Data Sharing For Linguists

To Deposit or Not to Deposit by Ainsley Seago, CC-BY
dx.doi.org/10.1371/journal.pbio.1001779.g001

Guest Lecture, Data Science for Linguistics, February 13, 2020
Slides available: http://bit.ly/dsfl2020

Lauren B. Collister
Director of Scholarly Communication
University of Pittsburgh

Dominic Bordelon
Research Data Librarian
University of Pittsburgh

Who are you? (Lauren)

Scholarly Communications Professional = someone who helps scholars talk about and share their research, no matter the form.

Open Access Advocate = someone who believes that we all have the right to research, that the results of scholarship should be free to read, reuse, and remix.

Sociolinguist = a person who researches how people use language with each other.

Language Data Nerd = Co-editor of the forthcoming MIT Press *Open Handbook of Linguistic Data Management* w/ Andrea Berez-Kroeker, Brad McDonnell, & Eve Koller. Member of the [Linguistics Data Interest Group of the Research Data Alliance].

Who are you? (Dominic)

Research Data Librarian

Research Data Management = supporting researchers in organizing, documenting, sharing, and reusing their data and software

Related: code/data literacy

Open Access/Science/Data Advocate

Happy Love Data Week!

Love Data Week exists to “raise awareness and build a community to engage on topics related to research data management, sharing, preservation, reuse, and library-based research data services. We will share practical tips, resources, and stories to help researchers at any stage in their career use good data practices.”

This talk is on the Love Data Week theme of “Open Data”. Read more here: http://lovedataweek.org/about/open-data/

http://lovedataweek.org/

What’s the Aim for Today?

For current you (in this class):

- Will you be able to publicly and legally share your data with your final project?
- What pitfalls might you encounter with your data for your project?

For future you (a data professional):

- Where can you find data that are openly available and easy to reuse and repurpose?
- What steps can you take for all your future projects to make your work as reproducible as possible?

Outline for today

- What does it mean to “own” data?
    - Q for you: Who owns the data you’re using for this class?
- What is “open data” and why should we share our data?
    - Q for you: Is the data set you’re using open? Can you make it that way?
- Where can you find and share open data?
    - Q for you: Where did you find your data? Where would it be appropriate to share data like what you’re working with?
- General Questions / Discussion

What’s the big deal about sharing data?

Video by TROLLing, visit them here: dataverse.no/dataverse/trolling

Who Owns Data?

What is copyright?

- US Copyright Office: “Copyright protects ‘original works of authorship’ that are fixed in a tangible form of expression.”
- Copyright is the right to do the following to these works:
  - Reproduce & distribute
  - Make derivative works
  - Perform and display
- “Copyright Basics” from the US Copyright Office, Circular 1. https://www.copyright.gov/circs/circ01.pdf

What is not subject to copyright?

- Work that is not fixed in a tangible form (e.g. speech that is not recorded)
- Titles, names, short phrases, slogans, familiar symbols
- Ideas, methods, processes, discoveries, devices, contents
  - Distinguished from the **expressions of these**, e.g. an article about your method of cake baking is subject to copyright, but not the actual method or the recipe (contents) itself.
- Common property, e.g. measurements of the state of the world
  - Most **quantitative data** falls into this bucket.
  - “Common property” can also cover some aspects of Traditional Knowledge

Copyright and Linguistics

- Spoken and written language is usually an ‘original work of authorship’.
- A lexicon is a list of the contents of language - not subject to copyright
  - Though your organization, layout, and description that may accompany this wordlist *are* subject to copyright.
- Vowel measurements are not subject to copyright
  - But your awesome plot of them *is*.
- Text mining newspapers, books, online language *may be subject to copyright*.
  - Read Terms of Service carefully. Some allow for use of these corpora for “research” or “academic/educational purposes.”
- Traditional Knowledge, e.g. folktales? Special case:

UNESCO Universal Declaration on Cultural Diversity (2001)

- Traditional knowledge (including language, stories, history) is a public good and should not be subject to intellectual property right.
- However, the individual performer (of a story, language, etc.) should be acknowledged, credited, and their rights protected.
- Therefore, be **clear** about your intended use of their works and get their permission for your use.
- Profits made off of this kind of work should be directed back to the community that the work came from.
- Read the whole thing: http://unesdoc.unesco.org/images/0012/001271/127160m.pdf

Doctrine of Fair Use

Fair Use is a doctrine of copyright law that allows for reuse of copyrighted works in ways that are considered fair--such as **criticism**, **comment**, news reporting, **teaching**, **scholarship**, and **research**. There are 4 factors:

- The purpose and character of the use, including whether such use is of commercial nature or is for nonprofit educational purposes, and whether the work is *transformative.*
- The nature of the copyrighted work (e.g., whether it is factual or creative in nature)
- The amount and substantiality of the portion used in relation to the copyrighted work as a whole
- The effect of the use upon the potential market for or value of the copyrighted work

More info: http://pitt.libguides.com/copyright/fairuse

Copyright and Text and Data Mining (TDM)

TDM has been considered “Fair Use” in several recent court cases and the Fair Use case for TDM is considered strong because:

- Purpose (Factor 1): TDM is transformative - it creates a new thing for new purposes.
- Nature (Factor 2): For TDM that uses creative works (e.g. novels), slightly weighs against Fair Use.
- Amount (Factor 3): Because of transformative nature of TDM, the whole text is needed.
- Market (Factor 4): TDM does not substitute for sale of original (may create more market for original).

