THE ETDPLUS PROJECT

Supporting Complex ETD Data

Martin Halbert
Dean of Libraries
University of North Texas

ETD 2016 Conference
Tuesday, July 13, 2016

EDUCOPIA INSTITUTE

Presentation Overview

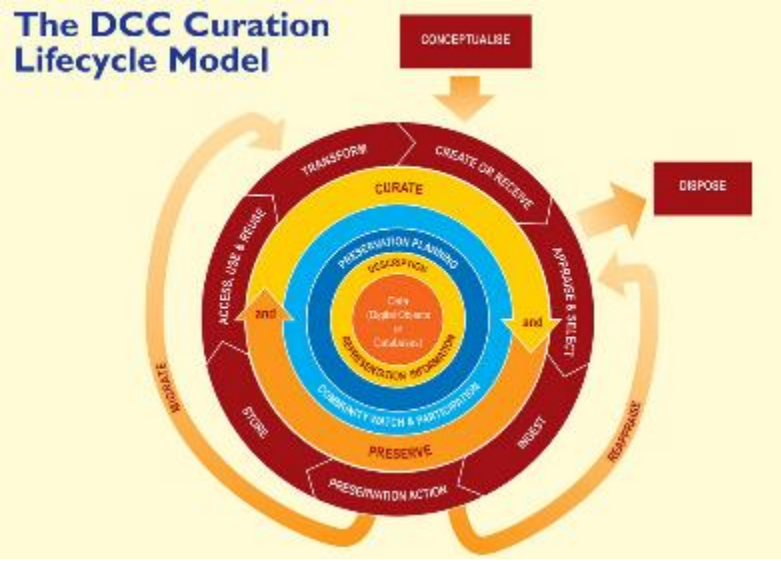
1. ETD Lifecycle Management Project Series - collaborative background
2. Survey findings
3. Guidance documents and training materials
4. Hydra-Sufia reference implementation tool for complex ETD materials submission/ingestion workflow
5. Brief thoughts going forward

SECTION 1:

ETD LIFECYCLE MANAGEMENT PROJECT SERIES - COLLABORATIVE BACKGROUND

Key Challenges of all ETD Programs

- How will institutions address the entire life cycle of ETDs?
- Can we ensure that ETDs acquired from students today will be available to future researchers? In 10 years? In a century?
- How will libraries identify and institutionalize the best long-term curatorial practices for this important genre of digital content?



Progression of Projects on these Issues

1. MetaArchive Cooperative ETD preservation activities 2007-2011
2. NDLTD/Educopia surveys conducted with Virginia Tech in 2011 and 2013
3. ETD LifeCycle Management Project funded by US IMLS 2011-2014
4. ETDplus Project on data/supplementary files funded by US IMLS 2014-2017

https://educopia.org/research/electronic-theses-and-dissertations

Research

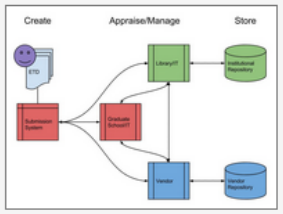
Continuing Education
- Nexus
- Mapping the Landscape

Digital Preservation
- Aligning National Approaches to Digital Preservation (ANADP)
- Chronicles
- Distributed Digital Preservation (DDP)
- Identifying Continuing Opportunities for National Collaboration (ICONC)
- Electronic Theses and Dissertations

Scholarly Communication
- Chrysalis

Incubating Ideas
- OSSArcFlow
- Project Meerkat
- Scholarly Publishing Management Suite

Electronic Theses and Dissertations



Websites:
Project Wiki

Exploring the evolution of research products by authors on the cusp of their careers and the cutting-edge of their fields.

Colleges and universities are steadily transitioning from traditional paper and microfilm formats for graduate theses and dissertations. While this move from print-based to electronic theses and dissertations (ETDs) greatly enhances the accessibility and sharing of graduate student research, it also presents significant challenges for academic libraries seeking to preserve digital content.

To promote best practices and to increase the capacity of academic libraries to reliably preserve ETDs, project participants will develop and share a toolkit of guidelines, educational materials, and a set of software tools for ETD life-cycle data management and preservation.

