Implementing and Promoting Data Citation for CLARIN Resources at FIN-CLARIN

Mietta Lennes
Department of Digital Humanities
University of Helsinki
Finland
mietta.lennes@helsinki.fi

Ute Dieckmann
Department of Digital Humanities
University of Helsinki
Finland
ute.dieckmann@helsinki.fi

Martin Matthiesen
CSC - IT Center for Science
Espoo, Finland
martin.matthiesen@csc.fi

Tommi Jauhiainen
Department of Digital Humanities
University of Helsinki
Finland
tommi.jauhiainen@helsinki.fi

Jussi Piitulainen
Department of Digital Humanities
University of Helsinki
Finland
jussi.piitulainen@helsinki.fi

Krister Lindén
Department of Digital Humanities
University of Helsinki
Finland
krister.linden@helsinki.fi

Abstract

Citation instructions are provided for all corpora in the Language Bank of Finland to promote good reference practices in the use of research data. By utilizing persistent identifiers (PIDs) and curated metadata, uniform citations can be constructed automatically, even for forthcoming resources. This work outlines how citation components such as authors, publication year, versions, and publisher are currently managed in the Language Bank of Finland, and discusses challenges related to metadata accuracy and interoperability. Data citation practices could be harmonized across CLARIN repositories by supporting data citation more directly via CLARIN compliant metadata.

1 Introduction

Today, the use and reuse of shared data is commonplace in scientific research. In order to ensure that studies can be verified and replicated, researchers should provide persistent references to their data. Data referencing is technically possible, but the practices of citing data are not yet well established in all fields. Moreover, additional effort may still be required from the researcher to gather the necessary details and to construct a data citation in the format requested by the publisher.

In this article, we describe how data set citation practices are encouraged and supported by the Language Bank of Finland (hereafter "the Language Bank")[1]. The Language Bank is a national research infrastructure (RI) service offered by the FIN-CLARIAH RI[2] through the University of Helsinki and CSC - IT Center for Science. It is partly funded by the Finnish Ministry of Education and Culture through various funding instruments. The Ministry and its affiliated organizations are committed to promoting the FAIR principles (data should be Findable, Accessible, Interoperable, and Reusable; see Wilkinson et

[1] https://www.kielipankki.fi/language-bank/

[2] https://www.kielipankki.fi/organization/fin-clariah/

al. (2016)). The aforementioned organizations are also strongly committed to develop Fairdata services together and to support the curation and storage of research data.

2 The current situation of data citation

Data repositories, including the Language Bank, should support researchers by systematically offering citation information for the data sets distributed by them, as recommended in the EU data citation guide (Publications Office of the European Union, 2022). Various guidelines have been published over the years, including those by Conzett and De Smedt (2022), the Tromsø recommendations for citation of research data in linguistics (Andreassen, Berez-Kroeker, Collister, Conzett, Cox, De Smedt, McDonnell, & Research Data Alliance Linguistic Data Interest Group, 2019), and the recommendations by CESSDA (Bornatici, Jernung, Alaterä, Tveit Sandberg, Strand, Štebe, & Trtíková, 2025)[3]. The recently published CLARIN Data Citation Guidelines (Matthiesen & Lenardič, 2025) reflect the aforementioned recommendations and are also supported by the Language Bank.

In the past, it has been difficult to cite data properly. Without well-established data citation conventions and persistent identifiers, it can be challenging to determine whether two similar-looking data descriptions appearing in two different contexts would in fact involve the very same data. Without trustworthy repositories and stable download locations, it is difficult to estimate whether a given data set would still remain accessible some years later. As a side-effect, researchers have also been inclined to make copies of all the potentially relevant data they were able to get their hands on.

For a long time, researchers have worked around these limitations by citing the papers describing the data, rather than citing the data sets directly. For derived data sets and data collections, the paper describing the original data set is often cited. However, unlike published papers, digital data sets may change over time for various reasons, and thus it is not possible to rely on research papers alone as the ultimate documents of the research data.

