Symposium & Panel Discussion:

Data Citation and Attribution for Reproducible Research in Linguistics

Andrea L. Berez-Kroeker (U Hawai‘i Mānoa)
Gary Holton (U Hawai‘i Mānoa)
Susan Smythe Kung, (U Texas Austin)
Geoff Nathan (Wayne State U)
Peter L. Pulsifer (U Colorado Boulder)
Anthony Woodbury (U Texas Austin)
Keren Rice (U Toronto)
Stanley Dubinsky (U South Carolina)
David Beaver (U Texas Austin)
Shobhana Chelliah (U North Texas)
Ruth Duerr (Ronin Institute)
Richard Meier (U Texas Austin)

Overview: Today’s Panel

- Introductory presentation
- Five 7-minute “mini-presentations” on aspects of the Data Citation question
- Followed by 40 minutes of facilitated discussion (led by Richard Meier) — We want to hear from you!

And tomorrow:

- Poster session, 10:30-12:00, Lone Star Foyer, #75-84

Introductory presentation:

Reproducible Research in Linguistics: Toward a data-driven science of language

Andrea L. Berez-Kroeker, University of Hawai‘i at Mānoa
Gary Holton, University of Hawai‘i at Mānoa
Susan Smythe Kung, University of Texas at Austin
Geoff Nathan, Wayne State University
Peter L. Pulsifer, University of Colorado at Boulder

Overview

- This symposium is part of an NSF-funded project to develop data citation and attribution standards for linguistics
- One of 8 projects funded under the “Supporting Scientific Discovery Through Norms and Practices for Software and Data Citation and Attribution” (SciSIP)
- We hope to engage the field of linguistics in the wider dialogue of *open science*
- What would it take to facilitate reproducible research in linguistics?

Goals of the project

- Develop and promote standards for
  - citing linguistic data sets
  - academic attribution for data set creation, curation, storage and sharing

...shifting the field toward a more scientific, data-driven model which results in reproducible research.

How do we get there?

By bringing together stakeholder communities

Workshop 1: Boulder, Sept 18-20, 2015

Workshop 2: Austin, April 8-10, 2016

LSA Symposium: Austin, Jan 5, 2017

LSA Poster Session: Austin, Jan 6, 2017

Workshop 3: Austin, January 8-9, 2017

40+ international participants so far

Felix Ameka, Leiden U

Helene Andreassen, TROLLing, UiT

Anthony Aristar, U of Texas at Austin

Helen Aristar-Dry, U of Texas at Austin

David Beaver, U of Texas at Austin

Andrea Berez-Kroeker, U of Hawai'i at Mānoa

Hans Boas, U of Texas at Austin

Brian Carpenter, American Philosophical Society

Shobhana Chelliah, U North Texas

Lauren Collister, U of Pittsburgh

Tanya E. Clement, U of Texas at Austin

Megan Crowhurst, U of Texas at Austin

David Carlson, World Climate Research Programme

Meagan Dailey, U of Hawai'i at Mānoa

Stanley Dubinsky, U of South Carolina

Ruth Duerr, U of Colorado Boulder

Colleen Fitzgerald, National Science Foundation

Lauren Gawne, School of Oriental and African Studies

Jaime Perez Gonzalez, U of Texas at Austin

Ryan Henke, U of Hawai'i at Mānoa

Gary Holton, U of Hawai'i at Mānoa

Kavon Hooshiar, U of Hawai'i at Mānoa

Tyler Kendall, U of Oregon

Susan Smythe Kung, U of Texas at Austin

Julie Anne Legate, U of Pennsylvania

Richard P. Meier, U of Texas at Austin

Bradley McDonnell, U of Hawai'i at Mānoa

Geoffrey S. Nathan, Wayne State

Peter Pulsifer, U of Colorado Boulder

Keren Rice, U of Toronto

Loriene Roy, U of Texas at Austin

Mandana Seyfeddinipur, ELDP

Gary F. Simons, SIL International

Maho Takahashi, U of Hawai'i at Mānoa

Nick Thieberger, U of Melbourne

Sarah G. Thomason, U of Michigan

Jessica Trelogan, U of Texas at Austin

Paul Trilsbeek, The Language Archive, Max Planck Institute for Psycholinguistics

Mark Turin, U of British Columbia

Laura Welcher, Rosetta Project, Long Now Foundation

Nick Williams, U of Colorado Boulder

Margaret Winters, Wayne State

Anthony C. Woodbury, U of Texas at Austin

Project outcomes

- Submit a proposal for a Resolution on citation and attribution to the LSA.
- Position paper on standards for citation and attribution in linguistics.
- Work toward an international body supporting open data practices in linguistics.

Reproducible research

- Scientific claims must be falsifiable, verifiable, and *reproducible*.
- Reproducibility is similar to, but distinct from, replicability
- *Replicability* applies original methods to generate new data in order to confirm (or disconfirm) original conclusion

Reproducible research

- *Reproducibility* applies original analyses to existing (original) data in order to confirm (or disconfirm) conclusions

- Valuable when faithfully reproducing the original research methods is not possible
  - Behavioral sciences, like linguistics

- **Crucially, reproducibility requires open, accessible data**

The reproducible research movement

- Berlin Declaration on Open Access (2003) encourages open data and now has more than 500 institutional signatories
- Open Science Project aims for “public availability and reusability of scientific data” (openscience.org/blog/?p=269)
- FORCE11 developing a new form of scholarly publication with accessible data (force11.org)
- Increasingly mandated by funding agencies
- Approx 20% of major journals now have a data policy

e.g., “PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction….”

