

# Workshop on Data Management Plans for Linguistic Research



LSA Summer Institute, University of Kentucky, July 29-30, 2017

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

Lauren Collister, University of Pittsburgh

Susan Smythe Kung, University of Texas at Austin



Sponsored by the National Science Foundation SMA-1447886

# Info Session 1: Welcome and workshop overview



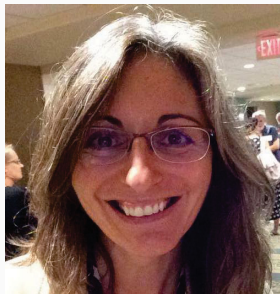
Sponsored by the National Science Foundation SMA-1447886

# Who we are

Andrea Berez-Kroeker

U Hawaii at Manoa

[andrea.berez@hawaii.edu](mailto:andrea.berez@hawaii.edu)



Lauren B. Collister

U Pittsburgh

[lbcollister@pitt.edu](mailto:lbcollister@pitt.edu)



Susan Kung

U Texas at Austin

[skung@austin.utexas.edu](mailto:skung@austin.utexas.edu)



Sponsored by the National Science Foundation SMA-1447886

# Aims for this workshop

To understand what a DMP is and why it's important.

To understand the components of a DMP.

To learn where to turn to fill in the gaps in our knowledge.

To draft a DMP for a real (or imaginary) linguistic research project.



Sponsored by the National Science Foundation SMA-1447886

# What this workshop is not

This workshop is not a basic data collection class.

This workshop is not a basic data management class.

However, we *will*

remind you of digital best practices and

show you how to describe your plans for your digital data to satisfy a DMP.



Sponsored by the National Science Foundation SMA-1447886

# Code of Conduct

We include in our definition of harassment the deliberate “scooping” or stealing of ideas from your colleagues without express written agreement and appropriate attribution.

This means that if you hear a great idea from another workshop participant and you want to use it, ask first. If this work is their intellectual property, they reserve the right to deny your request. For those concerned about their intellectual property, you are not required to share any such information and are free to use placeholders or dummy text in any documentation. However, idea sharing as happens in this workshop can not only improve the project, but also lead to fruitful collaborations and partnerships in the future.

[Link for more info & reporting info.](#)



Sponsored by the National Science Foundation SMA-1447886

# Schedule - Overview

7 Information Sessions followed by Working Sessions in small groups...

Overview of DMPs

Data collection

Legal and ethical considerations

Backup and storage

Documentation and metadata

Selection and preservation

Resources and responsibilities

...With plenty of time for questions and discussion.

Writing and peer review at end of day 2.



Sponsored by the National Science Foundation SMA-1447886

# Schedule - Detailed

## SATURDAY

9:00	Start
9:15-10	1: Overview
10-10:15	Break
10:15-11:15	2: Data collection
11:15-11:45	Questions / Discussion
11:45-1:15	Lunch
1:15-2:15	3: Legal and ethics
2:15-3:15	4: Backup & Storage
3:15-3:45	Break
3:45-4:45	5 Documentation & Metadata
4:45-5:00	Questions / Discussion
5:00	End

## SUNDAY

9:00	Start
9-10	6: Selection and Preservation
10-10:15	Break
10:15-11:15	7: Resources and Responsibilities
11:15-11:45	Questions / Discussion
11:45-1:15	Lunch
1:15-2:15	Writing
2:15-3:00	Peer Review
3:00-3:15	Break
3:15-4:15	Revisions/discussion
4:15	Sharing, including final questions / discussion
5:00	End



# What is a DMP?

A DMP is

- a written document
- outlining plans for handling (collecting, describing, organizing, processing, analyzing, preserving, sharing)
- all of the data resulting from a research project
- in the short term and the long term.



Sponsored by the National Science Foundation SMA-1447886

# What is a DMP?

A DMP includes

- Detailed procedures for data collection
- All aspects of organization and processing *before your data leaves your lab*
- *A plan to have data leave your lab* so that others can find and access it in perpetuity
  - with proper attention to legal and ethical concerns



Sponsored by the National Science Foundation SMA-1447886

# Why do we need DMPs?

Because digital data has  
a few problems with longevity.



Sponsored by the National Science Foundation SMA-1447886

# Digital data problems with longevity

Three central problems need to be solved:

The **media** problem

The **format** problem

The **storage and access** problem



Sponsored by the National Science Foundation SMA-1447886

# The media problem

The more advanced our technology becomes, the more ephemeral it is:

- Hard drives: 5 years <
- CDs/DVDs: 10 years <
- Cassette tapes: 30 years <
- Paper: 100-200 years (+) <
- Stone tablets:  $\infty$



Sponsored by the National Science Foundation SMA-1447886

# The media problem

Not only do media degrade...

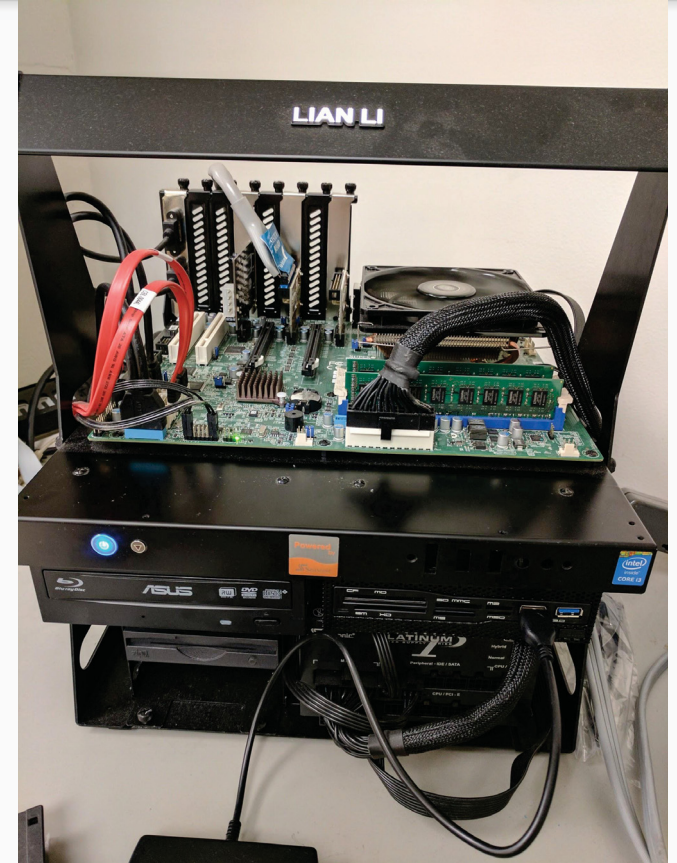


...devices for reading them become obsolete!



Sponsored by the National Science Foundation SMA-1447886

...requiring data rescuers and archivists to use machines like “Frank”



Sponsored by the National Science Foundation SMA-1447886

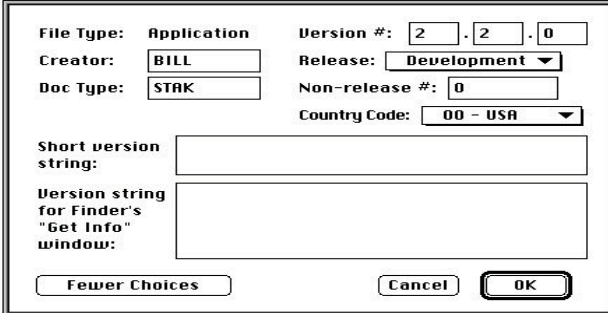
# The format (or encoding) problem

**Proprietary formats** are controlled by intellectual property law and are subject to the whims of the developers

- Cease development or support
- Charge fees to access data

Example: Hypercard dictionaries (eg Gwich'in)

- Data now ostensibly lost



A screenshot of a Hypercard dictionary metadata dialog box. The dialog contains the following fields and controls:

- File Type:** Application
- Version #:** 2 . 2 . 0
- Creator:** BILL
- Release:** Development (dropdown menu)
- Doc Type:** STAK
- Non-release #:** 0
- Country Code:** 00 - USA (dropdown menu)
- Short version string:** (text input field)
- Version string for Finder's "Get Info" window:** (text input field)
- Buttons:** Fewer Choices, Cancel, OK



Sponsored by the National Science Foundation SMA-1447886



# The storage & access problem

Data **cannot be effectively stored for longevity by individuals**, who

- Lack expertise in data migration to new formats
- Inevitably lose interest, retire, or die

Only an **archive** with an institutional commitment to migrating and backing up data is an effective locus of long-term storage



Sponsored by the National Science Foundation SMA-1447886

# The storage & access problem

Data must be **discoverable** and (correctly, ethically) **accessible**.