For example, a text corpus created before 1990 would not originally have been encoded in Unicode. However, in the research paper describing the corpus, ISO 8859-1 might have been stated as the character encoding of the texts. Given that the data was still in use a few years later, it would undoubtedly have been converted to Unicode for technical reasons, but the details of the conversion might not have been mentioned in any new scientific paper. Previous research papers can be cited by newer ones, but the original paper would not provide information about the changes that take place at a later point. In case a metadata record is made available, curated, and used systematically as a reference from the beginning, researchers will have consistent access to information on the original data set, including the details of the change in character encoding. In case cross-references are maintained between the metadata records, the researchers will also be able to find any newer, derived, or otherwise related versions of the data (see Matthiesen & Dieckmann, 2019, for details).

Persistent references to data cannot be taken for granted, which can be a replicability issue. For instance, in the data availability survey by Jauhiainen, 2024, the references to data sets were investigated in research articles published between the years 2018 and 2023 on the topic of automatic date detection in texts. The relevant datasets were available and sufficiently described in only three out of 22 articles. Of these three, two data sets were in GitHub repositories without any proper data citation instructions. They were not properly cited in the corresponding articles, either, apart from the links to the repositories. The relevant data set was properly referenced in just one article (by Grabovoy et al., 2021, who had stored their data set in the Mendeley Data service offered by Elsevier[4]).

Habits are difficult to break. As long as supervisors and publishers consider it sufficient to cite scientific articles on research data, proper data citation remains optional. The authors who submit their research to major scientific publications and conferences are usually required to deposit any relevant new data sets to an appropriate repository and to cite the data sets used in their research, but in practice, many researchers do not follow the recommendations.

[3] See also the CESSDA Data Citation Guide, https://datacitation.cessda.eu/

[4] https://doi.org/10.17632/95xt9sngzc.

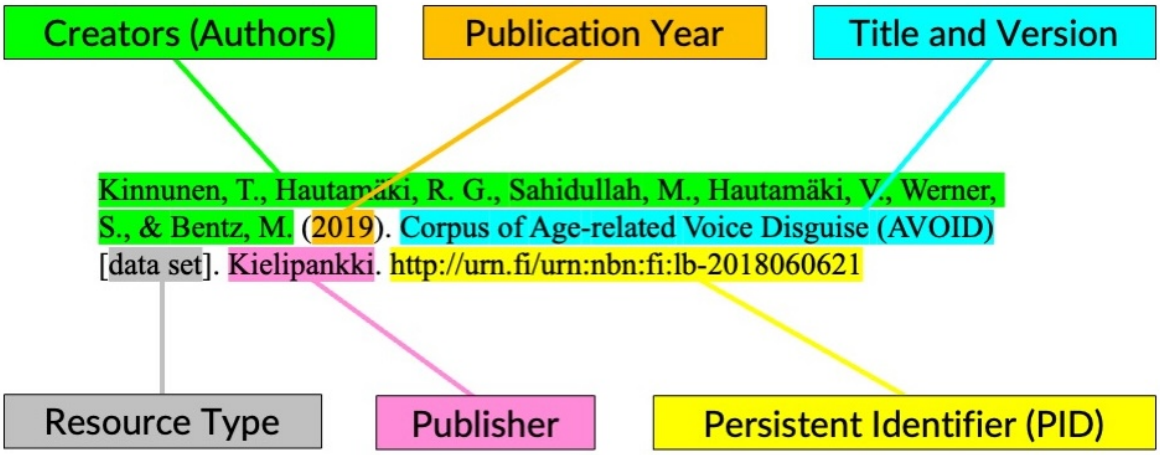

Figure 1: The structure of an example reference to a resource in the Language Bank of Finland.

The adoption of data citation may be slowed down due to the ways of measuring scientific impact. Publishing a data set may not yet count as scientific merit in a similar way to traditional publications.

By providing consistent data citation information, we cannot solve all the above-mentioned issues regarding data citation. Nevertheless, by making citation as easy as possible for researchers, we encourage the adoption of new practices. Let us now describe data citation at the Language Bank of Finland in more detail.

3 Citation Instructions at the Language Bank of Finland

The Language Bank maintains a public website (the Portal)[5] with information about the services of the Language Bank, including lists of the hosted language resources, e.g., corpora and tools. The list of corpora[6] is automatically generated from a separate resource database. The citation instructions for each corpus in the Language Bank are generated from the resource metadata stored in the database.

