# Track to the past: tracking workflows, versions, and citations of legacy language data

Tobias Weber
*Graduate School of Language & Literature, Ludwig-Maximilians-Universität*
München, Germany
tobias.weber@lrz.de

*Abstract*— **This paper discusses three issues encountered with legacy language data in archives: First, the provenance of an artefact containing the data may be unclear, as well as all procedures that shaped its form(at) or contents. Second, legacy language data are often orphan data with opaque links to other versions, or texts providing more information on them and their contents. Third, these data predate methods of data citation, thus requiring retroactive ways of citation tracking. With a few adjustments to their infrastructures, digital archives can be used as a platform to track workflows, versioning, and citations of legacy language data.**

***Keywords— legacy data, language documentation, linguistics, archiving, versioning, citation tracking, scientometrics, graph theory, anthropology***

## I. Introduction

Legacy materials are the outputs of past documentation projects [2]. As a result, working with these materials requires the researcher to understand the contexts of their creation, the history of their subsequent transformations, and the scientific as well as socio-cultural impact of the documentation project and its outputs. At the same time, relevant information is often scattered across different parts of archives and libraries, for example publications and raw data are kept in different repositories, while especially older documents might not be fully digitised or interoperable and metadata can be faulty. Digital language archives can offer tools and infrastructures to discover missing links, present data in context, and keep track of the histories of each artefact. This paper outlines three desiderata for language archives of legacy data. To ensure brevity, some issues pertaining to legacy materials are omitted from the discussion (for an in-depth discussion, see [19]).

## II. WORKFLOWS

The necessity to keep track of workflows is not just linked to professional conduct or the aims of making research intelligible or even 'reproducible' [3]. Contend that full reproducibility of linguistic analyses cannot be reached. The goal should rather be to enable readers and future researchers to understand the – occasionally subjective – decisions we made and to give them the necessary information to assess, discuss, and evaluate them on the basis of the data. Computational tools can still support this endeavour, yet not in a mechanistic replication of results [14] The goal is to have data and research papers in archives or libraries which are still understandable '500 years from now' [20]. Moreover, it is an ethical consideration to allow a review of our methodology in data collection. This involves acknowledging all individuals who contributed to a dataset or a publication [1, 8]. Yet, despite a clear requirement for transparent workflows and complete a metadocumentation, some points of metadata may be missing for legacy materials. This may be due to unknown provenance of an artefact, metadata loss during copying or transcribing, lossy artefact types (including physical media like manuscripts or wax cylinders), changes in professional standards, or idiosyncratic workflows. It might appear easy to discredit past researchers whose datasets lack sufficient records of metadata but this is not always a sign of unprofessional conduct. On the contrary, there are settings where privacy concerns or insecure socio-political circumstances have had an impact on the metadata recorded by a researcher [17–19] – we can only interpret the legacy materials and the accompanying metadata if we know about the historical contexts of their creation.

One solution to these obstacles can be found in the curation of language datasets. This process can be facilitated by archive structures that make texts findable and accessible, so that curators can reestablish links between artefacts. The curation process itself should be informed by contextual information from history, anthropology, or sociolinguistics, and is less technologically focused than data curation in other disciplines. Certainly, computational tools can support the process, although it is more about the speakers [6], or the 'human in the loop' [4], and less about the data as such. Thus, individual artefacts can even be approached from the perspective of forensics [9]. Due to the textual nature of the artefacts, we can also apply skills from philology [13], a text-based science that aims to understand texts in their historical and socio-cultural contexts. It involves comparing, commenting, and questioning texts and learning more about the circumstances of their creation. Furthermore, this approach is not just occupied with real-world contexts but also with the 'linguistic context' – the *cotext* [5] – the text surrounding a word, sentence, or paragraph. As a result, the restoration of links between artefacts is necessarily involved in intertextual networks: Field diaries help us to understand audio recordings, manuscripts support their transcription, references to previous documentation frames research projects and their objectives. With the same view, we can also link raw data and publications, or individual publications as a part of the same abstract workflow [16].

Digital archives can support the tracking of workflows through several means. For recent additions, ontologies of contributor roles and persistent identifiers for individuals and artefacts should be used for the metadata [11]. These should be necessarily thorough and supply information even on less prominent individuals, e.g. assistants who helped with transcription can leave noticeable traces in an artefact. This can also include knowledgeable scholars who are invited as external curators [21] and enrich a collection with their contextual awareness and information about research history. Community members may also be invited as curators, e.g. if they or their family members have been involved as consultants. In either case, the curator must, in turn, be credited for their work, as they leave their own traces in the dataset. These knowledgeable scholars can help to establish links between datasets and publications, and offer commentary based on the current state of research. Hyperlinking and referencing all relevant texts enables holistic treatment with a philological approach. The necessary requirements for this is transparency, including accessibility and findability of data.

