Graphs vs trees: encoding stemmata in TEI

Simon Gabay, Jean-Baptiste Camps, Gustavo Fernández Riva

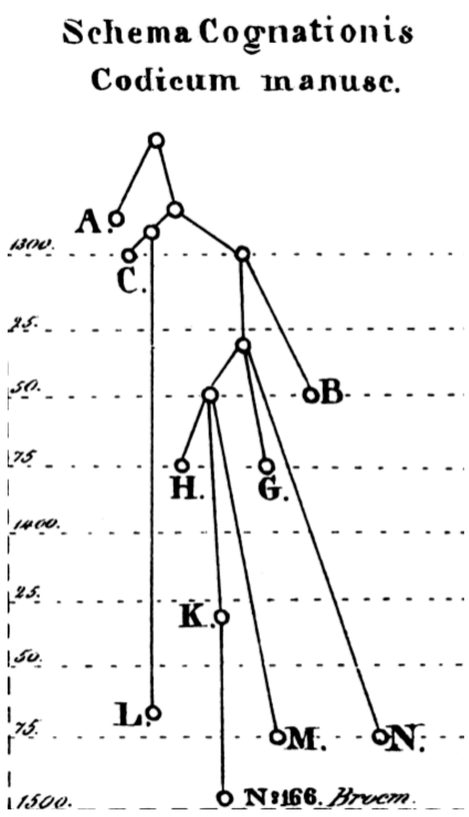
Graphs vs trees: encoding stemmata in TEI

Simon Gabay[1], Jean-Baptiste Camps[2], Gustavo Fernández Riva[3]

[1]Université de Genève (Switzerland). [2]École nationale des chartes | PSL (France)
[3]Universität Heidelberg (Germany).

Introduction

To understand textual traditions containing more than one witness, it is necessary to establish the relationships between the surviving manuscripts and to express them efficiently. To this end, philologists have developed a genealogical method known as stemmatology. For almost two centuries, they have used tree-shaped diagrams, called *stemma codicum* (plur. *stemmata*).



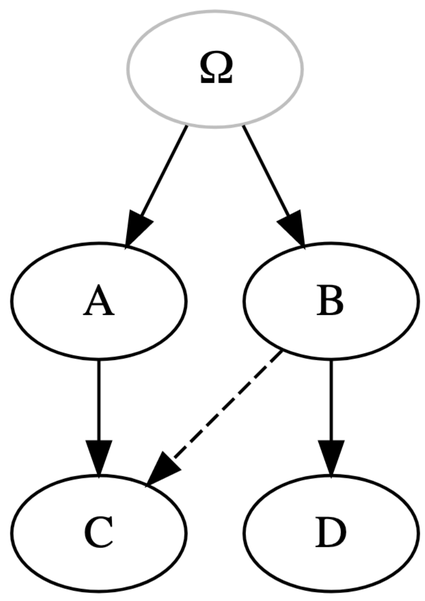
First known *stemma*, in Schlyter 1827 [1]

A challenge

The structure of a *stemma* is sometimes more complex than a tree, because of phenomena like lateral transmission or multiple ancestry (contamination). That two manuscripts from two different families are related would mean, following the arboreal metaphor, that two specific leaves of two different tree branches are connected to one another – a problem that forces us to rethink their nature.

Tree vs graph

As the shape containing nodes with more than one parent does not fit the graph theory definition of a "tree", it would therefore certainly be more appropriate to talk about "directed acyclic graphs" [2]. Such a decision is not only terminological, but could have an impact on some suggestions of the TEI guidelines (chap. 19.4), since these propose a tree-like encoding using the `<eTree>` element, which is unable to express efficiently abnormal configurations such as contaminations. For this reason, it is worth considering alternatives.



Example of a contamination B→C.

One possibility would be to use a `<ptr>` element with a specific type attribute inside `<eTree>` (see example below) to record all contaminations. However, `<ptr>` is a very general element and the specific use in this context would not be clear.
An alternative is to use the `<graph>` element, in combination with `<node>` and `<arc>`. The `<graph>` element is able represent contaminated traditions and it is in line with standard graph encoding formats such as DOT [3] or GraphML [4], that deal separately with the description of nodes and edges. The only custom addition required for a precise encoding would be a `@type` attribute for the `<arc>` element, created in an additional namespace.