The instructions appear in the Portal via dynamic links using the resource PID as the citation key. The citation instructions for corpora consist of the following elements: **Creators (Authors)**, **Publication Year**, **Title and Version**, **Resource Type**, **Publisher**, and **PID** (see figure 1).

It is essential to provide a persistent identifier (PID) for each data set (Wittenburg, 2019). In the Language Bank, the PID used in the citation instruction always points to the metadata record of the resource in question. The metadata record provides the necessary details and can be updated when required, while the citation text can remain short and efficient. The citations work in a persistent and uniform way for all resources, regardless of the platform where the resource contents are stored.

In case a new resource is known to have been completed by a researcher and the deposition with the Language Bank is expected to take place in the near future, the Language Bank can provide citation instructions for the forthcoming resource, even before it is officially available.[7] The citation instruction can be reliably generated in the Portal as soon as the necessary details, including the PID, are included in the database.

[5] https://www.kielipankki.fi/language-bank/

[6] https://www.kielipankki.fi/corpora/

[7] "Data should be properly cited, even if not generally available, not fully available, or available only on restrictive terms." (Publications Office of the European Union, 2022)

License information is not included in the resource-specific citation instructions of the Language Bank, since licenses may be updated at a later point and this can be documented in the metadata. Moreover, the date of last access or retrieval has been excluded from the citation instructions. The individual resources that are centrally maintained by the Language Bank are typically not expected to change significantly after their publication date, and if they do change, the changes can be transparently logged in the metadata. Thus, the PID already ensures that a referenced dataset can be reused later, regardless of when it was accessed[8].

Accurate resource metadata are essential for the automatic generation of reference instructions. In this section, we describe why and how certain fields of information are included in the citation instructions of the Language Bank. We also point out some issues that researchers and repositories need to consider.

3.1 Creators (Authors)

According to the current practices in the Language Bank, the ordered list of authors or **creators** is set in the deposition agreement of the resource. When negotiating agreements, the Language Bank encourages depositors to list individual people as creators, as recommended by, e.g., Lenardič and de Maiti Tekavčič (2024), as this is good practice in terms of merit, provenance, and scientific replicability. However, in some cases, it can be appropriate to attribute organizations instead of or in addition to individuals. The depositors make the final decision on authorship.

For each resource in the Language Bank, information about the creators is stored in the database as the first and last names, listed in the order of appearance in the citation. For a more general solution, the list of authors/creators including their contact details, affiliations and the order of citation, should be stated in the metadata record.

Determining reference instructions for older datasets can be challenging, since resource metadata has not always been collected in a consistent way in the past, and the original contributors may no longer be available for consultation. For example, when reviewing the older resource metadata records of the Language Bank, it turned out that some pieces of information had quite frequently been recorded in the wrong places and some details were often missing. For some resources, the publisher had been erroneously mentioned as the creator of the resource, even when the publisher had not really contributed to collecting, organizing, formatting or enriching the data. When trying to track down the parties that should be attributed for specific resources, it is fortunate if the original deposition agreements and/or some relevant email archives are available. For instance the 'IPR holders' or 'Licensors', defined in some metadata records, were not necessarily the same parties that should be cited as the 'creators' or 'authors'. In some cases, where the source version is available under a public license requiring attribution, the attribution information might not be explicitly stated at the source. Thus, a lot of additional effort may be required in case the information about authorship is missing or incomplete and if there is no written agreement about the data.

3.2 Publication Year

For citation instructions, the publication year is interpreted as the year the corpus was published in the Language Bank (even if the collection was previously made available elsewhere), since this is the publication time of the specific resource edition over which the repository has control. The publication year is retrieved from the database and formatted as a four-digit number. For resources that have not yet been published, the text "forthcoming" is shown instead.

3.3 Title and Version

The title of the resource is agreed with the depositor. The same title is stored in the database and in the corresponding metadata record in English and Finnish. The resource title usually includes extensions describing the variant (e.g., Korp or Download[9]) and the version of the resource. Like books, data sets might be published again as new versions or editions. More recent versions can contain more data, or some annotation layers may have been added or updated, e.g. by using an improved annotation method.

[8]For a discussion on assigning new PIDs when datasets change over time, see Matthiesen and Dieckmann (2019).

[9]See https://www.kielipankki.fi/support/corpus-location/

3.4 Resource Type

The title of a corpus or a data set may sometimes be similar to the title of a published paper or research project. To avoid confusions, it is good practice to include information about the type of resource. Square brackets are often used to separate this element in the citation.