## III. Versioning

A central benefit of digital archives is their accessibility through the internet. Yet, before the internet allowed for global access to data, they have been disseminated on physical media. As a result, a recording of the same event or a transcription of the same narrative might be archived in different locations. But are those copies actually the same? Even if they were created from the same original file, they are not identical [12]. On the one hand, different technological solutions or media have an influence on the data (e.g. loss rate, localisation), and receiving archives or researchers may have contributed further edits, annotations, or transformations to the data. On the other hand, considering the importance of co- and context, seemingly identical copies of the same data in different archives cannot be identical if we consider their archiving context relevant. Different contextualisations in the respective archives might arise from tags about the dataset, the compilation into overarching collections or thematic units, or the presentational formats of each archive (e.g. scanned copies by different archives in different resolutions or with different defective pixels). If we continue this line of thought, the observation about identity also affects data in publications, thus also the work of libraries [16] – each publication has a version of the dataset that is unique to this publication. Consider, for example, formatting rules, different layers of analysis, transcription rules, translations in to various languages [7], the surrounding interpretation (in the cotext).

Versions of textual artefacts need to be collated and compared, in order to establish the contexts in which changes were administered. This comparative task is well-known as part of textual scholarship, e.g. applied to medieval manuscripts, versions of literary texts. Stemmatology creates graphs of different versions, with each node an (actual or hypothetical) original version from which all its child nodes derive. If we think of the nodes as individual research papers, publications, compiled datasets and corpora – all in their own contexts – the image of all versions becomes opaque. Yet, we cannot separate the task of identifying relationships between individual versions from their concrete use. In other words, the version of the data has an impact on their analysis and their interpretation, and, in the spirit of replicability, we need to know which version produced a result or conclusion. We need each version to contain information on its position in the tree graph, and its relationship to other nodes. This means, it needs to be aware of preceding and subsequent versions, and the transformations from the original to this version; inheritance of metadata across time (horizontally) and sub- and supersets of the fragment (vertically) [15]. In digital environments, this can be dynamically generated and displayed, yet with the change of the medium (e.g. to print), we lose access to the history behind the data.

Digital archives offer several opportunities to support the tracking of versions. First, they hold the original data and often citable with persistent identifiers. Second, archives have the infrastructure to keep full accounts of metadata – although it is debatable whether archives should bear the onus of tracking versions, they have the capacity to do so. Third, the display of data citation and different versions alongside the original data can be beneficial to the scholarly community who can access and assess different interpretations of data and identify potential discrepancies. A necessary requirement for offering this function is the availability of digital copies on the side of the publications, which may be facilitated through the inclusion of digital libraries and publishing houses [14].

## IV. Data Citation tracking

As already mentioned in the previous sections, keeping track of versions is closely connected to tracking citations of data. On top of original citations, we also need to consider secondary citations, i.e. instances where data was copied from a publication and not from the original. Version tracking can help with this task and, considering its scope, highlight an important challenge to data citation tracking. Besides, we are potentially facing versions of data which are published in locations that are not accessible for citation tracking (e.g. community materials, blogs).Regarding online resources, Altmetrics [10] offer a possible approach to tracking citations in social media and on the internet. On the other hand, there are instances which may predate our infrastructure, i.e. citations before persistent identifiers were added to a dataset, publications which are not digitised or not included in databases. Since it is in the interest of the archive to keep track of the use of its data, a case for citation tracking by archives can be made. However, publishers and repositories need to support this endeavour by granting access to texts and cited references, especially for older publications which might not be fully digitised. This shows that the requirements are similar to those for version tracking, and that both procedures can be implemented alongside each other. For citation tracking, graphs can also be used to represent relationships between texts; combined with a copy of the version, its metadata, and all changes to the data itself, this becomes a powerful tool for researchers. At the same time, access to a holistic image of data use can prevent biases and misrepresentations, and allows all individuals who were part of the workflow to have their contributions appreciated and properly attributed.

## V. Conclusion

Legacy data poses different challenges to archives than recently deposited datasets. Apart from ethical concerns about their provenance, the history of the artefact can be unclear, including processes of its creation, subsequent use, and citation. Yet, omitting legacy data from research or restricting access to them due to their unclear history should be the last resort, as it means the loss of valuable knowledge and disregard to the creators' efforts, not least to that of the consultants' and communities'. The value of legacy materials needs to be appreciated through careful reconstruction using philological, anthropological, and historical knowledge and skills. Some of the required steps can be supported by computational methods, where the collaboration of digital archives and libraries is essential. At the same time, archives, libraries, and publishers stand to gain from transparent workflows, versions, and (data) citations of their resources. Legacy data must not be ignored and can, on the contrary, inform the design of tools that do not only work on recent data and metadata but also on historical records of our discipline. Creating this 'backwards compatibility' of legacy data with modern standards is a sign of our appreciation – the same appreciation we would want from future generations for our present-day deposits.

## REFERENCES

[1] Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics. https://doi.org/10.15497/rda00040.