Without proper metadata, we don't know anything about the data...

...or even that it exists!

Data that isn't accessible by anyone is useless.



Sponsored by the National Science Foundation SMA-1447886

# A DMP addresses these problems

A DMP is your plan to protect your digital data.

It helps YOU.

It helps YOUR FUNDER.

It helps YOUR RESEARCH.



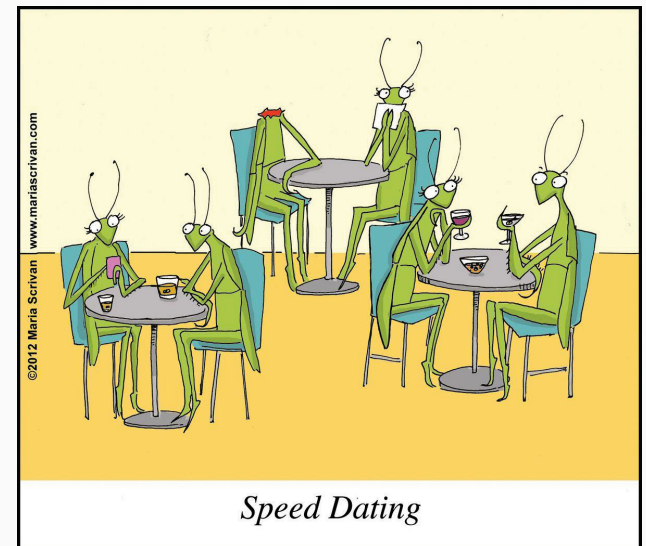
Sponsored by the National Science Foundation SMA-1447886

# Working Session 1: Speed data-ing!

Take 5-10 minutes to think about your project.

- What is the nature of your research project?
- What research questions are you addressing?
- What kind of data will your project produce?

Then be ready to briefly summarize your study to your speed data-ing partners in **one minute!**



# Info Session 2: Data Collection



Sponsored by the National Science Foundation SMA-1447886

# Thinking about your data for your DMP

DMPs are less about the *contents* of your data, and more about the *digital parameters of your data*.

NOT:

“I will collect sentences, paradigms, and grammaticality judgments.”

(That goes in your project description!)



Sponsored by the National Science Foundation SMA-1447886

# Thinking about your data for your DMP

BUT RATHER:

“I will collect WAV audio recordings (44.1Khz/16bit) and TIFF images of my field notebooks.”

DMPs are about data types, file formats, resolution, quantity...

...and **convincing the funder** you know the current best practices for digital data.



Sponsored by the National Science Foundation SMA-1447886

# Thinking about your data for your DMP

Helps to think about the *content* of your data to determine the digital parameters.

Language documentation → narratives, elicitation, grammaticality judgments, dictionaries, IGT → audio files, video files, images, text files, XML files, databases, metadata...

Laboratory phonetics → audio files, Praat text grids, R code, metadata...

Others?



Sponsored by the National Science Foundation SMA-1447886



# Describing digital parameters of your data

What type(s) and format(s) will your data be in?

TYPES	FORMATS
Audio files	.wav, .mp3, .aiff...
Video files	.raw, .mpg, .mp4...
text	.txt, .xml, .pdf, .eaf...
database	.db, .fp7...



Sponsored by the National Science Foundation SMA-1447886

# Describing digital parameters of your data

Selecting file formats:

You need to know which formats comply with the *best practices in your field*.

- Formats need to be selected with **longevity** in mind
  - Nonproprietary and/or open-source
  - High-resolution
  - As uncompressed as possible
  - Standards-compliant (eg Unicode)



Sponsored by the National Science Foundation SMA-1447886

# Describing digital parameters of your data

Selecting file formats:

- Formats and software also need to facilitate **access**
- Balancing longevity with ease of access
  - Will you create both high-resolution and low-bandwidth versions? WAV and MP3?
  - What is current for archives? Eg Video formats



Sponsored by the National Science Foundation SMA-1447886

# Describing digital parameters of your data

What is the approximate *volume* of your data?

- In GB?
- In time?
- In number of files?
- Pages, tables, etc.



Sponsored by the National Science Foundation SMA-1447886

# What standards/methods will you use to collect data?

Does your subfield have an accepted standard or method for data collection?

If so, cite it! **Demonstrate you know the standards.**

For example, in language documentation:

Bird, Steven, & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3):557-582.

Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36: 161-195.



Sponsored by the National Science Foundation SMA-1447886

# What equipment will you use?

Be as specific as possible!

*At least list specs* of your equipment.

- NOT “Digital recordings will be made.”
- RATHER: “Digital recordings will be made with a solid-state digital recorder with external XLR condenser lavalier mic, at minimally 44.1 kHz / 16bit resolution.”

If you can, list brands and model numbers.



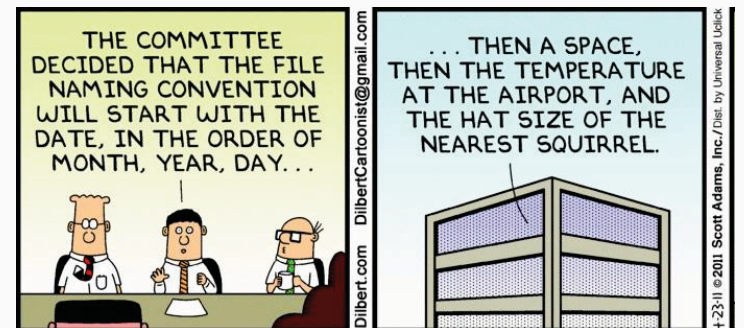
Sponsored by the National Science Foundation SMA-1447886

# How will you name and structure your files?

In file naming:

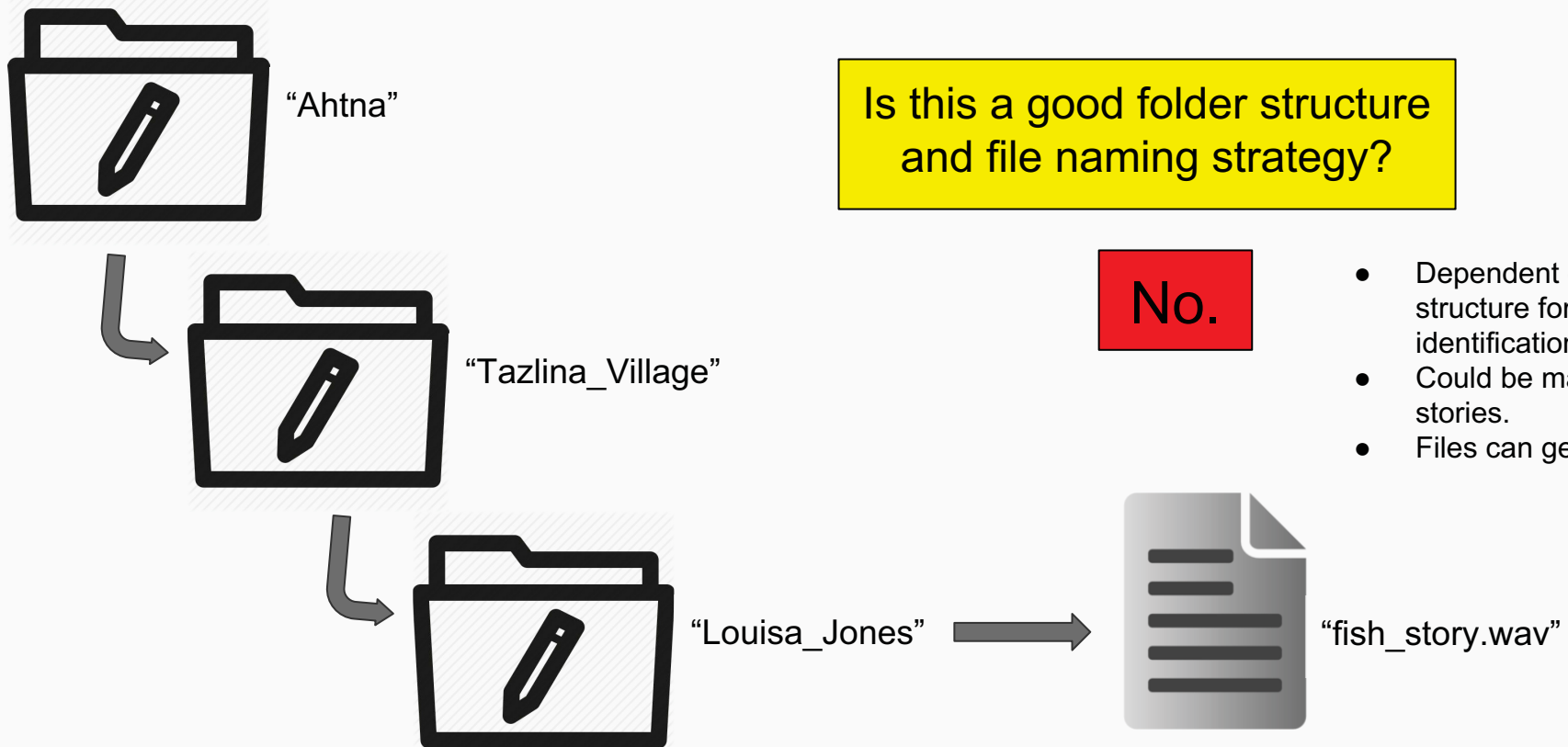
- Use a **unique ID** that is not dependent on file structure
- File names can be *semantic* or *non-semantic*
- No spaces or funny characters
- Select a convention and stick with it.
- **Check with your archive!**

Folder structure is for your convenience, but should not be used for file identification.



