

# Data citation and metadata standards in linguistics

## To our participants

- Feel free to add thoughts and comments in our collaborative notes document:  
<http://bit.ly/2GKI6xa>
- Help the organisers with their tracking report, and check in to the session via  
the [RDA11 Guidebook](#)
- All documents will remain available at our [RDA11 session page](#)
- Contact person for this meeting: [helene.n.andreassen@uit.no](mailto:helene.n.andreassen@uit.no)

# Data citation and metadata standards in linguistics

**Linguistics Data Interest Group Meeting @RDA Plenary 11, Berlin  
Thursday March 22, 2018, 13.30-15.00 (room C04)**

Helene N. Andreassen (UiT The Arctic University of Norway)

Philipp Konzett (UiT The Arctic University of Norway)

Koenraad De Smedt (University of Bergen & CLARIN)

Andrea Berez-Kroeker (University of Hawai'i at Manoa)

Lauren Gawne (La Trobe University)

# Meeting agenda

13.30-13.40

Welcome & introduction to RDA Linguistics Data Interest Group (LDIG):  
Objectives, outputs and directions.

13.40-13.50

Rationale for citation and metadata standards in linguistics.

13.50-14.15

Synthesis over citation and metadata standards recommended/in use by main actors in the field & needs and challenges revealed by the extended LDIG community (asynchronous meeting).

14.15-14.45

Feedback from the floor and discussion.

14.45 -15.00

How to proceed next: Development of a plan for forming a working group.

# Linguistics Data Interest Group (LDIG)

- Endorsed in 2017.
- Co-chaired by Helene N. Andreassen (UiT), Andrea Berez-Kroeker (Hawai'i at Manoa). Lauren Gawne (La Trobe).

## Main objectives

1. Development and adoption of common principles and guidelines for data citation and attribution (researchers, professional organizations, academic publishers, archives).
2. Education and outreach efforts (practical training and awareness of principles/sociological change).
3. Greater attribution of linguistic data set preparation within the linguistics profession (value “data work” as scholarly output at all career stages).

# Linguistics Data Interest Group (LDIG)

## Output 1 and directions

### *Austin Principles of Data citation in Linguistics*

- set of guidelines that enable linguists to make informed decisions regarding the accessibility and transparency of their research data.
- based on the FORCE11 Joint Declaration of Data Citation Principles.
- available at [linguisticsdatacitation.org](http://linguisticsdatacitation.org) (Berez-Kroeker et al., 2017a).

### *Implementing the Austin principles*

- integration into the [Unified Style Sheet for Linguistic Journals](#) (Joseph et al., 2007) and/or the [Generic Style Rules](#) (Haspelmath, 2014).
- adoption of the style sheets by publishers, archives, organisations and individuals.
- adoption of recommended metadata standards by archives to enable data citation in line with the style sheets.

# An editor's thoughts on data collection, use and reuse in linguistics

I have come to think of the **accuracy of data as a serious problem** for all linguistics journals, and for the field at large. I do not see any way to deal with it unless we constantly remind ourselves and our students of the importance of working to maintain the quality of the data that we use (p. 409)

[Keren Rice] notes that theories are often based on a **misunderstanding** of the primary sources or on an **inappropriately restricted** subset of the data available in the primary sources (p. 412)

it is vital for all authors to ensure '**clarity and replicability of the chain of evidence**' so that it will be as easy as possible for other scholars 'to evaluate the solidity of the various steps in the chain, and then to replicate and extend the work the claim is based on, if they choose to' (Mark Liberman, via email, 1993) (p. 410)

(Thomason, 1994, bolding ours)

# Rationale for citation and metadata standards in linguistics

Position statements:

1. Berez-Kroeker et al. *Reproducible research in linguistics: A position statement on data citation and attribution in our field*. *Linguistics* 56:1, 2017 (<https://doi.org/10.1515/ling-2017-0032>)
2. Ted Pedersen: *Empiricism Is Not a Matter of Faith*. *Computational Linguistics* 34:3, 2008 — *The sad tale of the Ziggiebottom tagger*

Estimates in chemistry and medicine: only 20–25% of research can be replicated.

Cf. studies for linguistics which will be mentioned later.

# Rationale for citation and metadata standards in linguistics

CLARIN aims to overcome fragmentation and make data FAIR

F: PIDs, catalogued metadata (VLO)

A: Standard protocols (HTTP/SAML), vocabulary (OpenSkos)

I: Community-based standards for data, metadata (CMDI) and vocabulary

R: Documentation, CLARIN license categories, provenance



# Rationale for citation and metadata standards in linguistics

## **Q1: What are the main purposes of metadata in the field of language and linguistics?**

- no part of language exists free of its context. Metadata provide that context and turn language into data available for analysis.
- metadata contributes in relating linguistics to other fields of social sciences.
  - descriptive metadata facilitate identification and discovery of linguistic resources.
  - structural metadata systematise and connect research data.
  - administrative metadata inform potential reusers about e.g. methods, access and authorship.
- linguistics is a vast field. Metadata should allow searching for data on the language of interest, across subfields.

