A Philological Perspective on Meta-Scientific Knowledge Graphs

Tobias Weber
Ludwig-Maximilians-Universität München

weber.tobias@campus.lmu.de

Dear colleagues,

Thank you for your interest in my presentation – and thank you to the organisers of this workshop for giving me the chance to present my topic in this venue. As I want to keep my talk short, I will focus on the two central aspects of my topic and refer you to the paper for more information. You can also contact me outside of this presentation – my email address is included here.

Philology

- textual curatorship and interpretation (Gurd 2015)
- the multifaceted study of texts, languages, and the phenomenon of language itself (Turner 2014)

My talk will present a philological view of Scientific Knowledge Graphs. As you are researchers and practitioners in this field, I do not need to tell you what Knowledge Graphs are or how they function – but, put simply, they are relationships between *uniquely identifiable entities* or *unique abstract concepts*. In science, these may be our objects of study, authors, papers, institutions, or various groupings thereof. Creating relationships internally, on our research objects, or externally, on the academic structures, is generating knowledge as a part of the scientific endeavour. As philology is, by definition, a text-based science, I would like to argue for the link of knowledge generation and texts in science.

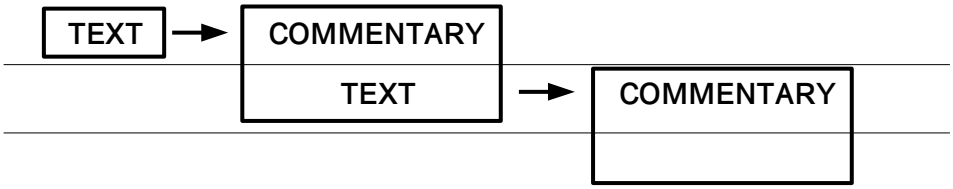
Philology

- textual curatorship and interpretation (Gurd 2015)
- the multifaceted study of texts, languages, and the phenomenon of language itself (Turner 2014)

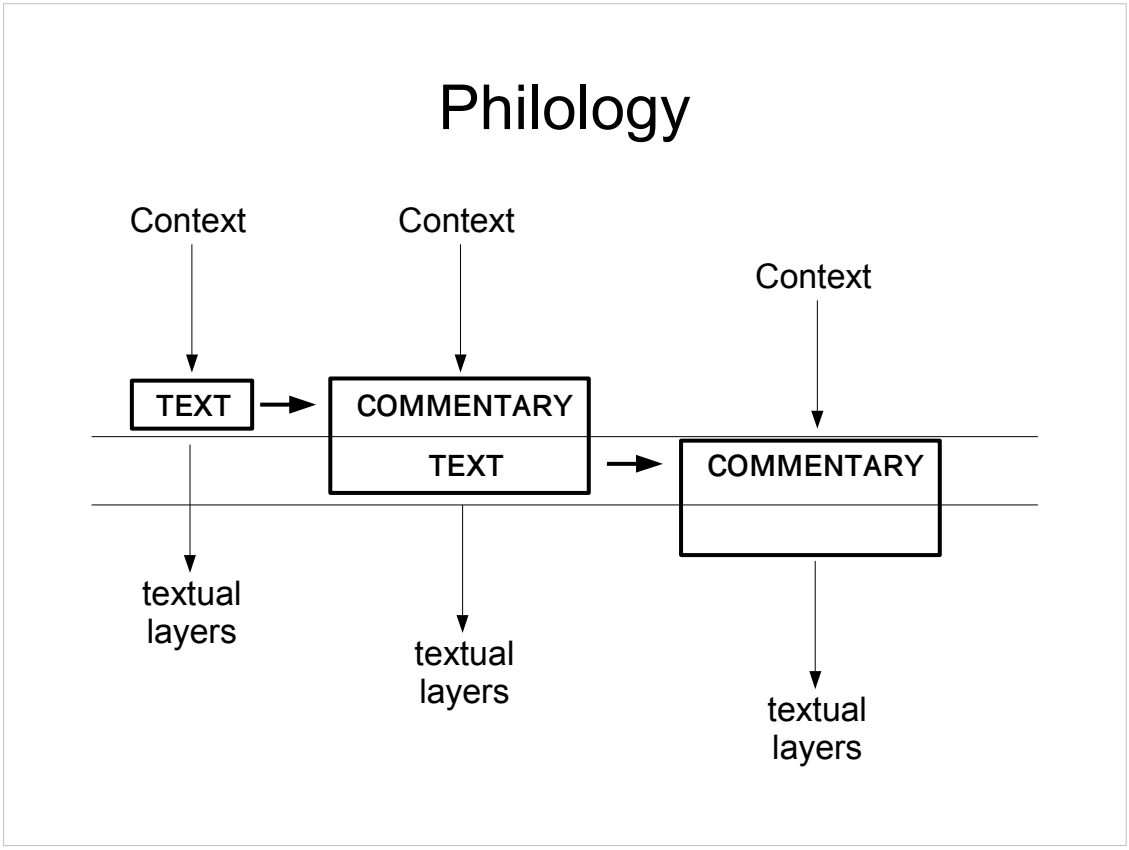
- Original TEXT
- Interpretative COMMENTARY