The resource type should be expressed in generally comprehensible terms that can be localized in many different languages. As we previously pointed out, excessive or potentially changing details should be avoided in data citations, since further information can be provided via the metadata record. Previously, the resource type *corpus* was used by default in the citation instructions provided by the Language Bank. However, there are many resources that cannot be categorized as 'corpora'. A common vocabulary for all types of resources has not been established. In addition, it is rather typical for language resources to include various combinations of different data types and structures, in which case selecting the resource type would not be straightforward. We therefore decided to use the more generic resource type *data set* (in Finnish, *aineisto*).

3.5 Publisher

In the citation instructions generated by the Language Bank, the publisher of every data set is the Language Bank, identified as "Kielipankki", since the Language Bank is responsible for maintaining the deposited data and for curating the metadata (see Publications Office of the European Union, 2022). Moreover, only Language Bank staff members can make changes to the metadata records.

3.6 Persistent Identifier (PID)

At the Language Bank, the PID of a resource points to the metadata record of the resource in question[10]. Consequently, it is possible to provide unambiguous citation instructions even before the resource has been officially published by the repository. This approach also allows the Language Bank to implement persistent tombstone pages after the data itself is no longer available, according to the FAIR principle A2[11]. In the citation instructions of the Language Bank, the PIDs are also made "actionable", i.e., clickable, as recommended in the data citation guide by the Publications Office of the European Union (2022). The link to the data access location is not included in the citation instructions. The metadata record provides the data access location in CMDI[12] compliant format for humans and machines.

The Language Bank uses URN-based PIDs supported by the National Library of Finland[13]. Handles are also supported as a backup system[14]. In line with Broeder et al. (2009), our PIDs do not contain semantic information. For practical reasons, new PIDs are constructed from dates and running numbers, but even the minting date can not be reliably derived from the identifier.

The PID of the resource points to the metadata record that provides information on how to access and use the resource (including license restrictions, technical documentation or guidelines for the user, information about annotation anomalies, etc.). The metadata can be updated by the Language Bank, e.g., to fix typos or to add new or previously missing information. Minor updates of the content of the resource itself do not result in a new metadata record and a new PID as long as the content is expected to be compatible with the resource version prior to the update. A minor update can be described just by editing the metadata record and the version number after the resource title that appears in the citation. For a more in-depth discussion, see Matthiesen and Dieckmann (2019, section 6).

3.7 Formatting the Output

The citation instructions are offered via the list of corpora in the Language Bank Portal in two languages, English and Finnish. Links to the instructions for individual resources can also be found in the metadata records. The links can be manually constructed if the resource PID is known[15].

[10] For a discussion of the special status of a "citable PID", see Matthiesen and Dieckmann (2019, section 7).

[11] https://www.go-fair.org/fair-principles/a2-metadata-accessible-even-data-no-longer-available/

[12] https://www.clarin.eu/content/cmdi-component-metadata-infrastructure

[13] https://urn.fi/URN:NBN:fi-fe2024051430651

[14] The PIDs can be resolved if one of the systems is unavailable: https://urn.fi/urn:nbn:fi:lb-201710212#Persistent_identifiers

[15] For example, https://www.kielipankki.fi/viittaus/?key=urn:nbn:fi:lb-2024051601&lang=en.

The plain text citation instructions are constructed roughly according to the APA style[16] with one exception: the version information is included in the title. The general structure is in line with other CLARIN centres, e.g., CLARIN DSpace at LINDAT/CLARIAH-CZ[17]. Unlike the Language Bank, CLARIN DSpace does not use the resource type attribute (e.g., "[data set]").

In addition to the citation instructions in plain text, the BibTeX and Zotero versions are also provided. The most commonly used bibliographic styles for BibTeX do not include an entry type for data sets especially. The reference text is generated at the Language Bank with the BibTeX entry type `@misc`, which is also used by CLARIN DSpace and produces reasonable-looking reference items. For Zotero, the format instruction is rendered in BibTeX syntax as `@techreport`, since this produced the best result (first tested in 2015, still holds true in 2025). Finally, an additional link is generated to search for references to the resource from Google Scholar.