# Rationale for citation and metadata standards in linguistics

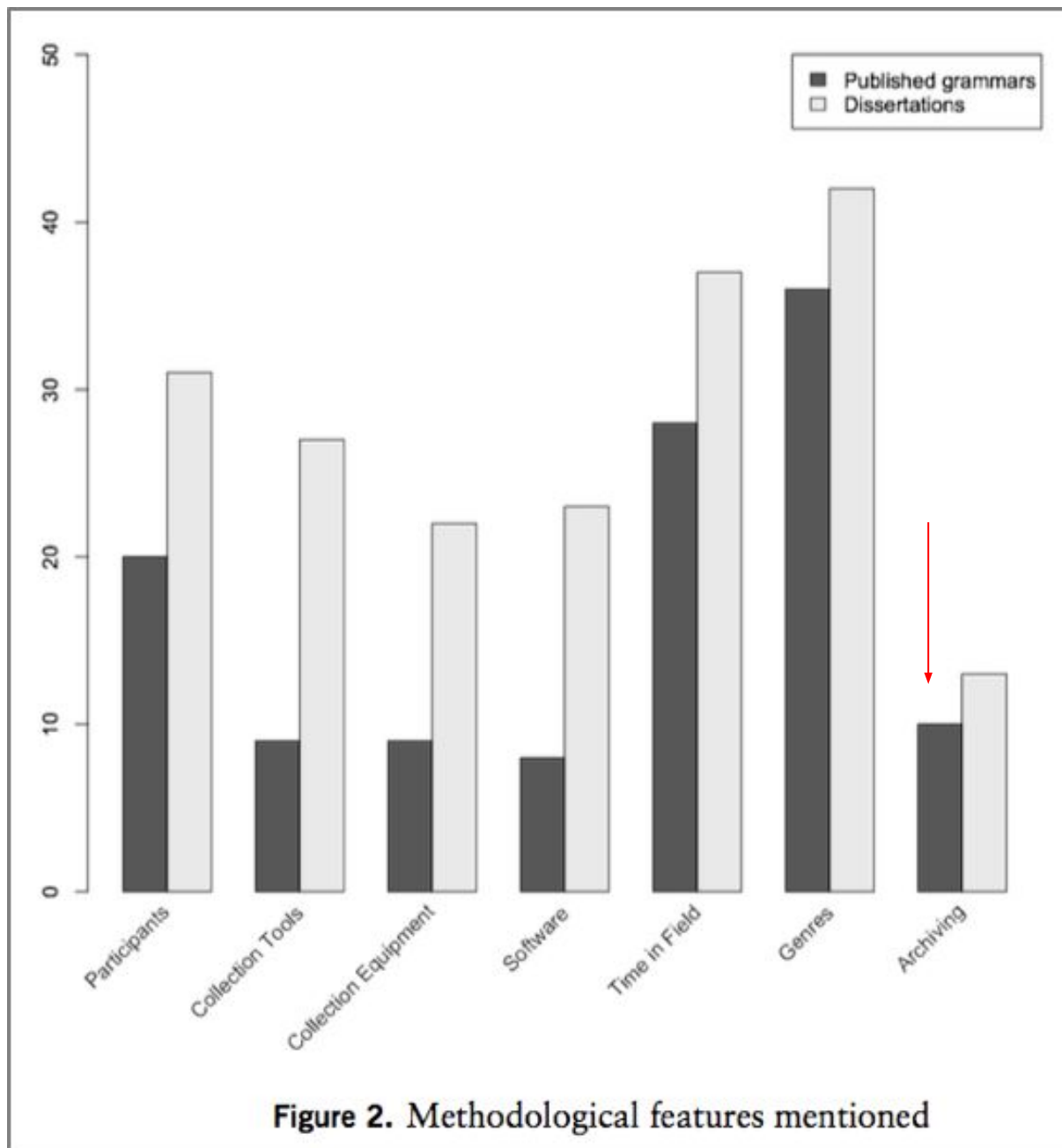
**Q6: In which respects does reproducibility put extra demands on citation practices and metadata?**