However...

Copyright and Contracts

Copyright law is the default and can be overridden by contracts.

These can include **Work for Hire** contracts (in which the material you create under employment belongs to the employer - important when you get a job!), **grant requirements**, or **Terms of Service** (contracts for using a particular platform or tool).

Know in advance what contracts you are working under & where to find them.

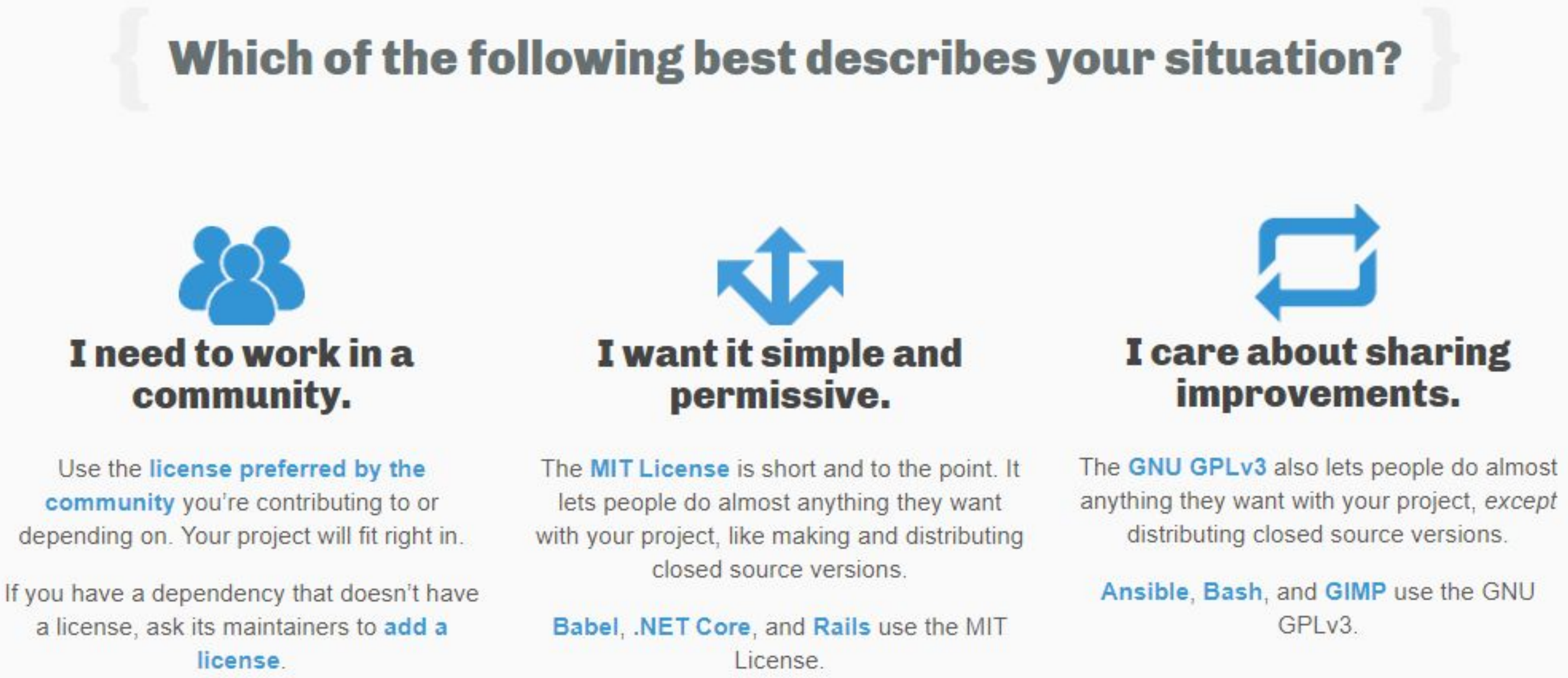
Copyright and Licenses

A license is a kind of contract that copyright owner can put on their work to allow other people to use it under specific rules.

If you own copyright to something you created, you can license it!

Common licensing tools are **Creative Commons** (for text), **GNU** (open source), **MIT** (code, very permissive).

For licensing software and code, we recommend using the Choose A License tool.



Which of the following best describes your situation?