All Research Grants

ETDplus

Dates:
October 2014 to September 2017
Funder: IMLS
Principal Investigator(s): Katherine Skinner, Matt Schultz, Sam Meister
Project Manager(s): Nick Krabbenhoeft

Lifecycle Management for Electronic Theses and Dissertations

Dates:
November 2011 to October 2014
Funder: IMLS
Principal Investigator(s): Katherine Skinner, Martin Halbert
Project Manager(s): Matt Schultz

ETD Lifecycle Management Project Partners 2011-2014 (Funded by US Institute of Museum and Library Services)

1. Networked Digital Library of Theses and Dissertations (NDLTD)
2. Educopia Institute and MetaArchive Cooperative
3. University of North Texas
4. Virginia Tech
5. Rice University
6. Boston College
7. Indiana State University
8. Pennsylvania State University
9. University of Arizona

Formats, Complex Content Objects, and ETDs - Format Repositing and Migration Issues

- Found that many ETD programs now mandate that the primary item deposited be some form of PDF, sometimes with format checking of the specific characteristics of the PDF.
- However, there are often many sorts of non-textual formats associated with theses and dissertations that are not preserved
  - Datasets (Excel, statistics, etc.)
  - Images (Photographs, scans, etc.)
  - Computer Programs
  - Audio and other multimedia formats
  - This information may be integral to the work of the thesis/dissertation

Critical Content is often Lost if only PDF is Deposited for an ETD

- The actual, relevant content of theses and dissertations may not be captured, or *capturable*, in a simple PDF
- Performative or functional works are often not preserved in ETD repositories
- Trying to address this issue led to the ETDplus Project

http://educopia.org/research/grants/etdplus

EDUCOPIA INSTITUTE

Search

About Us | Communities | Research | Consulting | Events | Downloads | Contact Us

Research

- Continuing Education
  - Nexus
  - Mapping the Landscape
- Digital Preservation
  - Aligning National Approaches to Digital Preservation (ANADP)
  - Chronicles
  - Distributed Digital Preservation (DDP)
  - Identifying Continuing Opportunities for National Collaboration (ICONC)
  - Electronic Theses and Dissertations
- Scholarly Communication
  - Chrysalis
- Incubating Ideas
  - OSSArcFlow
  - Project Meerkat

Research | All Grants | ETDplus

ETDplus

October 2014 to September 2017

Proposal:
Grant Narrative

Principal Investigator(s):
Katherine Skinner
Matt Schultz
Sam Meister

Project Manager(s):
Nick Krabbenhoeft

Supporting the evolution of ETD research products

ETDplus builds on the momentum of the Lifecycle Management of ETDs project to research and build tools to help manage a growing challenge in ETD programs: the creation and submission of materials beyond the PDF of a thesis or dissertation. Ranging from research data sets to video installations, from websites to music recitals, these digital objects are pieces of intellectual work that cannot be captured in words alone.

The project is producing guidance documentation, workshop materials, and software tools for students and staff to use in managing these complex digital objects. It is a partnership between Educopia Institute, bepress, Carnegie Mellon University, Colorado State University, Confederation of Open Access Repositories, Indiana State University, Morehouse School of Medicine, Oregon State University, Penn State University, Purdue University, ProQuest, University of Louisville, University of North Carolina School of Library and Information Science, University of North Texas, University of Tennessee Knoxville, and Virginia Tech University.

The Guidance Briefs are under public review from May 2-June 30, 2016! Please review and comment on these briefs, drawing our project team's attention to any components that need to be edited, revised, broadened, or narrowed.

The project is generously funded by the Institute of Museum and Library Services (IMLS) and led by the Educopia Institute, in collaboration with the NDLTD, HBCU Alliance, bepress, ProQuest, and the libraries of Carnegie Mellon, Indiana State, Morehouse, Oregon State, Penn State, Purdue, University of Louisville, University of Tennessee, the University of North Texas, and Virginia Tech.

ETDplus Project Group 2014-2017 (Also funded by US Institute of Museum and Library Services)

1. NDLTD
2. Educopia Institute
3. MetaArchive Cooperative
4. ProQuest
5. Carnegie Mellon University
6. Colorado State University
7. HBCU Library Alliance
8. Indiana State University
9. Oregon State University
10. Penn State University
11. Purdue University
12. University of Louisville
13. UNC School of Library and Information Science
14. University of North Texas
15. University of Tennessee Knoxville
16. Virginia Tech University

Core ETDplus Project 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?

SECTION 2:

