Sarah Dorpinghaus: Director of Digital Services and member of the UK Libraries Research Data and Scholarly Communications Committee

Workshop topic: Preservation of research data; my experience with and knowledge of data preservation is from the archives and special collection perspective- goal: preservation forever
Overview

- Why preserve research data?
- What should be preserved?
- How to preserve research data?
  - Ensuring discovery and access
  - Where to preserve research data
  - Tools and resources
- Fixity activity

Topics to be covered include (see slide)
First, clarifications:

- Definition of data preservation (see slide)

- Amount of time data needs to be preserved can vary
  - As long as necessary
  - As long as possible
  - Indefinitely

- Discussion today will cover the spectrum between backups (not really preservation) and a full preservation system
- Depends on one’s needs and resources
Ask AUDIENCE for ideas
Enabling **re-analysis** of the same products to determine whether the same conclusions are reached

Enabling **re-use** of the products for new analysis and discovery

Enabling **restoration** of original products in the case that working datasets are lost (backups!)
Preservation of research data supports UK’s mission: specifically, *serves a global community by disseminating, sharing, and applying knowledge*
Why?

- Enables
  - Re-analysis
  - Re-use
  - Restoration
- Mandated by granting agencies

Often mandated by granting agency for a certain amount of time

So, in order to get the $$$, one must have a data preservation plan
Need to **preserve** research data (preservation goes beyond backups) because **passive** but **dangerous** challenges occur over time

These are passive because you don’t have to take any action for the to happen

- Physical format
  - Obsolete: 3.5” diskette
  - Degrade: old CDs that no longer play; hard drives that fail

- File format obsolescence
  - Companies usually good about new versions of software running old formats, but often ask to upgrade file (requires you to access file!)

- Software obsolescence
  - Think Claris works, Multiplan (old version of Excel); old operating systems

- Hardware obsolescence
  - No longer have CD drivers on many laptops, zip disk drivers, etc.
Ask AUDIENCE for ideas and examples
What?

- Raw data and data sets
- Intermediate products
- Documentation of protocols
- Software or algorithms
- Results of analysis

Really depends on your research and research products
Ask AUDIENCE for examples of research data they have preserved.
Considerations ($)

- Storage footprint

You do need to think through what should be preserved— in some cases you can’t preserve *everything*

It boils down to **COST**
- Usually not a one-time fee; annual subscription or renewal fees for storage and systems
- Cost of storage generally decreases over time, but not quick enough!

Storage footprint:
- how much are you storing?
- How are you storing it? S3 (quick access) vs Glacier
Considerations ($)

- Storage footprint
- Complexity of data
- Supportive & contextual data

Complexity, supportive and contextual data:
- Just the files or is there software or code?
- Does it need to be organized in a particular way?
- Does it need to reference other data in a particular way?
Considerations ($)

- Storage footprint
- Complexity of data
- Supportive & contextual data
- Resources to recreate, reprocess, redo quality control

Resources to recreate, preprocess, etc.
- Software or code?
- Hardware?
- Is it less expensive (money and time) to store the data rather than reprocess or recheck quality?
HOW?

How do you preserve data?
Three major goals of preservation:

**Accessibility**—the content must be available for use to the appropriate community (you want the **RIGHT** people to be able to find it)

**Discoverability**—the content must have logical bibliographic metadata so that it can be found by end users through time (you want people to be able to find it)

**Usability**—the intellectual content of the item must remain usable via the delivery mechanism of current technology (you want people to be able to use it)
Integrity complete and unaltered in all essential respects;

Aims to “ensure data is recorded exactly as intended, and upon later retrieval, ensure the data is the same as it was when it was originally recorded”/ "Integrity". **Glossary.** Consortium of European Social Science Data Archives (CESSDA). cessda.net. Retrieved 2016-11-02.