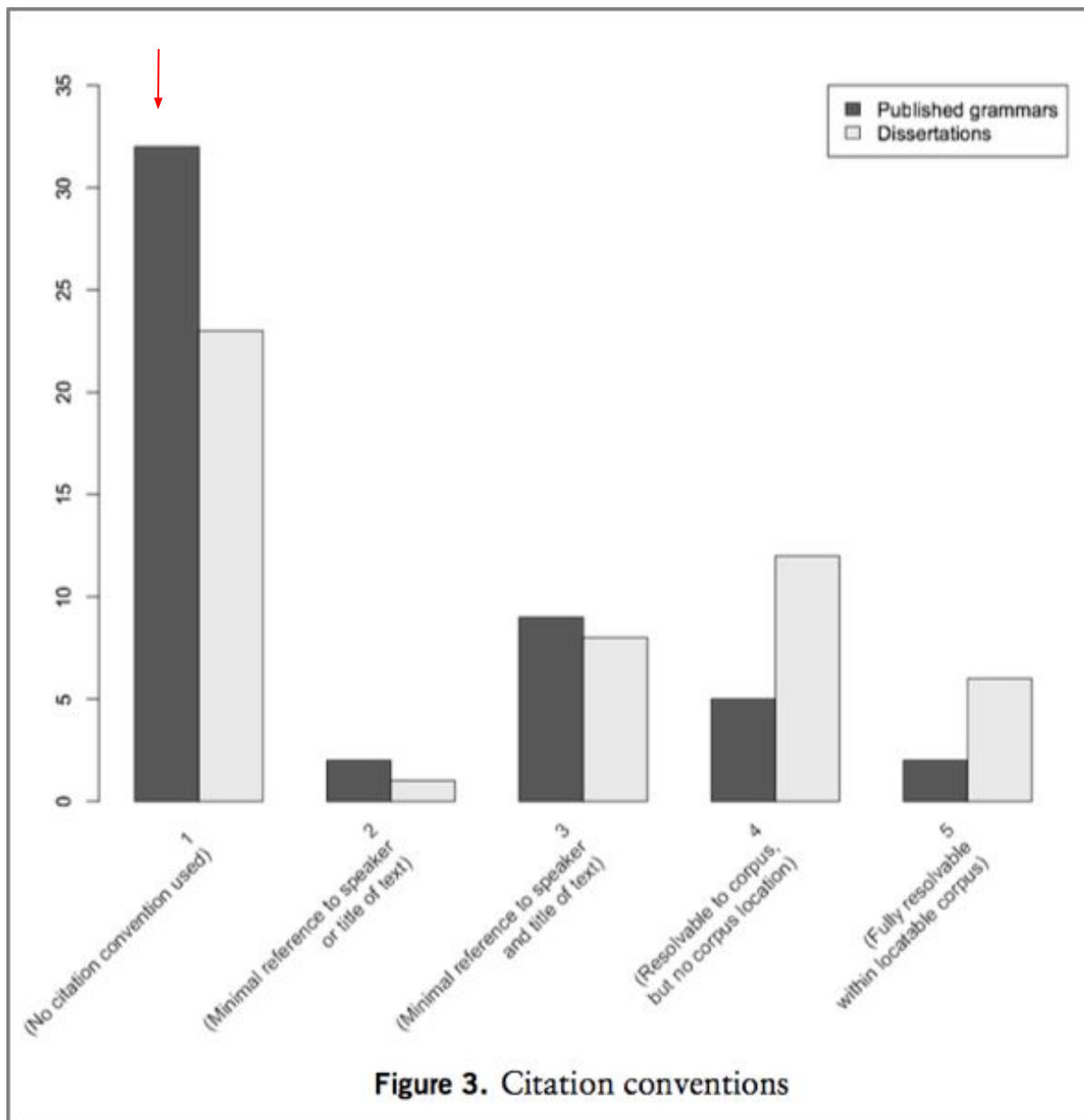
- citation is useless if it doesn't point to anything accessible, be it research data themselves, or open metadata for (temporarily) restricted research data.
- citing data directly should be the norm. PID and version control guide the reader to the correct dataset.
- granularity: it may be useful to refer to specific (arbitrary) parts of datasets, e.g. parts of a corpus.

# Citation standards for linguistic resources

A survey of 100 descriptive grammars (2003-2012) reveals that few authors make collection method and data source explicit (Gawne et al., 2017).

- methods well reported on time of collection, type of data, participants.
- methods less reported on tools, equipment, software, archive.