I need to work in a community.

Use the **license preferred by the community** you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to **add a license**.

I want it simple and permissive.

The **MIT License** is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

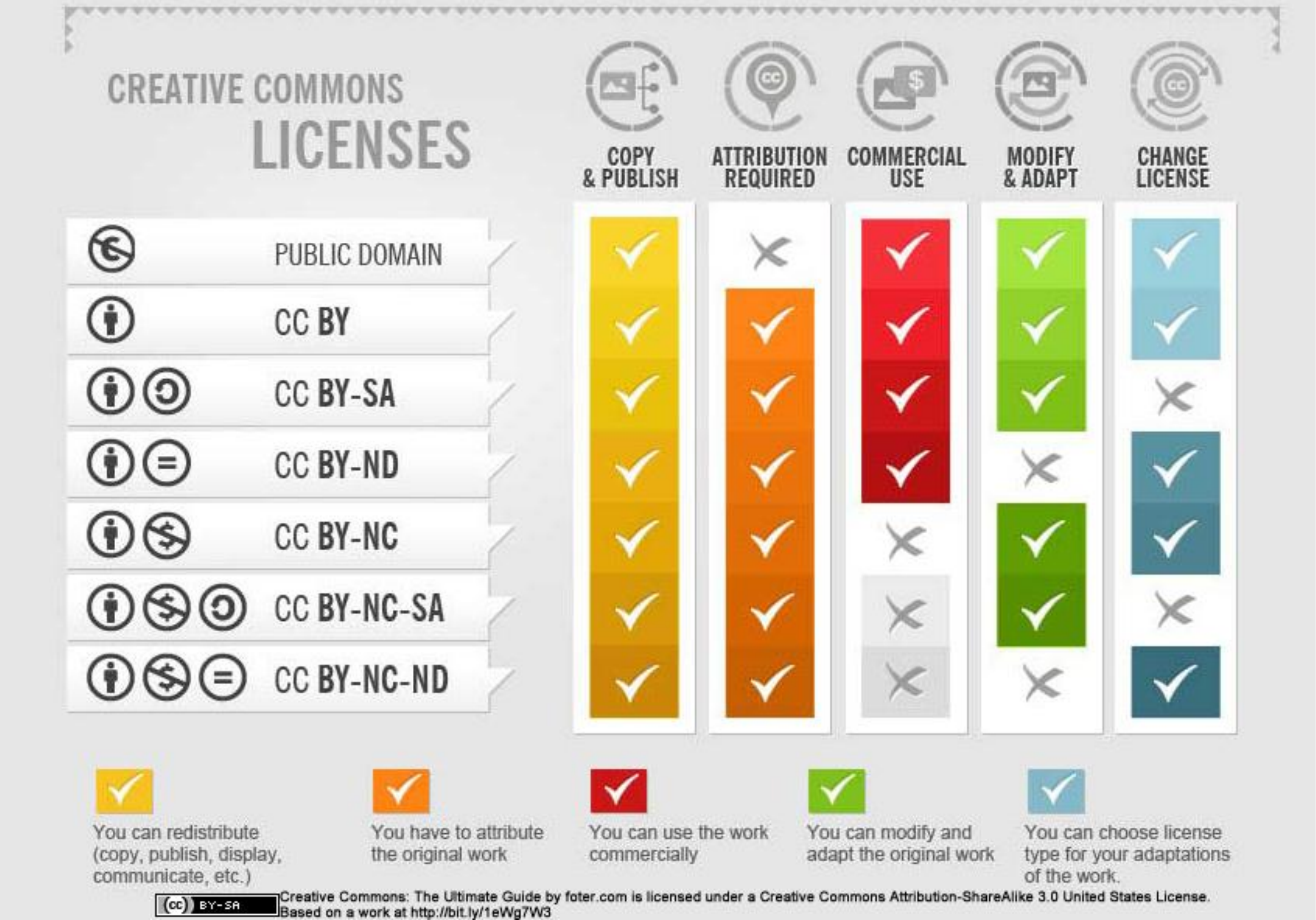
Babel, **.NET Core**, and **Rails** use the MIT License.

I care about sharing improvements.

The **GNU GPLv3** also lets people do almost anything they want with your project, *except* distributing closed source versions.

Ansible, **Bash**, and **GIMP** use the GNU GPLv3.

from http://choosealicense.com



CREATIVE COMMONS LICENSES

	COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
PUBLIC DOMAIN	✓	✗	✓	✓	✓
CC BY	✓	✓	✓	✓	✓
CC BY-SA	✓	✓	✓	✓	✗
CC BY-ND	✓	✓	✓	✗	✓
CC BY-NC	✓	✓	✗	✓	✓
CC BY-NC-SA	✓	✓	✗	✓	✗
CC BY-NC-ND	✓	✓	✗	✗	✓

✓ You can redistribute (copy, publish, display, communicate, etc.)