[2] Peter K. Austin. 2013. Language documentation and meta-documentation. In Keeping Languages Alive. Documentation, Pedagogy, and Revitalisation, Mari Jones and Sarah Ogilvie (Eds.). Cambridge University Press, Cambridge, 3–15.

[3] Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics 56, 1 (2018), 1–18.

[4] Steven Bird. 2020. Decolonising Speech and Language Technology. In Proceedings of the 28th International Conference on Computational Linguistics, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, 3504–3519. https: //doi.org/10.18653/v1/2020.coling-main.313

[5] John Cunnison Catford. 1965. A linguistic theory of translation: an essay in applied linguistics. Oxford University Press, Oxford.

[6] Lise Dobrin, Peter K. Austin, and David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. In Language Documentation and Description, Peter K. Austin (Ed.). Vol. 6. SOAS, London, 37–52.

[7] Jan Engh. 2006. Norwegian examples in international linguistics literature. An inventory of defective documentation. Universitetsbiblioteket i Oslo, Oslo.

[8] Alex O. Holcombe. 2019. Contributorship, Not Authorship: Use CRediT to Indicate Who Did What. Publications 7, 3 (2019), 1–11. https://doi.org/10.3390/publications7030048

[9] Gareth Knight. 2012. The Forensic Curator: Digital Forensics as a Solution to Addressing the Curatorial Challenges Posed by Personal Digital Archives. International Journal of Digital Curation 7, 2 (2012), 40–63. https://doi.org/10.2218/ijdc.v7i2.228

[10] Jean Liu and Euan Adie. 2013. Five challenges in altmetrics: A toolmaker's perspective. Bulletin of the American Society for Information Science and Technology 39, 4 (2013), 31–34. https://doi.org/10.1002/bult.2013.1720390410

[11] Steve Pepper. 2011. Ontologies in language documentation. In Language Documentation and Description, Julia Sallabank (Ed.). Vol. 9. SOAS, London, 199–218.

[12] Allen H. Renear and Karen M. Wickett. 2010. There are No Documents. Proceedings of Balisage: The Markup Conference 2010 5 (2010). https: //doi.org/10.4242/BalisageVol5.Renear01.

[13] Frank Seidel. 2016. Documentary linguistics: A language philology of the 21st century. In Language Documentation and Description, Peter K. Austin (Ed.). Vol. 13. SOAS, London, 23–63.

[14] Tobias Weber. 2019. Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics?. In 2nd Conference on Language, Data and Knowledge (LDK 2019) (OpenAccess Series in Informatics (OASIcs), Vol. 70), Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski (Eds.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 26:1–26:8. https://doi.org/10.4230/OASIcs.LDK.2019.26

[15] Tobias Weber. 2020. Metadata Inheritance: New Research Paper, New Data, New Metadata?. In Reframing Research Workshop Accepted Papers, Andrea Mannocci (Ed.). Zenodo. https://doi.org/10.5281/zenodo.4155362

[16] Tobias Weber. 2020. A Philological Perspective on Meta-scientific Knowledge Graphs. In ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium, Ladjel Bellatreche, Mária Bieliková, Omar Boussaïd, Barbara Catania, Jérôme Darmont, Elena Demidova, Fabien Duchateau, Mark Hall, Tanja Merčun, Boris Novikov, Christos Papatheodorou, Thomas Risse, Oscar Romero, Lucile Sautot, Guilaine Talens, Robert Wrembel, and Maja Žumer (Eds.). Springer International Publishing, Cham, 226–233. https://doi.org/10.1007/978-3-030-55814-7_19

[17] Tobias Weber. 2021. Consultant Identity in Historical Language Data: Anthroponyms as a Tool or as an Obstacle? In Proceedings of the International Onomastic Conference "Anthroponyms and Anthroponymic Researches in the Beginning of 21st Century", Anna Choleva-Dimitrova, Maya VlahovaAngelova, and Nadezhda Dancheva (Eds.). Bulgarian Academy of Sciences, Sofia, 165–175.

[18] Tobias Weber. 2021. Mind the Gap: Language Data, Their Producers, and the Scientific Process. In 3rd Conference on Language, Data and Knowledge (LDK 2021) (Open Access Series in Informatics (OASIcs), Vol. 93), Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, 6:1–6:9. https://doi.org/10.4230/OASIcs.LDK.2021.6

[19] Tobias Weber. forthcoming. Philology in the folklore archive: Interpreting past documentation of the Kraasna dialect of Estonian. In Language Documentation and Description, Lise M. Dobrin and Saul Schwartz (Eds.). Vol. 21. ELPublishing, London, forthcoming.

[20] Anthony C. Woodbury. 2003. Defining documentary linguistics. In Language Documentation and Description, Peter K. Austin (Ed.). Vol. 1. SOAS, London, 35–51.

[21] Anthony C. Woodbury. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving, David Nathan and Peter K. Austin (Eds.). SOAS, London, 19–36.