A central concept in philology is the distinction *text* and *commentary*; an *original source* and its *interpretation*.

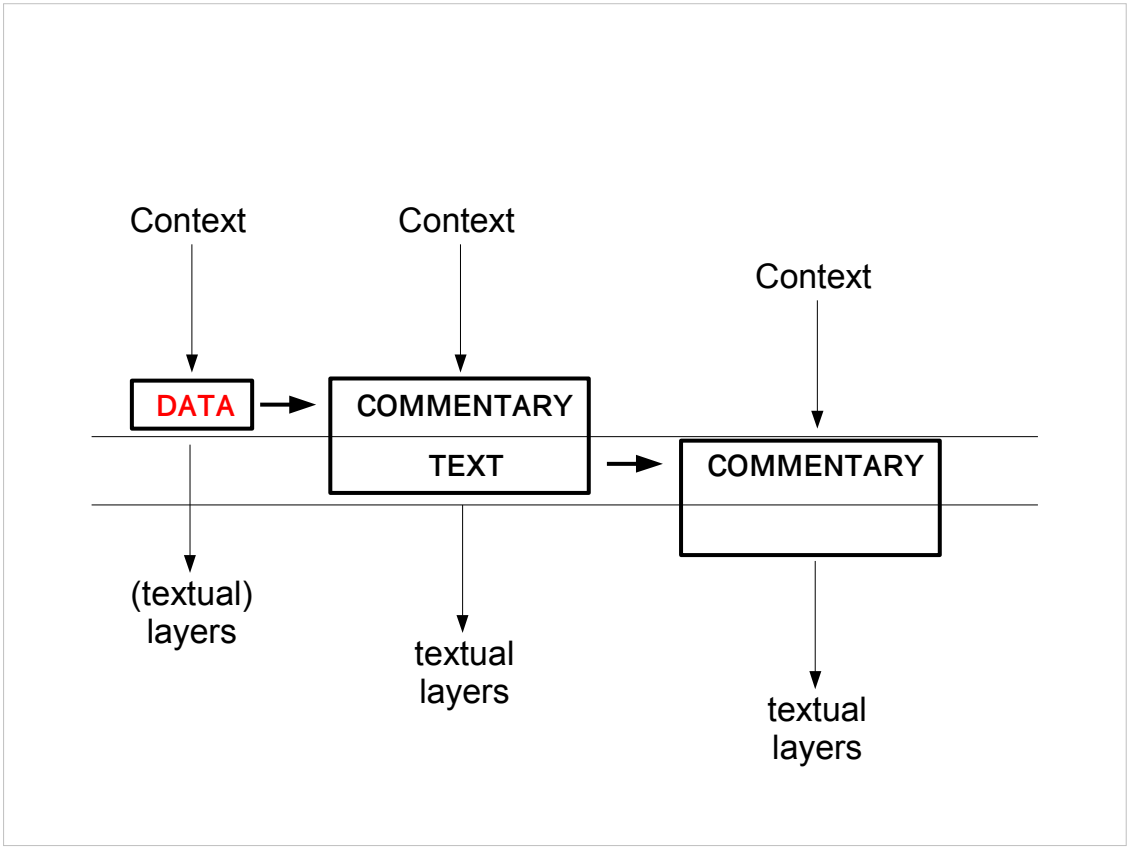