SURVEY FINDINGS

Surveys

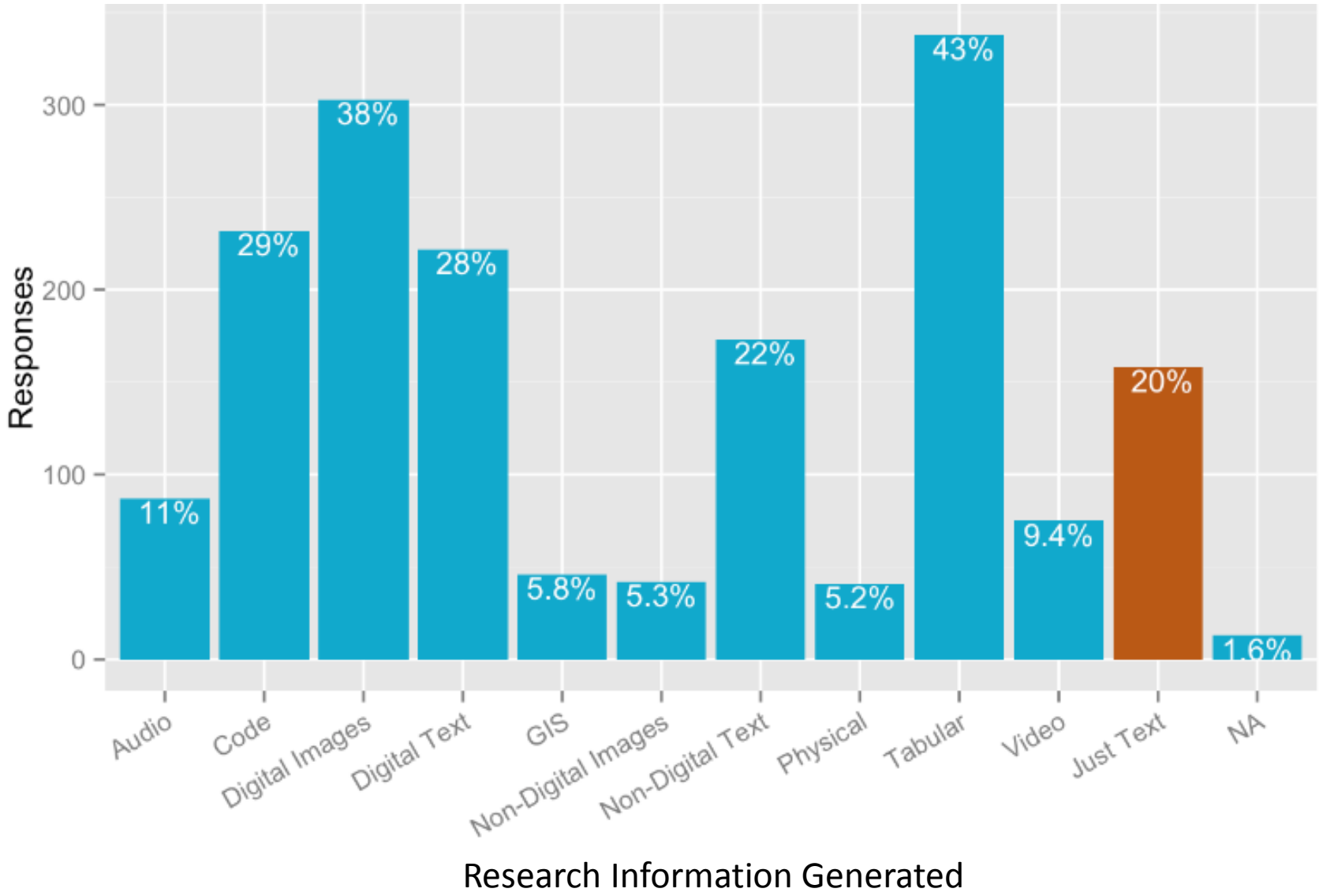
Deployed surveys at beginning of project to gather information regarding current stakeholder community needs :

1. **What research outputs** are students creating as part of their thesis/dissertation research process?
2. Which of these research outputs do the students consider **valuable or essential** for understanding and building upon their findings?
3. Which of these research outputs are they currently **planning to or able to submit** as part of their theses/dissertations packages?
4. What are some of the **common barriers** institutions report in accepting complex (non-PDF) submissions of theses and dissertations?

