Journal of Open Humanities Data

The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers

SPECIAL COLLECTION: LANGUAGE DOCUMENTATION

RESEARCH PAPER

TOBIAS WEBER

ubiquity press

ABSTRACT

This position paper outlines curation as a distinct area of linguistics which supports the assessment of existing linguistic data sets. The starting point of the discussion lies in the work with legacy materials and necessary skills for reviewing them. The discussion of skills and knowledge relevant for this task lays out requirements for training and preparing academics and professionals for curation in ways that creates career options. Due to the importance of properly maintained data sets and the time-consuming nature of curation, this specialisation deserves the same merit as other areas of linguistics and should be treated equally.

CORRESPONDING AUTHOR:
Tobias Weber
Graduate School Language & Literature, Ludwig-Maximilians-Universität München, Munich, Germany
weber.tobias@campus.lmu.de

KEYWORDS:
legacy materials; philology; computer literacy; documentary linguistics; data citation; curriculum development

TO CITE THIS ARTICLE:
Weber, T. (2021). The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers. *Journal of Open Humanities Data*, 7: 28, pp. 1–10. DOI: *https://doi.org/10.5334/johd.51*

1 CONTEXT AND MOTIVATION

Curation is a value-adding procedure. It broadly refers to all activities following the point of creation of data that facilitate their (continued) use and reuse. For linguistics, it can be defined with a focus on maintaining datasets, the 'development of indices, annotated linguistic corpora, and digitally encoded texts' (Muñoz, 2013), especially 'digital information that is produced in the course of research in a manner that preserves its meaning and usefulness as a potential input for further research' (Muñoz & Renear, 2011). Similar to the curation of artefacts for museums or galleries, it involves itemisation, restoration, preservation, and contextualisation which, nowadays, constitutes a full review of historical circumstances in their creation and dissemination (Kreps, 2003). The process requires curators and users to reflect on their own position towards the artefact and to form an evaluation. Consequently, curation is a method for (re-)assessing artefacts, whether those are paintings, ancient pottery, or linguistic data sets. This paper discusses curation of language data, especially historical legacy data, and considers the position of curation and those curating within academia.

Several important points follow from these defining characteristics: First, value-adding is tied to the workers or institutions adding the value through their labour, knowledge, and skills. As a consequence, these workers deserve credit, merit, or remuneration for their efforts – this can be understood in terms of academic recognition of scholarly work, opportunities for hiring and tenure, and the wages for an appropriately-compensated workforce. We must therefore consider the different agents in curation and the recognition they receive for their work. While publishing was often the main activity for scholars that would allow them to earn academic merit, scholars in language documentation have long been calling for the recognition of documentary activities and creation of data sets (Andreassen et al., 2019; Berez-Kroeker et al., 2018; Linguistic Society of America, 1994, 2010, 2018). The present paper follows their argumentation but aims to frame the activities worthy of recognition to also comprise work on pre-existing materials and not just new documentations (cf. Thessen et al., 2019). While the focus of the paper lies on academics, I agree with one of the reviewers that archive professionals deserve the same recognition and visibility of their work; neither linguists nor archivists should face adverse labour conditions when engaging in curation.

Second, curation and the value added are more than just the archiving and preserving of language documentation outputs or artefacts. Although documentary linguistics emphasised the role of community needs in archiving in the last decade (Bird, 2020; Woodbury, 2014), the idea behind archiving is still tied to discourses of citability, reproducibility, and the prevention of data loss (cf. Berez-Kroeker et al., 2018). While these are important aspects of archiving and preservation, this focus on technical details embodies the mechanistic orientation in documentary linguistics which, at times, distracts from the needs and interests of the language communities (Bird, 2020; Dobrin et al., 2009). At the same time, reducing the scope and requirements for archiving implies that the linguist's responsibility stops at the point where data is handed over to archives and repositories. Instead, academics should bear the responsibility for their data and contribute to their maintenance; reviewing and interacting with archived materials can improve their quality. We must subsequently ask: Who is responsible for curation? Which grey areas may exist between the linguist's and the archivist's purview?

