A Philological Perspective on Meta-Scientific Knowledge Graphs

Tobias Weber[0000−0003−1530−3568]

Ludwig-Maximilians-Universität München, Germany
weber.tobias@campus.lmu.de

Abstract. This paper discusses knowledge graphs and networks on the scientific process from a philological viewpoint. Relevant themes are: the smallest entities of scientific discourse; the treatment of documents or artefacts as texts and commentaries in discourse; and links between context, (co)text, and data points. As an illustration of the micro-level approach, version control of linguistic examples is discussed as a possible field of application in this discipline. This underlines the claim for data points to be treated like unique entities, which possess metadata of the datum and any text generating knowledge from it.

Keywords: Version control · Meta-documentary linguistics · Identifier · Knowledge graphs · Scientific discourse.

1 Why Philology?

At first glance, it might appear strange that a philologist would be involved in meta-scientific discussions among IT specialists and arguing for the philological perspective in the discourse. These topics would be expected to attract sociologists, philosophers, and data scientists - while the stereotypical philologist's place is in a library or archive, covered in the dust of old books and manuscripts. In this paper, principles of philology are presented with a view to ways in which they may be applied to support the creation of sound and thorough databases, ontologies, and knowledge graphs.

Philology comprises a range of definitions, usually including 'textual curatorship and interpretation' [12] or 'the multifaceted study of texts, languages, and the phenomenon of language itself' [29, p.IX]. Independent of its linguistic or literary orientation, it has a strong history of working with graphs. These graphs and their modern, digital representations help us to understand relationships between languages (like a phylogenetic tree), manuscripts, or versions of writings [8]. They form genealogies to help us understand aspects of our research objects in relation to other artefacts and abstracta, and investigate the history behind these artefacts. As such, the philological approach is characterised by 'an attitude of respect for the datum' [32, p. 18]. It provides the basis for investigating intertextuality and the human factor in the creation and reception of scientific discourse. As such, it can be seen as the 'fundamental science of human memory' [17, p. 345], with (textual) artefacts at the core of our endeavour.

If we consider science as a discourse, the exchange of texts and commentaries, facilitated by creating, discussing, and reviewing textual artefacts and constantly developing through a growing body of literature, the need for a philological stance becomes clear. As scholars, we are accessing this 'memory' of our disciplines by 'recalling, iterating, reading, commenting, criticizing, and discussing what was deposited in the remote or recent past' [1, p. 97]. This holds true for all disciplines, especially those in which discourse forms the basis of knowledge generation. These disciplines cannot rely on a 'transcription device' [14] which turns raw data (e.g. a sound recording) into primary data (e.g. a transcription) [15] reliably and independently of the researcher who abstracts data to create a text. Thus, wherever humans are involved in the selection, analysis, and interpretation of the artefacts, we can make a case for philological enquiry [25]. The documents we produce as evidence of our membership in the scientific community [6] are the starting point for charting science [3].

Philologists are increasingly aware and capable of applying computational methods in their research. Those may be applications of computational linguistics, automatic annotations, or the encoding of our artefacts in XML formats [4] such as the TEI standards [27, 22, 9]. These formats do not only enable digital processing but allow for the inclusion of metadata which enrich information as thick metadata [19] for presenting and storing data [28], ideally to keep it usable '500 years from now' [33]. The philological stance is, therefore, independent from any practical application or method and functions as a guiding principle throughout the methodology.

2 Layers of Knowledge

As mentioned earlier, philologists use knowledge graphs to represent relationships between artefacts based on textual and contextual clues. These are contained inside every text, or to an increasing degree in the metadata. It should be noted that commentaries about texts, which are peripheral to a particular artefact inasmuch as they are not contained by default, can form new texts in their own right with a set of new metadata. The transfer from data to text must not lose metadata on the underlying data sets while coming with a set of metadata on the text itself. Likewise, there are no identical copies, as every artefact bears traces of the technology used to (re)create it [26], they originate in different contexts. For example, a digital copy of a document comes with a set of individual metadata different from those of a printout, with its aesthetics and presentational formats as a potential point in a meta-scientific discussion. Similarly, excerpts of text form new entities in an information science interpretation of text [23].

