Using the ETD+ Toolkit to Teach Data Management Skills

Gabrielle V. Michalek
Director, Connected Scholarship Program
Learning Objectives

• Broaden and deepen familiarity with emerging methods and tools to support students’ digital object management.
• Anticipate the impact of expanding research formats and materials on lifecycle management and research workflow practices for ETD programs.
• Understand what other institutions are doing to support students in managing (or learning how to manage) their research outputs.
• Future direction of using the Toolkit on campus to support digital literacy and library instruction.
Outline

• Welcome, Introductions
• ETD Hand-on Activity
• Overview of ETD+ Workshop Modules
• Copyright Module and Other Scholarly Communications Questions
• Research Workflows
• Storage, Version Control and Use of Other Data and Research Management Tools
Let’s Get Started!
Exercise:

On the stickies provided, please answer the following question:

What content types does your institution’s ETD program currently accept? (one type per YELLOW sticky note)
Exercise:

On the stickies provided, please answer the following question:

What content types does your institution’s digital preservation program currently support? (one type per Seafoam green sticky note)
Workshop Background

• ETD Project, 2014-2017, funded by IMLS
• Addressed need for research data maintenance for ETDs and their related digital files
• Created a series of workshop modules, Guidance Briefs and Workbench software
ETD + Team:

- Educopia Institute
- MetaArchive Cooperative
- NDLTD
- ProQuest
- Carnegie Mellon University
- Colorado State University
- HBCU Library Alliance
- Indiana State University
- Oregon State University
- Penn State University
- Purdue University
- University of Louisville
- UNC School of Library and Information Science
- University of North Texas
- University of Tennessee at Knoxville
- Virginia Tech University
Core Research Question:

How can institutions best ensure the longevity and availability of ETD research data and complex digital objects (e.g. software, multimedia files) that comprise an integral component of student theses and dissertations?
ETD+ Toolkit Modules

- Module 1: Copyright
- Module 2: Data Organization
- Module 3: File Formats
- Module 4: Metadata
- Module 5: Storage
- Module 6: Version Control
The ETD+ Toolkit helps the academic community train students to ensure the longevity and accessibility of their research outputs.
Each Module Includes:

1. Learning Objectives
2. One-page Handout (Print)
3. Guidance Brief (Online & Customizable)
4. Slideshow with presenter notes
5. Evaluation survey
ETD+ Workshop Series at CMU

• Piloted with undergraduates in 2017
• All six modules presented to graduate students 2017-18 academic year
• Two modules presented to additional undergraduate groups in summer of 2018
• Updated modules will be taught in 2018-19 academic year
What Did We Learn?

• Students will ask questions that are tangential, but related
• Important to know key scholarly communication issues including author’s rights
• Important to understand the Research Lifecycle Management Approach
• Important to understand the research workflows and infrastructure used on campus
So...how do we use the Toolkit to teach data management skills in a relevant way to students??
Copyright and Scholarly Publishing

From the Understanding Copyright 2009
UNDERSTANDING COPYRIGHT By Renee Hobbs, Katie Donnelly and Sandra Braman
Module 1: COPYRIGHT

How can students gain appropriate permissions and how can students signal copyright for their own works?
Key Take Away(s)

Students learn from this module:
When and how to seek appropriate permission to use existing works.

Learn ways to establish and signal copyright for their works.

How to recognize what research outputs are and are not copyrighted automatically.
Understanding Copyright

• **US Copyright**: a legal tool authors and creators use to signal what other people can – or cannot – do with their works.

• **Public Domain**: works not limited by copyright that can be freely used by anyone. Usually, a work enters the public domain once a defined period of copyright protection has lapsed.
• **Fair Use:** If use a copyrighted work meets certain “fair use” criteria, courts have found that no formal permission is needed. Criteria taken into account include:
  - the purpose
  - the type
  - the amount
  - the effect
Understanding Copyright

Giving credit is no substitute for asking permission!
Using Copyright

• Copyright for certain types of works you author automatically belongs to you.
  o literary works
  o musical works, including accompanying words
  o dramatic works, including accompanying music
  o pantomimes and choreographic works
  o pictorial, graphic, and sculptural works
  o motion pictures and other audiovisual works
  o sound recordings
  o architectural works
Signaling Copyright

- However, copyright may not extend to research outputs you produce as part of the thesis or dissertation.
  - Data is only thinly protected by copyright;
  - Consider designating a CC license to accompany datasets (e.g., CC0)
Signaling Copyright

- Creative Commons
  - CC0: a waiver (no license)
  - CC-BY: attribution
  - CC-BY-ND: attribution, no derivatives
  - CC-BY-NC: attribution, non-commercial
  - CC-BY-SA: attribution, share alike

More: [https://creativecommons.org/](https://creativecommons.org/)
Copyright

**US Copyright** is a legal tool authors and creators use to signal what other people can – or cannot – do with their works.

