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

Tobias Weber[1 [0000−0003−1530−3568]]

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

Abstract. The paper highlights obstacles in discovering hidden stories and biases within the scientific endeavour which are caused by obscured knowledge generating procedures and the inaccessibility of lower data layers. Vertical and horizontal Metadata Inheritance is proposed as a solution to reveal traces of temporal and artefactual relationships.

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

1 Framing Science, Texts, and Data

In a recent paper, I argued for a philological view on metascience, especially in using computational methodology [11]. 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 starting point for this paper is the understanding of science as 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 the features of an 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 discussions or 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. As a result, we should treat data sets and various types of texts as equal as regards their status as a scientific artefact.

This raises two relevant points for 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. Likewise, any scientific text can become part of a data set for interpretation, for example in the analysis of scientific language use, or stylistic characteristics. Computational meta-science has, since its inception, tried to chart and 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, data collectors should receive the same academic merit as authors do [2].

Second, 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. This may appear logical, as we are not just recompiling information which already exists but adding our own interpretations, representations, or new insights to the existing literature or data sets. While some disciplines use a 'transcription device' [5] meant to automate the transfer from the data plane to the textual plane, the selection and application of transcription methods always requires decisions to be made. These could be in the selection of a topic or data set, the number or types of variables for an experiment, or the scope of the study, even for disciplines with transcription devices. While decisions may consist of simple adoption of existing data sets, methods, or representations (e.g. in a replication study), the artefacts are not identical. Therefore, I assume the data used in a commentary to be a new instance or version of the data set than was used in the original text. Texts evolve and with them their data sets change as well. 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. In a more radical understanding of document identity, two digital artefacts would only be identical if they contained the same sequence of bytes [7] - a point which reinforces the claim about creating new versions even through minor interaction with a text or data set.

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 of the same artefact but include derived artefacts and commentaries, as well, i.e. tracking changes after the creation of an instance. Using computational methods and resource identifiers, this is not a difficult task [10]. The difficulty arises 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 [4, 12], 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, this meta-documentation is 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 any other discipline, where data is analysed, manipulated, and presented. This represents 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 text. While we can cite sentences and words from a 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 tools 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, and metadata beyond information specified by the researchers. Within different parts of a scientific text, the metadata attached to a sentence or word may be different from the indicated data of its container, the article. Consider quotes from the literature, cited examples from external sources, or data which is attributed to particular individuals, e.g. an interviewee. Thus, while being able to track an author, we cannot necessarily track interview partners or other individuals supplying data (e.g. internet users, patients in medical records), without a sufficient archive structure. While there are privacy concerns about full disclosure of identity, 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. It important to investigate whether there are biases in our data - we can only rule out biases if we can compare all layers of metadata through a text and its related artefacts. 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.

The presented criteria of horizontal and vertical Metadata Inheritance are not yet fully implemented in any solution. The term ‘inheritance’ stems from programming languages, where features and functionalities of an instance are maintained for derived lower level sub-classes (vertically), as well as for duplicates or copies of the object (horizontally). The goal here is to clearly identify any (smallest) entity or constituent of the text or data set with respect to its position within a text and various layers of text, while clearly stating how this entity relates to preceding or derived versions of itself. Thereby, we could trace changes through time and be more precise in referencing and citing and overcome the obstacles of missing change logs, lacking access to lower levels of our artefacts, and the temporal boundedness to the point of creation beyond which we cannot incorporate derived versions into the artefact.

There are three possibilities to enable Metadata Inheritance. First, we may rely on external storage spaces for our metadata, in the forms of standardised databases of knowledge, e.g. names of authors, publications, institutions, ontological databases, typologies of scientific methodology. This strategy is already employed in the form of DOIs, author identifiers like ORCID, or ISBN / ISSN codes. As such, it works well for particular entities and artefacts on higher levels, without tracing changes. Furthermore, these identifiers do not grant access to the constituent level; we can only access documents or subsets of data through them, but not individual data points or words in a text. If radically employed, this method would require an enormous amount of handles and identifiers for the number of scientific publications or large data sets in academia nowadays. Potentially, this could be mitigated through use of slice functions or similarly generated versions of handles which, if consistently applied, could allow tracking of subsequent artefacts relating to the original.

Second, we may aim for an internal storage of metadata, for example as annotated texts in XML. This solution would include vectors or matrices for each point of metadata instead of singular information points which are replaced rather than amended by the addition of metadata for new versions. While these change logs would work well for the horizontal Metadata Inheritance, file size may become an issue if each word or data point required an individual change log. Especially for larger texts or data bases with many editors or long lists of citations, this may not be feasible. Furthermore, the publication format may filter the annotated texts and data sets, as printed publications do not allow for accessing the underlying codes or XML-marked documents.

Lastly, we could ignore either solution and aim for the retrospective reconstruction of links [10], which would enable us to also include texts and artefacts created before the introduction of any identifier or change log system. This philological work on textual artefacts also transcends boundaries of the medium, as the analysis does not rely on the inclusion of information in the original artefact but recreates these links at the time of the query. Yet, this system is not infallible either due to its reliance on data availability and access to necessary points of metadata, while requiring enormous capacities of computing power to collate and compare the large amount of texts and data sets at every time of a query.

A feasible solution for the proposed criteria of Metadata Inheritance is the combination of the aforementioned solutions. A sensible and potentially more extensive use of identifiers and handles allows researchers to access different levels of text and data, while providing a basis for tracing the subsequent use in other works. These links may then be included in change logs, as more complex ways of storing metadata which does not overwrite but add to existing sets of metadata. Combining change logs and identifiers can reduce file sizes, e.g. if all authors or institutions are referenced through unique handles rather than explicit data entries. In creating change logs, and for artefacts that predate the creation of the aforementioned systems, a computational approach to reconstructing relationships may be used, complementing the horizontal level [10]. The creation of transparent links and trajectories of our work, as outlined under the criteria of Metadata Inheritance, requires a sensible combination of different strategies for a quick, precise, and cost-efficient solution.

3 Conclusion

Keeping track of our actions and interactions with scientific artefacts allows for insights to the workings of disciplines and the scientific community, in general. In meta-scientific enquiry, we aim to understand how we as researchers work and which links exist in our work to that of others, seeking to understand the underlying networks. Thus, modern computer-assisted meta-science should aim to uncover the holistic image of knowledge generation and show the unseen links in our work. Metadata Inheritance plays an important role in this process, as it shows horizontal and vertical relationships in the network. Through ensuring horizontal and vertical Metadata Inheritance, through sound meta-documentation (or data set theorisation), identifiers for all artefacts (and all layers within them), or computer-assisted reconstruction of links, we can discover biases and the hidden stories within science.

Acknowledgements

I am grateful to the organisers of the *2nd Workshop on Reframing Research 2020* for the invitation to present this paper at the workshop, as well as to all reviewers and the audience for their useful comments on my work.

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. Weber, T.: A philological perspective on meta-scientific knowledge graphs. In: Bellatreche, L., et al. (eds.) ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. pp. 226–233. Springer International Publishing, Cham (2020)
12. 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)