✓ You have to attribute the original work

✓ You can use the work commercially

✓ You can modify and adapt the original work

✓ You can choose license type for your adaptations of the work.

Discussion / Work Time

Figure this out for the dataset(s) you’ve been working with for this class:

1. Is the data subject to copyright?
2. Who owns the data?
3. Did you have to sign a license, contract, or other agreement to use it?
    a. If so, where is that?
4. Are you using the data in compliance with these terms?
    a. No judgments if not! But how can you make it better?

Sharing

Be as open as possible, as closed as necessary.

Reproducible Research

“consistent results using the same input data; computational steps, methods, and code; and conditions of analysis” (NASEM)

Reproducibility is a core part of science and research; without verifiable findings (including shared data sets, code, and methods), results are unbelievable at best.

Sharing

Sharing data helps to preserve the health of our field.

Data sharing helps with

- Replication of scholarly work.
- Advancement of new studies without the need to collect new data.
- Collaboration on future projects.
- Cross-disciplinary work.
- Citations and overall elevation of your scholarly profile.
- Creating accessible cultural and historical resources for the community.
- Students can use it in their classes (e.g. YOU!).

What information gets published?

Argument:
Data - even data that aren’t part of a paper, or ones that don’t really show anything new or significant or support your hypothesis - should be shared.

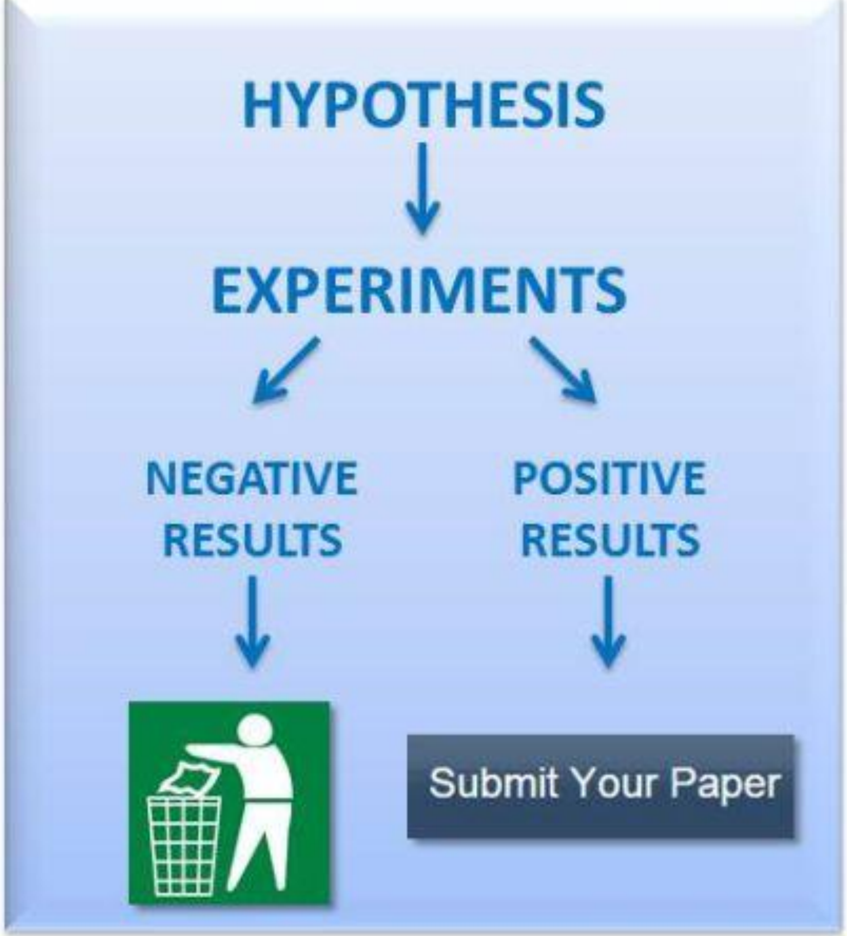


Figure 1: The most common approach taken by journals, in which only those experiments yielding positive results end up as publication material.

What is “Open Data”?

Open Data are data that fit the following specifications:

- Free to access for anybody
- Free to use for any reason
- Free to add to and build upon
- Free to share with others
- Findable (well described, preferable in a community repository)
- Interoperable with community-standard tools

Data that have restrictions, cost $ to access, and can only be used for certain purposes are not “open data” (even if you found it on the internet).