Third, if curation is broader than archiving and includes community-oriented tasks, the initial definition implies that curation be interdisciplinary in nature, and not preoccupied with technical details. This leads to the central questions of this paper: Which skills are required for curation and how do they relate to existing conceptualisations of linguistics? As I argue in the title of the paper, understanding curation as a distinct and fully-fledged specialisation in linguistics helps us to consider the importance of curating for the entire endeavour of documentary linguistics and to chart the grey areas mentioned above. The goal is not to separate curation from the other activities, as we are dealing with a dynamic process informed by various disciplines. It does not summon Himmelmann's (1998) strict distinction of language documentation and language description (cf. Austin & Grenoble, 2007) but aims to highlight an overlooked yet important field of interdisciplinary action with its distinct needs for training and funding.

The critical reader may ask why the tasks of curation should not be divided between the researcher and an archivist. I would like to illustrate this point with an anecdote from my personal experience, which also serves as motivation for this paper. In my work, I have focused

DOI: 10.5334/johd.51

on legacy materials of a South Estonian variety called *Kraasna* (Glottocode *kraa1234*), which has no speakers and is not a symbol of identity for any present-day community – in the terms of the Expanded Graded Intergenerational Disruption Scale (Lewis & Simons, 2010), it is *extinct*. Consequently, there are not many researchers working on Kraasna, with the last records of coherent, active language use (i.e. phonograph recordings and their transcriptions) dating to 1914. Digitising, transcribing, collating, and editing these legacy materials took several years (Weber, 2016). Yet, a reviewer for a recent article outlining this process (Weber, 2021e) asked in their comments: "I guess it wasn't only your [curating] work that was the aim of the project?". This is certainly not meant to sound disrespectful to the reviewers whose comments I highly value, but this particular question bears witness to the tacit assumption that curation itself *cannot* be a project: some additional use or subsequent outcome must be the real aim. Admittedly, the work on the legacy materials led to the publication of a linguistic description of the recordings (Weber, 2021c), yet I would not want to consider the prior work of transcribing and digitising as an ancillary activity. This view would devalue and relegate the knowledge and skills to an inferior position in the toolbox of a linguist. Likewise, unsuccessful submission to a journal does not render the efforts in curating legacy materials useless – the value of these activities is independent from further use of the curated materials. A special role in this discussion is occupied by corpus reviews or overview articles (see Linguistic Society of America, 2018), where the focus of the publication lies on reporting about a resource and its stage of curation. Yet, the main bulk of linguistic publications consists of theoretical and descriptive work where data collections are important resources but not the core of the publication. Admittedly, since the value-adding processes in preparation of a project or an article are very specific and particular to the respective methodology, it is difficult to detach them from envisioned goals of the project or needs of its target audience. Researchers often need to invest time ahead of their projects to adapt and prepare data for their needs – but updating and revisiting these materials is already a valuable activity in itself because it assesses and improves the quality of data sets.

Furthermore, this example highlights an issue with handing over responsibility to archivists. In many disciplines where data is highly standardised, data managers or data stewards at research institutions or archives (incl. libraries, databases, repositories) can take care of maintaining data sets and curating them. Yet, with linguistic data being very diverse and difficult to standardise at times, finding a curator at every institution to cover all types of data and all languages is not a feasible approach. Considering the frequently cited figure of over 7,000 languages in the world, external curators might be hired for the major languages with larger databases to be maintained but not for the majority of less-widely spoken languages, not to mention extinct, peripheral South Estonian dialects. For these smaller languages, curating remains a task for individual researchers or even community members (Woodbury, 2014). While individual researchers might feel the pressure to publish research papers and consider curation a less important or ancillary task (Thessen et al., 2019), community members need training and support with curation tasks and methods. There are already some projects recognising the importance of communities' participation and collaboration by providing special training in accessing and using language data.[1] Ultimately, the work of community members must be acknowledged (Andreassen et al., 2019), and enable paths into scientific work and research by the communities (see Grinevald, 2003). As such, the boundaries between researcher, curator, and consultants become blurred. This should not impact the quality of research and/or curation, as it would reinforce the vicious circle to the detriment of smaller languages (Weber, 2021d). Importantly, curation is a collaborative task that cannot be performed in its breadth by a single individual (Muñoz & Renear, 2011). Making use of individual skills and knowledge of different curators, the process can incorporate multiple approaches and outcomes. In light of supporting and training curators for endangered languages, initiatives focusing on existing resources like the Legacy Material Grant by the Endangered Languages Documentation Programme,[2] the Michael Krauss Archiving and Revitalization Legacy[3] or the HELP for Endangered Legacy

1 The Myaamia Center at Miami University in Ohio hosts a training programme with the National Breath of Life Archival Institute for Indigenous Languages (*https://miamioh.edu/myaamia-center/breath-of-life*).