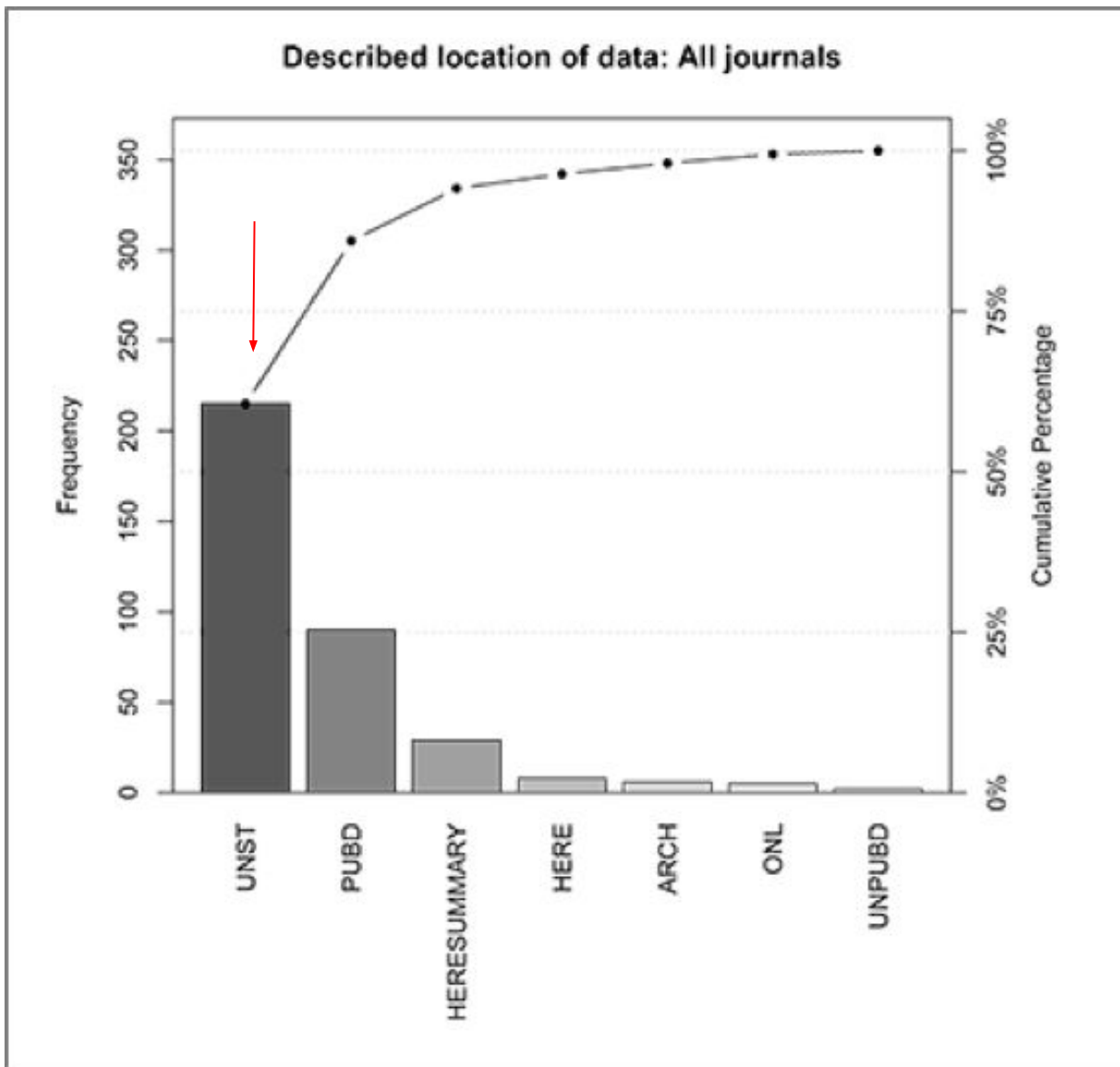


– most do not cite numbered examples.

# Citation standards for linguistic resources

A survey of 270 articles from 9 international journals reveals a similarly poor picture (Berez-Kroeker et al., 2017b)

- sub-disciplines excel in various facets of research transparency
  - phonetics and instruments
  - L2A and experimental method
  - documentary linguists information on informants and context of data collection



– more than half does not state placement of data

# Citation standards for linguistic resources

## Q5: What should editors/publishers do in order to improve citation practices for linguistic resources?

- Require publication of dataset (as supplement or in an open archive).
- Guide authors toward recommended archives to facilitate publication.
- Include *dataset* as a category in the style guide; in-text and bibliography.
- Collaborate on a uniform style guide for data citations, to be integrated into the Unified Style Sheet or the Generic Style Rules. The reference format of datasets should approximate conventional contributions to make it clear how they fit into the reference list.
- Reject papers without proper data citation; this includes citing in-text examples.



# Citation standards for linguistic resources

## **Q5: What should editors/publishers do in order to improve citation practices for linguistic resources? (cont.)**

- *Archives* should provide ready-to-use citations, including PIDs, formatted in major citation styles; “cite this”-buttons that interact with reference managers (which in turn need to contain a proper template for the document category *dataset*).
- *Archives* should provide guidelines for how to cite in-text examples.

# Citation standards for linguistic resources

**Q4: How would you like to see bibliographic references and in-text citations of linguistic resources (including complete datasets, excerpts, search results, examples, etc.) in the papers that you read?**

- Preciseness and meticulousness of citation is important as it helps research seekers find the most appropriate references for their research or manuscripts.
  
- Concerning granularity: What should the in-text data examples point to?
  - complete dataset = normal citation incl. PID
  - excerpt = PID part identifier or normal citation (incl. PID) + defining the subset
  
- Concerning crediting: Citing archives is important as these depend on funding!

# Metadata standards for linguistic resources

## What do we need to improve citation practices for linguistic resources?

- ❑ Good standards and requirements

## How is citation related to metadata?

- ❑ Good citation standards and practices depend on good metadata.
- ❑ But metadata are necessary for other reasons than citation, e.g. discoverability.

## Specific needs for metadata about linguistic resources?

- ❑ Q2: Are there special needs or demands for the metadata description, citation and attribution of linguistic resources compared to other fields of research?
- ❑ Q3: Do different types of linguistic resources need different elements of metadata?
- ❑ Also some answers to other questions

# Metadata standards for linguistic resources

## Existing standards

- ❑ General/basic standards for all(?) types of resources:  
Dublin Core, MARC, MODS, ...
- ❑ Specific standards for different types of linguistic resources:  
OLAC, IMDI, TEI header, ...?
- ❑ The specific ones usually build upon the general ones.