Usually done through checking fixity (using checksums to see if file altered or removed)
**Authenticity:** record is what it appears to be

the provenance of the content must be proven and the content an authentic replica of the original
**Cornerstones**

- Integrity
- Fixity
- Authenticity
- Sustainability

**Sustainability**

- Needs to counter challenges outlined before, like media-software-hardware obsolescence
- Funding and administrative support during period of preservation (including “preserve forever” collections)
Many organizations, standards, and best practices to help you achieve these goals and build upon these cornerstones

Too many to go through each one, so I’ll highlight some of the leading standards and best practices
OAIS: Open Archival Information System **REFERENCE MODEL** for preservation system
- For **long term** and **indefinite** preservation
- Doesn’t address a particular platform, just how a system should work & what the activities are

- Outlines 3 main roles and how they interact with the OAIS
  - Producer of content
  - Management of OAIS
  - Consumer of content
You’ll see these 3 roles here but with more detail

Major steps:
Ingest: Submission information package (SIP)- includes content and descriptive information

Turns into an (AIP) Archival information package; sits in archival storage which receives ongoing data management and preservation planning

If access is intended (internal or external, public or private) a (DIP) dissemination information package is created and communicates to access layer (digital library) and thus the consumer
Standards & Best Practices

Trusted Digital Repository (TDR) Checklist, ISO 16363
AKA Trusted Digital Repositories and Audit Checklist (TRAC)

Checklist that outlines the components of an OAIS compliant digital repository (put together by OCLC and NARA- National Archives and Records Administration)

Allows institutions to assess their repository

UK Libraries has used this to help identify and address gaps in their preservation system
OAIS is very detailed and impractical for an individual or small group of researchers trying to take the first step for basic data preservation.

There are more basic standards too, including LOCKSS: Lots of Copies Keeps Stuff Safe
- Backups + geographic distribution
- Many cloud storage providers offer this without expressly stating so
- UK home-grown example: store content on LTO tapes, LTO tape exchange with another university in SE United States
These seem a little overwhelming for someone getting started

Rest of the presentation focused on things YOU can do + resources and tools to help
**Refreshing**—to copy digital information from one long-term storage medium to another of the same type

Refreshing is a necessary component of any successful digital preservation program, but is not itself a complete program

Ensures you can get to the file (hopefully) but not that you have the software or operating system to open the file
Strategies: Replication

Make copies!

- Includes Lots of Copies Keeps Stuff Safe (LOCKSS)
- Also includes bit stream copying
Creating an exact duplicate

UK Special Collections Research Center: creates disk images of physical media of high research value: does not ensure we have the software or operating system to access the files, but the bits (1&0, DNA of files) is preserved
Can mean two things:
-- Moving data from one system to another
-- Converting data from one technology to another (or from an older file format to a newer one

– Example- Word: .doc vs. .docx
Normalization is an example of migration, but it’s moving from a proprietary format to an open format. Ideally, you want to capture as completely as possible—both the content and the look and feel, but sometimes that’s not possible. Later, I’ll mention some resource that list recommended open formats that offer ongoing data access.
**Emulation**: re-creation of the technical environment needed to view and use data

Need to have accurate information about both hardware and software so that it’s a true re-creation
One of the most fun examples is the Internet Archive’s MS-DOS games, which is part of the Software Library

Another example is from Emory University—Salman Rushdie papers: to access, computer workstation that emulates his machines

Speaking of access...
These are also critical components of data preservation. If you cannot find it or cannot open the file, what is the point of preserving it?
Metadata! We (information professionals) love the power of properly implemented metadata.

Metadata Librarian Kathryn Lybarger did an entire workshop on metadata (which can be accessed via UKnowledge), so this workshop won’t dig in too deep.
Metadata provides important contextual information:

**Descriptive:** what it is, where it’s stored, subjects

**Administrative:** who owns it, intellectual rights

**Structural:** how the items relate to each other (for compound or paged objects)

**Technical:** how the data was produced, what hardware and software is need to fully use the data

**Preservation:** what steps have been taken to preserve the data (was it migrated to a new system, to a new format); checksums
Questions / aspects to consider recording when generating your metadata
Tips:

Consistency is key

Many different metadata standards to help you be consistent

• Some are specific to the data type/genre (scientific dataset vs. collection of artistic pieces)
• Some are specific to the format
• Some are general, some are specific
Another way to ensure discoverability is through file naming conventions

Anyone who has done any major editing of files, especially with a group of people, understands how file names can get very messy very quick.
Before creating files, consider establishing a basic naming convention—something simple and brief.