2 *https://www.eldp.net/en/our+grants/legacy+material+grants*.

3 *http://www.endangeredlanguagefund.org/*.

Collections project by the Linguistic Society of America's Committee on Endangered Languages and their Preservation[4] must be noted as positive examples.

As seen from the previous discussion, curating activities are done by archivists, researchers, and community members, while relying on the support of funding programmes and training in relevant skills. The next section will outline the distinct profile of language data curation.

2 CURATION AS A DISTINCT ACADEMIC ACTIVITY

The curation of language data is as equally important as language documentation, archiving, linguistic analysis or description; all of these specialisations come with practical skills and theoretical foundations which need to be understood by students, scholars, and practitioners. There is an overlap between curation and the other fields, mainly due to the arbitrary points of inception and termination of each field. However, this overlap does not justify the subsumption of curation as an ancillary activity within either language documentation, archiving, or descriptive linguistics. In my understanding, curation functions as an external locus of reflection and negotiation between the communities of documentary and descriptive linguists.[5] It reviews procedures and outcomes, checking for accuracy or "replicating" results, and "mediates" between the needs of various groups (on the concept of mediation, see Holton, 2014). As such, curation bridges the gap between creating or depositing language data sets and the subsequent uses of the data sets. In doing so, it connects the present-day work to past documentation efforts which the curators must interpret and assess. Thus, curation puts the challenges posed by pre-existing materials into the focus of the discourse in documentary linguistics (Austin, 2017). The focus of my own research has been on legacy materials, although the relevant methodology can also be applied to data sets which were created more recently. In the curation of the Kraasna legacy materials, I used philological methods which link language documentation and linguistics to literary studies (Weber, 2021e), within a broader view of the humanities (for other examples of philological methods applied to language data see e.g. Broadbent, 1957; Goddard, 1973; Helimski, 1997; Winkler, 1994, 1997). The differences lie in the genre of text and the age of the source but critical literacy, text analysis, editing, and linguistic and socio-cultural knowledge form a core of the competences – the outcome would be a 'critical edition of legacy materials' (Linguistic Society of America, 2010). This does not mean that curation would not require digital literacy, on the contrary, computational skills support (philological) curation (Weber, 2019; 2020a, 2020b). In combination, this set of skills and knowledge can help curators to overcome medial, technological, linguistic, and intellectual obstacles. The use of a trained linguist as a curator who has sociocultural as well as practical knowledge of the language contained in artefacts or metadata files can support archives, which provide their expertise in physical restoration and digitisation. At this point, I feel obliged to echo a reviewer's plea to assess the physical state of a medium before attempting to digitise it. Even playing an audio tape, not to mention the Kraasna wax cylinders, can severly damage a recording. These tasks must be left for archive professionals with the correct, professional equipment.

For the Kraasna legacy materials, I needed to work with different types of artefact and media formats, apply digitisation methods such as optical character recognition, make informed decisions on representational formats (e.g. encodings), and make sense of the textual artefacts. While the medial and technological challenges can be solved by all trained archivists and data managers, the linguistic and intellectual challenges require linguistic training, such as knowledge of languages, scripts, handwriting, and textual comparison. These link curation with neighbouring disciplines in anthropology, historical science, and possibly also forensics (Knight, 2012), bibliography, or archival science, including skills such as palaeography, chronology and textual genealogies, editing (Seidel, 2016), or onomastics (e.g. anthroponyms in archival sources, Weber, 2021a). These steps render curation time-intensive, yet sometimes a community member or scholar with relevant language or contextual knowledge may be better equipped for handling these tasks than an archivist. This stance does not depreciate the invaluable work done by archivists and research data managers – external input can close gaps

4 *https://www.linguisticsociety.org/content/help-endangered-legacy-collections*.