**Public Domain:** Works are in the public domain once a defined period of copyright protection has lapsed, at which point they are no longer governed by copyright and can be freely used by anyone.

If you are using a work that is within copyright, but meets certain “fair use” criteria, courts have found that no formal permission is needed. The criteria that are taken into account include the purpose (e.g., educational and research uses favor fair use while commercial uses do not); the type (e.g., factual or nonfiction-based information may favor fair use; highly creative work likely will not); the amount (e.g., small quantities vs. a significant portion of the original work); and the effect (e.g., no negative impact on the copyright holder). [http://copyright.gov/circs/circ01.pdf](http://copyright.gov/circs/circ01.pdf)

Giving credit is no substitute for asking permission!

**Creative Commons (recommended)**
- CC0: a waiver (no license)
- CC-BY: attribution
- CC-BY-ND attribution, no derivatives
- CC-BY-NC: attribution, non-commercial
- CC-BY-SA: attribution, share alike

More: [https://creativecommons.org/](https://creativecommons.org/)

**What can copyright protect?**
1. literary works
2. musical works, including accompanying words
3. dramatic works, including accompanying music
4. pantomimes and choreographic works
5. pictorial, graphic, and sculptural works
6. motion pictures and other audiovisual works
7. sound recordings
8. architectural works

**What about my copyright?**
Copyright for a work you author automatically belongs to you. However, copyright may not extend to research outputs you produce as part of your thesis or dissertation. For example, data is only thinly protected by copyright; specifically designating a CC license to accompany datasets (e.g., CC0) is a good approach for simultaneously sharing and protecting these outputs.

**Resources**
- Cornell University “Fair Use Checklist” [http://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf](http://copyright.cornell.edu/policies/docs/Fair_Use_Checklist.pdf)

Source - [Guidance Briefs: Managing Your ETD Research Files](#)
Activity

• Select a chapter from your dissertation, at whatever stage it is now, and identify all the other works you cite, reference, or borrow from.

• Now, using the Fair Use Checklist, the Best Practices in Fair Use in Scholarly Research, and the Code of Best Practices in Fair Use for the Visual Arts, determine which (if any) works may require additional research to determine if permission is needed.
Students will ask indirectly related questions.

Questions answered here, even the silly ones.

Additional comments regarding this photo can be found on my blog @ travelinlibrarian.info/2006/08/silly-questions/
Scholarly Publishing

- How to get a work published?
- How to negotiate publisher contracts?
- How to self-archive or place in IR?
- How to determine authorship credit and order?
- How to evaluate scholarly journals?
Using the ETD+ Toolkit
Within the Research Ecosystem

From Wikimedia Commons, the free media repository
Lifecycle Management, Research Workflows and the ETD+ Toolkit

Available from mechanics gear gears blue way of thinking free from Pixabay
Progression of Digital Objects

DigitalNZ: http://makeit.digitalnz.org/
Example research workflows: traditional to experimental

Can be found at '101 Innovations in Scholarly Communication' webpage: [https://innoscholcomm.silk.co/](https://innoscholcomm.silk.co/)
Research Workflows, Lifecycle Management Approach and the Open Science Framework
ETD+ Toolkit Modules

MODULE 1: COPYRIGHT
How can students gain appropriate Permissions and how can students Signal copyright for their Own works?

MODULE 2: DATA ORGANIZATION
How can students structure, describe, store, and deposit data and research files for reuse and/or future access?

MODULE 3: FILE FORMATS
How will the formats students choose make future access to their research easier or more difficult?

MODULE 4: METADATA
How can students store information Describing their files to make sure they can tell what they are in the future?

MODULE 5: STORAGE
How can students make well informed Choices about where to store their Research materials?

MODULE 6: VERSION CONTROL
What mechanisms can students use to make it easier to see the history of a file with multiple versions?
Learning Objectives

• Understand the range of threats to storage stability in the digital environments you are using for your research outputs.
• Learn how to weigh the pros and cons of different approaches to research content storage.
• Gain familiarity with storage fundamentals, including the current and emerging storage options for digital research outputs.
Why Storing Copies in Multiple Locations Matters...