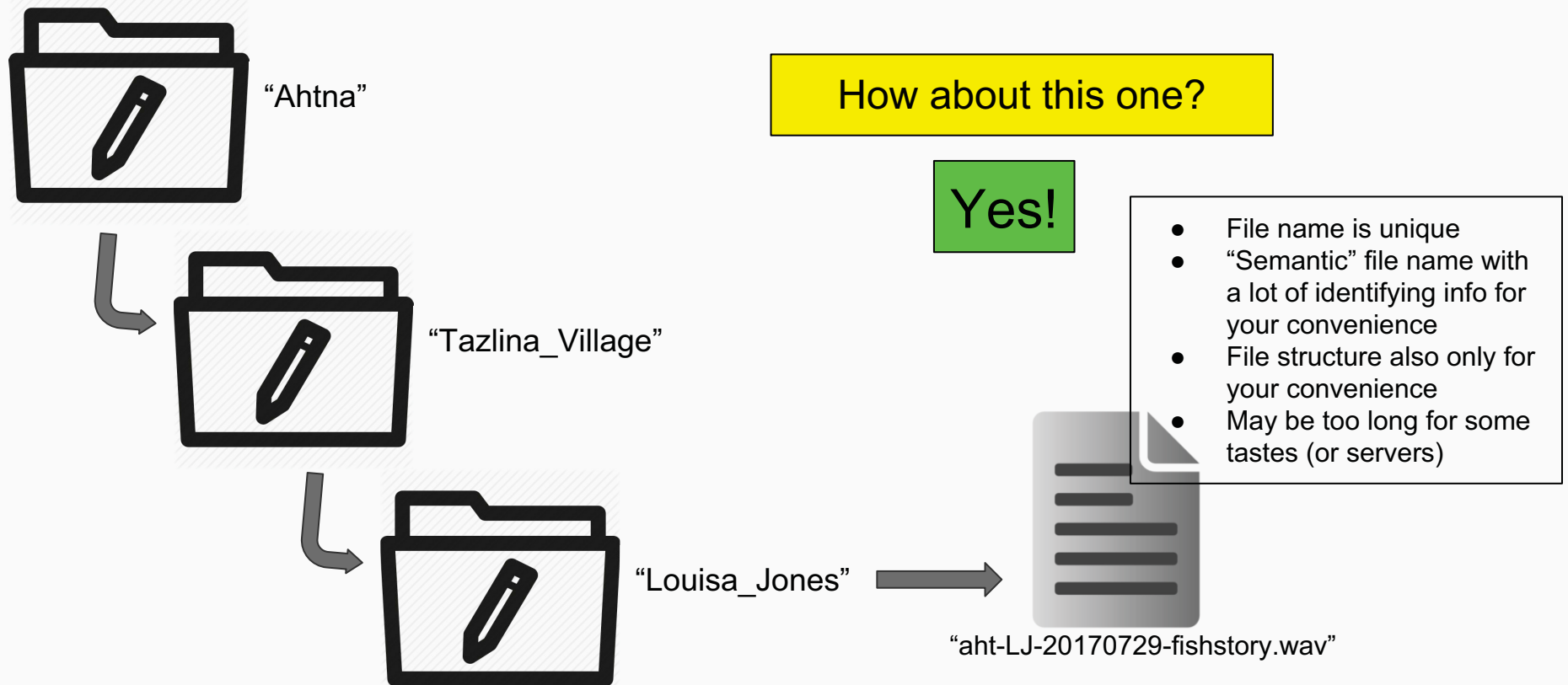
Sponsored by the National Science Foundation SMA-1447886

## Folder structure and file naming





## Folder structure and file naming



## Folder structure and file naming

Or this one?

Yes!

- Embedded file structure not necessary
- Files will order alphabetically

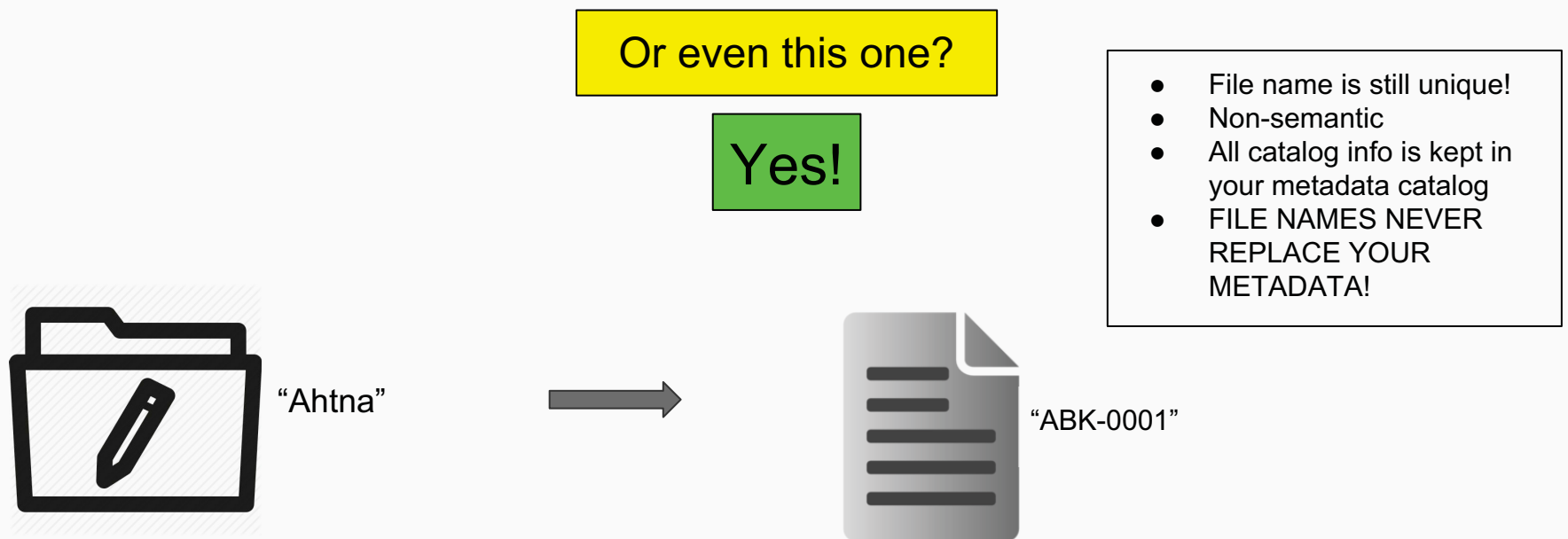


“Ahtna”



“aht-LJ-20170729-fishstory.wav”

## Folder structure and file naming



# How will you handle data versioning?



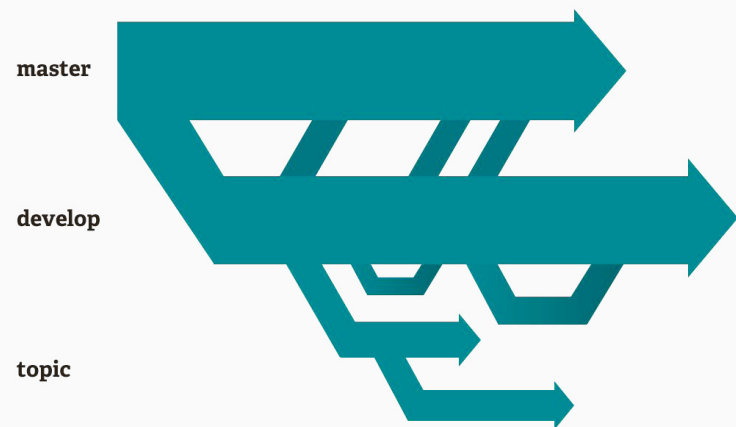
**Git** is a great tool for version control!