5 For a comparison of conceptualisations and graphic representations of the scientific field see Weber (2021b).

and amend existing archival records, yet these contributions need to be acknowledged. Ideally, all associated individuals, archivists and academics, are listed with their contributions to the process and are able to build a career on this work. To provide an example from the Kraasna legacy materials about the importance of knowing about curators: The phonograph recordings (for more information see Weber, 2021c) held by the archives of the Finnish Literature Society (SKS) bear labels like *puhetta* 'speech, talk', or *ohjelma* 'programme' which are non-descript and do not contain information on the actual contents of the recordings. An unknown archivist who digitised the recordings in the 1980s read out the labels before each recording, yet without adding the value of descriptive titles (*sisältää pesemistä ynnä muuta sellaista* 'contains Pesemistä [laundry] and so on'). The transcription (and understanding) of this three and a half minute recording took me several hours and allowed the identification of the other narratives, as well as linking them to an edited volume of related manuscripts where the laundry story is titled 'weekend' (*Nädalavahetusest,* Mets et al., 2014, 281–282). This example shows the importance of using philological care in restoring data and metadata, an area where a linguist may be best equipped to process the data. Furthermore, what may seem like a routine task, like transcribing a text, requires the careful recording of metadata; changes, no matter how minute, need to be attributable to their authors (Austin, 2017). This ties the merit for curating to the responsibility which each curator needs to accept.

On top of these activities and skills, curation can include (or should at least maintain ties with) pedagogy, arts, media, communications and publishing, and technology. While there are experts in these (distinct) fields, there is a reciprocal relationship between their work and the curator's. Decisions in curation are informed by the needs of these communities (e.g. the creation of learning materials, applications in Natural Language Processing). Likewise, the added value through different envisioned uses or versions is directly linked to the variety in curated data sets. This is not a one-way process (Weber, 2021d); curators can learn from different use cases of their data sets and use the final products of their colleagues for their own work (e.g. using applications, including examples in training). Their work forms a dynamic and constant review of the data collections and their colleagues' interactions with them. Yet, even in a collaborative approach (Fenlon, 2020; Wasson et al., 2016), the curators must not be seen merely as service providers – their work deserves equal mention and credit to other project outcomes, with funding available for their work (cf. Andreassen et al., 2019; Nowviskie, 2011). On the contrary, curation may form the core of all endeavours related to the data, in a "curation-centric" approach (Hedstrom, 2012). As advocated for by Hedstrom, curators must be enabled to build careers on their work and receive merit for their activities, whether they are professionals, researchers, or community members (see also Pryor & Donnelly, 2009). This links the discussion of giving credit for curation to training and curriculum development in linguistics and the humanities.

Using activities of curating language data sets for advancing one's career must be understood in different settings. On the one hand, there are professionals whose main occupation is managing data, e.g. at archives or research institutions. On the other hand, researchers and scholars who have engaged in curation, for example as part of their research projects, also need these activities to be recognised. The latter case also includes students or members of the public who engage with language data. Training for curation is thus not exclusively tied to vocational training or geared towards educating specialists in a narrow field of work. Curation is rather to be included as one component of academic programmes in linguistics or the humanities, due to its interdisciplinary focus – students in anthropology or history may contribute their expertise and skills to a curation project, while training in the relevant linguistic subjects allows them to work on language data, as well. As such, it is not replacing core elements of linguistic or language degrees (such as language practice, subjects in linguistics, literature and culture studies) but amending the profile of a linguist with skills focused on language data management and curation. Considering the workload for students, this suggestion does not add more courses or duties to the linguistics curricula. Instead, shaping existing modules and combining them in new ways gives educators the opportunity to populate the roles and profiles in between core specialisations in linguistics. A student will not be expected to fill all roles at the same time but to respond to the changing needs of various graduate profiles. While sensitive tasks requiring special knowledge and training must remain with experts (e.g. restoration of wax cylinders), there is a demand for graduates combining core linguistic training and applied

linguistics, computer science, or archival science. This need has been noticed by educators and policymakers; a graduate of academic programmes in linguistics needs to know how to handle language data. A specialisation in this area can open career paths as a language data specialist in the labour market (see Petrović et al., 2021), while the relevant knowledge and skills are applicable in many academic careers. Besides language skills and linguistic knowledge, graduates of linguistics degree programmes increasingly need IT skills – but their training should also include interpersonal skills, knowledge of ethics, rights, intercultural communication, and theories of (meta-)documentary linguistics (Bernardini & Petrović, 2021). The latter define the profile of a linguist working in language data curation or data management. The goal in training is not to turn students in linguistics and the humanities into computational linguists or computer scientists – instead, educators should emphasise the roots of the disciplines and highlight the benefits of humanistic education for a curator, archivist, or data specialist (Weber & Bradley, 2018). Consequently, the view on IT skills is slightly different, informed by community needs (Bird, 2020), and reflective in nature. This orientation away from minute technical details or standards towards the language data at hand and careful negotiation of interests defines the profile of curation. Yet, certainly, the goal is always to achieve the best possible outcome of a curated data set, but this stance does not disregard or ignore collections which do not or cannot fulfil the requirements set by (meta)data frameworks or ontologies (see Weber, 2021d). As regards the Kraasna legacy materials, there were gaps in the metadata which curation was able to close, while others still remain (e.g. the identity of the consultants remains unknown). This example shows that curation can increase the usability of language data sets, yet cannot solve all issues with legacy materials. At the same time, this work valorises these data sets, their contents, and their producers – a strict focus on technical details and standards might omit these data and go against the appreciatitive stance propagated in documentary linguistics. In other words, curation reconciles researchers and communities, various stances encountered in different scientific communities, and ensures that all data sets receive appropriate attention and are (re-)used adequately. Therein lies the necessity of a linguistic profile of the language data specialist. The combination of different skill sets beyond IT literacy, with a focus on the disciplinary roots in the humanities (including archival science), shape this profile and support the case for curation as an academic, as well as a professional, career path.

