Data Management (ANTHRO 400)

Online at http://bit.ly/dm-anthro - cengel @ stanford

Last updated: February 03, 2017

Contents

1 Create

1.1 Transcripts

- Express Scribe http://www.nch.com.au/scribe/index.html

Reminders. Use...

..a unique **identifier** that labels an interview either through a name or number
..a **header** with brief interview or event details such as date, place, interviewer name, interviewee details
..a **consistent** header style
..a **uniform layout** throughout a research project or data collection
..**speaker tags** to indicate turn-taking or question/answer sequence in conversations
..a **lexicon** of symbols to use and stick with them (overlapping talk, comments, pronounciations...)
..**line breaks** between turn-takes

1.2 Web content

- Webscraping:
  - with graphical interface, e.g: http://import.io
  - with scripting, e.g.: https://scrapy.org
- Internet Archive wayback machine: https://web.archive.org/
- Stanford Web archiving service: https://library.stanford.edu/projects/web-archiving
- Make your own website "archivable":

  Maintain stable links either through redirects or by not changing web addresses.

  Conform to web standards: To increase the chance that future browsers will be able to interpret today's code, validate against current web standards.

  Use durable data formats: The the tacit user base of common web formats tends to ensure those formats' continued support, making a format's popularity a good heuristic for durability. Prefer open formats or at least those that can be read using open-source software.

  Follow web accessibility best practices may be a legal mandate or organizational priority and critically ensures the usability. Providing equivalent text for non-textual content can also facilitate both search crawler indexing and later full-text search in the archive.

  Report media type and character encoding by adding a HTTP Content-Type response header or page-level http-equiv tags.

  Use responsive design ensures that archive users will continue to have a comparable experience of the original website, regardless of the platform they use for access.

2 Document and Organize

2.1 Data Management Plan (DMP)

Includes description of

- **Types of data**, samples, physical collections, software, curriculum materials, and other materials that will be produced over the course of the project
- **Standards** to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)
- Policies for **access and sharing**, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements
- Policies and provisions for **reuse**, re-distribution, and the production of derivatives
- Plans for **archiving** data, samples, and other research products, and for preservation of access to them

More on Stanford's DMP tool:

https://library.stanford.edu/research/data-management-services/data-management-plans

2.2 Document-level

- explanation or definition of codes and classification schemes
- definitions of specialist terminology or acronyms
- codes of, and reasons for, missing values

- derived data created after collection
- data listing of annotations for cases, individuals or items
- ...

Make use of a **README.txt** file

2.3 Metadata

A subset of core **standardized and structured** data documentation that explains

- creator (*Who*)
- content (*What*)
- time reference (*When*)
- geographic location (*Where*)
- purpose (*Why*)
- origin (*How*)

as well as:

- access conditions and
- terms of use of a data collection.

Typically used for **resource discovery**, providing searchable information that helps users to easily find existing data and as a bibliographic record for citation.

2.4 File naming

Best practices:

- Do not use generic file names that may conflict when moved from one location to another. Ensure filenames are independent of location and if you work on more than one computer ensure that your files are synchronized.
- File names should outlast the file creator who originally named the file.
- Consider how scalable your file naming policy needs to be e.g. if you want to include the project number, don't limit your project number to two digits, or you can only have ninety nine projects.
- Keep file names short and relevant - generally about 25 characters is a sufficient length to capture enough descriptive information for naming a data file
- Do not use special characters in a filename such as : & * % $ £] { ! @ as these are often used for specific tasks in different operating systems
- Use underscores instead of full-stops or spaces because, like special characters, these are parsed differently on different systems.
- The filename should include as much descriptive information as necessary to assist identification independent of where it is stored.
- Files should be distinguishable from each other within their containing folder
- Try to find a naming convention where files can be sorted in a logical sequence (for example by adding the date at the beginning as `YYYY-MM-DD`).
- Make file names unique, if possible.
- If including dates, format them consistently `2010-08-11_interview_Jane_Doe`
- Assume that `FILENAME`, `filename` and `Filename` are the same, even though some file systems consider them as different.

- Where possible, use file extensions (often defaults) to accurately reflect the software environment in which the file was created and the physical format of the file. Eg use `.xls` or `.xlsx` for Excel files, `.txt` for text files, etc.
- Mark different versions, if applicable.

Elements you might include in a naming system:

- Participant ID number (if appropriate)
- Type of data collection method
- Site of data collection (e.g. country, region, community, clinic)
- Interviewer or other relevant team member
- Date of data collection
- Characteristics you anticipate may be meaningful, like demographics

appropriate:	`2013-01-03-Rodin_BurghersOfCalais_Stanford.jpg`
not appropriate:	`Rodin Jan 13.jpg`

Renaming software:

Windows:

- Ant Renamer: http://www.antp.be/software/renamer
- RenameIT: https://sourceforge.net/projects/renameit/
- Bulk Rename Utility: http://www.bulkrenameutility.co.uk/

Mac:

- Renamer4Mac: http://renamer.com
- Name Changer: https://mrrsoftware.com/namechanger/

2.5 File Organization

Use a well designed folder structure, e.g. organized by

- by date
- notebook number
- data type
- by project
- site
- unique ID

For non-digital objects consider creating an index

2.6 Data repo tools (?) <<<<

- image repos
- document repos
- spatial data
- html/web stuff

3 Use

3.1 Search

To locally search your computer:

- Get to know Mac Finder and Spotlight: https://www.lifewire.com/use-mac-finder-2260739
- Get to know Windows Search: http://www.makeuseof.com/tag/top-7-windows-search-tricks-search-ninja/
- Explore alternative search tools: https://en.wikipedia.org/wiki/List_of_search_engines#Desktop_search_engines

3.2 Code

- NVivo
- Atlas.ti
- Dedoose

3.3 Count

- Spreadsheet
- R

3.4 Visualize <<<<

Some visualization tools for exploratory analysis...