Free, open source, small, fast.

Keeps track of version changes across users or just you.

Designed for software dev but works on many file types.

<https://git-scm.com/>



Sponsored by the National Science Foundation SMA-1447886

# How will you handle data versioning?

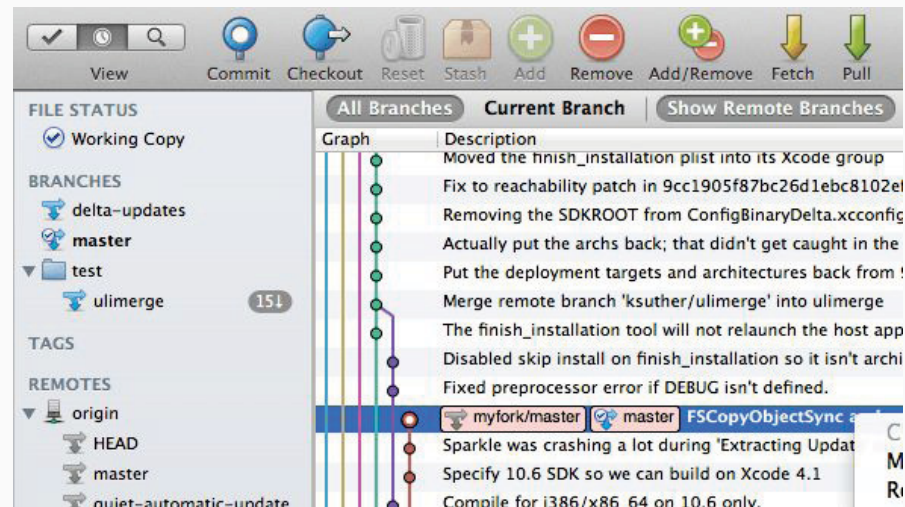


Use a GUI to make git even easier

- SourceTree [\(link\)](#)
- GitHub Desktop [\(link\)](#)
- TortoiseGit [\(link\)](#)
- GitKraken [\(link\)](#)

Also Pachyderm [\(link\)](#)

Like git, but for data.



Sponsored by the National Science Foundation SMA-1447886

# How will you handle data versioning?

Or a simple solution may fit your project:

Dates are a good versioning tool.

Use ISO format: YYYY-MM-DD

`dataset_2017-07-15.txt`

`dataset_2017-07-16.txt`

`dataset_2017-07-17.txt`

But watch out for international collaborators!

Appending version numbers (`_v002`) is ok too (don't forget to change them).



Sponsored by the National Science Foundation SMA-1447886

# What about quality assurance?



For some research, quality assurance procedures should be described.

- Inter-rater reliability (for coding)
- Periodic self-audits (for databases, metadata)
- Periodic informal presentation of findings to colleagues for feedback (for analytical results)
- Data entry validation workflow



Sponsored by the National Science Foundation SMA-1447886

# Working Session

1. What **type, format, and volume** of data will you be producing?
2. Do your chosen formats and software enable **sharing, high-resolution preservation, and long-term access** to the data?
3. Are there any **existing data** you can reuse?



Sponsored by the National Science Foundation SMA-1447886



# Working Session

## Guidance:

- Give a **brief description** of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage.
- **Outline and justify your choice of format** and consider the implications of data format and data volumes in terms of storage, backup and access (more on storage and backup later).



Sponsored by the National Science Foundation SMA-1447886

# Working Session

4. What **standards and methods** will you use?
5. How will you name and structure your **folders and files**?
6. How will you handle **versioning**?
7. What **quality assurance processes** will you adopt?



Sponsored by the National Science Foundation SMA-1447886

# Working Session

## Guidance:

- Outline **how the data will be collected/created** and which **community data standards** (if any) will be used.
- Consider **how the data will be organized** during the project, mentioning for example naming conventions, version control and folder structures.
- Explain how the **consistency and quality of data collection** will be controlled and documented.



Sponsored by the National Science Foundation SMA-1447886

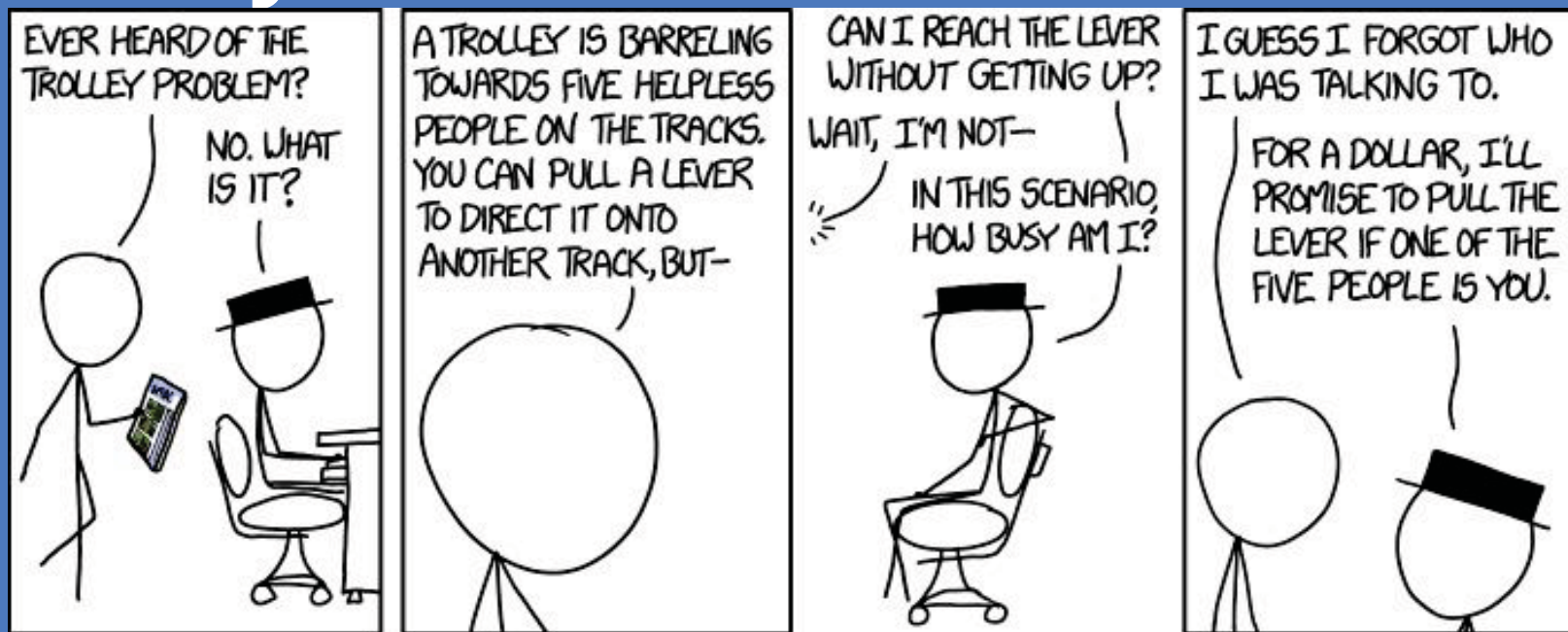
# Info Session 3: Ethics and Legal Compliance



Sponsored by the National Science Foundation SMA-1447886



# Why does this matter?



Sponsored by the National Science Foundation SMA-1447886

#### 4. Ownership.

All rights and title in and to the Service (including without limitation any user accounts, titles, computer code, themes, objects, characters, character names, stories, dialogue, catch phrases, locations, concepts, artwork, animations, sounds, musical compositions, audio-visual effects, methods of operation, moral rights, any related documentation, “applets,” transcripts of the chat rooms, character profile information, recordings of games) are owned by Blizzard or its licensors. The Game and the Service are protected by United States and international laws, and may contain certain licensed materials in which Blizzard's licensors may enforce their rights in the event of any violation of this Agreement.

“All rights and title in and to the service (including without limitation... moral rights, **any related documentation**, “applets,” **transcripts of the chat rooms**, character profile information, **recordings of games**) are owned by Blizzard or its licensors.”

[http://us.blizzard.com/en-us/company/legal/wow\\_tou.html](http://us.blizzard.com/en-us/company/legal/wow_tou.html)



Sponsored by the National Science Foundation SMA-1447886

# Don't be like 2007 Lauren.

Know your  
rights & responsibilities.