3.8 Data Provenance and Relations

Language resources often consist of texts or other materials that have been published elsewhere. Following the recommendations of Andreassen et al. (2019, chapter 1.1), the Language Bank offers citation instructions for entire data sets only. This is where the persistent identifier to the metadata record is useful and efficient. For instance, in case a large corpus of newspaper articles is used as a research data set, a reference can be provided to the entire corpus, without listing the bibliographic details of all the individual texts and their authors. If required, the original sources should be findable via the metadata.

A corpus may consist of several smaller corpora, each of which may have been created by different people. If using an individual text or a specific part of the larger data set, researchers can be instructed to add further references to the parts of the resource. In order to make this possible, sufficient information about the subcorpora should be readily available, e.g., via metadata relations or via external documentation. The necessary details should preferably be provided by the creator of the collection at the time of depositing or publishing the resource.

For instance, the *Uralic UD* resource group[18], available via the Language Bank, includes several versions of a collection of treebanks of Uralic languages from the Universal Dependencies project repositories. The Uralic UD corpus as a whole is provided, maintained and versioned by the Language Bank, but each original treebank has a different list of contributors. Therefore, the users of the resource are instructed to cite the entire corpus (this ensures persistent access), but it is also recommended to include additional references to the individual subcorpora if appropriate (for precise attribution). The details regarding the subcorpora were found in the source repositories on GitHub, and they are provided on the resource group page of the Language Bank, to which a link is available in the metadata records of the individual resource versions.

Resources can thus appear in several versions that may differ not only in the amount of primary data (e.g., original texts or speech recordings that accumulate from one version to the next), but also in the types and amounts of annotation and analysis files. Sometimes, the primary data set has been created by different people than the annotations that were added later. Since the contributions may be independent of each other, it is important not to include the creators of the original data set in the same reference instruction with the creators of the additional annotations without an agreement of all parties.

As an example, *ScotsCorr, the Helsinki Corpus of Scottish Correspondence (1540–1750)*[19] was originally published in the Language Bank in 2017. The suggested citation is as follows:

> Meurman-Solin, A., & Research Unit for the Study of Variation, Contacts and Change in English (VARIENG), University of Helsinki (2017). *Helsinki Corpus of Scottish Correspondence (1540-1750)* [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-201411071

However, new annotations were created later in another, unrelated project and published in the Language Bank as a different data set, the *Parsed Corpus of Scottish Correspondence (pcsc-src)*, included in

[16] https://apastyle.apa.org/style-grammar-guidelines/references/examples/data-set-references
[17] https://doi.org/10.17616/R30G6W
[18] Uralic UD resource group: http://urn.fi/urn:nbn:fi:lb-2022061003
[19] ScotsCorr resource group: http://urn.fi/urn:nbn:fi:lb-202104191

the same resource group. This data set has a citation instruction of its own:

> Gotthard, L. (2025). *The Parsed Corpus of Scottish Correspondence, source* [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2024070601

In addition, the users of the parsed corpus version are instructed to include a separate reference to the original corpus version as appropriate.

As seen from the examples above, the "web" of the different versions, variants, and combinations of a data set can become quite complex over time. It is likely that not all the relevant changes and possibilities of the intermediate data sets end up being documented in a research paper. However, if it were possible to automatically discover the contributions of the authors and creators of related resources on the basis of the metadata, and if this information could be made available in a uniform and transparent manner, researchers would not be obliged to keep citing both the datasets and the respective papers.

In the Language Bank, the relations between resources are described by using the vocabulary defined by DataCite (see DataCite Metadata Working Group, 2024). However, although the DataCite relations provide a means of expressing provenance information in a machine-readable way, the relations are currently not displayed or utilized by metadata services. Therefore, the landing page where the PID of a resource points may not be able to offer all the information included in the metadata. CLARIN does have a standardized way of accessing the underlying CMDI metadata[20], but this standard is not universal. In order to interpret the relations correctly, the metadata crawler would need to examine the schema.

4 Discussion

The metadata records of the Language Bank already contain almost all the information needed to create the citation instructions. Ideally, the citations could be generated directly from the CMDI metadata and provided to the users via the metadata catalogue. Unfortunately, the order of authors or creators is currently not encodable in the metadata for cases where both authors and institutions should be mentioned. In addition, reference instructions may also need to be generated for forthcoming resources, which are not yet officially displayed via metadata services. Therefore, the information required for the citation instructions is currently retrieved from a separate resource database at the Language Bank, as described above.