Reproducible research in linguistics

Recently discussed within the context of Language Documentation:

“[Language] documentation will ensure that the collection and presentation of primary data receive the theoretical and practical attention they deserve” (Himmelmann 1998:164)

“[...] it is our professional responsibility to provide the data on which our claims are based. [...] It enhances the scientific basis of the linguists’ work.” (Thieberger 2009:365-6)

“Establishing open archives for primary data is in the interest of making analyses accountable.” (Himmelmann 2006:6)

Reproducible research in linguistics

Relevant to all of linguistics:

- Thomason advises authors submitting to *Language* to “provide detailed information about sources of data and methodology of data collection” (1994: 413)
- Open data allows errors in data to be more readily discovered, leading to more robust science
- Not “gotcha” linguistics, but rather a way to make linguistics a more data-driven science.

What are data in linguistics?

Multimedia corpus (field notes, recordings, annotations) on which a descriptive grammar is based

Text corpus from which example sentences in a publication are drawn

Responses to experimental stimuli

Database codings on which typological generalization is made

4D ultrasound videos of articulation loci

F1 & F2 vowel measurements

Nasality measurements

Spectrograms

Praat text grids

Grammaticality judgments

…

Citation and *attribution*

Citation refers to the practice of identifying the source of linguistic data

- Can be more or less granular but crucially needs to identify data within a larger context.

Attribution refers to the practice of giving people credit for collecting (and providing access to) data

- Requires developing protocols for assessing and evaluating research outputs.
- What should people get credit for? And how?

Workshop 1 (September 2015)

- Archiving Community Working Group
    - Archives are the permanent hosts for data sets
    - Archives can provide the Persistent Identifier (PID) for data sets
    - Language Documentation is the subfield most versed in this kind of archiving
        - We can educate other linguists to evaluate Institutional Repositories, etc.
    - Archives can collaborate with Journal Editors
- Journal Editors Community Working Group
    - We support citing data sources in the best possible way
    - We need a unified stylesheet for citations
    - We need some guidance from the field re what constitutes the data set
    - We would appreciate citation templates (e.g. Zotero, Endnote, Mendeley…)
    - We support citation tracking when possible

Workshop 1 (September 2015)

Ordinary Working Linguist Community

- We need improved training of graduate students and early career researchers in data management and archiving
- We need a distributed culture that encourages proper data management at all stages
- We need to adopt metrics for evaluating data sets in tenure & promotion
- We need help evaluating (‘certifying’) archives/IRs as homes for our data
- We need solutions for managing the costs of archiving

IT/Big Data Community Working Group

- Our role is to make known to linguists what kind of technology is available
- Funding and sustainability are our biggest challenges
- PIDs (e.g., DOI) are key to citation tracking, which is new to linguistics
- OLAC needs continued support and development

Workshop 2 (April 2016)

1. Task Force: Data citation principles in linguistics
2. Task Force: Attribution and evaluation of linguistic data
3. Task Force: The role of journals and citation guidelines
4. Task Force: Outreach and education

…leading to the mini-presentations at toady’s symposium.

Today’s panel: Mini-presentation 1

Anthony Woodbury (UT Austin) and Nick Thieberger (U Melbourne) on

“Data Citation: Broad Principles and Guidelines”

- Data as important resources in their own right
- Need for established standards on reuse of data in linguistics

Today’s panel: Mini-presentation 2

Keren Rice (U Toronto) on

“Data Collections: Attribution for Academic Credit”

- Tradition of valuing datasets in linguistics
  ...but uneasy relationship with raw data
- “Big data” opens questions about the *scholarly merit* of data collection and preparation

Today’s panel: Mini-presentation 3

David Beaver (U Texas Austin) and Stanley Dubinsky (U South Carolina) on

“The Role of the Journal in Linguistic Data Citation and Attribution”

- As the main vehicle for dissemination, journals play a critical role
- Need to address issues of authorship, access, attribution, and dynamicity of data

Today’s panel: Mini-presentation 4

Shobhana Chelliah (U North Texas) on

“Education, Outreach and Resources: Effecting a Culture Shift in Linguistics via Data Management and Data Citation”

- How to create a broad culture shift among our colleagues
- Proposed curriculum modules

Today’s panel: Mini-presentation 5

Ruth Duerr (Ronin Institute) on

“Data Citation in the Sciences”

- What happens when cited sources are not accessible?
- Need to redefine the norms of the scientific process
- Each discipline needs to address this for its own community

Selected references

Himmelmann, N.P. 1998. Documentary and descriptive linguistics. Linguistics 36(1).161-95.

Himmelmann, N.P. 2006. Language documentation: What is it and what is it good for? Trends in Linguistics: Studies and Monographs 178, ed. by J. Gippert, N.P. Himmelmann and U. Mosel, 1-30. The Hague: Mouton de Gruyter.

Thieberger, N. 2009. Steps toward a grammar embedded in data. New challenges in typology: Transcending the borders and refining the distinctions, ed. by P. Epps and A. Arkhipov, 365-83.

Thomason, S.G. 1994. The editor's department. Language 70(2).409-13.

Mini-presentation 1

Data Citation: Broad Principles and Guidelines

Nicholas Thieberger & Anthony C. Woodbury

Focus of this talk

Valuing linguistic data. Linguistic data are important resources for a range of academic stakeholders, including community members as well as scholars.