For this paper, I construct a simple typology of scientific artefacts and their relationship to the processes of scholarly writing and publishing. This typology illustrates the different aspects of science which can be included, analysed, and represented using knowledge graphs (e.g. citation tracking, ontologies like DBpedia). Aside from time as a linear component (e.g. in genealogies or stemmata),

there are two basic distinctions. Firstly, the difference between the inherent and external attributes of the artefact, or concrete and abstract information about it. Secondly, we can refine the scope from the surroundings or container (macro-level) of the artefact or entity (meso-level) to its smallest components (micro-level). In other words, mirroring layers of philological and linguistic research on a text to the meta-level of the commentary. This mapping can either be structured in directed graphs by linear time (as versions) or symmetrically, requiring every 'commentary' to the 'primary text' to be linked back to this source in its own textual attributes.

Table 1. Typology of representable aspects of science.

	Concrete	**Abstract**
macro-level	context	meta-context
meso-level	text or cotext	meta-text or commentary
micro-level	constituent or 'component'	metadata

On the macro-level, we are dealing with the contexts in which a scientific artefact is embedded. The term context is to be understood primarily as the global extra-linguistic context (i.e. existing outside of the text), while the textual context [20] is particular to each text and relevant for the interpretation of semantics within it (i.e. its relation to the lower level, e.g. decoding of references). Examples of the global context are the inclusion in a particular journal or collection, as well as metadata about the artefact (format, size). The textual context might be exemplified by an article referring to 'the data' as a particular, identifiable set of data used for the study. The meta-level can be conceptualised as the position of the artefact among other artefacts, including its references to previous literature, the tracking of citations after publication, and possibly even a comparison of similarity with other documents [11]. The latter two are recorded by libraries and publishers and stored in databases [16].

The meso-level is the level of the actual text, or for its relation to the micro-level the co-text, as the linguistic environment of a text constituent [7]. This layer is charted by numerous scholars aiming to train machine learning algorithms for NLP and creating ontologies and databases of knowledge; it might be represented as n-grams of variable size around a particular word or concept. On its meta-level, we encounter 'commentaries', i.e. external texts referencing, discussing, or reviewing the original text. As a central scientific procedure, this has also been represented in digital formats. We might consider possible applications linking knowledge generated in one text to information contained in other scientific texts, as a representation of knowledge generation [21], or with a view to linking scholars, institutions, and datasets [13].

On the micro-level, the focus of our interest is the smallest entities or text constituents of science and the scientific text. These may be concepts or terms relevant for named entity recognition [18] in the creation of ontological databases [2], or data sets or examples which are discussed in the text. On the meta-level,

these smallest entities can create the biggest issues, as they come with large amounts of metadata. To provide an example from linguistics: Imagine an author citing a phrase from a longer narrative which is contained in a corpus, stored in an archive. The metadata should include the relation of the cited phrase to the superset or the narrative and the corpus, as a version with information on its creation. The language of the example, data about the consultant, and information on the documentation project (at minimum place and time of recording) are needed to identify and keep track of the source. Ideally, 'thick metadata' [19] cover all necessary information, which, in turn, may consist of full graphs itself (e.g. a genealogy of speakers). Furthermore, the micro-level relates to texts and commentaries across a range of contexts, and these may influence their representation, annotation, or interpretation. In other words, two instances may cite the same source but apply different means of analysis and reach different conclusions - as a result, we can find the same example arguing for *and* against a theory. Since this aspect is underrepresented in the literature, I will be briefly discuss it in the following section.

3 Keeping Track of Data

The ideal knowledge graph to chart scientific knowledge generation thoroughly links all concrete entities to the rest of scientific artefacts, while containing full accounts of (meta)data on all lower levels. That being said, all lower-level constituents shall be identifiable and contain information on their use throughout all (con)texts, at the same time linking the published version to its parent nodes (e.g. collections, full data sets), and allowing for enquiry through the metadata. Consequently, we need to treat data points like abstract entities collected in ontologies (i.e. unique named entities). Using metadata on the constituents and tracing the procedure of knowledge generation based on them, allows us to tell the story of science through its artefacts. This creates the opportunity to write the 'meta-documentation' [5] as a narrative of data use in science - the discourse we want to preserve in our (cultural) 'memory'.