There are some risks involved in generating reference instructions automatically from openly distributed metadata. Similarly to all descriptive metadata in CLARIN, the metadata of the resources in the Language Bank of Finland are not licensed, and the records are publicly available via the OAI-PMH[21] interface. Other metadata catalogues, such as the *European Language Grid* (ELG), can harvest and republish the metadata. The metadata may sometimes need to be re-processed to match the needs of the service in question. The external catalogue can then also generate reference instructions in a format of their own. Consequently, the citation instructions may look very different to end users.

Consider this citation of a Language Bank corpus, generated by ELG:

> The Suomi24 Sentences Corpus 2001-2017, Korp version 1.1 (2020, January 01). Version 1.0.0 (automatically assigned). [Dataset (Text corpus)]. Source: European Language Grid. https://live.european-language-grid.eu/catalogue/corpus/11759

Then compare the above reference to the recommended citation of the dataset, provided by the Language Bank:

> City Digital Group (2025). The Suomi24 Sentences Corpus 2001-2017, Korp version 1.3 [data set]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2020021803

Ultimately, both references point to the same data. The version numbers differ between the two records, since the metadata provided via ELG was manually imported and it is not synchronized with

[20] see Wittenburg et al. (2024, Chapter 7)

[21] https://www.kielipankki.fi/language-bank/oai-pmh/

the master metadata record that is maintained at the Language Bank. Moreover, during the import, an additional version number was generated by ELG on top of the existing version number, and the link to the catalogue item promoted by ELG has completely replaced the original persistent identifier.

By allowing the metadata to be added into various catalogues, the data sets can gain more visibility. The downside is that the metadata may appear ”out of context”, as shown by the ELG example. To avoid the risk of disseminating contradictory citation instructions for the same data set, new metadata records should always contain a clear link to the original metadata from which they are derived. This metadata should be regularly checked for updates and treated as master metadata. The master metadata should be curated at a single location in agreement with the creators and providers of the original data set. There are also ongoing efforts aiming to promote the propagation of the minimal metadata. For instance, the ”FAIR Signposting” approach to metadata publishing design, endorsed by the EOSC Association, can offer a transparent, compliant, and straightforward mechanism for helping automated agents navigate through metadata spaces to locate the essential FAIR elements: the globally unique identifier (GUID), the data records, and the corresponding metadata records (Wilkinson et al., 2024).

If a data set includes or is a version of other datasets, the authors/creators of the parts or older versions should also be attributed in an appropriate way. All the creators of the subsets should not be listed in the citation instructions of a larger collection, since the list of ”authors” would tend to grow longer and longer, and the individual contributors of the subsets are probably not to be attributed for the entire collection. It is therefore essential that researchers are able to find information about the relations between data sets in the metadata record.

Even a well-formed data citation can only fulfill its purpose if the dataset or at least its metadata remain accessible. It is not enough to provide the data online and mechanically assign a persistent identifier to it, unless the repository is supported and the identifier is curated in the long term. Thus, some infrastructure is required before scientific data citation can fully work.

5 Conclusions and Future Prospects

Citation indices may still reward researchers for their traditional publications rather than their published data sets. However, research data should not be cited by referencing related articles or books alone. A published article cannot be modified after publication and can easily become outdated, whereas the metadata record of the data set can be updated when necessary. Well-curated metadata can maintain and even add to the scientific impact of a data set after the data has been made available. The provenance of the data set can be made more transparent, if the relations of a data set with other resources are described systematically. By ensuring that all the metadata required for citation are available systematically and in an interoperable and user-friendly way, CLARIN can support the adoption of FAIR-compliant data citation practices.

The system for generating citation instructions at the Language Bank of Finland was developed in 2015. Small modifications have been implemented over the years, and the current citation instructions now also comply with the recently published CLARIN Data Citation Guidelines (Matthiesen & Lenardič, 2025). At the Language Bank, the information for all components in the citations is currently retrieved from the internal resource database. This allows the Language Bank to generate citations even for forthcoming resources, since the database includes information about the publication status of each resource. The database also supports the order of author/creator names, unlike the current CMDI metadata profile used by the Language Bank. Ideally, however, the details required for generating the citation instructions should be included in the public metadata records of all data sets in CLARIN.

