyperlinked authority data we will store corrected values or extended information in our TEI dataset. Chronological references are encoded in a simpler way, using the `<date>` element and a direct normalization within attributes of the `att.datable` class.

7 Our method of encoding references to entities is fairly similar to that employed in other TEI editions. For instance, in the edition of William Godwin's Diary (Myers, O'Shaughnessy, and Philp 2010), the more specific elements like `<persName>` and `<placeName>` are used instead of the `<rs>` element. In the Godwin edition, a `@ref` attribute points to an HTML website, which however is in most cases based on a TEI document consisting of a `<person>` element. Within this element, the normalized name of the person and biographical information are provided. Many other TEI projects use entity referencing mechanisms like this, even though the elements, attributes, and attribute values may vary. In many cases, tools for entity analysis and visualization can be applied to different data sources with minimal adaptation effort, so no standardization of the respective TEI source code is required. A problem we did encounter is that some TEI projects do not provide direct access to their XML files, which makes them harder to process automatically.

3. Semantic Analysis

8 Many TEI edition projects already encode references to entities, but semantic analysis entails far more than just knowing which entities have been mentioned by an author. Semantic analysis as we understand it is a methodological approach that builds upon such tagged entities. Applications query these entity data, aggregate them, possibly enrich them with or link them to external data, perform calculations on them (e.g., sorting algorithms), and generate different kinds of output. This output can be in the shape of text, images, moving images, or potentially any other medium, and can be either simple and concise or complex and rich, but may provide new insights into the data present in the TEI-encoded material. This kind of analysis shifts the focus from the written signifiers to the actual signified entities, and tries to make claims about the mentioned persons, places, dates, works, events and other entities, and their interrelations. This in turn allows conclusions about the TEI-encoded work in which these are referenced, as we show in the examples below.

9 Such semantic analysis applications are typically external to the TEI code, but need not be external to the TEI-encoded edition as a whole.[4] For instance, this kind of application might be offered as part of the edition website or might be run as a standalone tool by a third party, querying remote TEI data. These possibilities are discussed in the final section of this article.

10 The semantic analysis methods described here should not be confused with Linked Data processing (Berners-Lee 2006), which relies on explicitly qualified relations between entities (typically in the form of RDF triples). Such data can be derived from the TEI data discussed here, but no Linked Data as such is present in the datasets used in our project.

3.1 Timeline

11 As our first example of semantic exploration, we would like to take a look at the persons mentioned in Fontane's notebook C07, and ask: who were those people, or more specifically: were they Fontane's contemporaries or, from Fontane's point of view, historical figures? In the latter case, in which period did they live? Answering these questions will give us an idea of the different historical strata treated in this notebook. Our tool for this purpose will be the Timeline Widget[5] from the SIMILE (Semantic Interoperability of Metadata and Information in unLike Environments) collection of open source data visualization tools, originally developed at MIT. A SIMILE Timeline consists of events, associated with either a point in time or a duration, which are plotted on a chronological, in this case horizontal, axis. If we want to visualize the persons mentioned in the notebook on such a timeline, and use their lifespans as durations, where do we get their birth and death dates? As mentioned earlier, our references to entities are linked to authority records. In the case of persons, we found the Integrated Authority File (German "Gemeinsame Normdatei," GND)[6] by the German National Library to be the most useful and complete data source for our purpose. We have assigned a GND identifier to 37 of the 47 person entities referenced in the notebook C07. The remaining 10 entities are groups of people rather than individuals, such as "the German emperors" or "the Thuringian landgraves." Although the GND does provide records for some of these entities, they do not contain much useful information for further investigation.

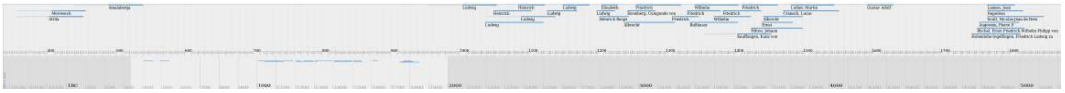
12 The SIMILE Timeline widget consists of an HTML document that uses JavaScript to process an XML file written in a simple XML markup language specific to SIMILE.