The position of curation in the curriculum, as advocated here, is not one of a detached specialisation or new kind of degree programme. On the contrary, the distinct role of curation can be achieved through a modular and interdisciplinary approach (cf. Petrović et al., 2021). Due to the diverse tasks in curation, skills and competences from various disciplines in the humanities can be required, like ethnography, philological editing, forensics, pedagogy, or (intercultural) communication. This being said, with a focus on language data, relevant skills also stem from computer science, technology, or IT – yet, with a distinct linguistic profile that does not forget about speech communities and consultants (Bird, 2020; Dobrin et al., 2009). Under this condition, curation can benefit from digital tools and methods (Knight, 2012; Weber, 2019). Overall, the need for more IT and data literacy is not unique to linguistics or the humanities. Studies in social sciences and natural science show that degree programmes in these disciplines also need to allocate more time for training their students in data management (Doonan et al., 2020; Tenopir et al., 2016). Thus, incorporating more modules of data and computer science into curricula is a universal task which supports scientists and graduates in the labour market to address the need for data specialists. Working with language data, especially from endangered languages, requires professionals who can adequately curate linguistic data sets, beyond the technical details. Curation as a distinct specialisation in linguistics prepares students for these needs and shapes the profile of a language data specialist with a linguistics background.

3 CONCLUSION

This paper highlighted the role of language curation as a mediator between different stakeholders in the creation and use of language data. Negotiating the needs of those stakeholders and preparing data sets accordingly is a time-consuming activity that needs the expertise of different specialists. Yet, if these activities are relegated to the grey areas of ancillary or preparatory activities, their importance becomes obscured. Especially for assessing and reviewing existing data sets and archived materials, curation can help to ensure the (re-) usability of data. In the process, corrections or adjustments to data and metadata can be

made, allowing language archives to become places of negotiation and reflection of academic practices. The promotion of curation as a distinct academic activity within linguistics, as advocated in this paper, emphasises the importance of adding value to existing data sets and makes these activities visible. There have been other approaches to increase the visibility of language data (collections) (cf. Woodbury, 2014), but curation still remains peripheral to the main occupations of scholars and graduates of linguistics. Integrating language data curation conceptually as a core element of linguistics helps us to react to the requirements and can create job opportunities within and outside of academia. It must be clear that anyone who bears the high responsibility for curating language data and invests time in existing collections must be able to claim credit and build a career on this work (rather than being dependent on the 'actual' project). This stance leads to a number of desiderata for the different stakeholders mentioned in the title of this paper.

Archivists and researchers, including students and citizen scientists, need to make their curating activities more visible. This implies spelling out and emphasising the 'preparatory' activities and the researchers' role in them. While researchers aim to follow guidelines, recommendations, or standards, there is a human factor in the creation and collation of any data set, even of automatically generated data sets (Bird, 2020; Thessen et al., 2019). Ignoring or failing to acknowledge the responsibility of the researcher in this process can be detrimental to the reliability or reproducibility of the data set (cf. Weber, 2019) – accepting the responsibility for a data set is the condition for claiming credit for its creation or curation. The same holds for archivists, as it is crucial to know who authored a change or made a decision that can influence our understanding or interpretation of a data set. Consequently, data collections need to be reviewed, commented on, critiqued and constantly improved as part of the curation process. It is the responsibility of the researcher, especially of senior researchers, to support the assessment of data sets and to consider their curation in reviewing applications and papers. Likewise, students and early-career researchers should be encouraged by their teachers and supervisors to take up the responsibility for a set of legacy materials, or a less well-curated data set (see footnote 4). This can also be realised in the form of an interdisciplinary project, or serve purposes in training. The engagement with more challenging, uncurated data sets over 'gold standard' ones can support minority languages and make historical records visible in contemporary research again (Weber, 2021d).