Metadata records should be widely accessible by default. It is often recommended that metadata be made available in the public domain or under a public license, with practically no limitations to what others can and cannot do with the information. However, if metadata records are copied from one platform to another, converted into a different metadata schema and re-published in isolation from the originals, with no method for retrieving potential updates, the copied metadata can easily become distorted. This is where CLARIN and EOSC could help by developing policies and best practices for displaying and using metadata outside of its original context.

The CMDI components supporting consistent citations should be designed by the CLARIN community and deployed over time by all repositories. This would allow for the development of central citation services, whereas the local consortia could focus on providing high-quality metadata. At the Language Bank, we are working on augmenting our CMDI schema to achieve this goal.

References

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

Bornatici, C., Jernung, A., Alaterä, T. J., Tveit Sandberg, L., Strand, K., Štebe, J., & Trtíková, I. (2025, March). CESSDA recommendations on data citation: Practical recommendations for key stakeholders. *Zenodo*. https://doi.org/10.5281/zenodo.15043854

Broeder, D., Dreyer, M., Kemps-Snijders, M., Witt, A., Kupietz, M., & Wittenburg, P. (Eds.). (2009). *Persistent and unique identifiers*. https://office.clarin.eu/pp/D2R-2b.pdf

Conzett, P., & De Smedt, K. (2022). Guidance for citing linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 143–155). The MIT Press. https://doi.org/https://doi.org/10.7551/mitpress/12200.003.0015

DataCite Metadata Working Group. (2024). *DataCite metadata schema documentation for the publication and citation of research data and other research outputs* (Version 4.6). DataCite e.V. https://doi.org/10.14454/mzv1-5b55

Grabovoy, A., Bakhteev, O., & Chekhovich, Y. (2021). The automatic approach for scientific papers dating. *2021 Ivannikov Ispras Open Conf. (ISPRAS)*, 107–113. https://doi.org/10.1109/ISPRAS53967.2021.00020

Jauhiainen, T. (2024). Data availability and evaluation reproducibility for automatic date detection in texts, a survey. *Digital Humanities in the Nordic and Baltic Countries Publications*, *6*(1). https://doi.org/10.5617/dhnbpub.11511

Lenardič, J., & de Maiti Tekavčič, K. P. (2024). The citation of language resource technologies in CLARIN. In V. Vandeghinste & T. Kontino (Eds.), *Proceedings of the CLARIN Annual Conference, 15 – 17 October 2024, Barcelona, Spain* (pp. 46–50). CLARIN ERIC. https://www.clarin.eu/sites/default/files/CLARIN2024_ConferenceProceedings_final.pdf

Matthiesen, M., & Dieckmann, U. (2019). A PID is a promise — Versioning with persistent identifiers. In I. Skadina & M. Eskevich (Eds.), *Selected papers from the CLARIN Annual Conference 2018* (pp. 103–112, Vol. 159). CLARIN ERIC. https://doi.org/10.3384/ecp159

Matthiesen, M., & Lenardič, J. (2025). CLARIN Data Citation Guidelines. https://doi.org/10.34733/DOC-189

Publications Office of the European Union. (2022). *Data citation – A guide to best practice*. https://data.europa.eu/doi/10.2830/59387

Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(160018). https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Sansone, S.-A., Grootveld, M., Dennis, R., Hecker, D., Huber, R., Soiland-Reyes, S., Van de Sompel, H., Czerniak, A., Thurston, M., Lister, A., & Gaignard, A. (2024). Report on FAIR Signposting and its uptake by the community. *Zenodo*. https://doi.org/10.5281/zenodo.10490289

Wittenburg, P. (2019). From Persistent Identifiers to Digital Objects to Make Data Science More Efficient. *Data Intelligence*, *1*(1), 6–21. https://doi.org/10.1162/dint_a_00004

Wittenburg, P., Van Uytvanck, D., Zastrow, T., Straňák, P., Broeder, D., Schiel, F., Boehlke, V., Reichel, U., & Offersgaard, L. (2024). *CLARIN B Centre checklist* (tech. rep. No. Version 7.4.1). CLARIN.