# Dublin Core

- ❑ A small set of vocabulary terms that can be used to describe
  - ❑ web resources (video, images, web pages, datasets etc.)
  - ❑ physical resources (books or CDs, and objects like artworks)
  
- ❑ DCMI Metadata Terms list: **15 Core terms (DCMES)** + qualified terms
  - ❑ abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, **contributor**, **coverage**, created, **creator**, **date**, dateAccepted, dateCopyrighted, dateSubmitted, **description**, educationLevel, extent, **format**, hasFormat, hasPart, hasVersion, **identifier**, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, **language**, license, mediator, medium, modified, provenance, **publisher**, references, **relation**, replaces, requires, **rights**, rightsHolder, **source**, spatial, **subject**, tableOfContents, temporal, **title**, **type**, valid

# An example from TROLLing (simplified and modified record)

- ❑ **Title:** Subject Placement in the History of Latin
- ❑ **Creator:** Danckaert, Lieven
- ❑ **Subject:** syntax, Latin, ...
- ❑ **Description:** The present dataset was used in a corpus study on ...
- ❑ **Publisher:** DataverseNO
- ❑ **Contributor:** Danckaert, Lieven
- ❑ **dateSubmitted:** 2017-04-15
- ❑ **Issued:** 2017-04-30
- ❑ **Type:** textual corpus
- ❑ **Format:** ---
- ❑ **Identifier:** doi:10.18710/V9D674
- ❑ **Source:** The LASLA corpus, ...
- ❑ **Language:** Latin
- ❑ **isReferencedBy:** To appear in Journal of Catalan Linguistics
- ❑ **Coverage:** ---
- ❑ **License:** CC0
- ❑ **Rights:** CC0 Waiver

# Metadata on the language(s)

- ❑ **Creator:** Danckaert, Lieven
- ❑ **Subject:** syntax, Latin, ...
- ❑ **Description:** The present dataset was used in a corpus study on ...
- ❑ **Source:** The LASLA corpus, ...
- ❑ **Language:** Latin
- ❑ **isReferencedBy:** To appear in Journal of Catalan Linguistics

## **Lauren Collister on Q1:**

“I think an important component of metadata in linguistics is finding all research about a particular language, regardless of the subfield of linguistics, so that language data can become data for research regardless of its subfield of creation. For example, a sociolinguist might collect a transcript that includes language data which could be useful for documentation, phonology, etc.”

## **Anonymous feedback on Q2:**

“... language is important (in a standardised form, like ISO 639-3)”

# Metadata on the language(s)

- ❑ **Creator:** Danckaert, Lieven
- ❑ **Subject:** syntax, Latin, ...
- ❑ **Description:** The present dataset was used in a corpus study on ...
- ❑ **Source:** The LASLA corpus, ...
- ❑ **Language:** Latin ← English
- ❑ **isReferencedBy:** To appear in Journal of Catalan Linguistics

## Anonymous feedback on Q2:

“... the various ways that linguists/language documenters need to think about the term "language". In standard Dublin Core metadata, language usually refer[s] to the language of the media content. [...] we also need to know the language(s) that is/are the focus or subject of the research, as well as information about the various languages that the data provider speaks, e.g., native language(s), other language(s), etc.”



# Metadata on the language(s)

## **Anonymous feedback on Q2:**

“... written material needs specification of what \*kind\* of language is involved (formal/informal, etc.). Likewise, spoken language needs the same, plus indication of the dialect in question.”

# Contributor(s)

- ❑ **Publisher:** DataverseNO
- ❑ **Contributor:** Danckaert, Lieven
- ❑ **dateSubmitted:** 2017-04-15

## **Anonymous feedback on Q2:**

“Another specific need is for data providers to get some sort of attribution for the data they helped to create; I'm thinking specifically of language consultants and data that are not subject to anonymization requirements.”

# Different types of linguistic resources

**Q3: Do different types of linguistic resources need different elements of metadata?**

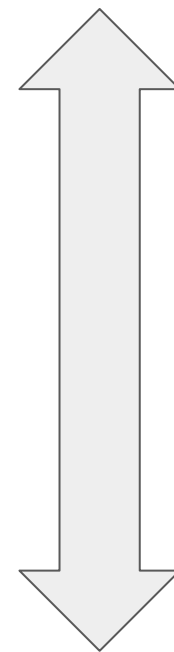
**Tom Honeyman:** “Yes. “linguistic resources” is an incredibly broad category.”

**Anonymous:** Absolutely. Any sort of experimental data should be accompanied by detailed information about the methodology, equipment and software used, and even specific settings for the software.

**Anonymous:** “Possibly a few (like size and its unit), but mostly not. The most important basic metadata is quite common.”