Philology



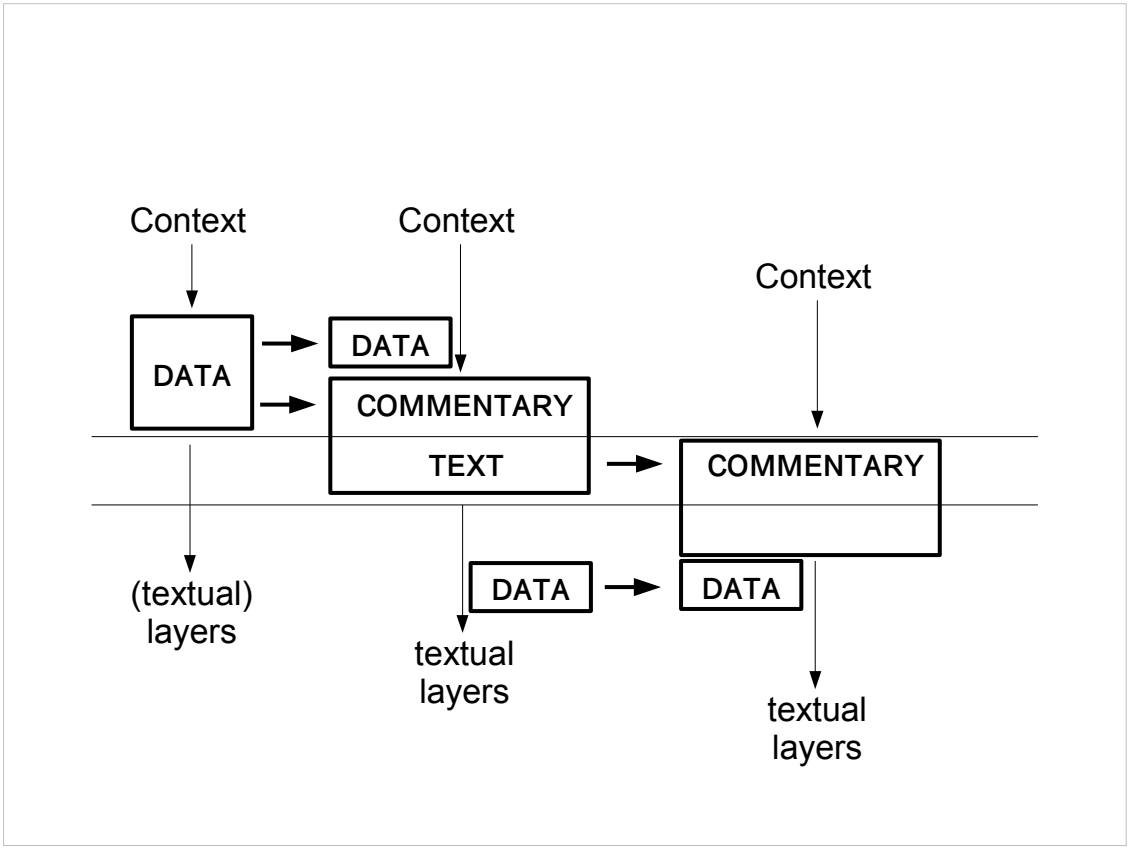
Commentaries can be texts in their own right, creating a chain of association, or a graph.