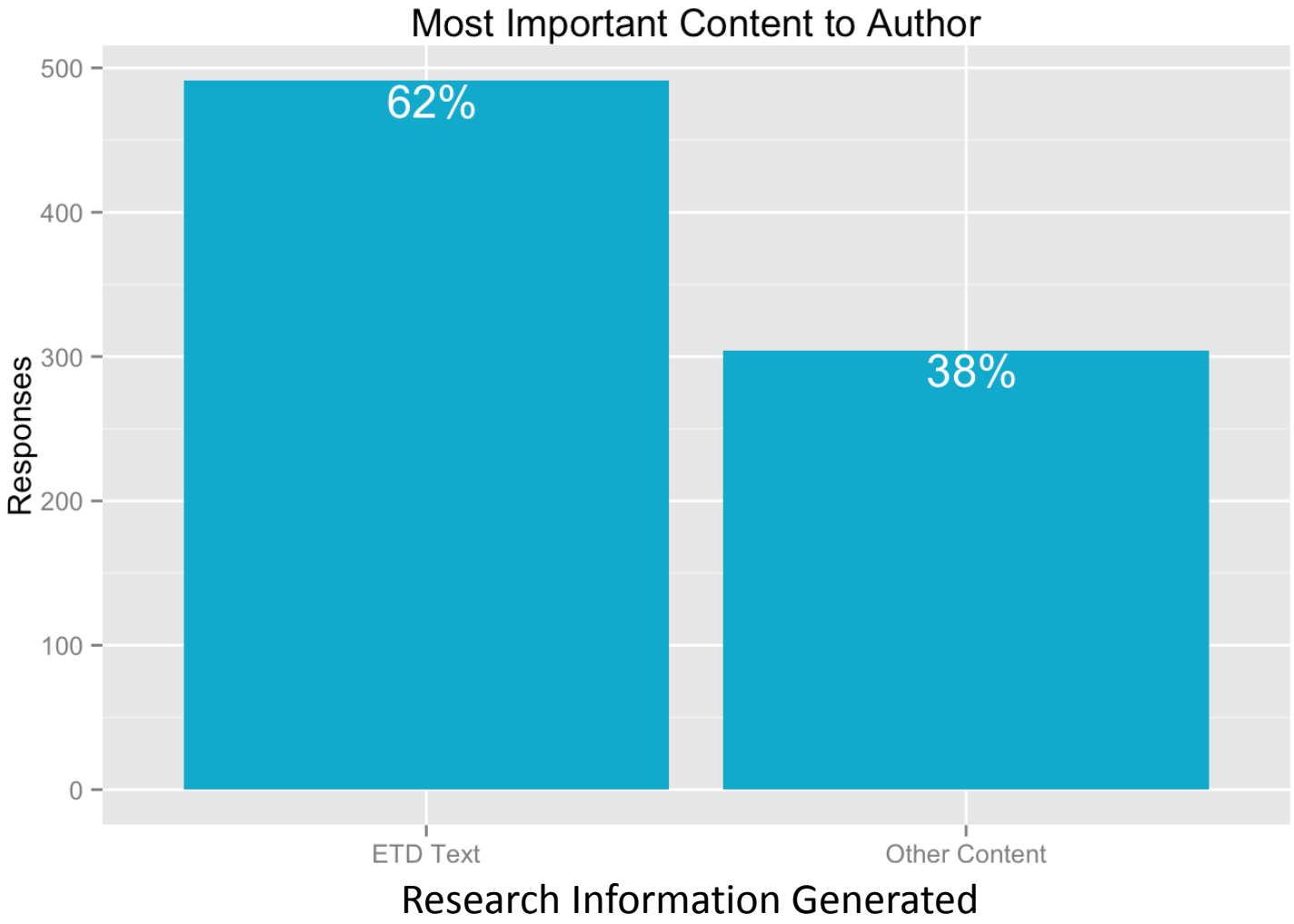
Surveys (cont.)

- Two surveys: 1) Graduate students; 2) Institutional ETD program staff
- 12 universities took part (w/ 12 IRBs…)
- March-May 2015
- Good response: 795 total graduate student responses