Licensing

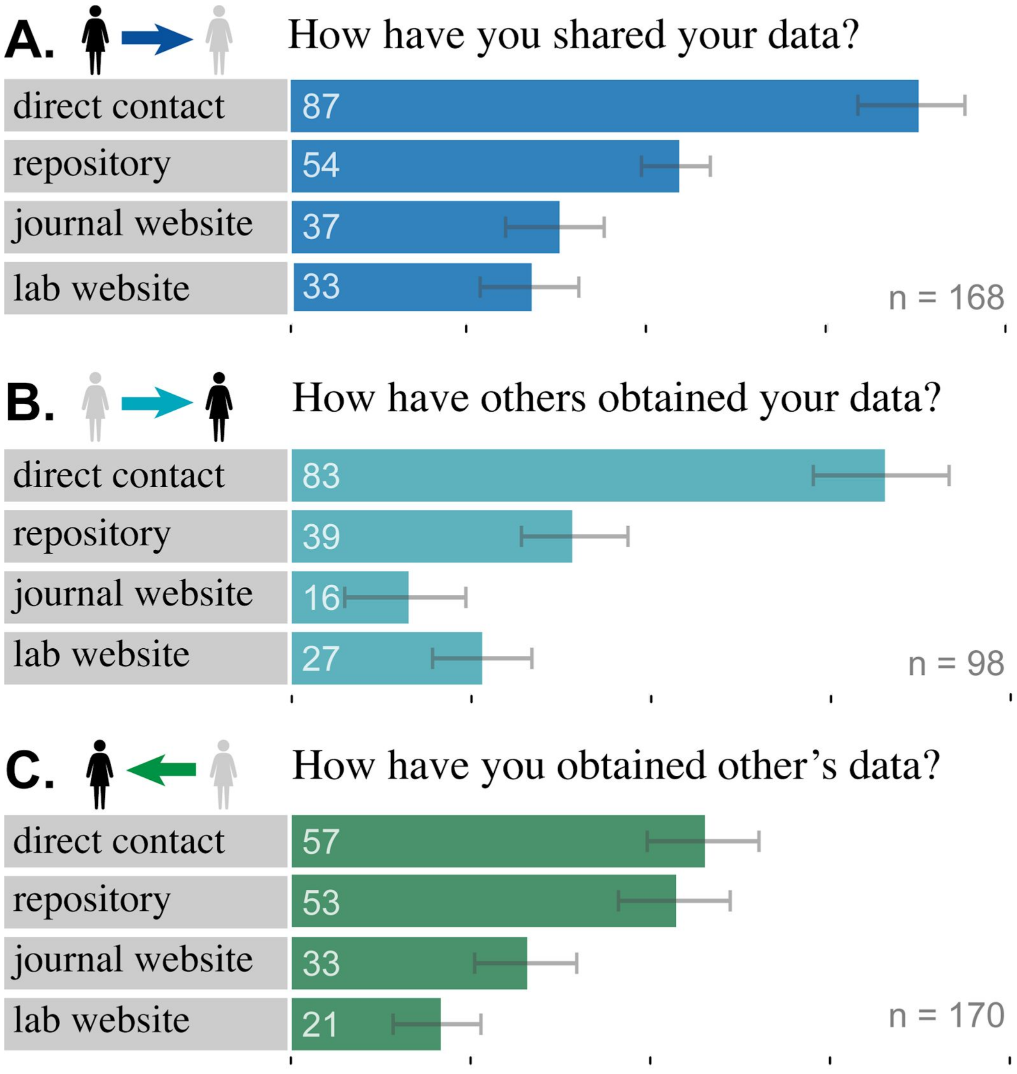
If copyright does not apply to data (e.g. quantitative data), then the dataset should be in the **public domain.**

-> Public Domain = Open Data

Some data that are subject to copyright (e.g. text files) may be licensed with **Creative Commons Licenses or Copyleft / GNU licenses,** which allows use with certain restrictions.

-> CC / Copyleft = Maybe Open / Semi-Open Data (depending on flavor)