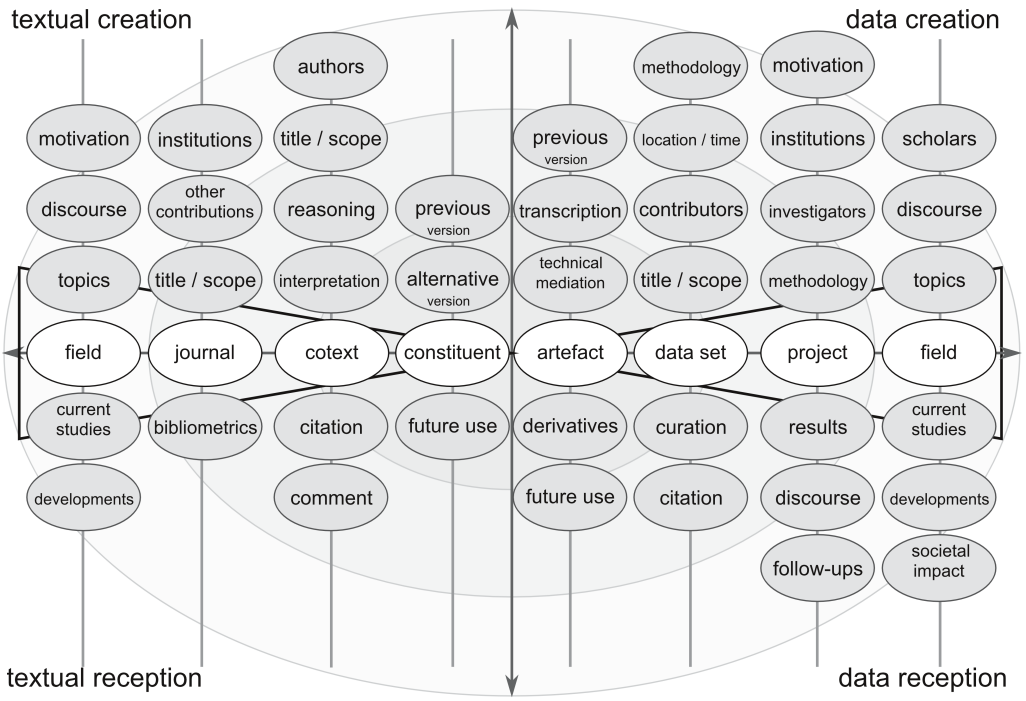
In order to understand interpretations and add to the knowledge contained within this line of thought, we check for textual clues, that is everything contained in the co-text, or non-linguistic external factors of the con-text. Thereby, we can always determine the position of a constituent within its cotext and context, or its position in the graph.



In data-driven sciences, data sets can be seen as an original source, as they are the basis for the interpretative commentary. As such, they should be treated equal to texts. Consequently, they have individual contexts and sub-layers, for example singular data points – all of them with corresponding metadata.