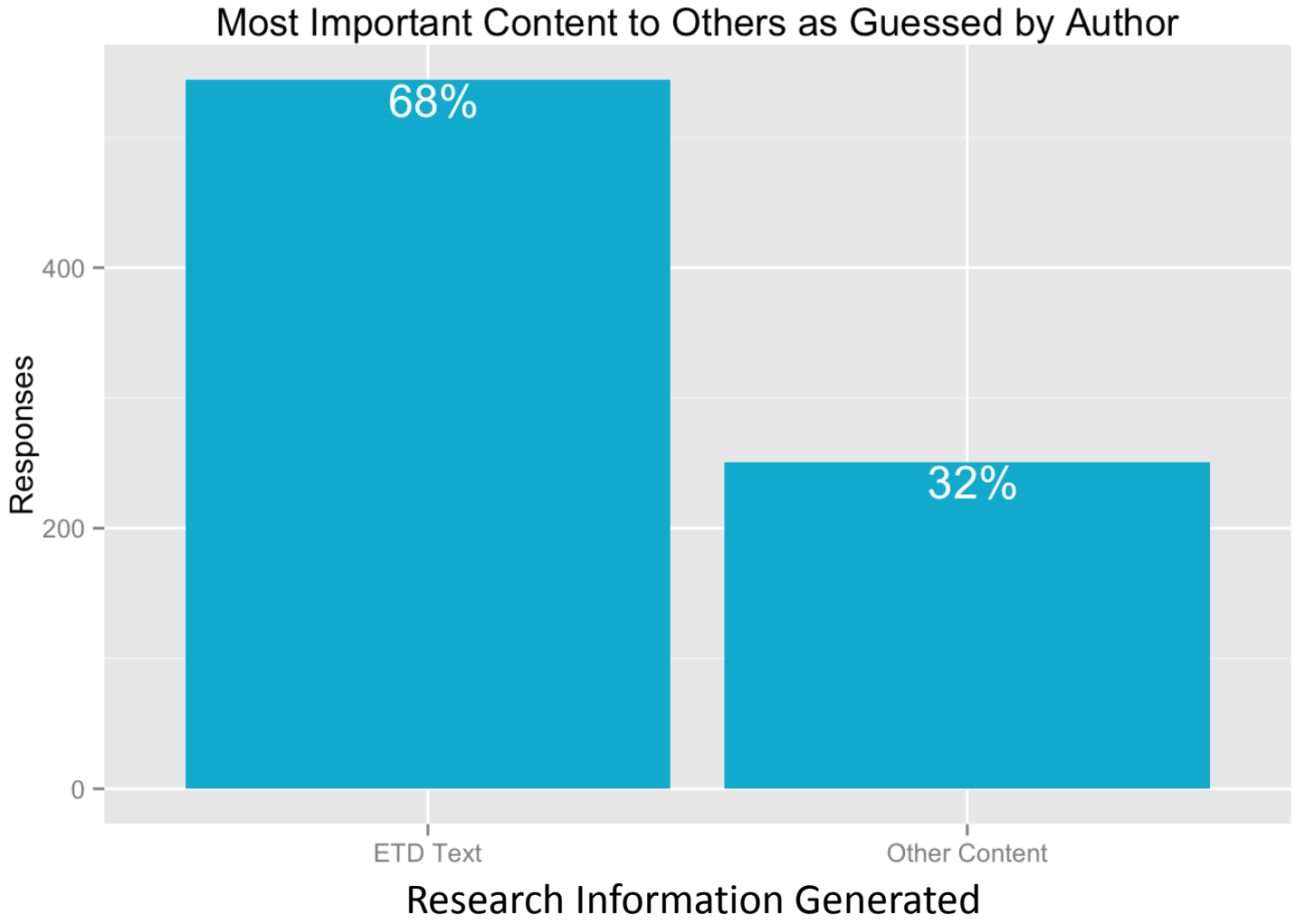
Types of Information/Materials Generated during Research Process as Reported by Graduate Students (respondents N=795)



Question to Graduate Students: What is the most valuable part of your thesis or dissertation research for you?



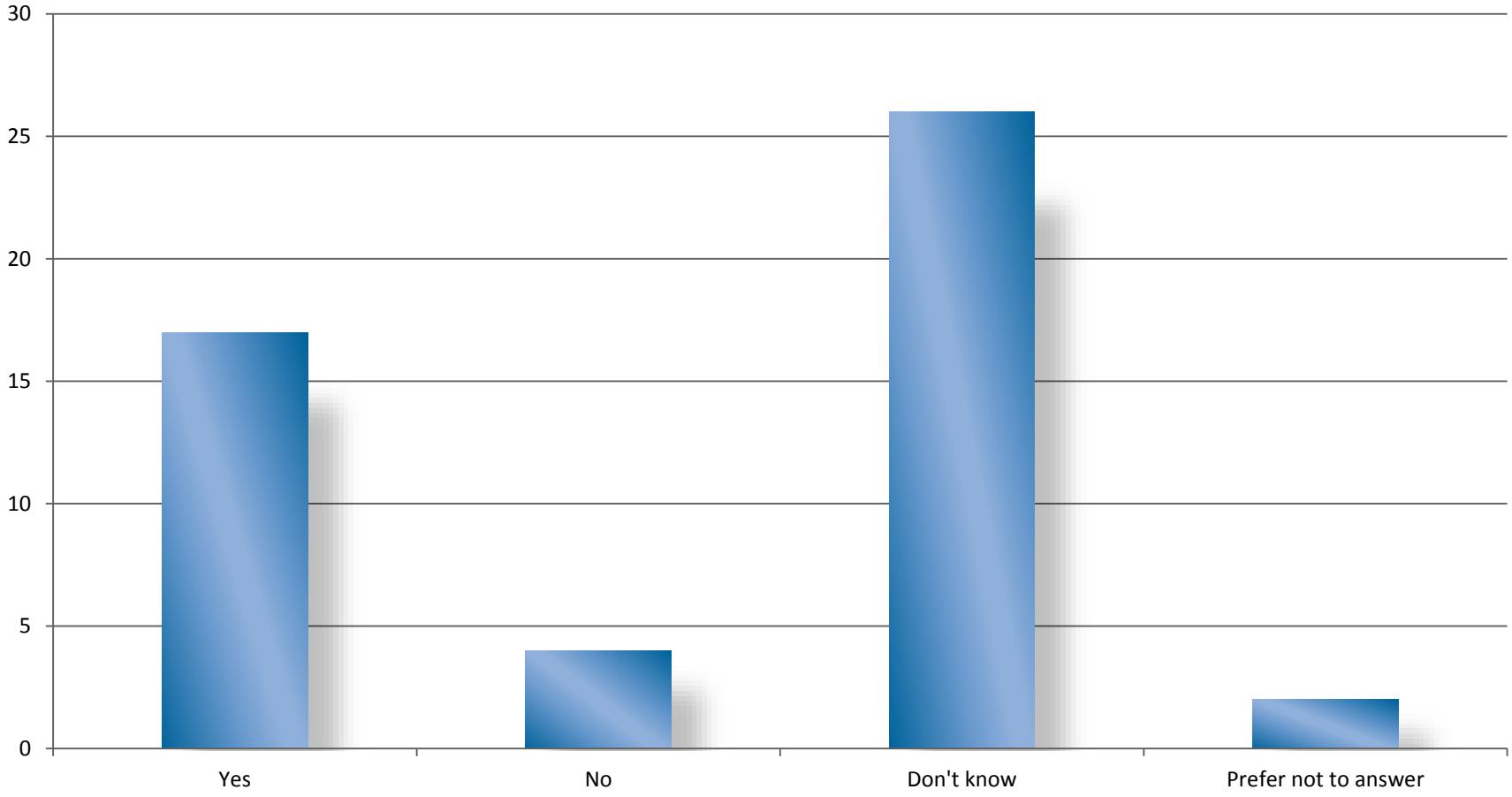
Question to Graduate Students : What is the most valuable part of your thesis or dissertation research for someone else that reads or accesses your work?