- What is copyrightable?
- What is the definition of copyright in your jurisdiction?
- Who owns the data you are using?
- What are you legally allowed to do with those data?
- What are the ethical considerations particular to your data?
- How will all of these impact your long-term plans?



# 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>



Sponsored by the National Science Foundation SMA-1447886

# 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



Sponsored by the National Science Foundation SMA-1447886

# 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
- 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>



Sponsored by the National Science Foundation SMA-1447886

# International Copyright



Make sure you know the law where you are.

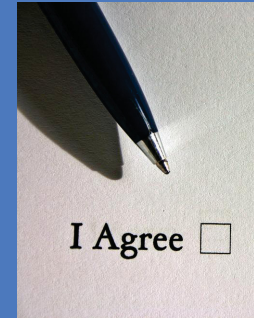
Doing fieldwork in another country? Check in with a copyright specialist before you go to make sure you don't run afoul of local laws.

- Search IP Laws and Treaties worldwide with WIPO Lex
  - <http://www.wipo.int/wipolex/en/index.jsp>
- Ask librarians at your institution or your host institution!
  - Scholarly Communications Librarian
  - Copyright Librarian
- General Counsel at your University is also a resource.



Sponsored by the National Science Foundation SMA-1447886

# 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 - check with your University!), **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.

If you hire a translator, transcriber, or other person to work with you, spell out who owns copyright in THEIR contract clearly!



Sponsored by the National Science Foundation SMA-1447886

# Copyright and Linguistics

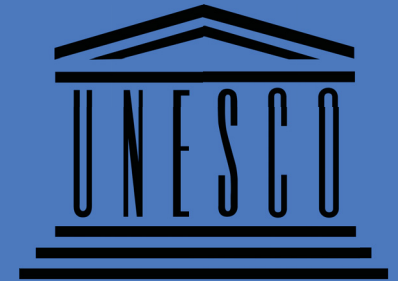


- Spoken and written language is usually an ‘original work of authorship’.
- A lexicon is a list of the contents of language - not subject to copyright
  - Though your organization, layout, and description that may accompany this wordlist *are* subject to copyright.
- Vowel measurements are not subject 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:



Sponsored by the National Science Foundation SMA-1447886

# 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>



Sponsored by the National Science Foundation SMA-1447886

# LSA Ethics Statement



"Some communities regard language, oral literature, and other forms of cultural knowledge as **valuable intellectual property** whose ownership should be respected by outsiders; in such cases **linguists should comply with community wishes regarding access, archiving, and distribution of results**. Other communities are eager to share their knowledge in the context of a **long-term relationship of reciprocity and exchange**. In all cases where the community has an investment in language research, **the aims of an investigation should be clearly discussed with the community and community involvement sought from the earliest stages of project planning.**"

(excerpt from the end of section 3, emphasis added)



Sponsored by the National Science Foundation SMA-1447886



# Sharing Data



Once you determine who owns the data, understand and plan for what you can **do** with the data. Data that isn't accessible by anybody is useless.

- Gain consent from participants at the outset for the preservation and sharing of data.
  - Include in your consent form that data will be archived and made publicly available, not just "used for research purposes" or "reproduced in scholarly works".
- Will you need to anonymize the data to share it? How will you do this?
- Are you collecting any sensitive data?
  - Security of storage and transfer - repository / archive staff can help you prepare for this.
- Will you be able to make your dataset open? Check [this decision tree!](#)



Sponsored by the National Science Foundation SMA-1447886

# More Resources

- Intellectual Property Issues in Cultural Heritage (IPinCH) <http://www.sfu.ca/ipinch/>
- Susan Smythe Kung's Bibliography: Consent, Copyright, IP, and Traditional Knowledge <http://bit.ly/IPTKbib>
- Newman, Paul. 2007. Copyright Essentials for Linguists. *Language Documentation & Conservation* 1(1): 28-43. Available from <https://scholarspace.manoa.hawaii.edu/bitstream/10125/1724/6/newman.pdf>
- Linguistic Society of America's Ethics resources <https://www.linguisticsociety.org/resource/ethics>
- Copyright Crash Course at UT Libraries <http://guides.lib.utexas.edu/copyright/>
- "Issues of consent, copyright, intellectual property and traditional knowledge: What they mean for digital language archives" Slides by Susan Smythe Kung <https://goo.gl/F3CXqM>



Sponsored by the National Science Foundation SMA-1447886

# Working Session

- For everyone: What parts of your data are subject to copyright or not?
- Choose one of the following, depending on your data situation:
  - Are you working under any contracts?
    - Identify any contracts that may apply, find them, and skim through for an “Ownership” section.
  - Are you working in different countries that may have different laws?
    - Identify a page to read at [WIPO Lex](#) and/or a person at your institution who you can ask for help.
  - Do you have any ethical considerations?
    - List any potential anonymization or privacy concerns and brainstorm on methods for working with them.



# Info Session 4: Storage and Backup



Sponsored by the National Science Foundation SMA-1447886

# Storage and Backup

Data must be **protected during collection and processing** (“in the field and lab”)

Protected for integrity

Protected for security and access

This is not the same as your plans to keep data safe, secure, and accessible after it leaves your lab

Although they may overlap in execution.



Sponsored by the National Science Foundation SMA-1447886

# Storage and Backup for data integrity

Your data is vulnerable during collection and processing!

Electronic/digital dangers:

Broken drives, power surges, viruses

Environmental dangers:

Water damage, fire damage, insects, mold

Human dangers:

Theft, loss, overwriting, dropping/crushing

Copyright 2005 by Randy Glasbergen.  
www.glasbergen.com



"We back up our data on sticky notes because sticky notes never crash."



Sponsored by the National Science Foundation SMA-1447886

A good rule to  
remember



**LOCKSS:**

**Lots Of Copies Keep  
Stuff Safe!**

<https://www.lockss.org/>

# Storage and Backup for data integrity: LOCKSS



Your DMP should describe how you plan to utilize LOCKSS during collection and processing:

*How* will you redundantly backup your data?

In the field:

Multiple hard drives? Flash cards?

Where will they be stored?



Sponsored by the National Science Foundation SMA-1447886



One in your cabin...



...one in your car...



...and one at the cultural center.



Sponsored by the National Science Foundation SMA-1447886

# Storage and Backup for data integrity: LOCKSS



In the lab:

Will you use IT-managed storage at your university?

IT-managed storage usually has built-in LOCKSS procedures. Confirm this.



Sponsored by the National Science Foundation SMA-1447886

# Another approach to LOCKSS: The 3-2-1 Principle



(At least) 3 copies on

(At least) 2 types of storage media\* with

(At least) 1 off-site

\*Different brands of hard drive, or a hard drive and flash storage, or a hard drive and DVDs, or....



Sponsored by the National Science Foundation SMA-1447886

# What about cloud storage?



Fine for convenience and sharing with collaborators.

Not to be considered primary backup, ever (not even worth mentioning in DMP)

Considerations: data ownership, cost, security, going out of business (eg Wuala).

Higher security: SpiderOak, Tresorit. Easier, more common: Google Drive, DropBox, iCloud



Sponsored by the National Science Foundation SMA-1447886

# Storage and Backup for data integrity: Other questions to think about

Do you need to include **money in your budget** for backup storage?

Purchasing hardware; paying fees

Who is responsible for regular backup?

What is your field-to-lab data transfer protocol?

How will data be recovered in the event of an incident?



Sponsored by the National Science Foundation SMA-1447886

# Storage and Backup for data security



What are the risks to security?

For language data: usually *confidentiality* is the biggest risk.

You should be guided by your IRB (see *Ethics*).

Plans to anonymize participants, if warranted.

Plans to secure the anonymization key.



Sponsored by the National Science Foundation SMA-1447886

# Storage and Backup for data security



If necessary, how will you control access to keep data secure?

Data can be password protected.

Don't forget to share passwords with the research team.

For more sensitive data, tougher standards can be followed, eg. [ISO 27001](#):  
"provides a model for establishing, implementing, operating, monitoring, reviewing, maintaining and improving an information security management system."



Sponsored by the National Science Foundation SMA-1447886

# Working Session

1. How will data be stored and backed up during the research?
  - a. Do you have sufficient storage, or will you need to include charges for additional services?
  - b. How will the data be backed up?
  - c. Who will be responsible for backup? For recovery?
  - d. How will the data be recovered in the event of an incident?





# Working Session

## Guidance:

- State how often the data will be backed up and to which locations. How many copies are being made?
- Storing data on laptops, computer hard drives or external storage devices alone is very risky. The use of robust, managed storage provided by university IT teams is preferable.
- Similarly, it is normally better to use automatic backup services provided by IT Services than rely on manual processes.
- If you choose to use a third-party service, you should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data.