Standards for citing linguistic data. Standards are needed as a means to verify claims made by researchers, to provide credit to data creators, and to facilitate transparency, discovery, critical evaluation, and the long-term use of data.

The challenges. The challenges are great because data takes many forms, it may raise ethical and proprietary issues, and it is used in many contexts including academic promotion.

Valuing linguistic data

Linguistic data are important resources in their own right and represent valuable assets for the field and for the people recorded, especially when they are recordings of high cultural worth

The recent emergence of neo-Boasian documentary linguistics, where autonomous language archives are forms of ethnography and creative intellectual communication

Making connections between data creation/archiving, and the curation and citation of data in linguistic scholarship

Citation standards: what are they and why are they needed?

For a linguist to be able to work with their own records or to use records created by others, there need to be established citation standards that can be adopted easily by all relevant members of the research community.

Quoting a sentence as an example relies on a form of identification of that sentence within a larger set of data, which could, for example, be a corpus of all that is recorded for a language. Agreed methods of pointing at that sentence will then allow others to cite back to the primary source, rather than referencing secondary or tertiary sources of decontextualised examples.

Concerns about proper citation and attribution of examples were raised in the context of the LSA as early as 1994, where Sally Thomason provides the following advice:

"always consult primary sources; use sources with care; consider all relevant data; and provide detailed information about sources of data and methodology of data collection." (Thomason 1994: 413)

Citation standards: broad goals

Citation implies a preservation strategy

Whatever is cited needs to be in a location to allow the curious scholar to access it, the work of the original scholar needs to be attributed properly, and citation formats have to allow for that attribution

Data should be made open as soon as possible, but with consideration of ethical exceptions

Data should be in usable formats, with sufficient machine and human readable documentation to allow informed re-use

These responsibilities are an integral part of linguistic research and must be shared by individual scientists, data stewards, research institutions, and funding organizations.

Challenges of data citation

Archived collections of data are aimed at wide academic and non-academic audiences and are therefore likely to be cited in many different contexts, for different purposes (e.g., their content, their linguistic characteristics), at different levels of presentation (whole pieces vs. individual sentences or words) and in different forms (e.g., as video, audio, transcripts, and or translations) or levels of analysis;

Challenges of data citation

Archived collections of data are aimed at wide academic and non-academic audiences and are therefore likely to be cited in many different contexts, for different purposes (e.g., their content, their linguistic characteristics), at different levels of presentation (whole pieces vs. individual sentences or words) and in different forms (e.g., as video, audio, transcripts, and or translations) or levels of analysis;

There is a range of ethical and proprietary issues associated with archived data, including copyright, intellectual property, fair use, protection of privacy, and restrictions on access placed by producers or persons portrayed;

Challenges of data citation

Archived collections of data are aimed at wide academic and non-academic audiences and are therefore likely to be cited in many different contexts, for different purposes (e.g., their content, their linguistic characteristics), at different levels of presentation (whole pieces vs. individual sentences or words) and in different forms (e.g., as video, audio, transcripts, and or translations) or levels of analysis;

There is a range of ethical and proprietary issues associated with archived data, including copyright, intellectual property, fair use, protection of privacy, and restrictions on access placed by producers or persons portrayed;

There is a range of academic contexts in which data citation arises, including journal articles, public presentation, use for pedagogical purposes, and use in academic promotion.

Precedents

We note that data citation standards are being developed in a number of disciplines and that we can adapt existing and well-considered documents such as the following:

Ball, A., & Duke, M. (2012). How to Cite Datasets and Link to Publications. Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/resources/how-guides#sthash.FmNPiXsU.dpuf

Lawrence, B., Jones, C., Matthews, B., Pepler, S., & Callaghan, S. 2011. Citation and Peer Review of Data: Moving Towards Formal Data Publication. International Journal of Digital Curation, 6(2), 4–37. http://doi.org/10.2218/ijdc.v6i2.205

Martone M. (ed.) 2014. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. San Diego CA: FORCE11 https://www.force11.org/group/joint-declaration-data-citation-principles-final

Thomason, Sarah. 1994. Editorial. *Language*, Vol. 70, No. 2: 409-413

Mini-presentation 2:

Data Collections: Attribution for Academic Credit

Keren Rice

With contributions from Felix Ameka, Meagan Dailey, Helen Dry, Ryan Henke, Julie Legate, Richard Meier, Nick Thieberger, Sally Thomason, Margaret Winters, Tony Woodbury

Some background

Long recognition of value of data in linguistics

Workbooks: Gleason, Merrifield, Halle and Clements, etc.

Uneasiness with data without theoretical analysis

Are grammars, dictionaries, and the like acceptable as dissertations?

Data today

Continued recognition of the value of ‘small’ data

Focus on ’big’ data

Some initiatives

Digging into data

LSA resolution on scholarly merits of language documentation

LSA resolution on cyberinfrastructure

Digging into Data

The Digging into Data Challenge aims to address how "big data" changes the research landscape for the humanities and social sciences. Now that we have massive databases of materials available for research in the humanities and the social sciences--ranging from digitized books, newspapers, and music to information generated by Internet-based activities and mobile communications, administrative data from public agencies, and customer databases from private sector organizations —what new, computationally-based research methods might we apply? As the world becomes increasingly digital, new techniques will be needed to search, analyze, and understand these materials. Digging into Data challenges the research community to help create the new research infrastructure for 21st-century scholarship.