The ETD Preservation Gap

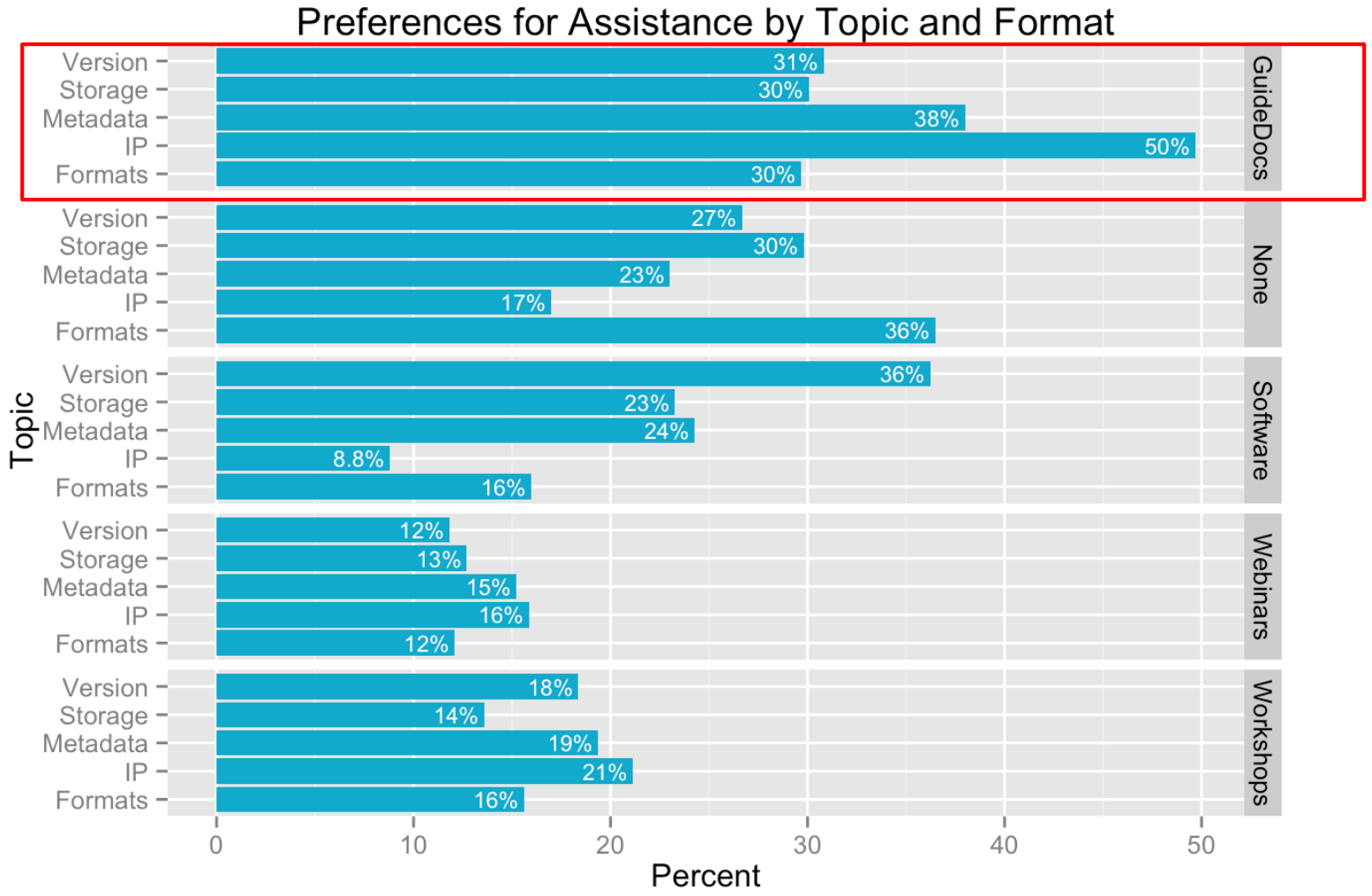
- Although more than a third of students said that the materials beyond their PDF are the most important, only 100 of the 795 surveyed students (13%) reported plans to actually submit those materials.
- An additional 521 (66%) specifically report that they will not submit materials beyond the PDF, and 174 (22%) reported that this question did not apply to their work.
- Students' perceptions of importance, in other words, seem not to be the key drivers for submitting their research outputs as part of their thesis/dissertation packages.