# Working Session

## 2. How will you manage access and security?

- a. What are the risks to data security and how will these be managed?
- b. How will you control access to keep the data secure?
- c. How will you ensure that collaborators can access the data securely?
- d. If creating or collecting data in the field how will you ensure its safe transfer into your main secured systems??



# Working Session

## Guidance:

- If your data is confidential (e.g. personal data not already in the public domain, confidential information or trade secrets), you should outline any appropriate security measures and note any formal standards that you will comply with, e.g. ISO 27001.

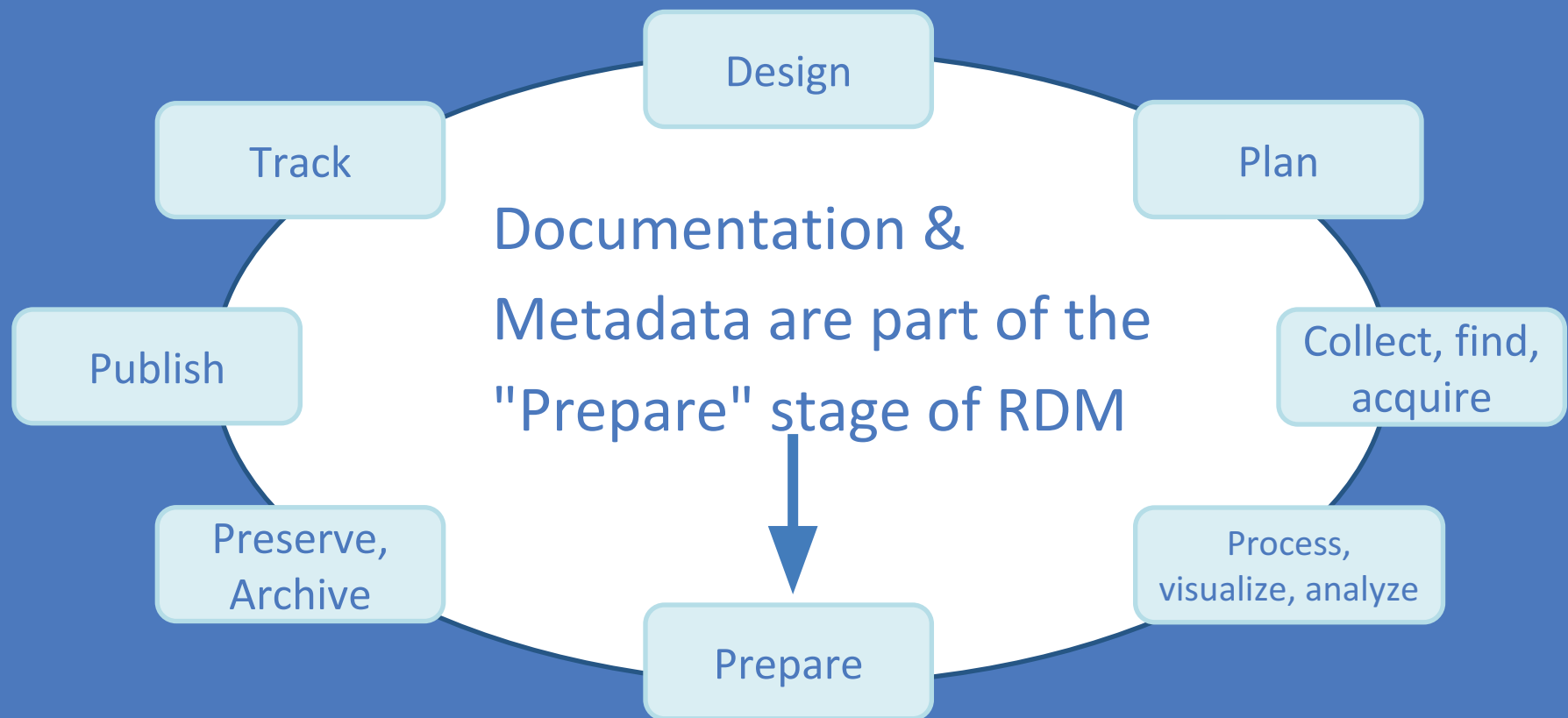


Sponsored by the National Science Foundation SMA-1447886

# Info Session 5: Documentation and Metadata



Sponsored by the National Science Foundation SMA-1447886



Sponsored by the National Science Foundation SMA-1447886

# Documentation



Documentation is any sort of digital or analog document(s), written or recorded, online or off, that provide important, contextual information. For example:

- user guides
- readme files
- white papers
- FAQs
- protocols
- manuals
- lab books
- hardware and software configurations
- workflows
- etc.

# Documentation

Documentation for a research project will include

- an overview of the project,
- the methodology for various components of the project,
- worklogs or logbooks kept during the project,
- protocols,
- lists,
- etc.

Document daily! Or at least as often as necessary. You will not remember the details forever (or even the next day!).



Sponsored by the National Science Foundation SMA-1447886

# Metadata

- Metadata is structured information that describes, explains, locates, or otherwise represents the research data.
- Metadata make it easier to find, retrieve, (re-)use, manage, understand, and cite the data. Metadata can be descriptive, technical, administrative, or structural.
- Metadata creation is best done by the data collector / creator at the time of data collection.

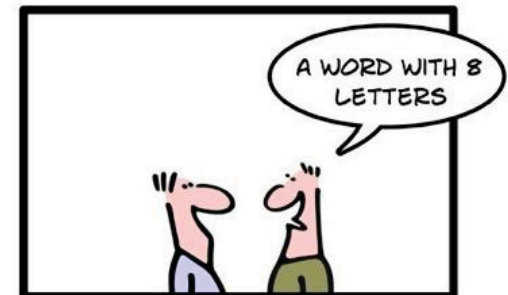
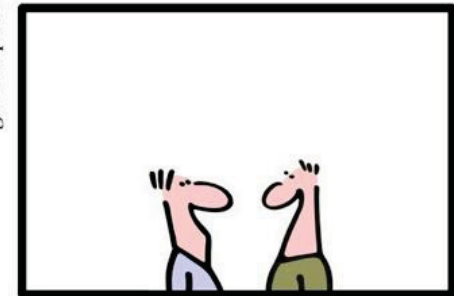


Sponsored by the National Science Foundation SMA-1447886

## SIMPLY EXPLAINED: METADATA



geek & poke





# Types of Metadata: Descriptive

Descriptive metadata are used for discovery, identification and retrieval (e.g., names, dates, languages, locations, keywords).



Sponsored by the National Science Foundation SMA-1447886

# Types of Metadata: Technical

Technical metadata are the technical details about a file (e.g., size) or the production of that file (e.g., sampling rate, recording equipment or programs).



Sponsored by the National Science Foundation SMA-1447886

# Types of Metadata: Administrative

Administrative metadata include details about how to manage the file (e.g., intellectual property, restrictions)

## **Level 1. Public access**

Users have full access to Level 1 files after agreeing to our [Terms and Conditions](#) and logging in.

## **Level 2. Password**

Level 2 files are password protected. They might have an associated hint that is meant to give access to the depositors' colleagues. These files cannot be made public access for a variety of reasons.

## **Level 3. Time limit**

Level 3 files are password protected until a specified date at which time access will change to Level 1. Example: 2050-01-01 (January 1, 2050.) This option is used for resources that are sensitive only for a period of time; for example, the lifetime of the narrator or the five years it takes a student researcher to finish his/her thesis.

## **Level 4. Depositor control**

Users must contact the depositor or a representative of the speech community directly to ask for permission to access Level 4 files. When you click on the file name, the contact person's email address will appear. If you write to the contact person and he/she does not respond, please [contact us](#).

# Types of Metadata: Structural

[Home](#) » [Collections](#) » [Hup Collection of Patience Epps](#) » [Adjective tone set \(P\)](#) » JUP004R007I001.mp3

## JUP004R007I001.mp3

[View](#) [Manage](#)

### Object Details

Language(s)	Hup
Language PID(s)	ailla:119685
Content type	primary text
Date Created	2004
Date Archived	2010-04-29
Technical Description	
Length	0:2:16
Encoding Specifications	kbps 64
Platform	pc, soundforge or audacity
Original Medium	audio:DAT

Structural metadata explains how files are organized in relationship to each other.

[Home](#) » [Collections](#) » [Hup Collection of Patience Epps](#) » [Adjective tone set \(P\)](#) » JUP004R007I001.mp3



Sponsored by the National Science Foundation SMA-1447886

# What information is needed for data to be read & interpreted in the future?

## Documentation

- **Programs:** title & version, proprietary vs. open source, settings, filters, template, etc. used for your study.
- **Methodology:** detailed description of what you did so you or someone else can reproduce (or understand) it later.
- **Experiments:** tools, tests, protocols used.