http://diggingintodata.org/about

2010 LSA Resolution Recognizing the Scholarly Merit of Language Documentation

Whereas the practice of linguistic fieldwork is shifting to a more collaborative endeavor firmly based on ethical responsibilities to speech communities and a commitment to broadening the impacts of scholarship; and

Whereas this shift in practice has broadened the range of scholarly work to include not only grammars, dictionaries, and text collections, but also **archives of primary data, electronic databases, corpora, critical editions of legacy materials, pedagogical works designed for the use of speech communities, software, websites, or other digital media**; and

Whereas the products of language documentation and work supporting linguistic vitality are of significant importance to the preservation of linguistic diversity, are fundamental and permanent contributions to the foundation of linguistics, and are **intellectual achievements which require sophisticated analytical skills, deep theoretical knowledge, and broad linguistic expertise**;

Therefore the Linguistic Society of America supports the **recognition of these materials as scholarly contributions** to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty. **It supports the development of appropriate means of review of such works so that their functionality, import, and scope can be assessed relative to other language resources and to more traditional publications**.

http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation

2010 LSA Resolution on Cyberinfrastructure

Whereas modern computing technology has the potential of advancing linguistic science by enabling linguists to work with datasets at a scale previously unimaginable; and

Whereas this will only be possible if such data are made available and standards ensuring interoperability are followed; and

Whereas data collected, curated, and annotated by linguists forms the empirical base of our field,

Therefore, the LSA encourages members and other working linguists to:

make the full data sets behind publications available, subject to all relevant ethical and legal concerns;

annotate data and provide metadata according to current standards and best practices;

seek wherever possible institutional review board human subjects approval that allows full recordings and transcripts to be made available for other research;

work towards assigning academic credit for the creation and maintenance of linguistic data bases; and

when serving as reviewers, expect full data sets to be published (again subject to legal and ethical considerations) and expect claims to be tested against relevant publicly available datasets

…

http://www.linguisticsociety.org/resource/resolution-cyberinfrastructure

Questions

How can the scholarly merit of data be assessed?

How can we recognize the value of data in a career?

On determining scholarly merit: creating a corpus is an intellectual undertaking

What is in a corpus?

Recordings, transcription, annotation, metadata

What is the data best structure for the data in order to be usable?

What information is included in the data?

What metadata (“information describing the constituent resources of a documentary corpus, including, for example, their content, creators, and any access restrictions” (Good 2011: 228) is included?

“Understanding how data collection and management fits into a documentation project is a kind of research. It, therefore, is amenable to all the requirements of research: keeping up with the field, knowing the limits of one’s expertise, tracking down outside sources, constantly evaluating and reevaluating one’s conceptual understanding and methodological practices, and instructing collaborators on appropriate practices. Just as analysing data requires research, so does working with the data itself.” (Good 2011: 233)

Jeff Good. 2011. Data and language documentation. In Peter Austin and Julia Sallabank (eds.), *Handbook of Endangered Languages*. Cambridge: Cambridge University Press. 212–234.

On determining scholarly merit: criteria for scholarly merit 1

Accessibility

Accessible (to the degree possible), available, long-term curation

Non-proprietary data format

In an archive with long-term curation

Nick Thieberger, Anna Margetts, Stephen Morey, and Simon Musgrave. 2015. Assessing annotated corpora as research output. *Australian Journal of Linguistics*.

Criteria for scholarly merit 2 (Thieberger et al)

Quality criteria

Background and corpus structure: how corpus came into being, preparer, funder, projects involved, types of material, abbreviations, orthography, ethnographic information geographic information

Metadata: contents, speakers, keywords, information on recording

Raw (recordings), primary (transcribed), and structural (annotated) data and their linking: transcribed, translated, time-aligned, annotated as basic, with further annotations (parts of speech, intonation contours, pause length, gestures, tagging of syntactic constructions, referentiality features, etc. plus basis for analysis [sketch grammar])

Content: range of speakers (age, gender, etc.), text types (narratives, conversation, songs, etc.)

Amount of data

On Recognition 1: Creating the environment for attribution through publications

(Thieberger et al)

Development of ways of assessing the quality of corpora: publications

Publication of articles describing corpora (e.g., Sophie Salffner. 2015. A road map to Ikaan language documentation. *Language Documentation & Conservation* 9: 237-267)

Reviews of corpora (like book reviews)

Reviews of archives

Corpus as an article (Martin Haspelmath and S.M. Michaelis 2014. Annotated corpora of small languages as refereed publications: a vision. http://dlc.hypotheses.org/691)

On Recognition 2: creating the environment for attribution through awards

Development of ways of assessing the quality of corpora: Awards

Awards for archived collections (DELAMAN)

On Recognition 3: creating the environment for attribution through education

Development of ways of assessing the quality of corpora: education

Appropriate citation of archival materials: culture of citation

Education of tenure and promotion committees

Education of people coming up for tenure and promotion about how to present corpora

Mini-presentation 3:

The Role of the Journal in Linguistic Data Citation and Attribution

David Beaver & Stanley Dubinsky

Editorial issues to be discussed

I. The need to reference data sources

II. Accessibility and citation format

III. The notion of authorship

IV. Citing dynamic data

V. Grain of citations

I – Need to reference data sources

“Whenever and wherever a claim relies upon data, the corresponding data should be cited.”

[Data Citation Synthesis Group: Joint Declaration of Data Citation Principles]

Benefits:

Supports replicability.

Increases visibility of data.

Tokenizes use of data.

Can linguists be encouraged to cite data sets?

Include data in an appendix? In a separate publication?