Yet, this preservation and curation of our scientific legacy is not exclusively useful for posterity. Gaining access to actual data use (in citation, analysis, commentary) and the underlying links in the textual matter of the discourse, can shed light on biases [24], trends, and epistemological foundations of science. And, while this endeavour is currently based on linking data from the macro- and meso-level (most visible in citation tracking [30]), we must not ignore the potential of the micro-level. At this point, where particular data points turn into knowledge, data becomes text (see Fig. 1). Yet, both sides are linked by attributes of text or data on any level of description (i.e. abstractness). Each aspect of metadata represented in Fig. 1 has its own attributes, e.g. individual have names, dates of birth, parents, employers, speak languages (all possible attributes in TEI, OWL, or schema.org), which may be included in other parts of the attributes. For example, different data points by mother and daughter could be linked through their genealogical affiliation; or researchers attributed to their

respective institutions with their projects, publications etc. The attributes in the periphery of the non-exhaustive diagram can be at the centre of other diagrams, e.g. different research articles may be at the centre, yet linked through their metadata. Importantly, text and underlying data are closely linked but their respective sets of metadata do not interfere. Even on the extreme abstract of 'field' (i.e. macro-level context), there can be different values for the attributes, e.g. this paper using philology on the data plane to generate knowledge on the textual plane of knowledge graphs.

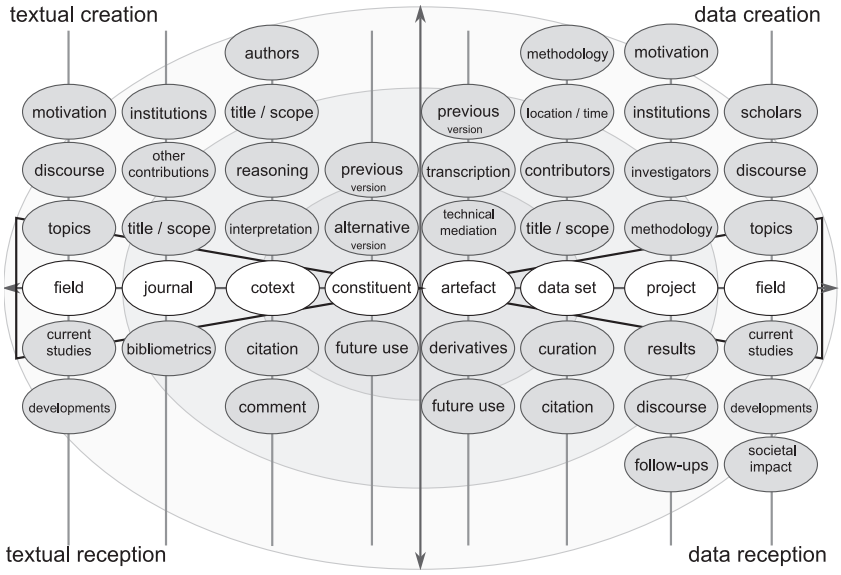

Fig. 1. Exemplary network of scientific text and data artefacts. The textual plane is on the left, the data on the right; the top half refers to attributes preceding the entity, e.g. relevant in its creation, while the bottom half lists aspects in the reception of the artefact, i.e. after its creation. The centre point (origin) is where scientific data is transferred into the realm of scientific knowledge with various degrees of abstractness around this point. The list of attributes for each instance is not exhaustive; other attributes or attributes of the presented attributes may be relevant.