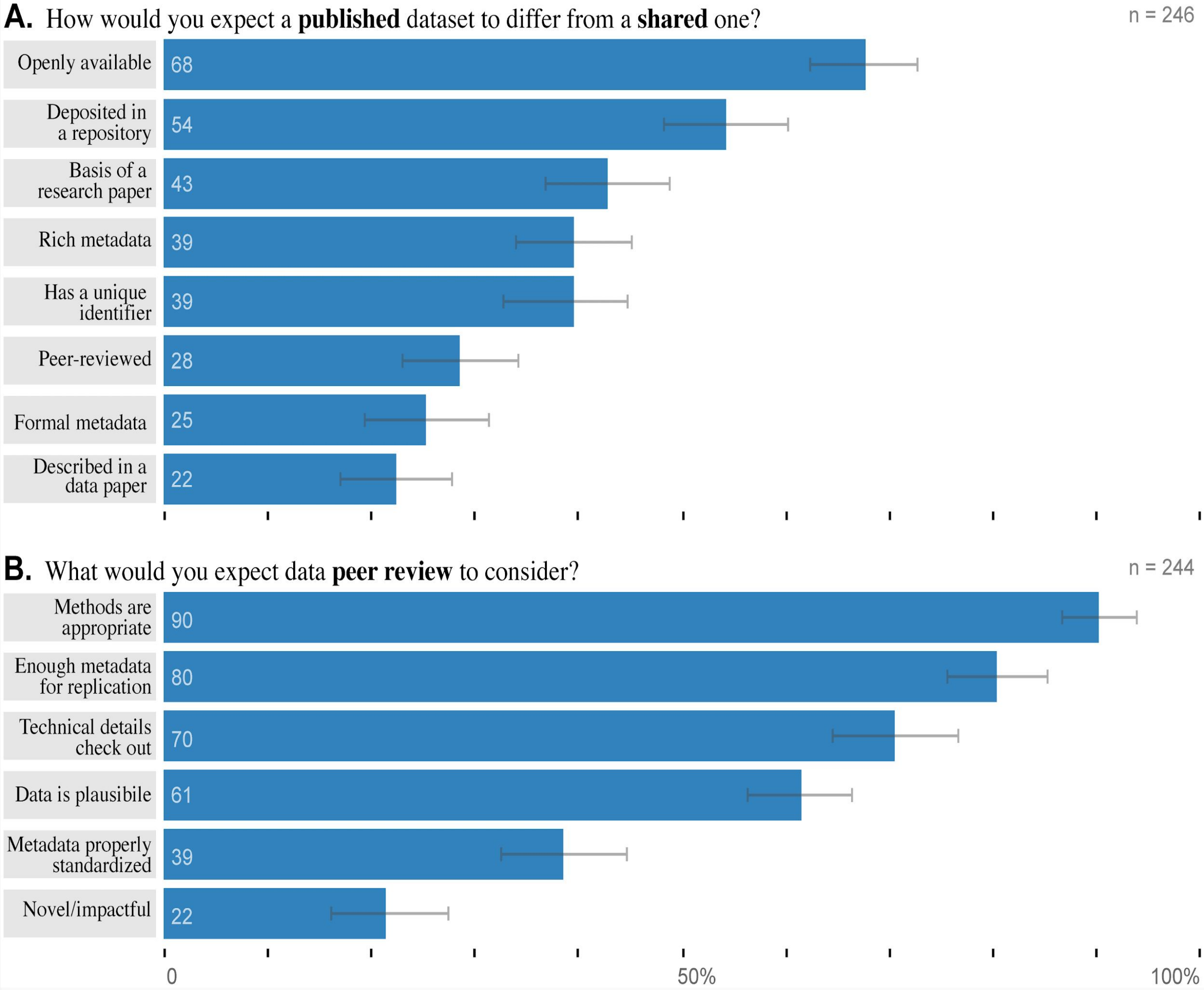
Researchers mostly share data through direct contact. Sharing ≠ Publication



Source: Kratz JE, Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. PLoS ONE10(2): e0117619. https://doi.org/10.1371/journal.pone.0117619

Nobody really knows what “data publication” means though

Source: Kratz JE, Strasser C (2015) Researcher Perspectives on Publication and Peer Review of Data. PLoS ONE10(2): e0117619. https://doi.org/10.1371/journal.pone.0117619



What about “as closed as necessary”?

Just because you (legally) can, doesn’t mean you (ethically) should.

Some data are sensitive, like

- Medical records
- Cultural heritage
- Sexual preferences (see 2016 OKCupid data shenanigans)

According to generally accepted research ethics for the social sciences, these should be protected: de-identified, access barriers, embargoes.

Have a question?
Found something unusual?

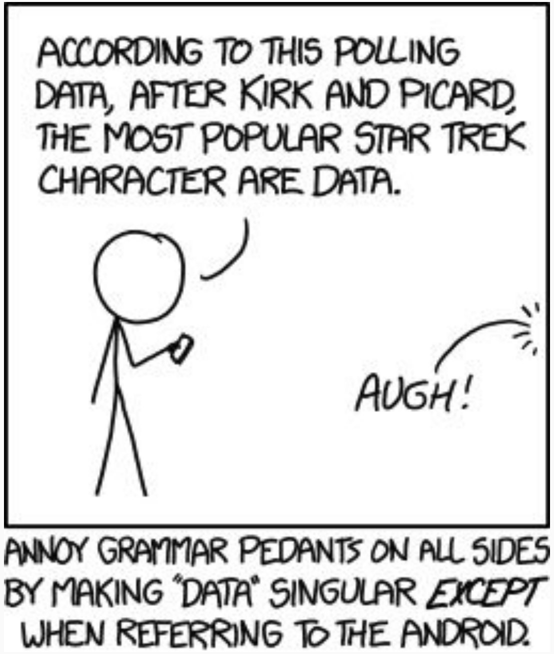
There are many librarians and specialists available to help you, whether it's about finding data sets, using tools to analyze data, or sharing a dataset that you've made! See our support services at these links for much more information to get started.