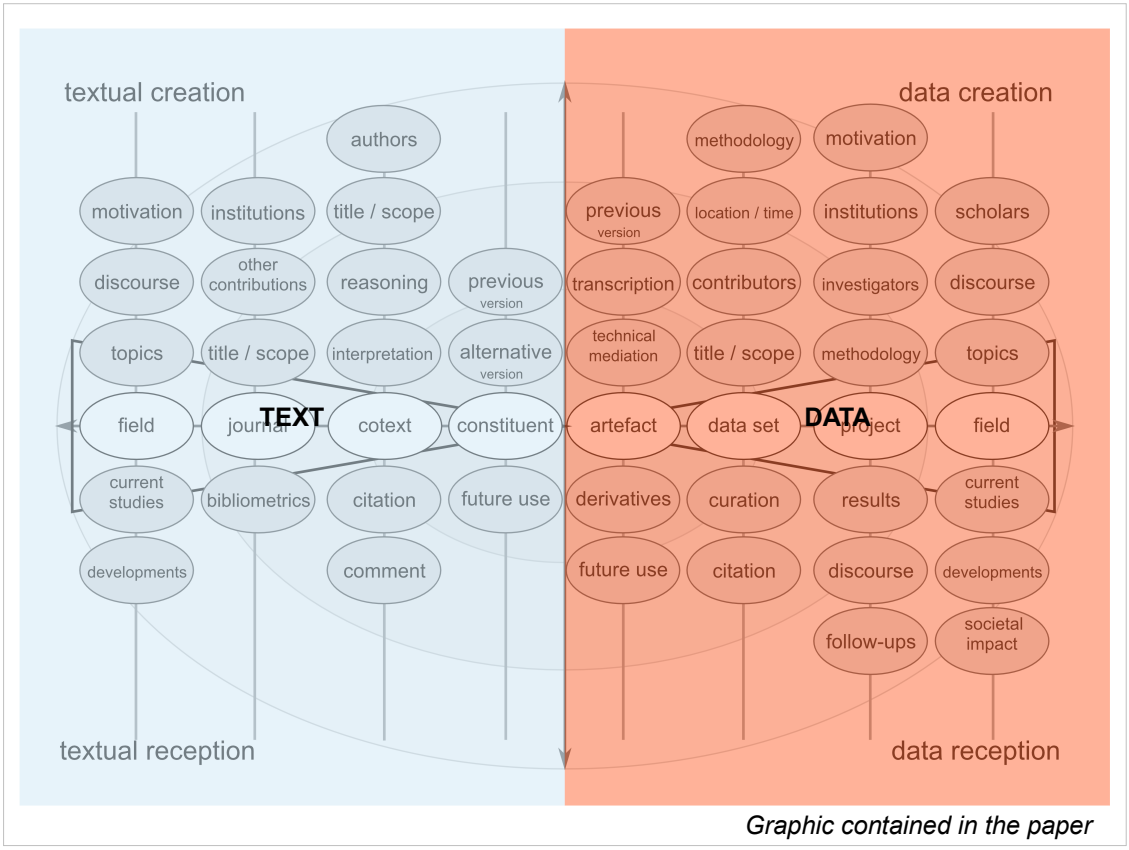


In the knowledge generating process, we excerpt, analyse, annotate, or cite data from a source. In doing so, we create a corresponding dataset as a basis for the commentary, which becomes a new reference; that is an original source or text, for future commentaries. As contexts and cotexts differ, and we possibly altered the data for knowledge generation, these datasets are not identical but new nodes on a graph. Ideally, we know from which parent node a new dataset stems, in which context it was generated and used, and which other graphs originate from it.