In language documentation, fieldworkers rarely find uncharted land, i.e. to any project there is a precursor, historical interactions between the researcher and the communities, prior research on a language, or existing artefacts in an archive. As a result, it is common to find recurring themes, interacting with the same families as other researchers, and interacting with their research, even by critically assessing their work and righting historical wrongs (e.g. colonialism,

missionary efforts). But for a depiction of the progress in science, we do not only require accounts of controversy and debate - we also need to understand links in data which are not immediately obvious. These are, for example, cases in which a family has supported several different projects through multiple generations, and the knowledge contained in the various collections is linked by the genealogy of the consultants; or where a cited example has later been changed, rejected, or edited for a variety of reasons (e.g. consultant's request, new transcriptions, publishing policies) which are not inherent to the datum but to its contexts of creation and reception. While readers may consider this to be a specific issue of linguistics, there are instances in other fields where such a micro-level approach to the scientific process may yield yet unknown potential. For example - if this was ethically acceptable - cross-linking of participants' data who have been on different medical experiments, provided data to separate socioeconomic enquiries, or answered questionnaires for multiple surveys. Only in certain, specific scenarios is it possible to cross-link such data, e.g. historical medical data [10]. The potential of creating a holistic image does not lie within the macroscopic view of science but within every single data point.

As far as linguistics are concerned, a strong case can be made for the introduction of standardised tracking codes for linguistic examples which grant access to the multiple links behind each utterance. This version control must link an excerpted version to the full version with all relevant metadata on the data plane, while containing information on the role of the example across all layers, in citation, analysis, reception, and recreation on the textual plane [31]. This would expand the networks of researchers' contributions on a paper, as we are reaching conclusions building on the analyses, transcriptions, and interpretations of other scholars. Furthermore, already the selection of an example over others can shape the conclusions, and, even if we copy examples without altering them, we do accept them as valid contributions in the discourse and interact with them. While the precise form and technical implementation of these codes can be debated, their tracking does not differ much from the information already gathered on articles, e.g. providing information on further citations, references cited, bibliographic information, figures and tables. There is no reason why cited examples and (parts of) data sets should not be identified and tracked accordingly, with the philological 'attitude of respect' for the micro-level constituents of knowledge generation. They are identifiable entities in their own right.

4 Conclusion

Philology offers a structure to analyse texts and commentaries on their artefactual level, as well as their contexts and the components constituting it. All of these layers can be presented using knowledge graphs, in themselves and through their relations to other artefacts or other levels of description. Constructing the network of science, as a document- or artefact-based discourse, requires the linking of all layers to uncover the narratives behind science. Independent of the technical details and requirements for this endeavour, the focus on the smallest

entities of scientific research, data points and examples, can help to paint a holistic image. Thus, an extended focus on the micro-level, apart from Named Entity Recognition and linking of concepts as in DBpedia, is required to understand how scientific knowledge is constructed from data points. And with a focus on these individual entities, all researchers should adopt the philological stance.

References

1. Assmann, A.: Canon and archive. In: Erll, A., Nünning, A. (eds.) Cultural memory studies: An international and interdisciplinary handbook, pp. 97–107. Mouton de Gruyter, Berlin (2008)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) The Semantic Web. pp. 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
3. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a Knowledge Graph for Science. In: WIMS '18 (2018)
4. Austin, P.K.: Data and language documentation. In: Gippert, J., Himmelmann, N.P., Mosel, U. (eds.) Essentials of Language Documentation, pp. 87–112. Mouton de Gruyter, Berlin, New York (2006)
5. Austin, P.K.: Language documentation and meta-documentation. In: Jones, M., Ogilvie, S. (eds.) Keeping Languages Alive. Documentation, Pedagogy, and Revitalisation, pp. 3–15. Cambridge University Press (2013)
6. Bond, G.C.: Fieldnotes: Research in Past Occurrences. In: Sanjek, R. (ed.) Fieldnotes. The Makings of Anthropology, pp. 273–289. Cornell University Press, Ithaca and London (1990)
7. Catford, J.C.: A linguistic theory of translation: an essay in applied linguistics. Oxford University Press, Oxford (1965)
8. Crane, G., Bamman, D., Jones, A.: ePhilology: When the Books Talk to Their Readers. In: Schreibman, S., Siemens, R. (eds.) A Companion to Digital Literary Studies. Blackwell, Oxford (2008)
9. Cummings, J.: The Text Encoding Initiative and the Study of Literature. In: Schreibman, S., Siemens, R. (eds.) A Companion to Digital Literary Studies. Blackwell, Oxford (2008)
10. Dong, L., Ilieva, P., Medeiros, A.: Data dreams: planning for the future of historical medical documents. Journal of the Medical Library Association **106**(4), 547—551 (2018). https://doi.org/10.5195/jmla.2018.444
11. Gipp, B.: Citation-based Plagiarism Detection. Springer Vieweg, Wiesbaden (2014)
12. Gurd, S.: Philology and Greek Literature (03 2015), https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935390.001.0001/oxfordhb-9780199935390-e-65
13. Henk, V., Vahdati, S., Nayyeri, M., Ali, M., Yazdi, H.S., Lehmann, J.: Metaresearch Recommendations using Knowledge Graph Embeddings. In: AAAI 2019 Workshop on Recommender Systems and Natural Language Processing (2019)
14. Latour, B., Woolgar, S.: Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton (1986)
15. Lehmann, C.: Data in linguistics. The Linguistic Review **3/4**(21), 275–310 (2004)