II – Accessibility and citation format

Data should be made available, along with meta-data needed for citation/access. Standard citation styles are insufficient.

> **Sherzer, Joel (Researcher) and Olowiktinappi (Speaker, Translator). (1970a). "Report of a curing specialist." Kuna Collection. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org. Type: primary text. Media: audio. Access: public. Resource ID: CUK012R004, File ID: CUK012R004I001.mp3. Accessed 9 Oct. 2015**

Which metadata to include in citation?

Should journals host data? Guarantee accessibility? Verify the accuracy of citations?

Should editors recommend use of standard repositories?

III – Authorship

The important issue is *attribution*, not *authorship*. Many have responsibility, and their roles are not uniform.

Data generator: someone who produces data for curation, description, or analysis. E.g.:

Native-speaker consultants (pronouncing words, telling stories, providing judgements)

Technical assistants (processing sensor/scanner outputs, generating experiments, or performing simulations)

Agreggator, editor, curator, analyst, translator…?

In which cases attribution appropriately cited?

Authorship (cont.)

“**All those who have made significant contributions should be offered the opportunity to be listed as authors.** [Others] … should be acknowledged, but not identified as authors.” [Guidelines from APS]

A 33-page article in *Physical Review Letters* with 9 pages of text and 22 pages of “authors”. All 5,154 of them made “significant contributions”?

Is this a reasonable/useful standard for linguistics?

A speaker of an endangered language being documented clearly makes a significant contribution.

Is the option of citation listing them automatic?

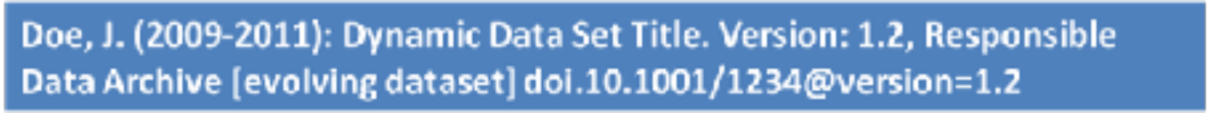
Editors need not decide, but can raise the question.

IV – Dynamicity of data

Datasets may change after being cited. Citations must allow retrieval of data in the form it had when article was written.

Ignoring this affects reproducibility or comprehension of the citing paper. Citation recommendation for evolving data set: Cite year of creation and time of last update (c.f. Duerr 2012).

Example (ANDS “Citing Dynamic Data”):



Doe, J. (2009-2011): Dynamic Data Set Title. Version: 1.2, Responsible Data Archive [evolving dataset] doi.10.1001/1234@version=1.2

Are editors responsible for ensuring stability of cited data?

V – Grain of citations

Replicability and attribution imply a need to uniquely identify the relevant part of a data set.

How fine-grained should the citation be? What additional information should be given in the text and/or notes?

Suggestion: Editors pass the buck to data set’s creator or publisher, citing at the lowest level for which there is a DOI.

Do we then provide all lower-level information required for unique identification and retrieval in text/notes?

Conclusion

Most authors focus on other aspects of their work; not on providing access to or attribution for data.

Editors find themselves in the Wild West as regards data: Few standards; evolving practice.

Editors can take a lead in establishing good practice, by discussing citation formats and establishing standards.

What changes are needed for better access and attribution?

How can editors support those changes?

References

Aad, G. et al., "Combined Measurement of the Higgs Boson Mass in p p Collisions at s= 7 and 8 TeV with the ATLAS and CMS Experiments." *Physical review letters* 114.19 (2015): 191803.

American Physical Society, 02.2 APS *Guidelines for professional conduct* (Adopted by Council on November 10, 2002) https://www.aps.org/policy/statements/02_2.cfm

Archive of the Indigenous Languages of Latin America. (2002). AILLA Citation Guidelines. http://www.ailla.utexas.org/site/citation.html

Australian National Data Service, Citing dynamic data http://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation/citing-dynamic-data

Duerr, R., *ESIP Data Citation Guidelines*, National Snow and Ice Data Center, 2012 https://nosc.noaa.gov/EDMC/documents/edmcon/2012_breakout_sessions/Duerr-ESIP_Data_Citation_Guidelines.pdf

Martone M. (ed.), Data Citation Synthesis Group: Joint Declaration of Data Citation Principles, San Diego CA: FORCE11; 2014

Mini-presentation 4:

Education, Outreach and Resources: Effecting a Culture Shift in Linguistics via Data Management and Data Citation