At the same time, funding agencies or institutions can support the curation of data sets by providing financial support for related activities, consider them in the hiring process, or count them towards tenure (Linguistic Society of America, 2018). To increase visibility, they can host events with the community or hold an exhibition of language data (Woodbury, 2014). Furthermore, awards for language data sets can create incentives for researchers, although the conditions should allow for legacy materials and historical records to also be considered along newly collected data, e.g. the Society for the Study of the Indigenous Languages of the Americas Archiving Award[6] or the Digital Endangered Languages and Musics Archives Network Award.[7] The crucial point is that preparing data sets for (potential) subsequent uses must be treated as an output in its own right; projects in linguistics occasionally contain the publication of a refurbished data set or corpus interface as an additional deliverable. This sells the time, effort, and craftsmanship short of their actual value and can potentially impact the quality of the outcome. Instead, emphasising the value of the curated data set, independent from other parts of the project or the successful publication of results, can signal the relevance and importance of well-maintained data in linguistics.

Ultimately, the responsibility lies with educators to train students in linguistics in more than the theories and methods of language documentation and description. The inclusion of curation, and possibly other areas of applied linguistics (computational linguistics, pedagogy, language policy), in curricula can support students in working with language data, whether in subsequent academic positions or in the labour market (Penfield & Tucker, 2011). This can increase the profile of students in linguistics and, with a view towards an interdisciplinary definition of curation, in the humanities, in general. At the same time, communities can

6 *https://www.ssila.org/en/archiving-award*.

7 *https://www.delaman.org/delaman-award/*.

benefit from researchers who have had sufficient training in the creation of sound pedagogical materials from documented language data, instead of by-products in projects aiming at scientific publications alone, for example. Incorporating curation as a distinct and valuable area in linguistics ensures that linguistics responds to its social responsibility, prepares students for working with language data, and creates awareness of existing language data sets. This valorisation of past language documentation projects, the work of peers, and the data they produced also affects the perceived value of legacy materials – and we can learn from legacy materials and their 'producers' through curation projects. In this view, curation should treat historical language data with the same respect that we would want from future generations of researchers for our own language data from present-day documentation projects.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR AFFILIATION

Tobias Weber
Graduate School Language & Literature, Ludwig-Maximilians-Universität München, Munich, Germany

REFERENCES

Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., De Smedt, K., McDonnell, B., & **the Research Data Alliance Linguistic Data Interest Group.** (2019). Tromsø recommendations for citation of research data in linguistics. *Research Data Alliance*. DOI: *https://doi.org/10.15497/rda00040*

Austin, P. K. (2017). Language Documentation and Legacy Text Materials. *Asian and African Languages and Linguistics, 11* , 23–44.

Austin, P. K., & **Grenoble, L.** (2007). Current Trends in Language Documentation. In P. K. Austin (Ed.), *Language Documentation and Description,* 4, 12–25. London: SOAS.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & **Woodbury, A. C.** (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics, 56*(1), 1–18. DOI: *https://doi.org/10.1515/ling-2017-0032*

Bernardini, S., & **Petrović, M. M.** (2021). Toward a new profile for twenty-first century language specialists: Industry, institutional and academic insights. *Zenodo*. DOI: *https://doi.org/10.5281/zenodo.5030873*

Bird, S. (2020). Decolonising Speech and Language Technology. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). Barcelona: International Committee on Computational Linguistics. DOI: *https://doi.org/10.18653/v1/2020.coling-main.313*

Broadbent, S. M. (1957). Rumsen I: Methods of Reconstitution. *International Journal of American Linguistics, 23*(4), 275–280. DOI: *https://doi.org/10.1086/464419*

Dobrin, L., Austin, P. K., & **Nathan, D.** (2009). Dying to be counted: the commodification of endangered languages in documentary linguistics. In P. K. Austin (Ed.), *Language Documentation and Description, 6*, 37–52. London: SOAS.