Example 3. Example of `SIMILE XML`.

```
<data>
 <event durationEvent="true" title="Luther, Martin" start="1483" end="1546"/>
 <!-- ... -->
</data>
```

13 We have written an XSLT stylesheet in order to produce the HTML document and at the same time generate the required XML data from the TEI code of our notebook. During the transformation, this stylesheet picks up the GND identifiers of all person entities in the TEI code and looks up the corresponding GND record online at the German National Library. It then fetches the date of birth and date of death of the persons from the GND RDF/XML record, as well as their normalized names, and uses these names as labels in the timeline. With only 100 lines of code, this XSL document is quite short and simple, but still checks for missing birth or death dates in the GND record to substitute a calculated estimate for the missing value, as we will show below. Person entities for which neither a precise birth date nor death date can be found in the GND are not included in the timeline at all, which is the case for one of the 37 persons referenced in the notebook (the Thuringian king Hermannfrit, whose birth and death dates are provided in a non-standard form as the text string "ca. ?–531,"[7] which is not recognized by the stylesheet used).

Figure 1. Complete SIMILE timeline for notebook C07[8]



14 The resulting timeline (figure 1) consists of two bands, the lower one aggregating the upper one on a larger scale. In the upper band, light blue bars indicate that the date of birth (that is, the beginning of the bar) is an estimate, and only the death date (the end of the bar) could be fetched from the GND record. Our timeline begins in the fifth century CE, when Thuringia was under Frankish rule and the Frankish king Merowech defended Thuringia against the attacks of Attila the Hun. In the Middle Ages we see the long line of Thuringian landgraves (most of them called either Ludwig or Friedrich) who ruled Thuringia until the year 1440 when it became part of Saxony. The timeline ends in the early nineteenth century with Napoleon and other military commanders who fought in the battles of Jena and Auerstedt during the Napoleonic Wars. This timeline enables us to tell which historic periods are covered in the notebook, if we assume that the mentioned persons are a suitable indicator for that. We can also see that, at least in this notebook, Fontane did not mention any of his contemporaries.

15 It can be very interesting to compare the timeline of one of Fontane's notebooks to a timeline created from a different TEI data source. For this purpose we again turn to the edition of William Godwin's Diary, as it resembles Fontane's notebooks in both the time of creation and the nature of the material. Of course, a diary is a different medium than a notebook, but both contain a sufficient number of references to person entities that we can display on a timeline. To narrow the material down, we selected a single year's worth of diary entries, for the year 1835, which is the last complete year within the scope of Godwin's diaries. The XSLT code to create the timeline has to be adjusted to the Godwin TEI code, though only slightly. The main difference is that birth and death dates are already contained within the `<person>` elements, and do not need to be retrieved from elsewhere. Again, a filter was applied for missing birth and death dates. Overall, this XSL document is even shorter and simpler at only 75 lines.

16 The resulting timeline (figure 2) shows us many people who were all either alive in 1835, or recently deceased, such as a John Curran. (Godwin refers in his diary to the removal of Curran's body from London to Ireland at the time of writing.)

Figure 2. SIMILE timeline for William Godwin's diary entries of the year 1835.

17 The comparison of the two timelines shows a marked difference: while Fontane exclusively mentions historical figures in his notebook C07, Godwin's diary entries of 1835 are concerned with the present, as can be expected from a diary. The advantage of this method is that it may give an overview of some aspects of the content of large amounts of textual data in a short time, without being prone to the bias of human annotators and indexers.

3.2 Geospatial-Temporal Data Aggregation

18 The referencing of places within our encoded documents is also realized with `<rs>` and a corresponding node within the `<teiHeader>` connected via `@ref`. Sixty-eight different places are mentioned in notebook C07.

19 We used the authority files GeoNames[9] and OpenStreetMap (OSM),[10] which provide the required data, and we manually selected identifiers from these databases and added them to the TEI dataset, similarly to the procedure for the personal data described above. Again, the use of automatic systems to identify the historical places would not be feasible, as they are sometimes referred to in the notebooks by uncommon phrases which require human interpretation to match them to corresponding modern-day identifiers. Another XSLT script is used to transform the TEI dataset to a Keyhole Markup Language (KML) file, the typical input format for geospatial visualization tools. The XSLT resolves the IDs and retrieves the coordinates from the respective database. OSM is able to deliver polygons instead of geographic coordinates from a single URL in the resulting file: for example, for All Saints' Church in Wittenberg or the Saint Augustine Monastery at Erfurt. Again the transformation is very simple, as the common format for spatial information—KML—is also expressed as XML. This approach tries to provide a possible combination of place names that appear in the neighborhood of `<date>` elements. The geospatial-temporal visualization tool of our choice is the DARIAH Geo-Browser,[11] which offers a timeline with a map interface together with different features for selecting data.