Shobhana Chelliah

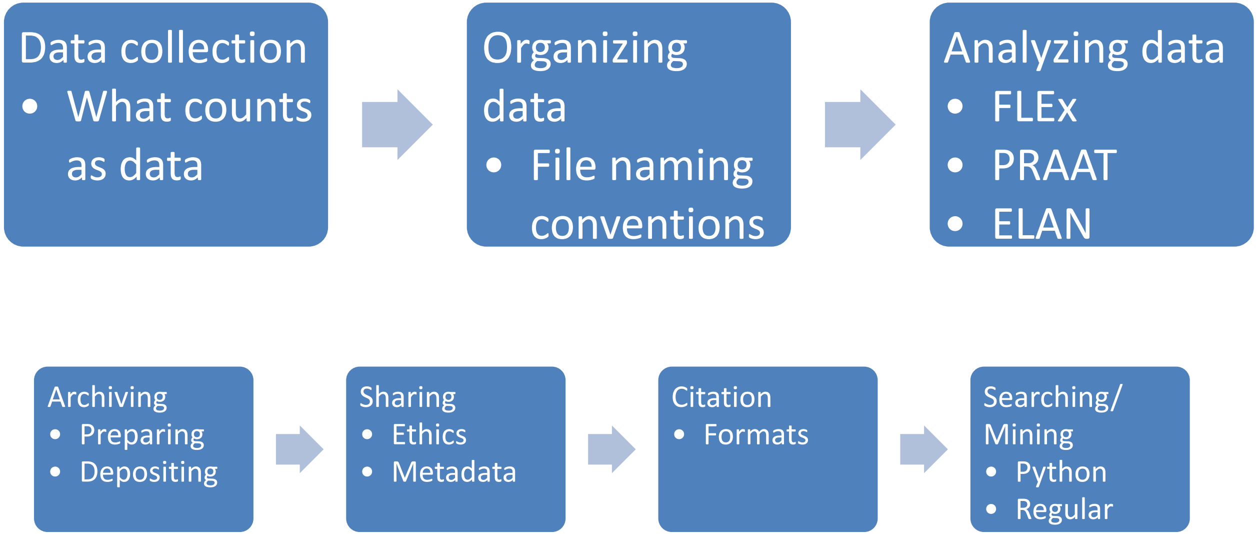
Suggested actions

Develop curriculum

Engage in outreach

Create resources

Content for Instructional Modules/Courses



Data collection
- What counts as data

Organizing data
- File naming conventions

Analyzing data
- FLEx
- PRAAT
- ELAN

Archiving
- Preparing
- Depositing

Sharing
- Ethics
- Metadata

Citation
- Formats

Searching/ Mining
- Python
- Regular

BA Level Courses

Three credit general education requirement course with research component in any discipline. Possible courses:

Endangered languages

Language and computers

One or two credit lab course perhaps with online delivery. Possible courses

Introduction to Linguistics

Endangered languages

Language and computers

MA/PhD level Courses

One or two credit lab course perhaps with online delivery. Possible courses:

Research Methods

Linguistic Field Methods

Sociolinguistics

Discourse Analysis

Language Typology

Dissertation preparation/Graduate Capstone/Professional Development

Funding to support new curriculum

NEH Digital Humanities

using archives to teach about culture

NSF DEL

resource creation (ELDP for legacy materials)

NSF EHR

learning technology applied to language archives, databases, individual corpora.

Private institutions, state government, casinos

training, building digital museums

Standalone modules for conferences and workshops

For standards **creation and updating**

Language Documentation and Conservation
CoLang
AILDI
LSA Institute brown bag

For **training**:

-LSA Institute academic writing workshops, LSA videos and webinars

For standards **sharing within the field**:

-presentations at language area conferences, e.g. Sino-Tibetan Languages and Linguistics

For standards **sharing with related fields:**

presentations at library science conferences, e.g., Institute of Museum and Library Science
presentations at Digital Humanities collections and Digital Library

Outreach to LSA

Aggregation of resources on Data Management

LSA graduate student mixer

LSA Salon on Data Practices

- Blogging and open discussion on citation practices

Outreach to “accidental”curators

heritage and diaspora language-data collectors

squib collections

legacy data on individual researchers

Outreach to university evaluators

Create rubric for evaluation of language data collections: annotated corpora, databases, archived collections →

Encourage scholarship in creating citable collections →

Letters from the LSA

to department chairs on society standards

to reviewers of tenure and promotion cases

to journal editors for review of articles and citation practices

Outreach early career scientists

budgeting time and start up funding for data management planning

early communication on storage and metadata needs

regular updating of data management skills

mentoring can come from university librarians

Outreach to digital repositories

Roadshow to digital libraries

Upgrading bibliography software

Outreach to the public

Provide communities of users with well-conceived and curated heritage and diaspora collections

include search facilities to encourage use by communities

facilitate creation of guides and questions for use in study groups

Resource Creation

Clearinghouse on data management such as the opencon.org with sample Data Management Plans (from various sub-disciplines)

Update E-MELD

Early Career Linguist educators working group

Create repositories and Match to need

LSA website: modules, slides, webinar, videos

Dissemination through university libraries

Mini-presentation 5:

Data Citation in the Sciences

Ruth Duerr

Outline

Citation status in the sciences

Major issues

Examples – On handout

Practice meets policy

- High level guidance is available
  - Joint Declaration of Data Citation Principles
  - Data Citation Implementer's Group Recommendations
- Many societies and organizations (AGU, AAAS, ESIP, RDA, etc.) have put/are putting domain specific policies and guidance in place
- Journals are beginning to require citation of data (and software)
- Repositories are working/re-working to implement citation as per guidance
- Repositories and journals are working together to move the field forward (e.g., Coalition for Publishing Data in the Earth and Space Sciences)
- Many researchers are still clueless
  - How to cite data
  - What even to do with their data