Doonan, A., Akmon, D., & **Cosby, E.** (2020). An Exploratory Analysis of Social Science Graduate Education in Data Management and Data Sharing. *International Journal of Digital Curation, 15*(1), 1–18. DOI: *https://doi.org/10.2218/ijdc.v15i1.671*

Fenlon, K. S. (2020). Sustaining Digital Humanities Collections: Challenges and Community-Centred Strategies. *International Journal of Digital Curation, 15*(1), 1–13. DOI: *https://doi.org/10.2218/ijdc.v15i1.725*

Goddard, I. (1973). Philological Approaches to the Study of North American Indian Languages: Documents and Documentation. In T. A. Sebeok (Ed.), *Linguistics in North America, 10*, 727–745. The Hague: Mouton. DOI: *https://doi.org/10.1515/9783111418780-023*

Grinevald, C. (2003). Speakers and documentation of endangered languages. In P. K. Austin (Ed.), *Language Documentation and Description, 1*, 52–72. London: SOAS.

Hedstrom, M. (2012). Digital Data Curation – Examining Needs for Digital Data Curators. *Paper presented at Cultural Heritage Online: Trusted Digital Repositories & Trusted Professionals*. Florence: Fondazione Rinascimento Digitale.

Helimski, E. (1997). *Die Matorische Sprache, 41*. Szeged: University of Szeged.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36 , 161–195. DOI:

https://doi.org/10.1515/ling.1998.36.1.161

Holton, G. (2014). Mediating language documentation. In D. Nathan & P. K. Austin (Eds.), *Language Documentation and Description 12: Special Issue on Language Documentation and Archiving* (p. 37–52). London: SOAS.

Knight, G. (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), 40–63. DOI: *https://doi.org/10.2218/ijdc.v7i2.228*

Kreps, C. (2003). *Liberating Culture: Cross-Cultural Perspectives on Museums, Curation and Heritage Preservation*. Milton Park: Routledge.

Lewis, M. P., & **Simons, G. F.** (2010). Assessing Endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique, 55*(2), 103–120.

Linguistic Society of America. (1994). *The Need for the Documentation of Linguistic Diversity.* Statement, 1 June 1994. Retrieved from https://www .linguisticsociety .org/resource/resolutions-statements-and-guides

Linguistic Society of America. (2010). *Recognizing the Scholarly Merit of Language Documentation.* Resolution, 8 January 2010. Retrieved from *https://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation*

Linguistic Society of America. (2018). *Statement on Evaluation of Language Documentation for Hiring, Tenure, and Promotion.* 25 September 2018. Retrieved from *https://www.linguisticsociety.org/resource/resolutions-statements-and-guides*

Mets, M., Haak, A., Iva, T., Juhkason, G., Kalmus, M., Norvik, M., Pajusalu, K., Teras, P., Tuisk, T., & **Vaba, L.** (2014). *Lõunaeesti keelesaarte tekstid. Eesti murded, IX*. Tallinn: Eesti Keele Instituut & Tartu Ülikool.

Muñoz, T. (2013). Data Curation as Publishing for the Digital Humanities. *Journal of Digital Humanities, 2*(3).

Muñoz, T., & **Renear, A.** (2011). *Issues in Humanities Data Curation. Whitepaper for Humanities Data Curation Summit*. Retrieved from *http://cirss.ischool.illinois.edu/paloalto/whitepaper/premeeting/*

Nowviskie, B. (2011). Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. *Professions,* 169–181. DOI: *https://doi.org/10.1632/prof.2011.2011.1.169*

Penfield, S. D., & **Tucker, B. V.** (2011). From documenting to revitalizing an endangered language: where do applied linguists fit? *Language and Education, 25*(4), 291–305. DOI: *https://doi.org/10.1080/09500782.2011.577219*

Petrović, M. M., Bernardini, S., Ferraresi, A., Aragrande, G., & **Barrón-Cedeño, A.** (2021). *Language data and project specialist: A new modular profile for graduates in language-related disciplines.* Zenodo. DOI: *https://doi.org/10.5281/zenodo.5030929*

Pryor, G., & **Donnelly, M.** (2009). Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career? *International Journal of Digital Curation,* 4(2), 158–170. DOI: *https://doi.org/10.2218/ijdc.v4i2.105*

Seidel, F. (2016). Documentary linguistics: A language philology of the 21st century. In P. K. Austin (Ed.), *Language Documentation and Description, 13*, 23–63. London: SOAS.