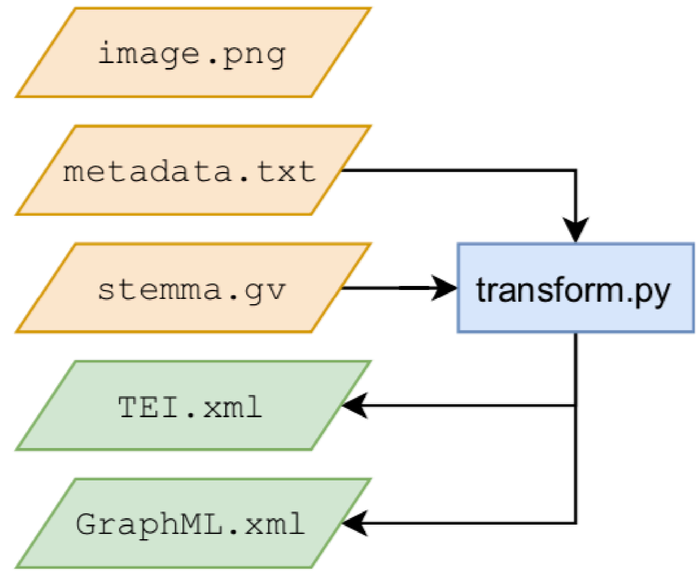
Open stemmata

With the emergence of computational philology, the use of *stemmata* is slowly drifting from "simple" ecdotical purposes to broader questions regarding textual variation [5] or the modelling of textual transmission [6]. The case of our newly created digital collection of textual genealogies, *Open Stemmata* [7], shows the importance to encode *stemmata* as graphs rather than trees for technical, but also philological reasons. More information can be found online at the following addresses:

- `https://openstemmata.github.io/`;
- `https://github.com/OpenStemmata/`.

Workflow

Submissions in *OpenStemmata* contain three files: a reproduction of the published *stemma* (`image.png`), a text file with metadata generated from an online form (`metadata.txt`) and the *stemma* encoded in the DOT format (`stemma.gv`). The DOT format has been chosen because it is easy to use and it allows to express all the required information. The metadata and graph files are transformed automatically, when pushed on our GitHub repo, into two standard formats, TEI and GraphML, to facilitate exchange, preservation and analysis.



Three encodings

- Graph in `DOT` format (left)
- Graph in `TEI` as a tree (center).
- Graph in `TEI` as a graph (right).

```
digraph {
    omega[label="Ω", color="grey"];
    omega ->A;
    omega -> 1;
    1 -> B;
    1 -> C;
    1[label="", color="grey"];
    A -> B [style="dashed", dir=none];
}
```

```
<eTree type="hypothetical">
   <label>Ω</label>
   <eTree type="extant">
      <label>A</label>
      <eLeaf type="extant"
             xml:id="C">
        <label>C</label>
      </eLeaf>
   </eTree>
   <eTree type="extant">
      <label>B</label>
      <eLeaf type="extant">
         <label>D</label>
      </eLeaf>
      <ptr type="contamination"
         target="#C"/>
   </eTree>
</eTree>
```

```
<graph type="directed">
   <node xml:id="omega" type="hypothetical" inDegree="0" outDegree="2">
      <label>Ω</label>
   </node>
   <node xml:id="A" type="witness" inDegree="1" outDegree="1">
      <label>A</label>
   </node>
   <node xml:id="B" type="witness" inDegree="1" outDegree="2">
      <label>B</label>
   </node>
   <node xml:id="C" type="witness" inDegree="2" outDegree="0">
      <label>C</label>
   </node>
   <node xml:id="D" type="witness" inDegree="1" outDegree="0">
      <label>D</label>
   </node>
   <arc cert="unknown" from="#omega" to="#A" od:type="filiation"/>
   <arc cert="unknown" from="#omega" to="#B" od:type="filiation"/>
   <arc cert="unknown" from="#A" to="#C" od:type="filiation"/>
   <arc cert="unknown" from="#B" to="#D" od:type="filiation"/>
   <arc cert="unknown" from="#B" to="#D" od:type="contamination"/>
</graph>
```

References

[1] Carl Johan Schlyter and Hans Samuel Collin, editors. *Corpus juris Sueo-Gotorum antiqui*. Z. Haeggström, 1827.

[2] Armin Hoenen. The stemma as a computational model. In *Handbook of Stemmatology. History, Methodology, Digital Approaches*, De Gruyter Reference, pages 226–241. De Gruyter, 2020. URL `https://doi.org/10.1515/9783110684384`.

[3] Emden R Gansner. The dot language, 2002. URL `www.research.att.com/~erg/graphviz/info/lang.html/`.

[4] GraphML Team. The graphml file format, 2002. URL `http://graphml.graphdrawing.org/`.

[5] Tara L. Andrews and Caroline Macé. 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):504–521, 2013. ISSN 0268-1145. doi: 10.1093/llc/fqt032. URL `https://doi.org/10.1093/llc/fqt032`.

[6] Jean-Baptiste Camps and Julien Randon-Furling. A Dynamic Model of Manuscript Transmission. In *Workshop on Computational Methods in the Humanities (COMHUM 2018)*, Lausanne, 2018.

[7] Jean-Baptiste Camps, Simon Gabay, and Gustavo Fernández Riva. Open Stemmata: A Digital Collection of Textual Genealogies. In *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities*, Krasnoyarsk, 2021. URL `https://halshs.archives-ouvertes.fr/halshs-03260086`.