LINGUISTICS

Data Sharing in Linguistics

There are many different types of data used in the field of Linguistics, including: documentary data (e.g., text, audio, video), grammaticality judgements, instrumental data (e.g., eye tracking, EEG measurements, spectrograms), experimental data, derived data (e.g., transcriptions, annotations, syntactic treebanks), metadata, lexical data (e.g., dictionaries), language catalogues, computational data, and interview data. Much of this data is shared through domain-specific repositories which support the use of metadata, citation formats, and licences specific to the field (e.g., Tromsø Recommendations, CLARIN licences), but the data may also be shared in national or general repositories. Some linguistic data is shared via other types of websites (e.g. project websites, personal websites) or in article supplementary files. The Open Handbook of Linguistic Data Management (MIT Press, 2022) provides a valuable overview of good practice in collecting, managing, archiving, sharing, and citing data in the field.

Where can I find resources and tools for...

Methods and Documentation

- FAIRsharing
- The Open Handbook of Linguistic Data Management
- The Austin Principles of Data Citation in Linguistics

- CLARIN Licences
- Tromsø Recommendations for Citation of Research Data in Linguistics

Learning about Languages

- Ethnologue
- Glottolog
- Arctic Indigenous Peoples languages and revitalization

Data Management and Processing

- SSH Open Marketplace
- SIL
- ELAN
- Praat

Finding and Depositing Data

- Virtual Language Observatory
- CLARIN
- LDC Catalog
- CESSDA DC Data Catalogue
- The Language Archive

EOSC Portal

The EOSC Portal is a gateway to many of the innovative services, tools, publications and data listed here, and it is constantly growing with additions from the community of Linguistics researchers and research-supporting organisations. Do you have a resource that you want to share with others? Consider onboarding into EOSC.

▶

- Language Archive Cologne
- Tromsø Repository of Language and Linguistics
- Child Language Data Exchange System (CHILDES)
- Endangered Languages Archive (ELAR)
- Tekstlaboratoriet
- Ortolang
- Computation Resource for South Asian Languages (CoRSAL)
- Pangloss Collection
- Kaipuleohone Language Archive
- Native American Languages Collections (NALC)
- Paradisec
- Archive of the Indigenous Languages of Latin America (AILLA)
- Center for Native American and Indigenous Research (CNAIR), American Philosophical Society
- California Language Archive (CLA)
- Repository and Workspace for Austroasiatic Intangible Heritage (RWAAI)

Community and Professional Supports

- CLARIN
- Collaboratory for Indigenous Data Governance
- DARIAH
- Digital Endangered Languages and Musics Archives Network (DELAMAN)
- Global Indigenous Data Alliance (GIDA)
- Linguistics Data Consortium (LDC)
- Linguistic Society of America
- Huma-Num

Learn more about the Research Data Alliance (RDA)

- Linguistics Data IG
- International Indigenous Data Sovereignty IG
- Psychological Data IG
- Sensitive Data IG
- Social Science Research Data IG

How to do FAIR and Open Science

- What is FAIR?
- FAIR Community Support
- What is the goal of Open Science?

What are the challenges for linguistic data in Open Science?

"Linguistics has a history of developing data practices in relative isolation by subfield, lab, and researcher, but as a discipline, we could strongly benefit from broader discussions about the role of data and best practices in research data management. Also, older, unpublished linguistic data often contains personal information, and it can be difficult to make this data openly available so long after the point of collection. Historical linguistics data is also often sensitive in nature, especially when it concerns the words and languages of historically marginalized peoples.

The RDA Linguistics Interest Group was formed to provide guidance about how to make linguistic research transparent and reproducible all while taking into account any relevant ethical and legal issues. Outputs of the Interest Group should be relevant to all subfields of the discipline and all types of linguistic data, while simultaneously acknowledging the diversity of sources, and should as such facilitate the researchers' decisions on whether and how to archive, cite and share their data."

How can EOSC help researchers working with linguistic data?

"There is infrastructure support for the field, particularly in terms of discipline-specific repositories and developments for metadata, as well as initiatives that are providing global support and consensus around aspects of Open Science like data citation and attribution. EOSC will improve repository support for FAIR data, and by creating a European network, it can also provide better connections across infrastructures available globally."

— **Helene N. Andreassen**
RDA/EOSC Future Domain Ambassador for Linguistics