- Viruses
- Theft
- Storage device malfunctions
- Physical disasters
- Overwritten files
- Accidental erasure
- File Corruption
- Lost password or key
- Hacking
- Malicious deletion

Why storing copies in multiple locations matters...
Back-ups

**Back-up:** A copy of your digital content, ideally stored in a different location from the original, usually made to prevent data loss.
Common storage options

- Laptop
- Desktop
- External hard drive (spinning or SSD)
- Flash drive
- The “cloud”
Storage recommendations

• Maintain at least one local (i.e., non-cloud-based) copy of your content.
• Maintain at least three separate complete copies of your research content.
• Maintain at least one of those copies in a different geographic location.
Storage recommendations (continued)

• Maintain a history of changes in at least one location (e.g., using a “Time Capsule” software package to automatically back up your content onto an external drive without deleting older copies).
**Preservation**: The “series of managed activities necessary to ensure continued access to digital materials for as long as necessary”

-Digital Preservation Coalition
Managed Activities (continued)

• Systematize your folder- and file-name conventions using human-identifiable names.
• Use naming conventions to mark versions of files (e.g., MusicofSocialChange-v12.csv).
• Make sure your filenames are followed by the correct file extension (e.g., .txt, .csv).
• Avoid using special characters in all file and folder names (e.g., \?:*<=>{][&$,:!).
Managed Activities and Preservation

• Produce and maintain an inventory of all of your content, documenting file names, sizes, locations, types, and “checksums”.

• Create and regularly check “checksums” for your most important research files.

• Employ a tool like “Fixity” to scan specified folders or directories on a regular basis and report changes to you via email.

https://github.com/avpreserve/fixity
Resources


• For general info on archiving and backing up content, see the Personal Digital Archiving resources. http://digitalpreservation.gov/personalarchiving/
Storage

Back-up: A copy of your digital content, ideally stored in a different location from the original, usually made to prevent data loss.

Preservation: The “series of managed activities necessary to ensure continued access to digital materials for as long as necessary”. —Digital Preservation Coalition

Basic recommendations:
1. Maintain at least one local (i.e., non-cloud-based) copy of your content
2. Maintain at least three separate complete copies of your research content
3. Maintain at least one copy in a different geographic location
4. Maintain a history of changes in at least one location (e.g., using a “Time Capsule” software package to automatically back up your content without deleting older copies)
5. Document in a text file how, when, and where you store and back up your materials
6. Systematize your folder- and file-name conventions using human-identifiable information
7. Use naming conventions to mark versions of files, e.g., using consecutive numbers to track a file through all edits and revisions that take place to it. (e.g., filename-v12.txt)
8. Make sure your filenames are followed by the correct file extension (e.g., .txt, .csv)
9. Avoid using special characters in all file and folder names (e.g., \?:*<>[]\$\,;\!)
10. Document the formats you are managing and the potential sustainability issues
11. Save a copy of your research files in non proprietary formats, so that you don’t need a software license to render and use them.

Advanced recommendations:
1. Produce and maintain an inventory of all of your content, documenting file names, sizes, locations, and types
2. Create and regularly check “checksums” or digital signatures for your most important research files. Checksums can be generated by several open source tools and utilities and they can be stored in your inventory.
3. Monitor your content to ensure missing, moved, and renamed files are automatically brought to your attention. A tool like “Fixity” can scan specified folders or directories on a regular basis and report changes to you via email.

Threats to storage environments:
- Natural disaster
- Human error
- Human malice
- Drive failure
- Format obsolescence
- Media obsolescence
- Bit rot
- Business failure
- Software or hardware error

Resources
1. For “back-up” advice, see Jesus Vigo, Best Practices to Back up Your Data
2. For more on cloud-based backups, please see Charles Beagrie Ltd. How Cloud Storage can address the need of public archives in the UK
3. For general information, see also Personal Digital Archiving

Source - Guidance Briefs: Managing Your ETD Research Files
Activity

• Take one project you are working on now, and develop a spreadsheet-based inventory for the associated files indicating file names, sizes, types, and storage locations. (Use http://www.cdlib.org/services/dsc/contribute/docs/submission.inventory.rtf as a guide).

• Establish a regular routine for backing up your content in at least one additional location. Make sure the routine includes a regular schedule, a way of storing content organized by the date of a backup, and a way to maintain multiple backups simultaneously.
Students will ask indirectly related questions
Local and Cloud Storage Options

From The Open Science Framework (OSF) at Notre Dame: Connecting the Workflow and Supporting the Research Mission
Version Control

VERSION CONTROL

a developer’s best friend
Learning Objectives

- Understand that your research files will change over time
- Learn to use naming/storage conventions to “version” your research files, making it easier to know your research history
- Gain familiarity with software tools that can help you with versioning, particularly when you’re collaborating on documents with others
Version Control

• The process of managing changes to your files over time (aka, revision control or source control).

