Metadata Inheritance: New Research Paper, New Data, New Metadata?

First Author[1][0000−1111−2222−3333]

Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com

Abstract. The paper highlights obstacles in discovering hidden stories and biases within science which are due to obscuring knowledge generating procedures and the inaccessibility of lower data layers. Vertical and horizontal Metadata Inheritance is proposed as a solution.

Keywords: Object identifiers · Anthropology of Science · Scientometrics · Knowledge Graphs · Algorithmic bias.

1 Framing Science, Texts, and Data

In a recent paper, I have argued for a philological view on metascience, especially in using computational methodology. With philology being a text-based science [8], the resulting view of the scientific endeavour is one of texts, documents, and discourses. This interpretation is not new [1], but a philological approach could help to reveal hidden stories within the scientific process.

The assumption is that science is a network of scientific artefacts (a concept borrowed from ethnography) - those being original texts and interpretative commentaries. More importantly, many disciplines use non-textual data sets as the basis for knowledge generation, which still bear all features of the original text in philology. As a consequence, texts and data sets, while formally different, fulfil the same function in providing the basis for interpretation and analysis. In the resulting commentaries, we cite from data sets, analyse information, or present data in a format which allows for knowledge generation, just like we would treat any textual source.

This fact contains two relevant points for the discussion: First, commentaries can become ‘texts’, when they provide the basis for further commentaries, i.e. they are not terminal nodes in the citation graph. Computational meta-science has, since its inception, tried to present these networks of texts and sources [6, 9] for researchers, institutions, and publishers. As an aside: The relationship between data and text is often seen as ancillary, with data preceding and supporting the text. However, at closer examination, the relationship is reciprocal [3] - we need descriptions to know which further data to elicit and must describe our collected data to derive new hypotheses. Under this premise, academic merit must be given to the collectors of data sets just as received by authors of the articles [2].

Secondly, we can see that researchers interact with their data. The data on which an analysis or interpretation is based, is not identical to the underlying, original data set, as it has been selected, excerpted, formatted, or transcribed. While some disciplines use a 'transcription device' [5] meant to automatise the transfer from the data plane to the textual plane, the selection and application of transcription requires decisions to be made. Therefore, I assume the data used for a commentary to be a new instance or version of the data set. As with the relationship between texts and commentaries, we are dealing with versions of underlying data sets which differ, if only by new contextualisation. Under this assumption, the idea of reproducibility needs to be re-examined [10].

As we can see, science is the repeated and reciprocal creation of data sets and texts, which are linked to other instances on the textual or the data plane, as versions, commentaries, or citations. Each new article is linked to previous texts, subsequent commentaries on itself, and an underlying data set as a source, which is in itself linked to other versions or subsets of the original data from which it was generated using a particular 'transcription device' or method. As we are trying to understand these links, we need a good (meta-)documentation, which draws from the metadata. Yet, each new instance brings its own, additional metadata, since identity of artefacts cannot be facilitated [7].

2 Metadata Inheritance

As mentioned, each artefact comes with metadata describing the artefact itself, including information on the internal structure, and relationships to other artefacts, data or texts. These metadata are sometimes explicitly spelt out (e.g. a bibliography), or can be inferred (e.g. position of a word within the text). Importantly, we should aim to construct the relationships not only to preceding instances but include derived artefacts and commentaries, as well. Using computational methods and resource identifiers, this is not a difficult task [10]. The difficulty comes from not keeping full accounts and change logs of all processes in generating knowledge, and the unavailability of identifiers for all layers of our data sets.

On the first point, linguistics has developed the concept of corpus theorisation and mediation [11, 4], as a part of the meta-documentation. In these supplemental texts, the contents of a text corpus (as a data set) are described, as are processes and decisions in the creation of the particular data set. Yet, these meta-documentations are only strictly valid for the original and not for derivatives. Ideally, we find a way of tracking changes to the data sets and processes applied in deriving underlying data sets for a paper. These important changes to the data cause changes to the metadata (e.g. additional editors) and need to be recorded. This idea can be extended to other disciplines, where data is analysed, manipulated, and presented. A horizontal Metadata Inheritance from previous versions to derivatives.

The second point relates to the lack of clear identifiers to reference parts of a data set or texts. While we can cite sentences and words from text, we have limited access to lower levels of singular constituents or data points. Not only are metadata on these instances often inaccessible to the public (e.g. for privacy reasons), we also do not have ways of citing these data points. While there may be handles or identifiers within each project, these are not unique beyond the project and get lost in citation. The lack of access to the lower layers prevents us from fully understanding links between projects, data sets, metadata beyond information specified by the researchers. Thus, while being able to track an author, we cannot necessarily track consultants who provided data. While there are privacy concerns about full disclosure, this lack of vertical Metadata Inheritance obscures the full image and may distort the view on a subject. For example, we cannot rule out biases due to over- or under-representation of particular social groups, if we cannot access these metadata. This also holds for applied sciences, where lack of information on consultants for the underlying data sets can create algorithmic biases favouring a particular worldview, language use, or behavioural pattern.

Facilitating horizontal and vertical Metadata Inheritance through sound meta-documentation (or data set theorisation) and identifiers for all artefacts (and all layers within them) can support the discovery of biases and the hidden stories within science. Modern computer-assisted meta-science should aim to uncover the holistic image of knowledge generation and show the unseen links.

References

1. Auer, S., et al.: Towards a Knowledge Graph for Science. In: WIMS '18 (2018)
2. Berez-Kroeker, A.L., et al.: Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics **56**(1), 1–18 (2018)
3. Himmelmann, N.P.: Language documentation: What is it and what is it good for? In: Gippert, J., et al. (eds.) Essentials of Language Documentation, pp. 1–30. Mouton de Gruyter, Berlin, New York (2006)
4. Holton, G.: Mediating language documentation. In: Nathan, D., Austin, P.K. (eds.) Language Documentation and Description 12, pp. 37–52. SOAS, London (2014)
5. Latour, B., Woolgar, S.: Laboratory Life: The Construction of Scientific Facts. Princeton University Press, Princeton (1986)
6. 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)
7. Renear, A.H., Wickett, K.M.: There are No Documents. In: Proceedings of Balisage: The Markup Conference 2010. vol. 5 (2010)
8. Turner, J.: Philology: The Forgotten Origins of the Modern Humanities, The William G. Bowen Series, vol. 70. Princeton University Press, Princeton and Oxford (2014)
9. Web of Science, http://wokinfo.com/
10. Weber, T.: Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics. In: Eskevich, M., et al. (eds.) 2nd Conference on Language, Data and Knowledge (LDK 2019). pp. 26:1–26:8. Schloss Dagstuhl, Dagstuhl (2019)

11. Woodbury, A.C.: Language documentation. In: Austin, P.K., Sallabank, J. (eds.) The Cambridge Handbook of Endangered Languages, pp. 159–186. Cambridge University Press, Cambridge (2011)