16. Leydesdorff, L., Milojević, S.: Scientometrics. In: Wright, J.D. (ed.) International Encyclopedia of the Social & Behavioral Sciences, pp. 322 – 327. Elsevier, Oxford, second edition edn. (2015), http://www.sciencedirect.com/science/article/pii/B9780080970868850308
17. McGann, J.: Philology in a New Key. Critical Inquiry **39**(2), 327–346 (2013)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticæ Investigationes **30**(1), 3–26 (2007)
19. Nathan, D., Austin, P.K.: Receonceiving metadata: language documentation through thick and thin. Language Documentation and Description **2**, 179–187 (2004)
20. Petőfi, J.S.: Text-grammars, text-theory and the theory of literature. Poetics **2**(3), 36 – 76 (1973)
21. Popping, R.: Knowledge Graphs and Network Text Analysis. Social Science Information **42**(1), 91–106 (2003)
22. Renear, A.H.: Text Encoding. In: Schreibman, S., Siemens, R., Unsworth, J. (eds.) A Companion to Digital Humanities. Blackwell, Oxford (2004)
23. Renear, A.H., Wickett, K.M.: There are No Documents. In: Proceedings of Balisage: The Markup Conference 2010. Balisage Series on Markup Technologies, vol. 5 (2010), https://doi.org/10.4242/BalisageVol5.Renear01
24. Risam, R.: Telling Untold Stories: Digital Textual Recovery Methods. In: levenberg, l., Neilson, T., Rheams, D. (eds.) Research Methods for the Digital Humanities, pp. 309–318. Springer International Publishing, Cham (2018)
25. Seidel, F.: Documentary linguistics: A language philology of the 21st century. Language Documentation and Description **13**, 23–63 (2016)
26. Shils, E.: Tradition. The University of Chicago Press (1981)
27. The TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange (2020), www.tei-c.org
28. Thieberger, N.: Research Methods in Recording Oral Tradition: Choosing Between the Evanescence of the Digital or the Senescence of the Analog. In: levenberg, l., Neilson, T., Rheams, D. (eds.) Research Methods for the Digital Humanities, pp. 233–241. Springer International Publishing, Cham (2018)
29. Turner, J.: Philology: The Forgotten Origins of the Modern Humanities, The William G. Bowen Series, vol. 70. Princeton University Press, Princeton and Oxford (2014)
30. Web of Science, http://wokinfo.com/
31. Weber, T.: Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics. In: Eskevich, M., de Melo, G., Fäth, C., McCrae, J.P., Buitelaar, P., Chiarcos, C., Klimek, B., Dojchinovski, M. (eds.) 2nd Conference on Language, Data and Knowledge (LDK 2019). OpenAccess Series in Informatics (OASIcs), vol. 70, pp. 26:1–26:8. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2019). https://doi.org/10.4230/OASIcs.LDK.2019.26
32. Wenzel, S.: Reflections on (New) Philology. Speculum **65**, 11–18 (1990)
33. Woodbury, A.C.: Defining documentary linguistics. In: Austin, P.K. (ed.) Language Documentation and Description, vol. 1, pp. 35–51. SOAS, London (2003)