## Metadata

- **Who:** data creator/researcher, other research participants like speakers, videographers, transcribers, etc.
- **What:** title of data set
- **When:** data of creation / experiment
- **Where:** location of data creation

# How will you capture or create Documentation and Metadata?

**Documentation:** Keep documentation notes as you go in a work log or log/lab book, clearly label all files.

**Metadata:** spreadsheet, text files, notebook or log/lab book, database program for metadata (e.g., SayMore, ToolBox, Arbil, CMDI Maker, FileMaker Pro)

**Capture techniques:**

- date stamps (cameras, databases & other digital files)
- audio/video record key metadata at the beginning of every track



Sponsored by the National Science Foundation SMA-1447886

# What Metadata standards will you use and why?

- Dublin Core (DC) >> OLAC
- Metadata Object Description Schema (MODS) of LOC
- Metadata Authority Description Schema (MADS) of LOC
- International Standard for Language Engineering (ISLE)
- ISLE Meta Data Initiative (IMDI) of the MPI

Each archive/repository uses a specific metadata schema that you should follow at the time of data creation - CHECK WITH YOUR INTENDED ARCHIVE!



Sponsored by the National Science Foundation SMA-1447886

# Working Session: Documentation & Metadata

Imagine that 15+ years from now, someone comes upon the data from your project in an archive, and with that in mind, answer these questions:

- What documentation and metadata need to accompany your data so that future person can understand what it is and the context of your project?
- What documentation is needed for the future person to replicate or reuse your data?
- How will you create or capture this documentation and these metadata now (at the time you create these data)?



Sponsored by the National Science Foundation SMA-1447886



# Discussion: Documentation & Metadata

Share results of work session.

Q&A



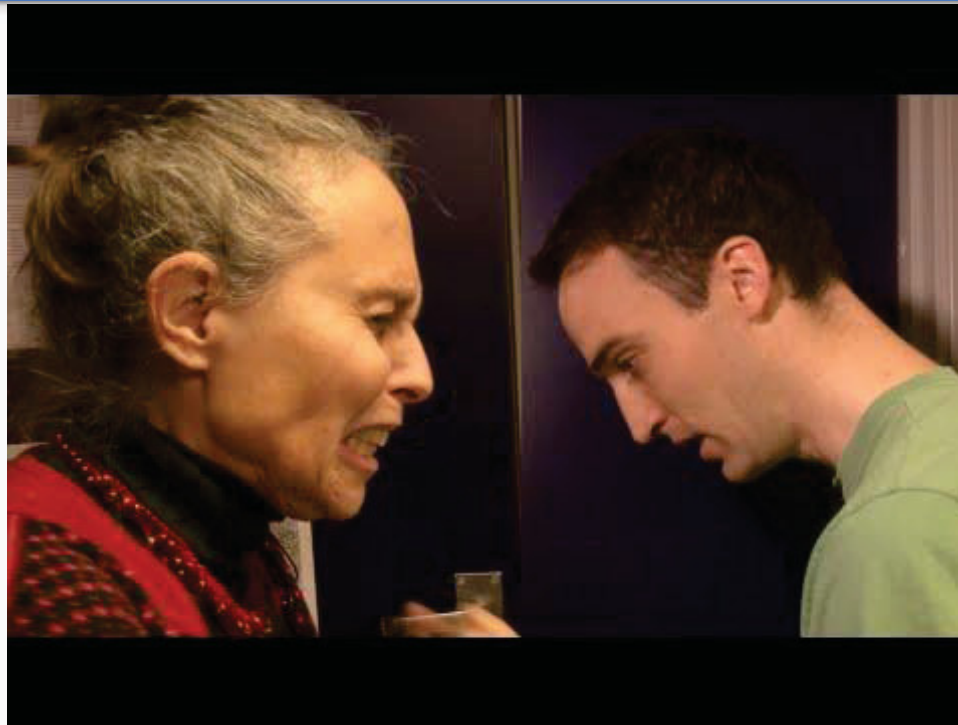
Sponsored by the National Science Foundation SMA-1447886

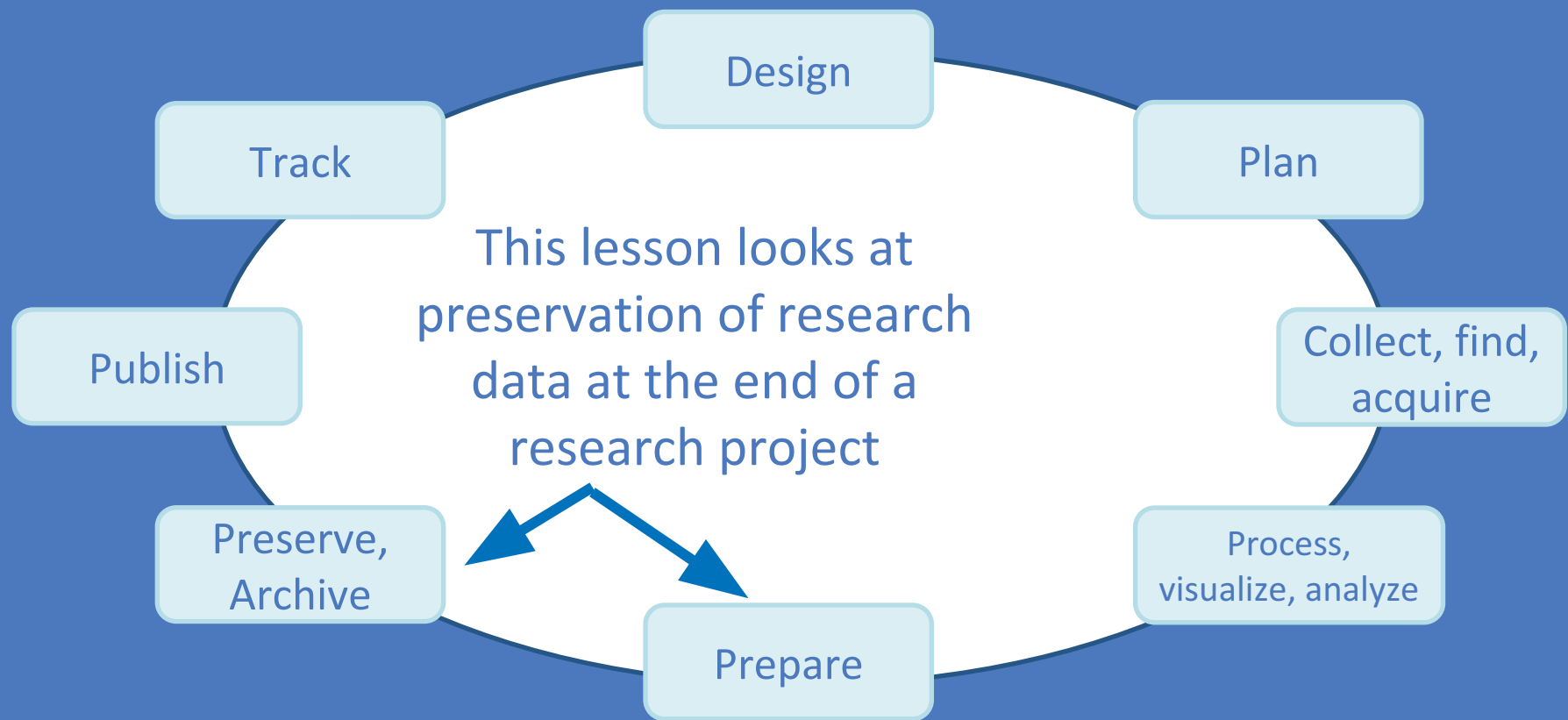
Sunday, July 30, 2017  
Info Session 6:  
Selection, Preservation, and Sharing



Sponsored by the National Science Foundation SMA-1447886

Librarians and repository staff can help!







Sponsored by the National Science Foundation SMA-1447886

# Selection

Which data should be retained, shared, and/or preserved?



Sponsored by the National Science Foundation SMA-1447005

# Selection

Some questions to consider:

- Do you need to retain or destroy any data due to contractual/legal/ethical obligations?
- How does your community (if applicable) want the data to be shared?
- How might other scholars reuse your data?