Joint Declaration of Data Citation Principles

Importance: Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

Credit and Attribution: Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

Evidence: In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

Unique Identification: A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

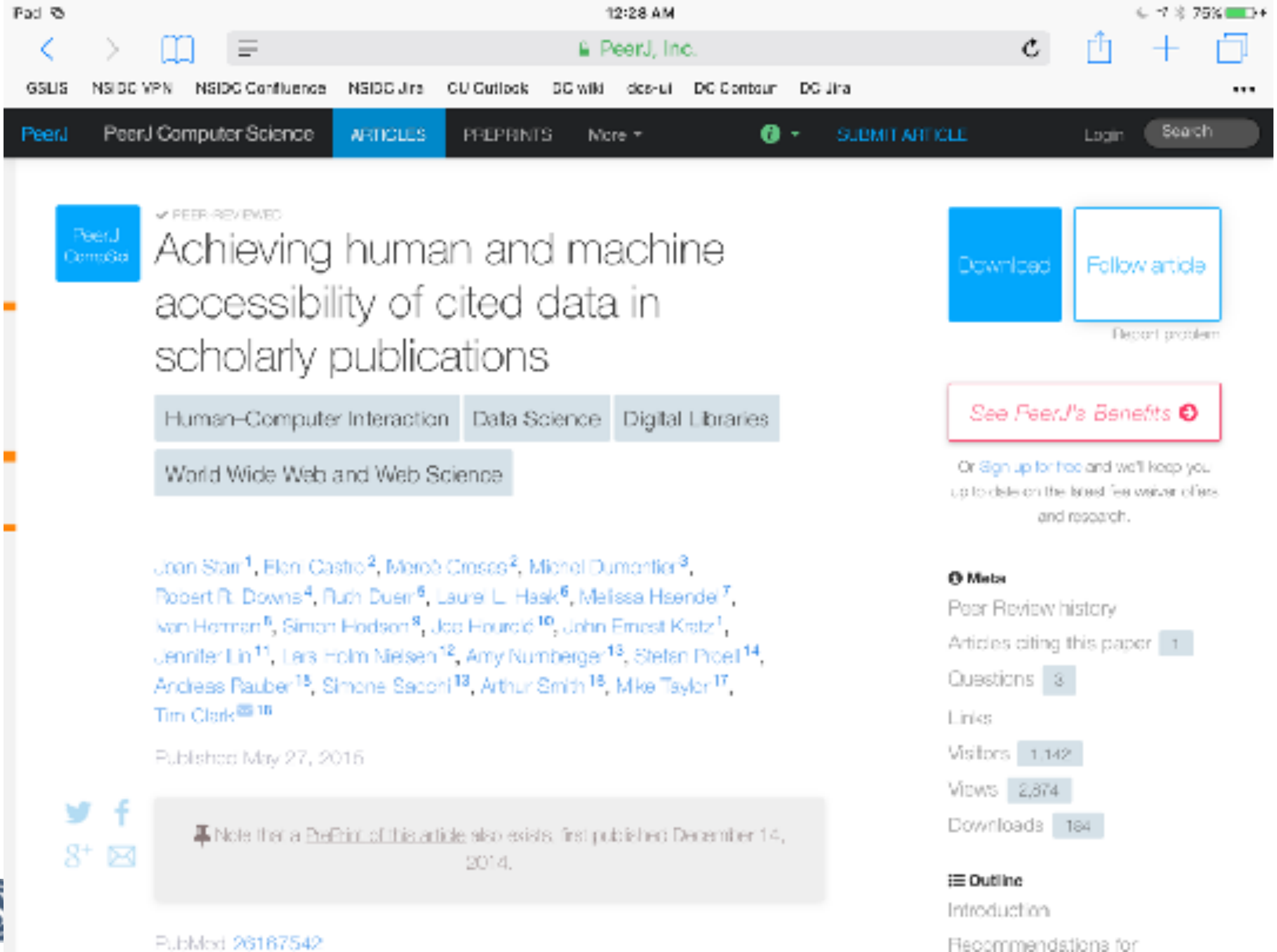
Access: Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Persistence: Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

Specificity and Verifiability: Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific time slice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Interoperability and flexibility: Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014; https://www.force11.org/datacitation



✓ PEER-REVIEWED

Achieving human and machine accessibility of cited data in scholarly publications

Human–Computer Interaction | Data Science | Digital Libraries | World Wide Web and Web Science

Joan Starr[1], Eleni Castro[2], Mercè Crosas[2], Michel Dumontier[3], Robert R. Downs[4], Ruth Duerr[5], Laurel L. Haak[6], Melissa Haendel[7], Ivan Herman[8], Simon Hodson[9], Joe Hourclé[10], John Ernest Kratz[1], Jennifer Lin[11], Lars Holm Nielsen[12], Amy Nurnberger[13], Stefan Proell[14], Andreas Rauber[15], Simone Sacchi[13], Arthur Smith[16], Mike Taylor[17], Tim Clark[18]

Published May 27, 2015

Note that a PrePrint of this article also exists, first published December 14, 2014.

PubMed 26167542

Or Sign up for free and we'll keep you up to date on the latest fee waiver offers and research.