20 A notable feature is the use of historical maps to provide better context. The selection of historical maps in the Geo-Browser is still limited, but it is also possible to load one's own overlays. Furthermore, the data are presented in tabular form with a search function. In addition to the mandatory data—at least one place name with latitude and longitude—HTML code can be inserted in the KML file. A useful method is to pass back links to the digital edition or, more specifically,

to the page where the selected place can be found in the manuscript. These hyper-references will appear directly in the geographical information system. We integrated an embedded version of this tool via `<html:iframe>` into our website which is built on an eXist database.[12] An XQuery script executes the transformation, stores the KML file in the database, and generates the required `<html:iframe>` element. The `<html:src>` attribute value contains the parameters to control the Geo-Browser. A URL-encoded string which points to the previously-generated KML file within our database is passed to the tool.

21 To find dates corresponding to the named places, we selected an interval from eight preceding or following elements starting from the matching `<rs>` of the place. We prefer the following dates if both the nearest preceding and the nearest following element are at an equal distance. The idea behind this matching criterion is that the proximity of place and time references in the notebooks suggests a semantic link. Ideally, they describe where and when one single event took place. The chosen interval of up to eight steps was determined by trial and error and yields the highest number of meaningful matches. In the future, users should be able to specify this interval as well as the priority of the date, if there is one with the same distance left and right from an entity. In the case of notebook C07, 24 items appear in the Geo-Browser's table. The Geo-Browser only displays 17 of our 24 items because the others refer to more abstract terms like the Holy Roman Empire ("HRR") or Thuringia, and the authority files we use are not able to provide the required data, especially not for the desired time.

22 The dates returned by this algorithm are also used to select a background map provided by the Geo-Browser. The arithmetic mean of the temporal data in this example is 1451.25 CE and 11 out of 17 dates with spatial reference specify the early sixteenth century. A suitable background map with the borders from 1492 can be selected via parameter (`&currentStatus=mapChanged=Historical+Map+of+1492`). Another possible solution is to use the information about the notebook's date of creation, which can be found on the book cover in most cases. Then the background map of 1880 is the best selection for notebook C07.

Figure 3. DARIAH-DE-Geo-Browser's timeline with the data from notebook C07.

23 Placenames corresponding to a selection in the timeline are displayed in a table below, where backlinks to the edition can be placed as well as any additional information. The table also includes the terms and values we are not able to show in the map, like the name "HRR" we described above.