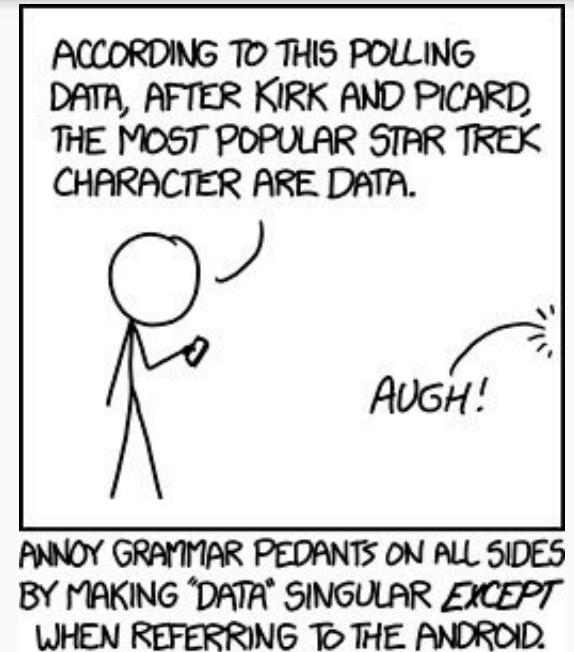


Sponsored by the National Science Foundation SMA-1447886

# Selection

Envision how the data may be reused by you and others.

- Validate/replicate research findings
- Conduct new studies (in the same or different fields)
- Build a corpus or other resource
- Teaching materials



Sponsored by the National Science Foundation SMA-1447886



# Selection

Think not just about linguists like you, but other disciplines!

- Could a syntactician find examples of a linguistic phenomenon in your language documentation data to use in their paper?
- Will you be collecting narratives that could be a source for sociolinguists?
- Would phoneticians like to use your sound recordings?
- If you collect information about local wildlife and plants, would a botanist, pharmacist, zoologist, biologist possibly find information about these useful? Could it be harmful?

Obviously you can't envision all possibilities, but preserving and sharing more kinds of data opens up more possibilities.



Sponsored by the National Science Foundation SMA-1447886

# Preservation

What is the long-term plan for the dataset?



Sponsored by the National Science Foundation SMA-1417005

# Preservation



Archives and repositories are the way to preserve your dataset.

These are usually found in university libraries or run by research groups.

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.

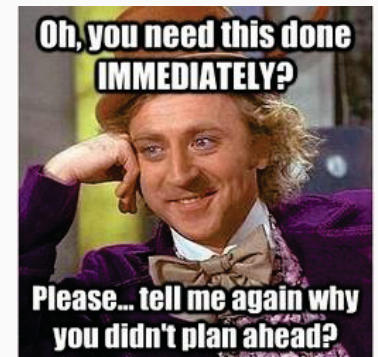
There are also Institutional Repositories, although these may or may not have the capacity for data.



Sponsored by the National Science Foundation SMA-1447886

# Advice for Archiving Data

- Repository staff won't process your data for you - build in time and resources into your DMP for this work!
- Find out the file and metadata requirements of your intended repository and attempt to integrate these as you collect and work with your data.
- Ask repository staff for any documentation or guides they might have available *early*.



Sponsored by the National Science Foundation SMA-1447886

# Sharing

Be as open as possible, as closed as necessary.



Sponsored by the National Science Foundation SMA-1447005

# 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.



Sponsored by the National Science Foundation SMA-1447886

# Sharing

Questions to consider:

- Are there ethical or legal restrictions on what you can share?
- Are any sensitive data anonymized properly?
- How will you license your data?
- Where (e.g. which repository) will you share your data?
- Will you need any access restrictions for privacy/ethics, or can your dataset be Open? Check [this decision tree](#).
- How will you track use of your dataset?



Sponsored by the National Science Foundation SMA-1447886

# Work Session

Find a repository for your data.



Sponsored by the National Science Foundation SMA-1447686



# Work session instructions

Find a potential repository for your data set using the resources on the next two slides.

Answer the following questions for your DMP:

- What is the repository name and why is it appropriate for your data?
- What considerations (e.g. file formats, naming) will you be able to implement during data collection?
- What are the potential costs for long-term storage or preparation of the dataset that you should figure in to your grant request?



Sponsored by the National Science Foundation SMA-1447886

## Types of DLAs

### 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](#))

### 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](#))

## Other Kinds of 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](#))

### Institutional Repositories

- These may or may not have infrastructure to support data. Contact a librarian at your university to ask about your Institutional Repository.
- Go to your library's website and search for the repository or archive.

Sponsored by the National Science Foundation SMA-1447886

# Info Session 7: Resources, Responsibilities & Timeline



Sponsored by the National Science Foundation SMA-1447886

# Data Management Responsibilities & Resources & Timeline

## **Responsibilities:**

Who will be responsible for implementing your DMP and for ensuring that it is followed, reviewed, and revised?

## **Resources:**

What resources will be required to implement your DMP?

## **Timeline:**

What is your timeline for managing--and especially for archiving--your data?

# Responsibilities

DM activities include:

- Data creation/capture
- Metadata creation/production
- Data quality control (archival master files vs. access copies)
- Version control
- Storage and backup
- Data archiving
- Data sharing



Sponsored by the National Science Foundation SMA-1447886

# Work session instructions: Responsibilities

- Brainstorm some responsibilities or steps for each of the DM activities from the previous slide.
- Who will be responsible for each activity? Name names!

For collaborations or group projects:

- How will responsibilities be split across collaborators/sites?
- Will data ownership and DM responsibilities be part of an agreement, contract or MOU between partners?



Sponsored by the National Science Foundation SMA-1447886

# Resources

Resources that you might need include:

- Software
- Hardware
- Physical storage
- IT-administered storage (university server, cloud, etc.)
- Data repository
- Technical expertise



Sponsored by the National Science Foundation SMA-1447886



# Resources

## Where to seek help:

- Your Library
- Your department or college
- Your Office of Sponsored Projects



Sponsored by the National Science Foundation SMA-1447886

# Work session instructions: Resources

- Brainstorm the types of resources you might need for your project.
- Justify why they are needed.
- Where will you find/get these resources?



Sponsored by the National Science Foundation SMA-1447886

# Timeline

- Contact your archive to determine when they can accept your data.
- Clearly lay out the timeline that you will (attempt to) follow as you collect, analyze and archive your data.
- Keep in mind that you might want to archive your data in phases (raw data, derivative data, etc.)



# Work session instructions: Timeline

Consider these questions in plotting your timeline:

- How long will your project last? (e.g. 1 year, 2 years, 3 years, etc.)
- Does the archive have limited times when it will accept data?
- When will you archive your data?
- Will you archive all of your data at once? Will you archive yearly or on some other schedule?



Sponsored by the National Science Foundation SMA-1447886

# References

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

DCC. (2013). Checklist for a Data Management Plan. v.4.0. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/data-management-plans>.

Oliver Widder. (2010, April 23). What's Metadata? Geek and Poke cartoon series. Available from <http://www.datamation.com/cnews/article.php/3878261/Tech-Comics-Whats-Metadata.htm>.



Sponsored by the National Science Foundation SMA-1447886

# Helpful Links

## Helpful Links

Stebbins, Michael. Feb. 22, 2013. Expanding Public Access to the Results of Federally Funded Research, posted on The White House Blog, accessed on Feb. 13, 2017 from

<https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research>

Piowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature*, 473(7347), 285–285.

<http://doi.org/10.1038/473285a>

Copyright and Intellectual Property Toolkit, by Lauren Collister

<http://pitt.libguides.com/copyright>

Pitt guide to data management, including modules

<http://pitt.libguides.com/managedata/DMP>



Sponsored by the National Science Foundation SMA-1447886

# Helpful Links

## Helpful Links

Sample NSF Data Management plans:

<https://www.dataone.org/data-management-planning>

DCC Checklist for a DMP

[http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP\\_Checklist\\_2013.pdf](http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf)

UT-Austin DMP guide:

<http://www.lib.utexas.edu/datamanagement/plan>

NSF - SBE Directorate DMP requirements:

[https://www.nsf.gov/sbe/SBE\\_DataMgmtPlanPolicy.pdf](https://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf)



Sponsored by the National Science Foundation SMA-1447886

# Helpful Links

## Helpful Links

NSF DMP FAQ

<https://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp>

Bibliography for Issues of Consent, Copyright, Intellectual Property and Traditional Knowledge:  
What They Mean for Digital Language Archives by Susan Smythe Kung

[https://docs.google.com/document/d/1zBt69YcCBunTSct\\_DdytNLqtE-IL3iK35VO6ld3zl54/edit](https://docs.google.com/document/d/1zBt69YcCBunTSct_DdytNLqtE-IL3iK35VO6ld3zl54/edit)



Sponsored by the National Science Foundation SMA-1447886