Meta

Peer Review history

Articles citing this paper 1

Questions 3

Links

Visitors 1,142

Views 2,374

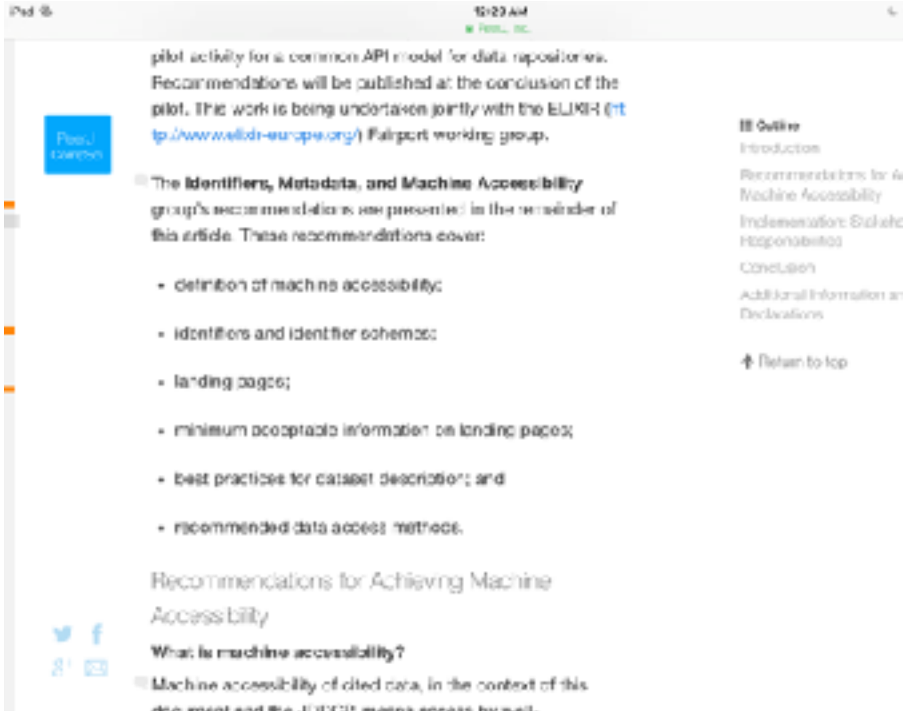
Downloads 184

Outline

Introduction

Recommendations for

Data Citation Implementers Group


iPad
10:20 AM
Read Comments
Outline
Introduction
Recommendations for Achieving Machine Accessibility
Implementation: Stakeholder Responsibilities
Conclusion
Additional Information and Declarations
Return to top

pilot activity for a common API model for data repositories. Recommendations will be published at the conclusion of the pilot. This work is being undertaken jointly with the ELIXIR (http://www.elixir-europe.org/) Fairport working group.

The **Identifiers, Metadata, and Machine Accessibility** group's recommendations are presented in the remainder of this article. These recommendations cover:

- definition of machine accessibility;
- identifiers and identifier schemes;
- landing pages;
- minimum acceptable information on landing pages;
- best practices for dataset description; and
- recommended data access methods.

Recommendations for Achieving Machine Accessibility

What is machine accessibility?

Machine accessibility of cited data, in the context of this

Issues

Scale – What is a data set? How does that relate to what you cite?

Citing parts of a “database”

Citing dynamic data

Citing samples and sample data

Linking data to publications to people, organizations, funding, etc.

Orange – Work remains to even define the principles, policies and/or standards
Yellow – General principles exist but there is yet no general agreement about how to start implementing them and they may even conflict with each other
Green – Principles actively being put into practice

Discussion

One Department Chair’s Perspective on Reproducible Research

Richard P. Meier

We are working with two notions, I think.

Archived data that will enable *reproducible research*. Includes data on endangered languages & quantitative data on sentence processing.

The subset of archived data that is itself a vital *record* of human language & culture and that is a *resource* for linguistic communities and for future research.

Thieberger & Woodbury: “Linguistic data are important resources in their own right.” T & W refer to recordings of “high cultural worth”.

Reproducible Research: The culture of research and publication

Beaver & Dubinsky: raise many questions about how data should be cited

Our discipline will need to work towards answers to these questions.

Departments and their larger universities will need to develop cultures of data citation and attribution

Faculty will need training, mentoring, and financial support.

Undergraduate & Graduate Education

Shobhana Chelliah: data management as part of undergraduate & graduate degree

Benefit: Increase marketability of our students. With training in stats or comp ling → successful applications to data scientist positions in industry?

But, a concern: will the time required for data management increase time-to-degree in doctoral programs?

Increasing pressure to bring time-to-degree down to 5 or 6 years. We need to work to ensure that documentary dissertations can be completed in a timely fashion.

Electronic Data-Sets & Faculty Promotion

Keren Rice reminded us of the 2010 LSA resolution on the value of language documentation. The membership affirmed that:

Archives of linguistic data are evidence of scholarly activity.

Such evidence should contribute to promotion decisions.

Promotion: Training Our Colleagues

Departments, deans, promotion committees—all must be “trained” to understand what goes into the construction & evaluation of a substantial data set.

Training can be successful!

Different promotion standards in language departments for linguists vs. literary critics

In comp ling, different publication standards than in much of linguistics.

These publication standards have been accepted by promotion committees, but department chairs must remind them.

Some administrators are receptive: UT’s recent president indicated that data sets can be one factor in a promotion case, “if published.”

But he wasn’t thinking of linguistics. And he didn’t define “published.”

We linguists want a crucial role in determining how data sets are evaluated in our own field.

What is a valuable data set?

What constitutes publication of a data set?

Accessibility, permanence, subject to peer review

K. Rice: Peer Review of Archived Materials

Publication of peer-reviewed articles describing corpora

Reviews of corpora: like book reviews, but could archives institute peer review procedures for accessioning corpora?

Awards for archived collections (DELAMAN)

Evidence that the archived data support new research? Citation of data by other scholars—

but I suspect that such citations will be few at tenure time.

Questions for Discussion

- What value do YOU think should be placed on electronic data sets?
- How would a culture of reproducible research play out in your own departments and universities?
  - In conducting research?
  - In faculty promotion?
  - In educating our students?

Discussion

We want to hear from you!

Thank you

Please visit our poster session tomorrow

10:30-noon

Lone Star Foyer, posters #75-84

Get involved!

We want to hear from you.

Please contact us at
lingdata@hawaii.edu

And visit our website at
http://bit.ly/LinguisticsDataCitation