Tenopir, C., Allard, S., Sinha, P., Pollock, D., Newman, J., Dalton, E., Frame, M., & **Baird, L.** (2016). Data Management Education from the Perspective of Science Educators. *International Journal of Digital Curation, 11*(1), 232–251. DOI: *https://doi.org/10.2218/ijdc.v11i1.389*

Thessen, A. E., Woodburn, M., Koureas, D., Paul, D., Conlon, M., Shorthouse, D. P., & **Ramdeen, S.** (2019). Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. *Data Science Journal, 18* , 54. DOI: *https://doi.org/10.5334/dsj-2019-054*

Wasson, C., Holton, G., & **Roth, H. S.** (2016). Bringing User-Centered Design to the Field of Language Archives. *Language Documentation & Conservation, 10* , 641–681. Retrieved from *http://hdl.handle.net/10125/24721*

Weber, T. (2016). *Kraasna – the state of documentation and description of an extinct South Estonian dialect*. Bachelor thesis submitted at the Institut für Finnougristik/Uralistik, Ludwig-Maximilians-Universität München. Retrieved from *https://kraasna.wordpress.com/*

Weber, T. (2019). Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics? In M. Eskevich, G. de Melo, C. Fäth, J. P. McCrae, P. Buitelaar, C. Chiarcos, B. Klimek, & M. Dojchinovski (Eds.), *2nd Conference on Language, Data and Knowledge (LDK 2019), 70*, 26:1–26:8. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. DOI: *https://doi.org/10.4230/OASIcs.LDK.2019.26*

Weber, T. (2020a). Metadata Inheritance: New Research Paper, New Data, New Metadata? In A. Mannocci (Ed.), *Reframing Research Workshop Accepted Papers. Zenodo*. DOI: *https://doi.org/10.5281/zenodo.4155362*

Weber, T. (2020b). A Philological Perspective on Meta-scientific Knowledge Graphs. In L. Bellatreche, M. Bieliková, O. Boussaïd, B. Catania, J. Darmont, E. Demidova, F. Duchateau, M. Hall, T. Merčun, B. Novikov, C. Papatheodorou, T. Risse, O. Romero, L. Sautot, G. Talens, R. Wrembel, & M. Žumer (Eds.), *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (pp. 226–233). Cham: Springer International Publishing. DOI: *https://doi.org/10.1007/978-3-030-55814-7*

Weber, T. (2021a). Consultant Identity in Historical Language Data: Anthroponyms as a Tool or as an Obstacle? In A. Choleva-Dimitrova, M. Vlahova-Angelova, & N. Dancheva (Eds.), *Proceedings of the International Onomastic Conference "Anthroponyms and Anthroponymic Researches in the Beginning of 21st Century"* (pp. 165–175). Sofia: Bulgarian Academy of Sciences.

Weber, T. (2021b). Curation as a distinct academic activity – a perspective from working with legacy materials. *Poster presented at the 7th International Conference on Language Documentation and Conservation (ICLDC)*. Retrieved from http://hdl.handle.net/10125/74501

Weber, T. (2021c). *A linguistic analysis of Heikki Ojansuu's phonograph recordings of Kraasna.* (in press)

Weber, T. (2021d). Mind the gap: Language data, their producers, and the scientific process. In D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, & B. Heinisch (Eds.), *3rd Conference on Language, Data and Knowledge (LDK 2021), 93*, 6:1–6:9. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum für Informatik. DOI: https://doi.org/10.4230/OASIcs.LDK.2021.6

Weber, T. (2021e). *Philology in the folklore archive: Interpreting past documentation of the Kraasna dialect of Estonian.* (in press)

Weber, T., & **Bradley, J.** (2018). Exploring Finno-Ugric linguistics through solving IT problems. In D. Fišer & A. Pančur (Eds.), *Proceedings of the conference on Language Technologies & Digital Humanities* (pp. 248–253). Ljubljana: University of Ljubljana.

Winkler, E. (1994). *Salis-Livische Sprachmaterialien, 21*. Munich: LMU Munich.

Winkler, E. (1997). *Krewinisch. Zur Erschließung einer ausgestorbenen ostseefinnischen Sprache, 49*. Wiesbaden: Harrassowitz.

Woodbury, A. C. (2014). Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. In D. Nathan & P. K. Austin (Eds.), *Language Documentation and Description, vol 12: Special issue on language documentation and archiving* (pp. 19–36). London: SOAS.

TO CITE THIS ARTICLE:
Weber, T. (2021). The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers. *Journal of Open Humanities Data*, 7: 28, pp. 1–10. DOI: https://doi.org/10.5334/johd.51

Published: 30 November 2021

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.