Survey of 65 administrators and ETD staff at the same 12 institutions asked about submission policies concerning non-pdf objects.



Does your institution accept objects in addition to the PDF of the thesis or dissertation?

What resources students would find most helpful in five curation tasks



SECTION 3:

GUIDANCE DOCUMENTS AND TRAINING MATERIALS

Guidance Briefs - foundation

Questions we asked ourselves:

- What do students need to know about digital content management as pertains to their research outputs?
- How might the ETD—as a common rite of passage in research careers—be used to help students learn how to structure, share, and manage their digital content appropriately?

http://educopia.org/deliverables/etdplus-guidance-briefs

EDUCOPIA INSTITUTE

Search

About Us | Communities | Research | Consulting | Events | Downloads | Contact Us

Downloads

- Publications
- Presentations
- Grant Deliverables
  - ETDplus Guidance Briefs
- Under Review

Part of

Electronic Theses and Dissertations
Exploring the evolution of research products by authors on the cusp of their careers and the cutting-edge of their fields.

Downloads | Grant Deliverables | ETDplus Guidance Briefs

ETDplus Guidance Briefs

Download(s):
ETDplus: Copyright
ETDplus: Data Structures
ETDplus: Format
ETDplus: Metadata
ETDplus: Storage
ETDplus: Version Control

Released:
02 May 2016

Author(s):
ETDplus Project Team

Preserving & Curating ETD Research Data & Complex Digital Objects

The *Guidance Briefs* have been authored by the ETDplus project team as short (3-4 page) "how-to" oriented briefs that will help ETD/IR programs build and nurture supportive relationships with student researchers.

These briefs will be released as open documents that colleges and universities can adopt and adapt to assist their own student researchers in understanding how their approaches to data and content management impact credibility, replicable research, and general long-term accessibility: knowledge and skills that will impact the health of their careers for years to come.

Interested ETD stakeholders can download and evaluate copies of the Guidance Briefs using the links on this page. We invite your feedback! Please draw our project team's attention to any components that need to be edited, revised, broadened, or narrowed.

We will gladly accept comments between May 3 and June 30, 2016. Please send an email with your suggestions to Courtney Vukasinovic and/or track your changes within the documents and email those back to us. We will integrate the community's feedback before formally issuing these Briefs later this summer.

The project is generously funded by the Institute of Museum and Library Services (IMLS) and led by the Educopia Institute, in collaboration with the NDLTD, HBCU Alliance, bepress, ProQuest, and the libraries of Carnegie Mellon, Indiana State, Morehouse, Oregon State, Penn State, Purdue, University of Louisville, University of Tennessee, the University of North Texas, and Virginia Tech.

Example of a Guidance Brief

Data Structures

While conducting your research, you begin to amass a significant amount of information - responses from surveys, image files, and geospatial data - that you plan to use in your thesis or dissertation, and that you may also want to reuse later in your career. What can you do to give yourself the best chance that your data will be findable and usable by both you and other researchers in the future? Considering the way you structure your data is a good place to start.