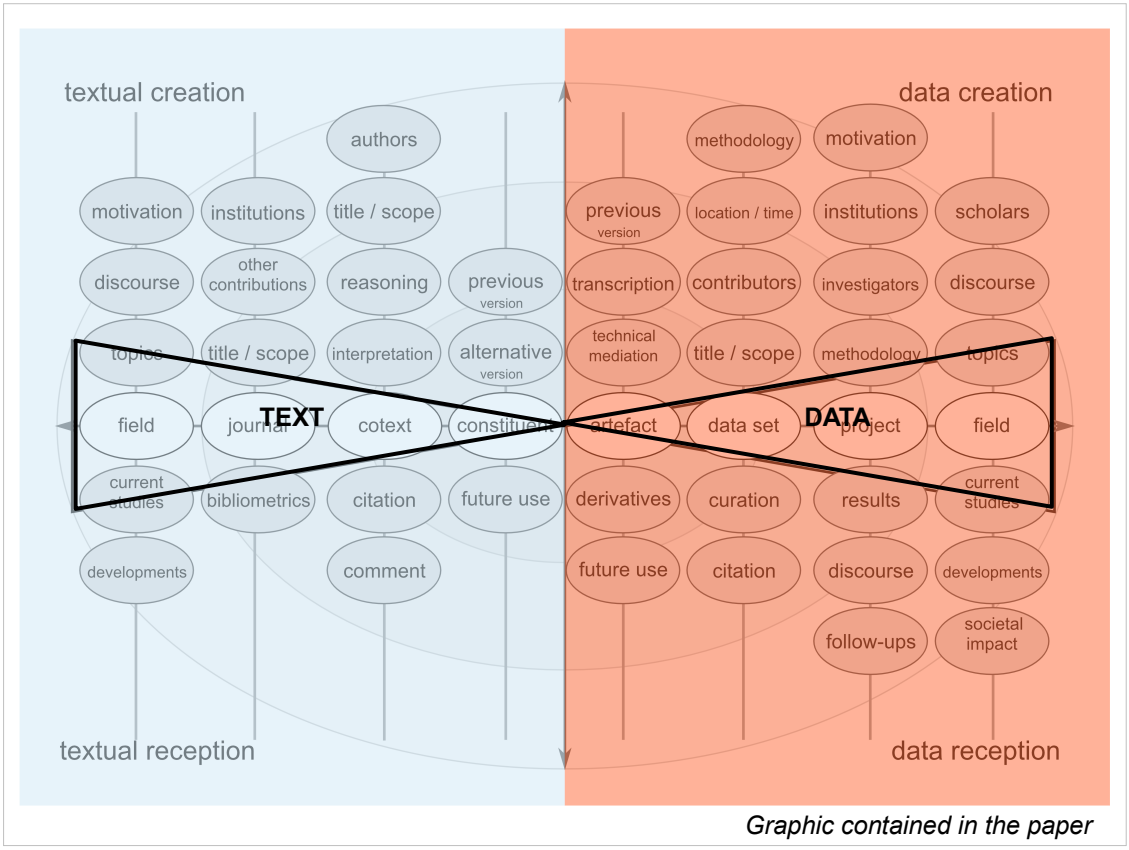
Graphic contained in the paper

To illustrate this further:



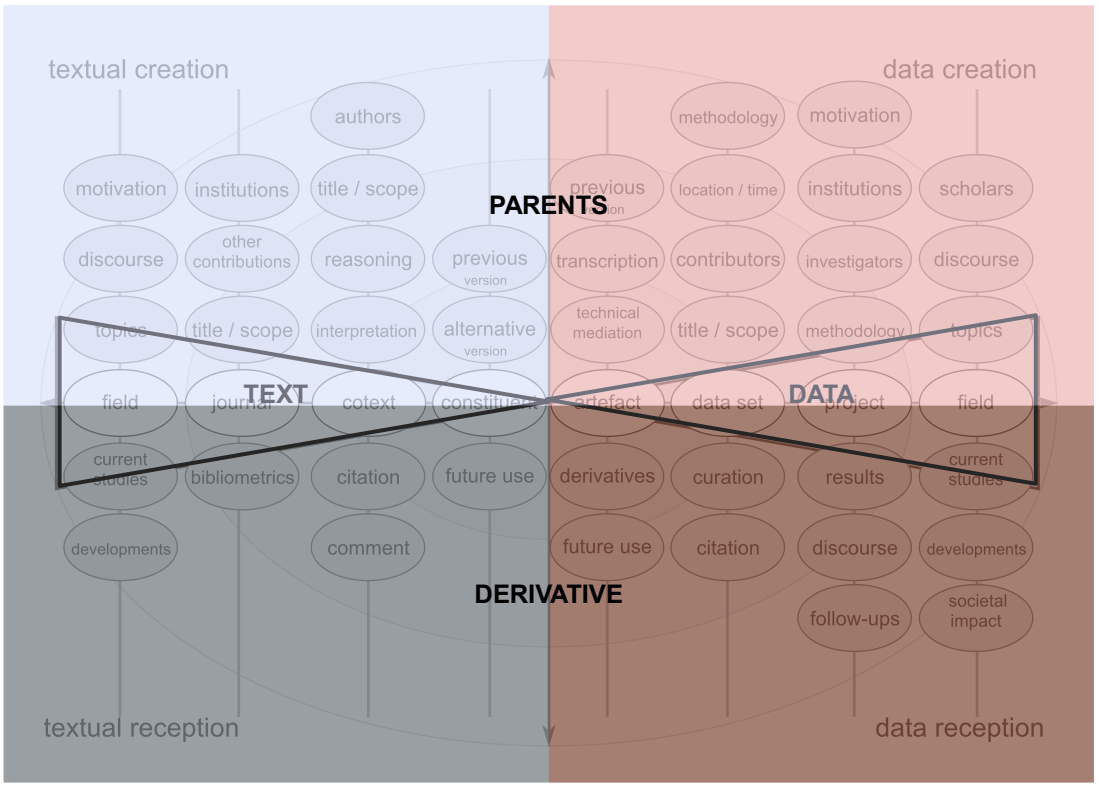
Graphic contained in the paper

We are dealing with two separate planes: data and text, including commentaries.