4 Store

4.1 Backup

What shoud I back up?
Depending on the characteristics of a file and how critical it is (master copy) you may only want to back-up particular files or the entire computer system (complete system image)

How often should I back up?
Depending how much you are willing to risk to loose you want to backup after each change to a data file or at regular intervals. There are two common types of backup:

- An *incremental backup* is one in which successive copies of the data contain only that portion that has changed since the **preceding** backup copy was made. When a full recovery is needed, the restoration process would need the last full backup plus **all the incremental backups** until the point of restoration.
- A *differential backup* is a cumulative backup of all changes made since the last **full** backup. The advantage to this is the quicker recovery time, requiring only the last full backup and **the last differential backup** to restore the entire data repository.

What file formats should I back up in?
Back-ups of master copies should ideally be in file formats that are suitable for long-term digital preservation, that is open as opposed to proprietary formats.

Where should I store my back-ups?
Depending on the form of back-up and the risks associated with data loss there are several options.

Good storage media	Acceptable, for certain data	Not acceptable for long term
Laptop, external HD, Stanford CrashPlan (faculty & staff)	CD, DVD Stanford Box, Google Drive	USB, (obsolete media)

Validation of back-up copies
It is important that you verify and validate back-up files regularly by fully restoring them to another location and comparing them with the original. Back-up copies can be checked for completeness and integrity, for example by checking the MD5 checksum value, file size and date.

How should I organise my backups?
If you are making your own back-ups on removable media, make sure they are well labelled and well organised. Without some management, achieving the ultimate aim of restoring lost data may prove difficult. It is also advisable to have strategies for all systems where data are held, including portable computers and devices, non-network computers and home-based computers.

Other recommendations:

- Follow the **3 - 2 - 1 Rule:** 3 copies - 2 different types of media - 1 offsite
- Copy or migrate data files to new media between two and five years after they were first created, since both optical and magnetic media are subject to physical degradation
- Backup non-digital assets by scanning. Create digital versions of paper documentation in PDF format for long-term preservation and storage
- Ensure that areas and rooms for storage of digital or non-digital data are fit for the purpose, structurally sound, and free from the risk of flood and fire

4.2 Data Transmission and Encryption

Computer security:

- software up to date, antivirus
- practice safe usage: browsing, downloads,
- strong passwords

Data access:

- secure environment (Internet cafe!)
- don't transfer sensitive data in email - duh.
- encryption: encrypted USB, encrypted files (Windows: winzip, 7zip, Mac: disk utility or command line)

5 Share

For example:

- share research data (sdr.stanford.edu, dataverse.harvard.edu)
- share dissertation (SDR)
- share supplemental, public data (SDR, Web)

Examples:

- M. Kohrman: Cigarette Citadels
- T. Mullaney: Grave Reform in Modern China (preview)
- C. Blevins: Geography of the Post

- Z. Frank: Slave Market in Rio de Janeiro: http://purl.stanford.edu/wt635jq5834 (Interactive Visualization: https://cengel.shinyapps.io/RioSlaveMarket/)
- N. Bauch: Enchanting the Desert

6 Preserve

6.1 File formats

Proprietary vs. open formats

File formats that are non-proprietary (e.g. open source, de facto standards), and/or in widespread use, will tend to retain the best chance of being usable in the long term.

When it is necessary to save files in a proprietary format, consider including a README.txt file in your directory that documents the name and version of the software used to generate the file, as well as the company who made the software.

Guidelines for choosing formats

- Non-proprietary if possible
- Unencrypted
- Uncompressed
- Common within the research community
- Interoperable among diverse platforms and applications
- Fully published and available royalty-free
- Fully and independently implementable by multiple software providers on multiple platforms without any intellectual property restrictions for necessary technology
- Developed and maintained by an open standards organization with a well-defined inclusive process for evolution of the standard

Some preferred file formats

Type	Format
Containers:	TAR, GZIP, ZIP
Databases:	XML, CSV
Geospatial:	SHP, DBF, GeoTIFF, NetCDF
Moving images:	MOV, MPEG, AVI, MXF
Sounds:	WAVE, AIFF, MP3, MXF
Statistics:	ASCII, DTA, POR, SAS, SAV
Still images:	TIFF, JPEG 2000, PDF, PNG, GIF, BMP
Tabular data:	CSV
Text:	XML, PDF/A, HTML, ASCII, UTF-8
Web archive:	WARC

Library of Congress recommended formats statement http://www.loc.gov/preservation/resources/rfs/

6.2 Stanford Digital Repository (SDR)

Preservation of scholarly work and reserach data in a robust, reliable, and secure environment, available from persistent URLs (PURLs) with optional access controls.

https://sdr.stanford.edu

6.3 Stanford Data Management Services

https://library.stanford.edu/research/data-management-services