Version 01
Version 02
Version 03
Version 04
Version 05
Why Versions Happen

• Saving a new draft of text for editing
• Refining raw survey results into a clean dataset for analysis
• Producing a transcript of an interview based on an audio recording
• Creating a smaller version of an image to post online
Version control is all about PROCESS
## Version Control

<table>
<thead>
<tr>
<th>OK</th>
<th>OOPS</th>
<th>BETTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>image1_v1.jpg</td>
<td>image1_v1.jpg</td>
<td>image1_20151021</td>
</tr>
<tr>
<td>image1_v2.jpg</td>
<td>image1_v10.jpg</td>
<td>image1_20151214</td>
</tr>
<tr>
<td>image2_v1.jpg</td>
<td>image1_v2.jpg</td>
<td>image1_20160123</td>
</tr>
<tr>
<td>image2_v2.jpg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Version Control – Collaborative Documents

- dataset1_20160402_KES
- dataset1_20160301_WTC
- dataset1_20160814_GSC
- ...
- ...
- ...
Many software packages do not support versioning.
Many of your students will be using tools that give them versioning functionality.

From The Open Science Framework (OSF) at Notre Dame: Connecting the Workflow and Supporting the Research Mission
Versioning Tools

- Examples: Git, Subversion and Open Science Framework
- Check-in and check-out file processes
- Usually you can only view and edit the working version of a file.
- Each time you change a file, you can save a “revision” and attach a short summary of your changes.
Resources

• MATRIX at Michigan State University gives file naming advice: http://ohda.matrix.msu.edu/2012/08/file-naming-in-the-digital-age

• Udacity offers a free online course on using Git and GitHub: https://www.udacity.com/course/how-to-use-git-and-github--ud775

• Hello World offers another helpful GitHub guide: https://guides.github.com/activities/hello-world/

• Version Control with Subversion is a free book authored by Subversion software developers: http://svnbook.red-bean.com/
Version Control

**Version Control:** The process of managing changes to your files over time (aka, revision control or source control)

**Manual Version Control**

A simple method to store the current revision is at the end of the file name. This way, files can be grouped by their names and sorted by version number:

- filename-v01.jpg
- filename-v02.jpg
- ...

You can also use dates to designate version numbers, using year-month-day (20150930) to help your computer sort versions in chronological order:

- filename-20160402.jpg
- filename-20160407.jpg
- ...

If the files you are using are created or edited collaboratively, incorporate names or initials so you know who updated which version:

- filename-20160402-KES.jpg
- filename-20160407-WTC.jpg
- ...

**Software-Assisted Version Control**

There are also software tools that can help you version your content. These tools store your content in such a way that they can remember its state from revision to revision. Usually, they also allow you to “check in” and “check out” your content, ensuring that revisions never happen simultaneously in two different locations (e.g., if collaborating researchers both attempt to revise the same file at the same time, or a researcher unwittingly tries to revise the same file on two different machines). Key differences between these software-assisted methods and the manual methods include:

1. You can only view and edit the working version of a file
2. When you change a file, you can save a revision and attach a short summary of your changes.

Research is active and iterative. You will edit and re-edit your research materials many times before finishing your thesis or dissertation. How will you know that you are working with the most current revision of your materials?

**Resources (For more information)**

- The digital humanities center MATRIX (Michigan State University) provides advice on how to structure file names based on oral history projects that is broadly applicable: [http://ohda.matrix.msu.edu/2012/08/filename-naming-in-the-digital-age](http://ohda.matrix.msu.edu/2012/08/filename-naming-in-the-digital-age)
- Udacity offers a free online course on how to use Git and GitHub with interactive exercises to familiarize you with using the tools. [https://www.udacity.com/course/how-to-use-git-and-github--ud775](https://www.udacity.com/course/how-to-use-git-and-github--ud775)
- Another helpful GitHub guide is available from Hello World. [https://guides.github.com/activities/hello-world/](https://guides.github.com/activities/hello-world/)
- The Subversion community provides free access to the book Version Control with Subversion: [http://svnbook.red-bean.com/](http://svnbook.red-bean.com/)

Source - Guidance Briefs: Managing Your ETD Research Files
Activity

- Find a folder of research materials that you have collected on your computer. Look through the materials and answer the following questions:
  - Are there multiple versions of the same materials (documents, images, etc.)?
  - How are the different versions labeled?
  - Can you quickly identify a file’s most recent version? Its authoritative version? Its original version?
Who can use the Toolkit?

Anyone may freely adopt and adapt this toolkit

http://educopia.org/etdplustoolkit
Questions?

Gabrielle V. Michalek

gabrielle@cmu.edu