**Anonymous:** “All the linguistic resources should exclusively yield to the same elements of metadata [...].”

**YES!**



**NO!**

# Open Language Archives Community (OLAC)

- ❑ Enriched version of Dublin Core
- ❑ Based on the complete set of Dublin Core metadata terms
- ❑ Extensions allow expression of community-specific qualifiers, e.g. covers some of the specific needs for metadata about linguistic resources discussed earlier, e.g. Dublin Core subject field:

```
<dc:language xsi:type="olac:language" olac:code="lat"/>
```

# Other specific metadata standards/schemes

- ❑ ISLE Meta Data Initiative (IMDI)
  - ❑ Metadata standard to describe multi-media and multi-modal language resources
  - ❑ Provides interoperability for browsable and searchable corpus structures and resource descriptions
  
- ❑ Text Encoding Initiative (TEI) header
  - ❑ Part of the Text Encoding Initiative P5 Guidelines for Electronic Text Encoding and Interchange
  - ❑ Addresses the problems of describing an encoded work so that the text itself, its source, its encoding, and its revisions are all thoroughly documented

# How specific metadata do we really need?


## Gary Holton on Q1:

“The intro to this document held up both OLAC and IMDI as gold standards for documentary linguistics, but I would actually argue against developing a general scheme with the hierarchy and granularity of the IMDI schema.

One of the reasons OLAC works so well is that it is relatively lightweight. It doesn't ask much of the metadata creator, but it's sufficient for most discovery purposes. Back in the early 2000's there were many discussion about how to integrate OLAC and IMDI, but they ultimately failed owing to differing perspectives on granularity. We can develop metadata schemes that allow people to search for all tokens of sibilant-initial words uttered by female L2 speakers between the ages of 25-25. But how many people are really going to search for this?”

# Component Metadata Infrastructure (CMDI)

- ❑ Profiles (e.g. for "corpus") are put together out of reusable components (e.g. "language", "address")
- ❑ Components are published in the component registry
- ❑ Metadata records are harvested to several catalogs, esp VLO
- ❑ If necessary, records are converted to CMDI
- ❑ Structure of metadata allows faceted search

newspaper Showing 1 to 10 of 24214 results within selection for newspaper FinnishResults per page: 10 

Use the categories below to limit the search results to those matching the selected value(s).

Language Finnish Collection Resource type 

Type to filter or search for more

TEXT (24207)

Analytic serial (24196)



Newspaper Issue (24196)

Newspaper Title (11)

Serial (11)

Corpus (4)

Written Corpus (1)

Modality Format Availability 

&lt;&lt; &lt; 1 2 3 4 5 6 7 8 9 10 &gt; &gt;&gt;

## Finnish **newspaper** subcorpus from 2012 (fin\_news\_2012\_300K)

(Part of Leipzig Corpora Collection)

 300.000 sentences of a Finnish **newspaper** corpus based on material from 2012

## The Karelian Finnish **Newspaper** Corpus

(Part of CLARIN Centres)

 The corpus contains issues of the Karjalan Sanomat **newspaper** published in 2012-2014. The corpus is available in Kielipankki - the Language Bank of Finland (<http://urn.fi/urn:nbn:fi:lb-2016112501>). In case you are not a member of an academic institution please read the access rights instructions at <https://www.kielipa...>


## The Karjalainen Corpus

(Part of LRT + Open Submissions Data &amp; Tools)

 computer corpus of Finnish **newspaper** texts of the 1990s (**newspaper** Karjalainen, Joensuu)

## The Karjalainen Corpus

(Part of CLARIN Centres)