[Digital Scholarship Services]

[Support for Research Data Management]

[Research Data Management @ Pitt: Planning for Data Sharing]

Data Types & Resources


ACCORDING TO THIS POLLING DATA, AFTER KIRK AND PICARD, THE MOST POPULAR STAR TREK CHARACTER ARE DATA.
AUGH!
ANNOY GRAMMAR PEDANTS ON ALL SIDES BY MAKING "DATA" SINGULAR *EXCEPT* WHEN REFERRING TO THE ANDROID.

Super mega thanks to Gesina Phillips and Tyrica Terry Kapral for their help with & ideas for the following slides!

Locations for Sharing and Finding

To Deposit or Not to Deposit by Ainsley Seago, CC-BY
dx.doi.org/10.1371/journal.pbio.1001779.g001

Archives and repositories are the way to preserve your dataset or find new data!

These are usually found in university libraries or run by research groups. They help with metadata, documentation, and accompanying materials.

Depending on your field, a Digital Language Archive (DLA), a Linguistics Data Repository, or a General Data Repository may be the right choice for you.

There are also Institutional Repositories, although these may or may not have the capacity for data.

And - lots of amazing resources for texts that you can use without worry!

Digital Language Archives

Dedicated digital language repositories

- The Archive of the Indigenous Languages of Latin America (AILLA)
- The Endangered Language Archive (ELAR)
- The Language Archive (TLA)
- Digital Endangered Languages and Musics Archive Network (DELAMAN)
- Open Language Archives Community (OLAC)
- The Tromsø Repository of Language and Linguistics (TROLLing)

Institutional repositories with Language collections

- Archivo Digital de Language Peruanas (Pontificia Universidad Católica del Perú)
- Kaipuleohone Language Archive (U of Hawai'i at Manoa)

Physical Archives with some digital collections

- Alaska Native Language Archive (ANLA)
- American Philosophical Society (APS)
- National Anthropological Archives (NAA)

Data Repositories

Subject-specific Data Repositories

- Tromsø Repository of Language and Linguistics: TROLLing
- Linguistics re3data repositories: re3data.org
- Open Access Directory's list of subject repositories
- ICPSR for Social Science Data
- Linguistic Data Consortium

Subject-agnostic Data Repositories

- Dryad
- FigShare
- Data Archiving and Networked Services (DANS)
- Zenodo
- Kaggle