3.3 Network Analysis

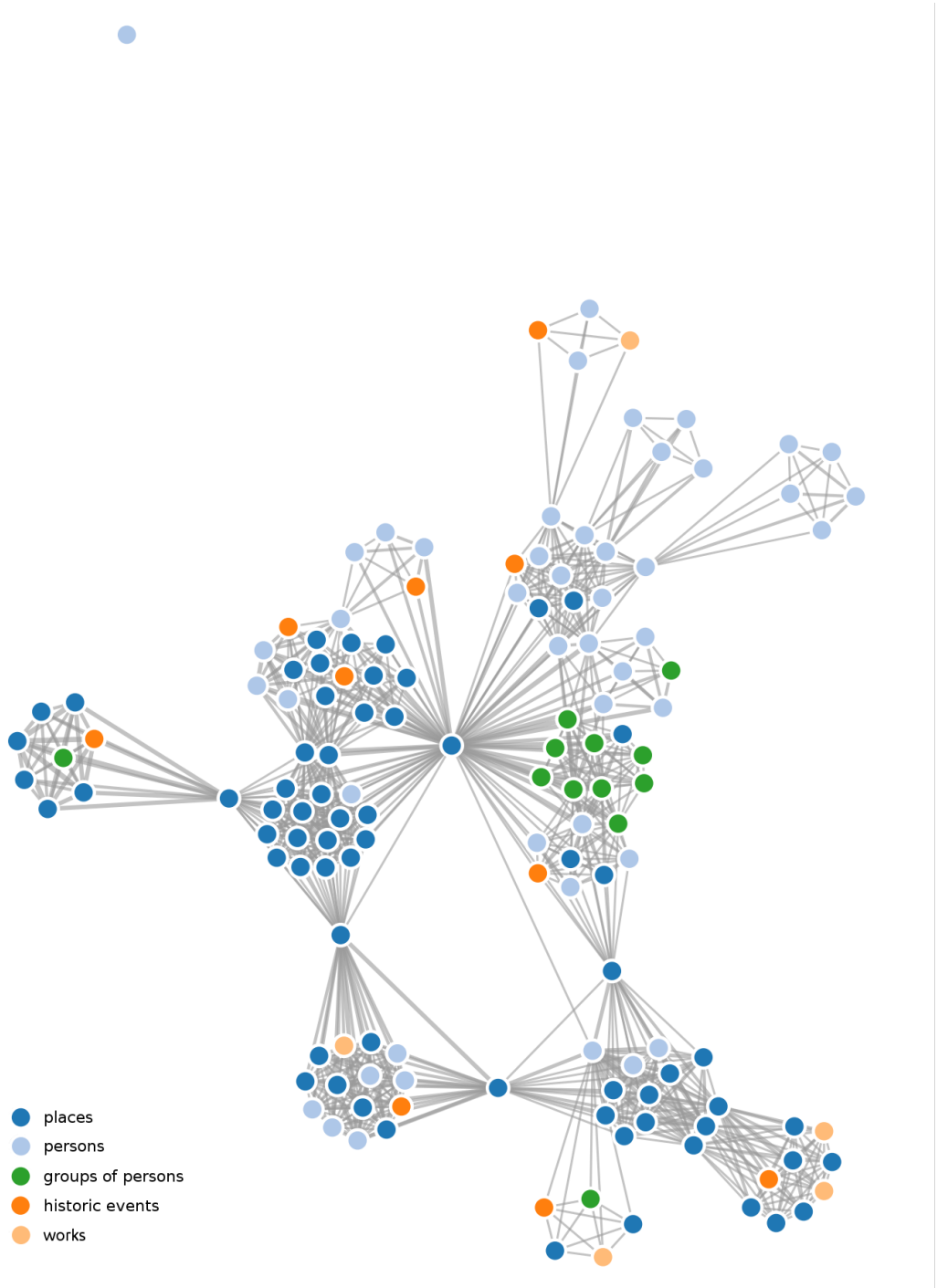
24 The examples above investigate the persons, dates, and places appearing in the text. As mentioned, groups of people have been excluded from our investigation, and there are even more kinds of entities that can be identified in the notebooks: historical events, organizations, and works. To bring them into context with the others in order to generate an aggregated view of all named entities, a network graph can be built. Network visualization is not new to TEI data. Bingenheimer, Hung, and Wiles (2011) introduce "a way of visualizing social networks extracted from a TEI-encoded corpus" (p. 271) consisting of biographic data. The interface is realized with a proprietary plug-in built upon the Prefuse[13] software library. One of our goals is to implement the aggregations within the digital edition, and for this we would like to use web technologies only. The D3.js (Data Driven Documents Javascript library) created by Mike Bostock provides a framework for different visualizations. The list of examples[14] is a good starting point and gives an introduction to the tool's functionality. The following implementations are reproductions of the Force-Directed Graph[15] and the Hierarchical Edge Bundling[16] examples. Again we use an XSL transformation to extract the data from our source.

25 The transformation script matches all entities and generates the required documents. The first document is the HTML file, which contains the needed JavaScript and a reference to the external D3.js library. The second is a JSON file, which contains one object per entity and one associated array per object that includes a list of connected entities. The tree-like structure of XML allows the transformation of any document to a network graph by selecting elements that share the same ancestor. The only requirement is that the XML input consists of at least two elements. For example, the `<surface>` element (which is used in the Fontane TEI data to encode a notebook page) is a common ancestor of the entities. This produces a network of co-occurrences; the (undirected) edges mean that the connected nodes are both descendants of a `<surface>`; both appear on one

page, if the condition is defined to match only the <surface> elements that are direct children of <sourceDoc>, because more than one surface may be part of a single page, for example where there are glued-in newspaper articles.