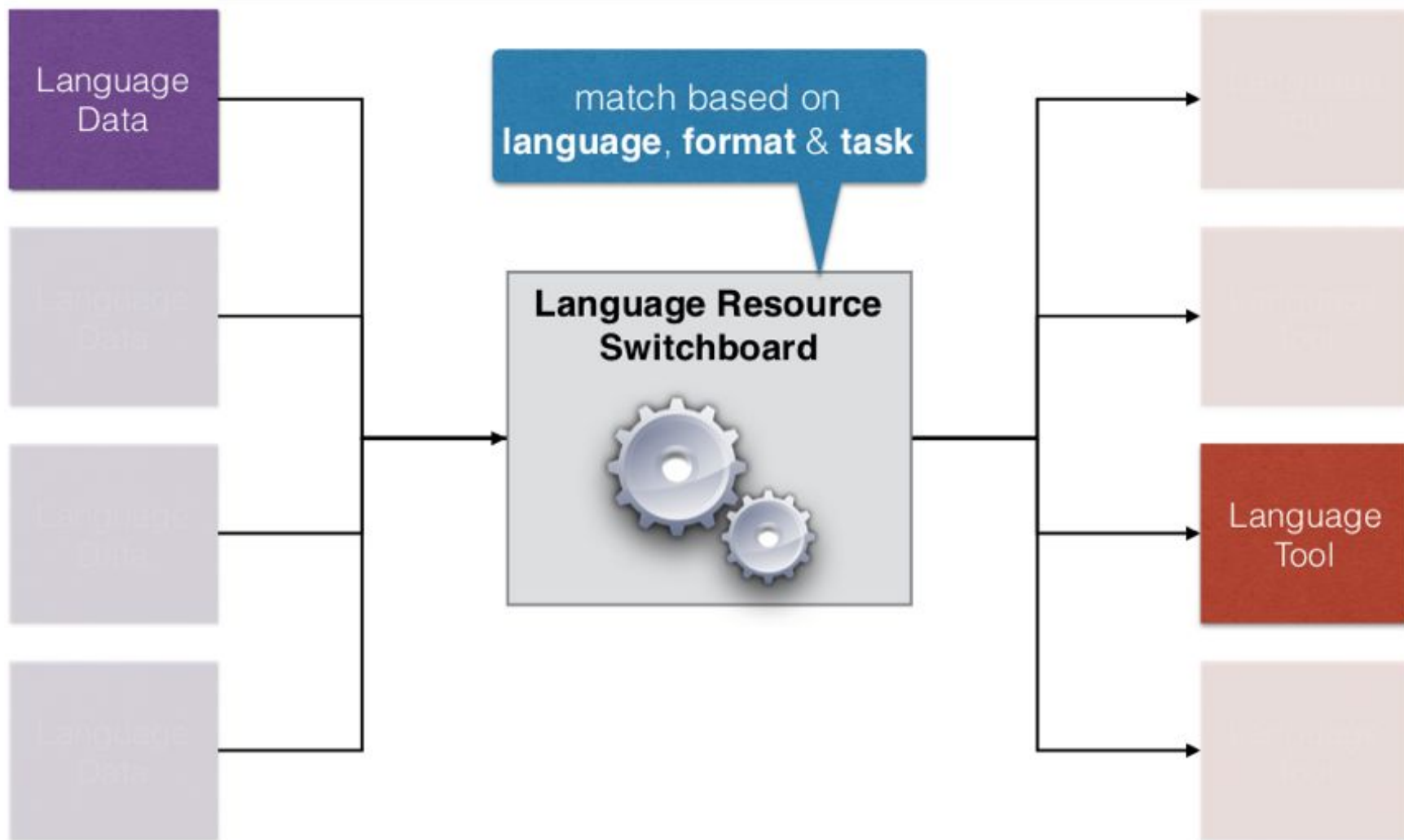
 Computer corpus of Finnish **newspaper** texts of the 1990s (**newspaper** Karjalainen, Joensuu). The purpose of the resource use must be outlined in a research plan.

## The **Newspaper** and Periodical OCR Corpus of the National Library of Finland (1771-1874)

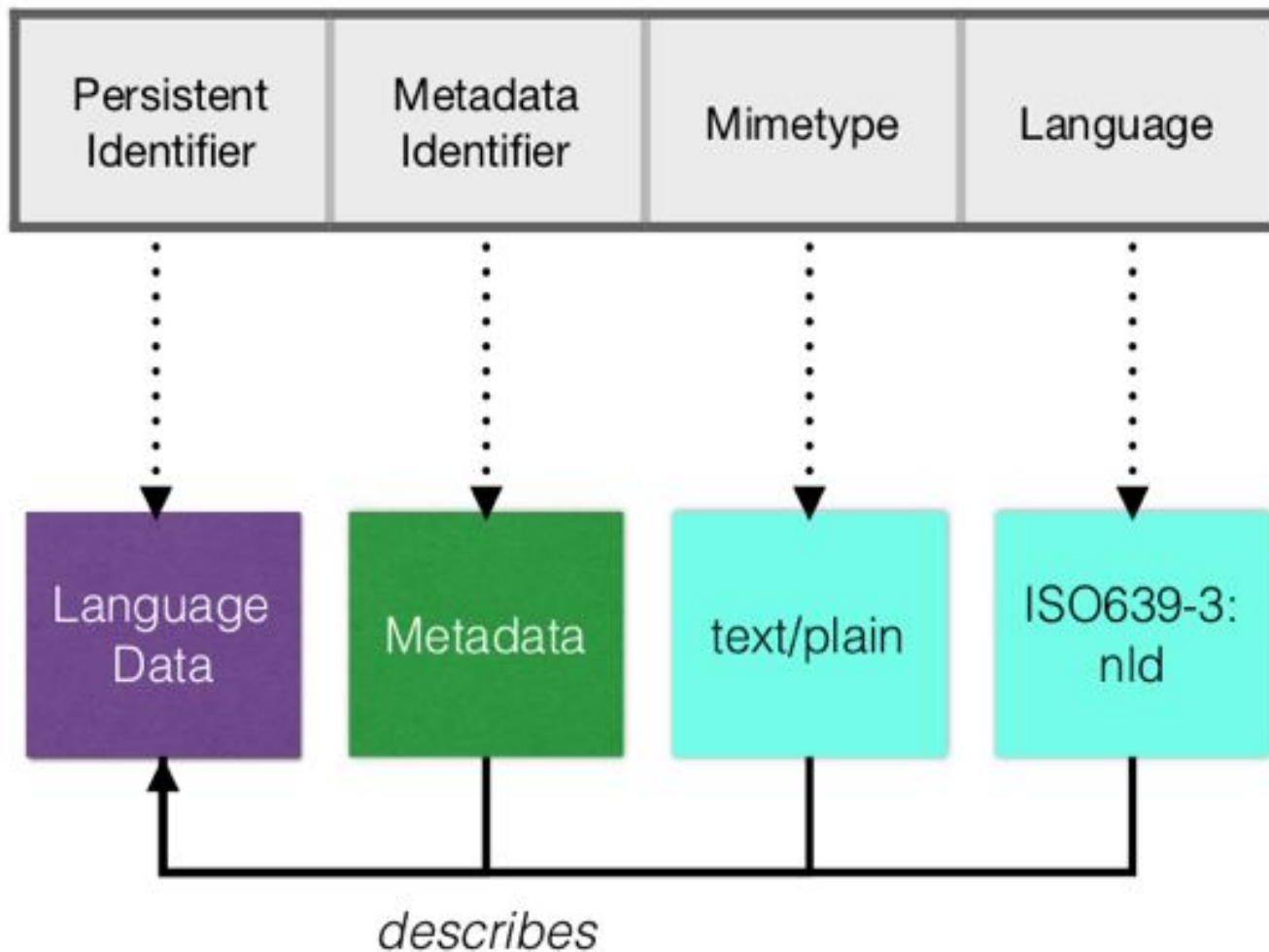
(Part of CLARIN Centres)