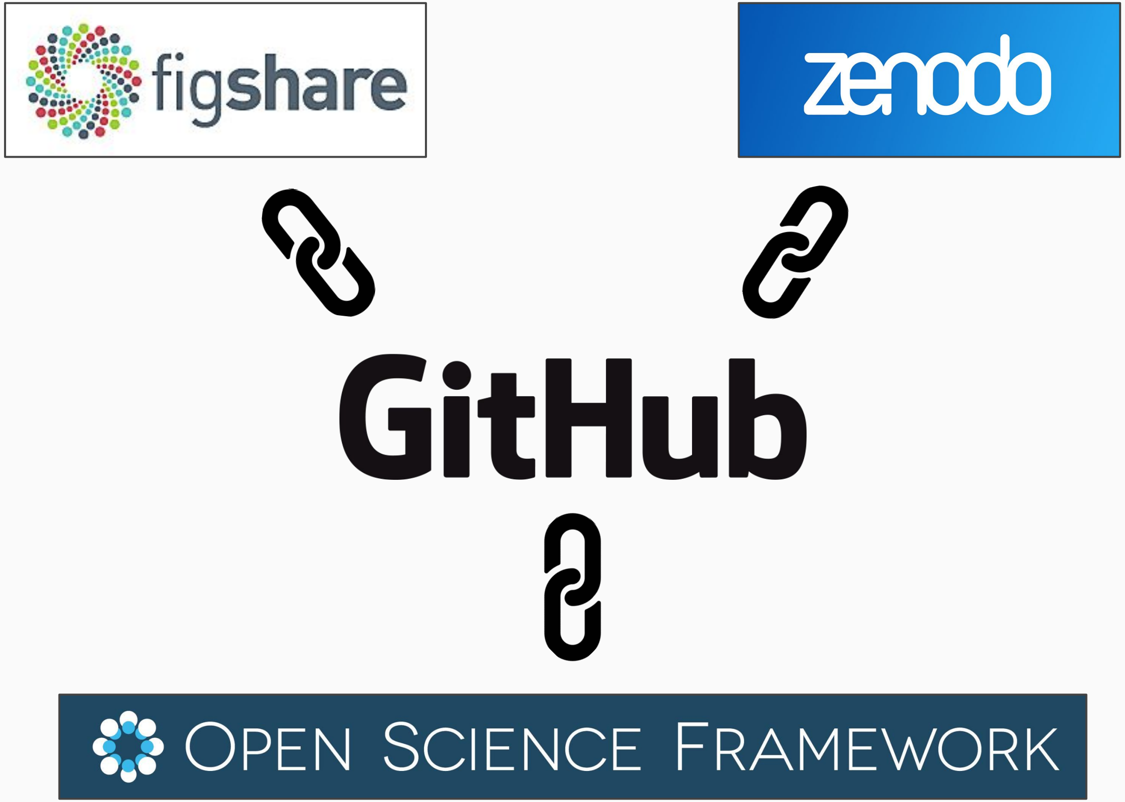
Institutional Repositories

- D-Scholarship@Pitt can support data up to 2GB per file. Bigger files? Ask us!
- The Harvard Dataverse is one of the biggest institutional data repositories in the world. It's not just for Harvard, too!
- Check out Deep Blue Data from the University of Michigan to find neat language datasets.

Github integration

These (generalist) repositories feature convenient integration with Github:

- figshare
- OSF (Open Science Framework)
- Zenodo



Icons made by [Creaticca Creative Agency]() from [www.flaticon.com]()

More Public Domain & Friendly Licensed Data Resources

- **BYU corpora:** https://corpus.byu.edu/ [Corpora developed by Mark Davies, Linguistics Professor at BYU]
- **Chronicling America:** http://chroniclingamerica.loc.gov/ [Database of historical newspapers dating from 1789 to 1922 provided by the Library of Congress]
- **Early English Books Online:** http://www.bodleian.ox.ac.uk/eebotcp/ [Corpus of Early English literature, from 1473 to 1700]
- **Folger Shakespeare Library:** http://www.folgerdigitaltexts.org/ [A digital collection of Shakespeare's plays, poems, and sonnets]
- **HathiTrust:** https://www.hathitrust.org/ [Collection of millions of titles digitized from libraries around the world]
- **JSTOR Data for Research:** http://about.jstor.org/service/data-for-research [A self-service tool that enables exploration of more than 7 million journal articles and a set of 19th Century primary resources]
- **The Oxford Text Archive:** http://ota.ox.ac.uk/ [A collection of electronic literary and linguistic resources]
- **Wikipedia List of Text Corpora:** https://en.wikipedia.org/wiki/List_of_text_corpora

Cite Your Data Properly

Follow the Tromsø Recommendations for Citation of Research Data in Linguistics (explanation of citation components and practices, lots of examples)

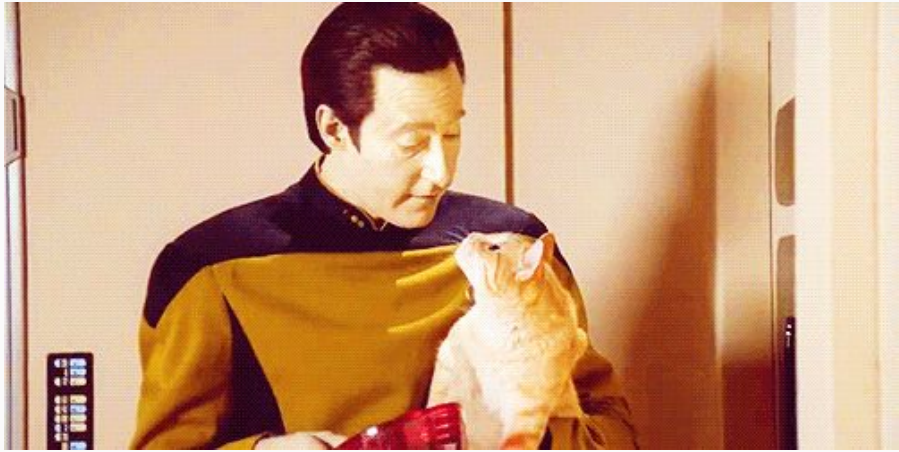
Full bibliographic reference (**green/bold = required**, *purple/italics = recommended*):

Author, *Other Attribution (Roles),* **Date, Title, Publisher, Locator,** *Version*, *Date accessed*, *Tag.*

Discussion / Work Time

1. Where did you get your data that you use for this class? Was it a repository, archive, or elsewhere?
2. What's the license on your data? Is it in the public domain, or CC/Copyleft?
   a. Is there no license on your data? Who can you ask about or report this problem to?
3. Is your data Open Data, Semi-Open Data, or Closed Data?
   a. Was it easy to find?
   b. Should these data be open at all? Check this decision tree.
4. How do you cite your data?

General Questions & Discussion



You can also contact us any time about this topic!

Lauren B. Collister, Ph.D.
E-mail: lbcollister@pitt.edu
Twitter: @parnopaeus
Consultations: bit.ly/TalkToLauren

Dominic Bordelon
E-mail: djb190@pitt.edu
Consultations: bit.ly/TalkWithDominic

Slides available: http://bit.ly/dsfl2020