26 As in the Geo-Browser example above, we assume that the proximity of two references to entities suggests a semantic connection. Naturally, such a connection may also exist between two entity occurrences separated by a page break. Therefore, better criteria for connectedness could be proposed, such as co-occurrence within one sentence, but this requires linguistic markup which is not part of our notebook edition. The notebook page as a unit of semantic coherence is still a relatively meaningful criterion, and the most feasible due to our usage of the <surface> element.

Figure 4. Force-Directed Graph of entities for notebook C07[17]



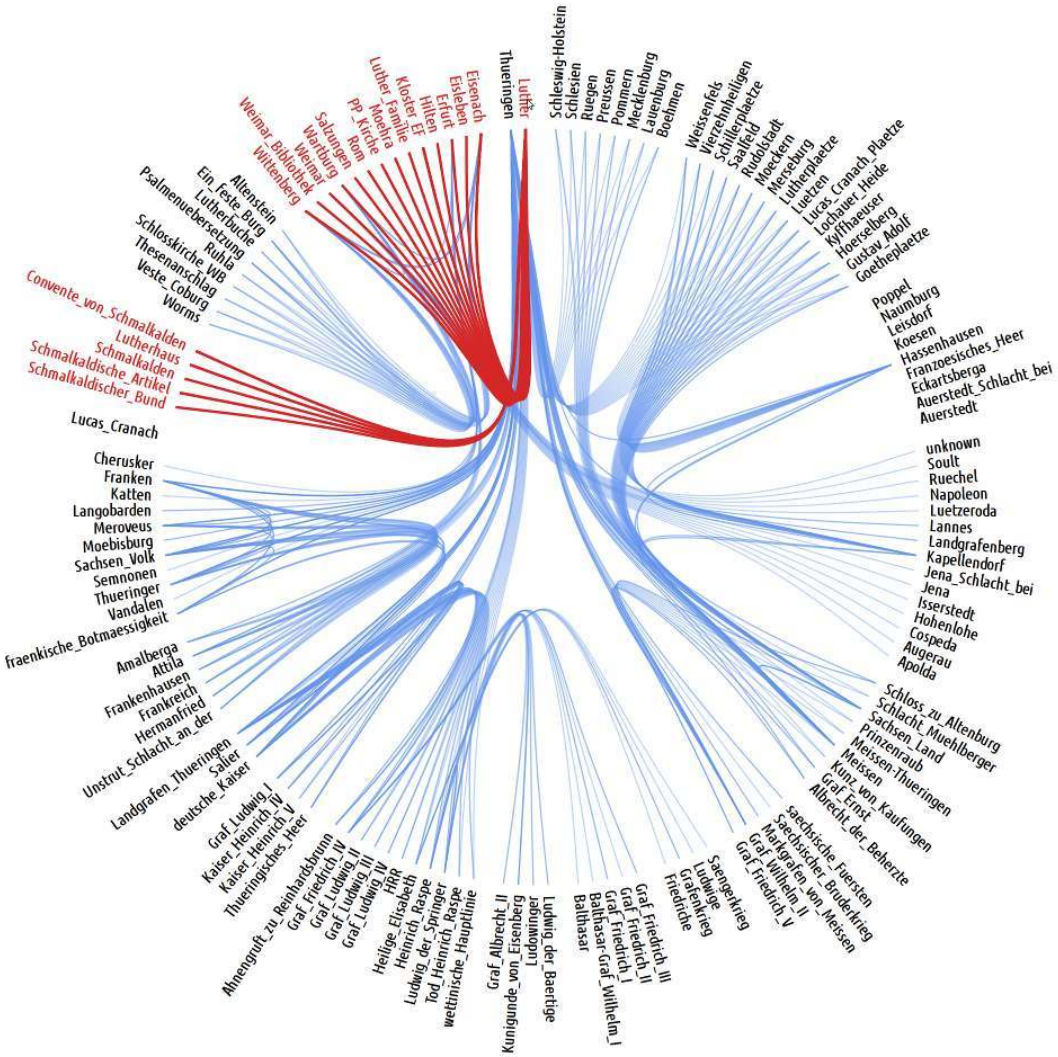
27 The Force-Directed Graph algorithm creates a network starting from randomized positions of the nodes and applies a weight for a single node and a link strength. Based on these values, a network is rendered. The result for C07 (figure 4) is a reliable network that shows Thuringia in a centered position with the most edges; it is the most frequently occurring entity and it is mentioned on several pages together with the connected ones. The other places with many links are Erfurt, Weimar, and Kapellendorf. Erfurt and Weimar were the cultural centers of Thuringia; Kapellendorf is a village where the last of the battles of Jena and Auerstedt during the fourth part of the Napoleonic Wars (War of the Fourth Coalition) took place. There is a part of the network where ethnic groups (green) appear together, which represents a page of notebook C07 (6 recto) on which Fontane describes the Thuringians in opposition to their neighbors, the Franks, Cherusci, Saxons, and others. This page's headline is "Thüringens Geschichte" (History of Thuringia), which is also the topic of the following pages. The benefit of the network is that a major topic can be identified with a single view.