# Digital object for language data



# Feedback from the floor

Q1: What are the main purposes of metadata in the field of language and linguistics?

Q2: Are there special needs or demands for the metadata description, citation and attribution of linguistic resources compared to other fields of research?

Q3: Do different types of linguistic resources need different elements of metadata?

Q4: How would you like to see bibliographic references and in-text citations of linguistic resources (including complete datasets, excerpts, search results, examples, etc.) in the papers that you read?

Q5: What should editors/publishers do in order to improve citation practices for linguistic resources?

Q6: In which respects does reproducibility put extra demands on citation practices and metadata?

# What next? Some task suggestions

- ❑ Developing templates for citing linguistic resources.
- ❑ Agreeing on mandatory metadata elements for linguistic resources.
- ❑ Recommendations for repositories
  - ❑ Provide (a) metadata schema(s) including mandatory elements for linguistic resources
- ❑ Recommendations for journals
  - ❑ Establish requirements
  - ❑ Citation Style sheet(s)
  - ❑ Point to repository/-ies for data deposit
- ❑ Developing and delivering data management training for linguistics.

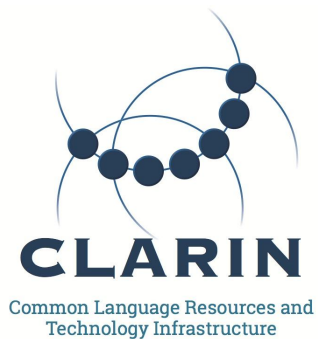
# How? RDA working group(s)?

- ❑ Working groups in RDA
  - ❑ Short-term (18 months).
  - ❑ Develop and implement data infrastructures (e.g. tools, policy, practices and products) that are adopted and used by projects, organizations, and communities.
  
- ❑ Join existing working groups?
- ❑ Create a dedicated working group on citation and metadata standards in linguistics?
  - ❑ Main task: Connect citation practices and metadata management in linguistics, by developing dataset citation standards in light of recommended metadata standards.

# Thanks to our institutions and sponsors!



UiT / THE ARCTIC UNIVERSITY OF NORWAY



# References

Berez-Kroeker, A. L., Andreassen, H. N., Gawne, L., Holton, G., Kung, S. S., Pulsifer, P., Collister, L. B., The Data Citation and Attribution in Linguistics Group, & the Linguistics Data Interest Group. (2017a). *Draft: The Austin Principles of Data Citation in Linguistics (Version 0.1)*. <http://site.uit.no/linguisticsdatacitation/austinprinciples/> Accessed 19.03.2018.

Berez-Kroeker, A., Gawne, L., Kelly, B. F., & Heston, T. (2017b). *Survey of Current Reproducibility Practices in Linguistics Journals, 2003-2012*. <https://sites.google.com/a/hawaii.edu/data-citation/survey>

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., . . . Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, 56(1), 1-18. <https://doi.org/10.1515/ling-2017-0032>

Component MetaData Infrastructure (CMDI). <https://www.clarin.eu/content/component-metadata>

Gawne, L., Kelly, B. F., Berez-Kroeker, A. L., & Heston, T. (2017). Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation*, 11, 157-189. <http://hdl.handle.net/10125/24731>

Haspelmath, M. (2014). *The Generic Style Rules for Linguistics*. <https://doi.org/10.5281/zenodo.253501>

ISLE Meta Data Initiative (IMDI). <https://tla.mpi.nl/imdi-metadata/>

Joseph, B. et al. (2007). *Unified Style Sheet for Linguistics*. <https://www.linguisticsociety.org/resource/unified-style-sheet>

OLAC Metadata. <http://www.language-archives.org/OLAC/metadata.html>

# References (cont.)

Pedersen, T. (2008). Empiricism is not a matter of faith. *Computational Linguistics*, 34(3).

TEI header. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html>

Thomason, S. G. (1994). The Editor's Department. *Language*, 70(2), 409-413.

TROLLing; editing status 2018-01-17; re3data.org - Registry of Research Data Repositories.  
<http://doi.org/10.17616/R3834T> last accessed: 2018-03-20