File Naming Conventions

- Simple and brief

2017_trial02
File Naming Conventions

- Simple and brief
- Consistent

2017-04_trial02
2017-04_trial02_results

Try to be as consistent as possible (start all with dates, format all the dates the same, whether)
File Naming Conventions

- Simple and brief
- Consistent
- Not sole source of description, but descriptive

*001.tif vs. 1995_picnic_001.tif*

Remember that a file name does not need to be (and should not be) the sole source of description, but it should be descriptive enough that you know what it is if the file is misplaced
File Naming Conventions

- Simple and brief
- Consistent
- Not sole source of description, but descriptive
- Avoid special characters
- (_) instead of (.) or ( )

Avoid special characters like * and !

Use underscores instead of periods and spaces
Last thing to mention regarding access draws back to what was said about normalization and using open formats.

Proprietary vs. open

- Open formats allow people to access the code for the files, which allows them to create viewers that open those file types.
- Reliance on proprietary formats require the company to maintain the format and viewers

There are many different types of formats that each have options for long-term preservation. These can be found online at this url, which is also listed on the resources document in the Google Drive folder.
WHERE?

Where can data be stored?
The world of data preservation is ... expansive to say the least. Off-campus options really depend on:

- The type of data you are generating
- Your sharing and privacy needs
- The complexity of your data and how you need to interact with it (some cases you need a safe place to put it and LEAVE IT while other cases that data needs to be referenced regularly)
- Your resources and familiarity with setting up a proper preservation system
- Enterprise options like AWS (S3 & Glacier), Microsoft Azure
- Open/non-profit options like the Internet Archive
- Need to poke around online and see what the appropriate options are for your field

Registry of Research Data Repositories includes many options
On campus resources:

- Most obvious place to get help storing data is UK ITS - have many different options but just signed a contract with Amazon, so can get significant discounts
- Center for Computational Sciences - has options for those doing high performance computing
- Research Computing Group - new group at UK: blend of UK ITS and folks at the Center for Computational Sciences; eager to help faculty build robust systems that meet research needs
- UKnowledge: institutional repository

Just ask for help! Strongly suggest setting up a consultation meeting to chat with folks from ITS, CCS, or Libraries to help you find whatever solution may meet your data preservation needs.
There are also plenty of tools out there to help you plan for and implement data preservation- DIY style!
DMP Tool: many templates to help you plan for data preservation, has many templates based on the type of research data you are generating and also based on granting agencies to help you meet their requirements. Many granting agencies require you to have a data preservation/management plan as part of the application process.
You can find additional information about DMP Tool and other research data management tips on our Research Guide! The Research Data and Scholarly Communications Committee will be updating the guide regularly, so be sure to review it often.
Tools & Resources

- Data Management Planning Tool (DMP Tool)
- Community Owned Digital Preservation Tool Registry (COPTR)
- AV Preserve
- Digital Curation Centre
- Library of Congress NDIIPP
- Much more!

Also many other tools and resources out there:

- Digital preservation basics
- Repository reviews
- Free digital preservation tools: transferring, generating checksums, extracting embedded metadata

See the Resources document in Google Drive folder for a more complete (but not comprehensive) list of resources and tools.
FIXITY ACTIVITY
Also many other tools and resources out there:

- Digital preservation basics
- Repository reviews
- Free digital preservation tools: transferring, generating checksums, extracting embedded metadata

See the Resources document in Google Drive folder for a more complete (but not comprehensive) list of resources and tools.
Images from:
Digital Bevaring.DK https://digitalbevaring.dk/illustration/