28 The output of this D3.js application is an SVG graphic which can be further transformed. `<svg:title>` elements are used to store the node names, which modern browsers should display on mouseover. To get a better overview of the entities in the notebook, the node names should actually be inserted as nodes, but since there is not much space available in the Force-Directed Graph, a different design might be a better choice. The Hierarchical Edge Bundling example (figure 5) provides a circular layout with the nodes in alphabetical order on each level of hierarchy. Again, the hierarchy is based on the `<surface>` element, on which the attribute `@n` determines the level and will be expressed as the first part of the object *name* within the JSON file. In our case this attribute contains the leaf number with a letter `"r"` for recto and `"v"` for a verso side and this attribute is transformed in a `html:id`, so we can go back from a single entity to the leaf of its first occurrence by generating a hyperlink with the help of JavaScript. If this part is left out, the objects will be sorted in alphabetical order and the network will contain more edges to link those entities that co-occur on one page. Applying the hierarchy allows these edges to be deleted, because the categorization lets the nodes appear together and a bigger gap between the categories marks the border. This is one of the rare cases in which adding more information to a visualization simplifies and refines the output at the same time.

29 The result is an interactive graphic in which the appropriate edges are highlighted when the cursor is placed over a node. If one selects the node "Luther," all links and nodes that appear together with a reference to Martin Luther on a page will be highlighted. When one does this, the items within the first and third clusters change their color to red. Thus we get the same information as in the Force-Directed Graph: the topic of Martin Luther is confined to the first part of notebook C07, while Thuringia is the central topic of the rest of the document.

30 Both networks show an outlier. The personal entity of Lucas Cranach, a German Renaissance painter and a good friend of Luther and his wife, is located outside of the network. He appears as the only entity on one page. Furthermore his name is the only inscription on this page at all and it is followed by three blank pages. One possible way to integrate this entity into the network might be the use of another apportionment, as the fact that two entities are referenced on the same page may be regarded as artificial. Only minor changes would have to be made to the XSLT code to use other divisions of text, such as chapters. As notebooks are not typically organized in chapters, we can use paragraphs (encoded with `<milestone>` here) instead. Instead of one specific element, a distinctive number of blank surfaces in series can be interpreted as a marker between two sections. Based on this information, we can build a different network. This is experimental and aimed at defining the best clusters according to the purpose of our examination.

Figure 5. Hierarchical Edge Bundling (mouse cursor placed over the node "Luther")[18]



31 The network graphs above take advantage of the precisely marked up text and focus on all named entities. We seem to take these entities out of their contexts, but actually we make their contexts more clearly visible as we group them together where Fontane has written about them on the same page in his notebook. The networks still represent every surface, or represent the notebook, but the visualizations just pay attention to entities. It would not be difficult to apply these scripts to any other element to aggregate or summarize given information with the option to browse to the respective lines of text where the entity occurs. As our dataset grows, we might be able to categorize the notebooks or to distinguish literary manuscripts from other notes. At some point we

will be able to apply methods from network theory, for example to measure the centrality of nodes and compare the values from different networks for a single entity. This could provide insights into the structure of the analysed texts, e.g., regarding the identification of topics.