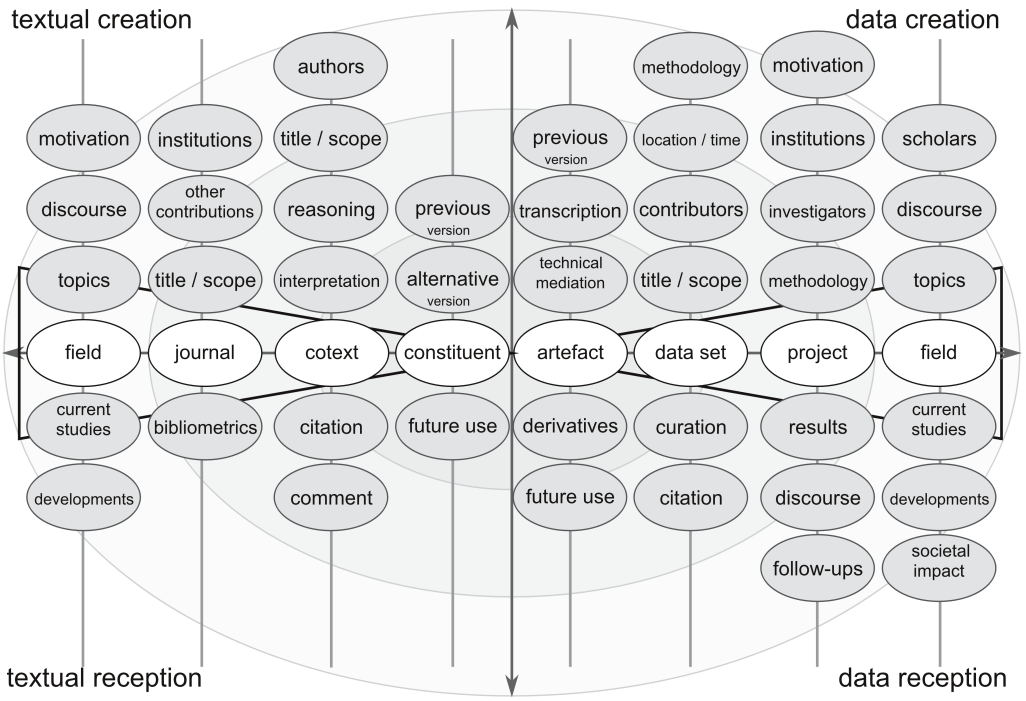
Graphic contained in the paper

Each has individual layers of abstractness.



Graphic contained in the paper

And for each instance we can consider parent nodes preceding it and children nodes as derivatives.



Graphic contained in the paper

Furthermore, each instance comes with different cotextual and contextual clues contained in the metadata. And with texts and data being new, unique instances, their metadata should be linked as well, across all planes.

Admittedly, this is an enormous task but the only way to uncover all stories behind knowledge generation in science, beyond citation tracking.

Solutions

- Unique identifiers
  - Version control
  - Lower level identifiers; below DOI, URI, ORCID
- Understanding links behind data and texts
  - Transparency and reliability
  - 'Alternative' solution (*http://bit.do/**WeberLDK***)

Ultimately, a practical example from the paper:

> One suggestion would be the introduction of version control for linguistic examples which link it back to original data sets in archives, intermediary sources and contexts, and changes in the data. This would allow access to the full history of data, texts and commentaries – as in the philological approach.
>
> Likewise, all data points could be given unique identifiers, not only data sets and articles. In a way, they are points on a vector of text, or nodes on a graph. Study of links between those instances shows us a holistic image as a potential means to achieve transparency and accountability.

Solutions

- Unique identifiers
  - Version control
  - Lower level identifiers; below DOI, URI, ORCID

- Understanding links behind data and texts
  - Transparency and reliability
  - 'Alternative' solution (*http://bit.do/**WeberLDK***)

Thank You

weber.tobias@campus.lmu.de

I hope you consider adopting a philological view on science yourself and am happy to answer any question.

Thank you.