4. Conclusion

32 All the example applications we have presented in this article build upon existing tools and services. The additional effort required to make them work with the TEI data from the forthcoming edition of Theodor Fontane's notebooks (and from the edition of William Godwin's Diary) was minimal. The necessary scripting (mainly XSLT) and code customization were easily carried out in addition to our regular work within the Fontane edition project. These efforts were facilitated by a spirit of openness shared by all parties involved: both the D3.js library and the SIMILE Timeline widget are open-source software released under a BSD license; the data sources GND, GeoNames, and OpenStreetMap have permissive licenses—Creative Commons Zero (CC0), Creative Commons Attribution (CC BY), and Open Data Commons Open Database License (ODbL), respectively; and the data from William Godwin's Diary is released under a Creative Commons Attribution-NonCommercial license (CC BY-NC) and, just as importantly, is offered directly in TEI/XML.

33 The benefits of such an approach are undeniable: it enables researchers who are concerned with the meaning within textual material to explore questions that would otherwise be difficult and/or tedious to tackle, thus fulfilling the promise of the digital humanities. Therefore, we wonder why this kind of semantic exploration is not applied more often within the TEI community (or the digital humanities community as a whole). The lack of similar work is particularly regrettable because the true power of this approach would become apparent if there were more results to compare, so that a scholarly dialogue might ensue. Likewise, the availability of more datasets would increase the validity of the analyses performed on them, while more available tools would increase the possible research problems that could be examined through semantic investigation. Furthermore, this approach is a possible answer to the call for markup analysis that Fotis Jannidis started at a panel discussion during the DH2012 conference (Bauman et al. 2012). The work of scholars in the field of literature is more and more digital, but rather than using the richly encoded TEI document, plain text is still the most common format for text analysis.

34 The underlying question here is: who is responsible for carrying out the required work to develop, maintain, and customize tools for semantic exploration? It is naïve to believe that ready-to-use applications would emerge from "the Web" or "the community" on their own. Rather, all practitioners in the field ought to ask themselves what their contribution to this chain of development, use, sharing, and re-use could be. Instead of waiting for someone else to develop a generic tool that fits all purposes, anyone can make a small effort towards such a development by providing interchangeable data, using common standards, and building on pre-existing work. And with the philosophy of open source and sharing ideas in mind, we provide all our scripts as well as interactive examples at http://fontane-nb.dariah.eu/tei-conf/.

BIBLIOGRAPHY

Bauman, Syd, David Hoover, Karina Van Dalen-Oskam, and Wendell Piez. 2012. "Text Analysis Meets Text Encoding." Panel discussion at Digital Humanities 2012 conference, Hamburg, Germany, July 16–22.

Berners-Lee, Tim. 2006. "Linked Data." http://www.w3.org/DesignIssues/LinkedData.html.

Bingenheimer, Marcus, Jen-Jou Hung, and Simon Wiles. 2011. "Social Network Visualization from TEI Data." *Literary and Linguistic Computing* 26, no. 3: 271–78. http://llc.oxfordjournals.org/content/26/3/271.full.pdf+html. doi:10.1093/llc/fqr020.

Le Boeuf, Patrick, Martin Doerr, Christian Emil Ore, and Stephen Stead. 2013. *Definition of the CIDOC Conceptual Reference Model*, version 5.1.2 (October). http://www.cidoc-crm.org/docs/cidoc_crm_version_5.1.2.pdf.

Myers, Victoria, David O'Shaughnessy, and Mark Philp, eds. 2010. *The Diary of William Godwin*. Oxford: Oxford Digital Library. http://godwindiary.bodleian.ox.ac.uk/

Radecke, Gabriele. 2010. "Theodor Fontanes Notizbücher: Überlegungen zu einer überlieferungsadäquaten Edition." In *Materialität in der Editionswissenschaft* (= Beihefte zu editio, vol. 32), edited by Martin Schubert, 95–106. Berlin: De Gruyter.

———. 2013. "Notizbuch-Editionen: Zum philologischen Konzept der Genetisch-kritischen und kommentierten Hybrid-Edition von Theodor Fontanes Notizbüchern." *editio* 27: 149–72. doi:10.1515/editio-2013-010.

Radecke, Gabriele, Mathias Göbel, and Sibylle Söring. 2013. "Theodor Fontanes Notizbücher: Genetisch-kritische und kommentierte Hybrid-Edition, erstellt mit der Virtuellen Forschungsumgebung TextGrid." In *Evolution der Informationsinfrastruktur: Kooperation zwischen Bibliothek und Wissenschaft*, edited by Heike Neuroth, Norbert Lossau, and Andrea Rapp, 85–105. Glückstadt: Hülsbusch Verlag. http://webdoc.sub.gwdg.de/univerlag/2013/Neuroth_Festschrift.pdf.

Wettlaufer, Jörg, and Sree Ganesh Thotempudi. 2013. "Named Entity Recognition in Historical Texts from the Natural History Domain." Poster presented at Mehr Personen – Mehr Daten – Mehr Repositorien, Tagung des Personendatenrepositoriums der BBAW, Berlin, Germany, March 4–6. http://www.gcdh.de/files/2013/6429/9184/Wettlaufer_Thotempudi_2013_NER_final.pdf.

Wüsten, Sonja, ed. 1973. *Reisen in Thüringen: Notiz- und Tagebuchaufzeichnungen aus den Jahren 1867 und 1873.* By Theodor Fontane. Potsdam: Theodor-Fontane-Archiv der Dt. Staatsbibl.

NOTES

1 Theodor Fontane–Arbeitsstelle, *Genetisch-kritische und kommentierte Hybrid-Edition von Theodor Fontanes Notizbüchern basierend auf einer Virtuellen Forschungsumgebung*, http://fontane-notizbuecher.de/

2 TextGrid: A Virtual Research Environment for the Humanities, http://textgrid.de/

3 Syncro Soft, http://www.oxygenxml.com/

4 A typical example can be found in the edition of William Godwin's Diary (Myers, O'Shaughnessy, and Philp 2010), in which plots of diary appearances of mentioned people are provided within HTML websites that make use of TEI data without being part of a TEI document themselves.

5 Timeline: Web Widget for Visualizing Temporal Data, http://www.simile-widgets.org/timeline/.

6 Deutsche Nationalbibliothek, http://www.dnb.de/EN/gnd.

7 Deutsche Nationalbibliothek, GND integrated authority file, http://d-nb.info/gnd/119444720

8 Scrollable version at http://fontane-nb.dariah.eu/tei-conf/simile/.

9 GeoNames geographical database, http://www.geonames.org/.

10 http://www.openstreetmap.org.

11 DARIAH-DE Konsortium, http://geobrowser.de.dariah.eu/embed/?.

12 http://exist-db.org/.

13 UC Berkeley Visualization Lab, http://prefuse.org/.

14 https://github.com/mbostock/d3/wiki/Gallery.

15 http://bl.ocks.org/mbostock/4062045.

16 http://bl.ocks.org/mbostock/7607999.

17 Interactive version at http://fontane-nb.dariah.eu/tei-conf/net/.

18 Interactive version at http://fontane-nb.dariah.eu/tei-conf/heb/.

ABSTRACT

Within the last few decades, TEI has become a major instrument for philologists in the digital age, particularly since a set of mechanisms has recently been incorporated which facilitates the encoding of genetic editions. Editors use the XML syntax while aiming to preserve the quantity and quality of old books and manuscripts and publish many more of them online, mostly under free licenses. Scholars all over the world are now able to use huge datasets for further research. There are now many digital editions available, but only a few tools to analyze them. This article explores how web technologies (XML and related technologies as well as JavaScript) can be used to enrich the forthcoming edition of Theodor Fontane's notebooks with data-driven visualizations of named entities and how at the same time applications can be built on these visualizations which are reusable for other edition projects in the TEI world. Because of the density and historical scope of references to named entities and the variety of entity types, Fontane's notebooks lend themselves to advanced methods of semantic analysis.

INDEX

Keywords: applications, authority files, visualization, scholarly edition, Fontane, Theodor, notebooks

AUTHORS

MARTIN DE LA IGLESIA

Martin de la Iglesia works for the project *Genetic-critical and annotated hybrid-edition of Theodor Fontane's notebooks based on a virtual research environment* as a librarian in the Metadata and Data Conversion group at Göttingen State and University Library. He earned a Magister Artium degree in art history and library science from Humboldt University Berlin in 2007 and has been working as a metadata specialist since then.

MATHIAS GÖBEL

Mathias Göbel is a research associate in the project *Genetic-critical and annotated hybrid-edition of Theodor Fontane's notebooks based on a virtual research environment* since 2012. As an IT specialist he is developing the forthcoming website based on a native XML database and the data created within the TextGrid environment. In 2012 he received a Diploma degree in economics, with a minor in German Language and Literature, from Göttingen University.