Rationale and Motivations (Why)

As you develop your research, you will have to consider how to structure and store any data you gather. Your choices will be based on the type of research you are doing, how you intend to analyze the data you collect, and the standards of your field. When you prepare and submit your work, either as a thesis or dissertation or as a publication, earlier decisions about how the data is structured may have implications for just how easy it will be for others to access and make use (or sense!) of the data you have worked so hard to generate.

Each field has specific methods of analysis, and potentially a range of software tools that have been developed to help researchers accomplish that analysis. As you begin to gather your data, and as you clean and organize it, consider not just how you will need to use the data today, but how to make sure it will be readable and understandable in the future.

Each discipline or study will have data needs that are specific to the questions being asked. Whether your data is organized in lists, arrays, hash sets, dictionaries, queues, trees, heaps, or relational databases, it is important to be aware of disciplinary norms, as well as institutional and funder requirements, that will make its deposit, storage, and long-term support more likely. Increasingly, the path for long-term support involves taking steps to make sure your data is deposited alongside data collected by others in your field or discipline.

The Basics (How to do it)

Researchers' particular data structures vary depending on disciplines and research questions. Still, there are general guidelines for structuring data that make it more likely to be usable in the future.

The following questions should be considered for any data project, first at the planning stage, again as data is being gathered and stored, and once more prior to final deposit into a digital archive or repository.

1. *What are the data structure standards for your field?* For example, there may be standard ways to label fields that will make your data machine-readable. There may also be specific variables and coding guidelines that you can use that will make your work interoperable with other datasets in or beyond your field. There also may be accepted hierarchies and directory structures in your discipline that you can build upon.
2. *What are the data export options in the software you are using?* If using proprietary and/or highly specialized software to analyze large data sets, export the data in a format that is likely to be supported in the future, and that will be accessible from other software programs. Remember that you may not have access to the same software platforms in

the future that you do today.

3. *What forms of the data will be needed for future access?* Consider the various forms the data may take, and the scale of the data involved. You may need to preserve not only the underlying raw data, but also the resulting analyses you have created from it.

There is also a range of general principles that apply across many data types and forms that you can use to guide your work. These include the following:

Structure

1. Use one variable per column.
2. Make one observation per row.
3. Use human-readable column names.
4. Include one table per tab.
5. If you are using multiple related tables, use an ID or key to indicate how the tables are related.

Example

Movie Title	Director	Distributor	Running Time	Budget	Released
Peter Pan	Herbert ~~Brenon~~	Paramount Pictures	105 minutes	40,000	Dec 29 1924
Girl Shy	Fred C. ~~Newmeyer~~ and Sam Taylor	~~Pathe~~ Exchange	82 minutes	400,000	Apr 20 1924
Greed	Eric Von Stroheim	Metro-Goldwyn-Mayer	140 minutes	665,603	Dec 4 1924

Context

1. Include a readme text file detailing the following information:
   a. Abstract – describe why the data has been collected and for what purpose
   b. Content – include a list of the files in your data package and a brief description of what each file is
   c. Basic Data Dictionary – for each table (file) in the data package, provide a list of the variables included in the file and a description of what each variable is.

Other spreadsheet best practices for data sharing:

Do:

- Consider what your NULL values are and how they are represented
- Consider whether a more robust data dictionary is required (e.g. with more in-depth description of methods, instruments, models, etc. used to generate data)

Do Not:

- Use formatting to convey information
- Place comments in cells
- Use special characters in field names

Tools (What to use)

The "Basics" guidance above belies a tangle of disciplinary-specific guidelines for data curation, including structuring. Consult with your advisors, peers, and campus data specialists at the library to make sure you know the current state of guidance for your field. Some organizations

EDUCOPIA INSTITUTE

Data Structures in brief

Structuring your data well enables you to:

-Reproduce results
-Reuse it in the future
-Share it with others
-Gain and retain credibility
-Comply with IRB/funder requirements

The decisions you make about how you organize and structure your data today will have implications for how you and others can access and make use (or sense!) of that data in the future.

Data Organization Principles:

1. Use one variable per column
2. Make one observation per row
3. Include one kind of data per column
4. Use human-readable column name
5. Use an ID or key to indicate the relationship between multiple tables (*If you apply this principle, you should be using a Relational Database*)
6. Include a readme text file detailing why the data has been collected, and what files comprise your data package.

Whether your data is organized in lists, arrays, hash sets, dictionaries, queues, trees, heaps, or relational databases, it is important to be aware of disciplinary norms, as well as both institutional and funder requirements, that will make its deposit, storage, and long-term support more likely. Increasingly, the path for long-term support involves taking steps to make sure your data is deposited alongside data collected by others in your field or discipline.

Questions to consider for any data project:

1. What are your field's (or funding agency's) data structure standards and requirements?
2. What are your university's policies relating to your data
3. What are your data export options?
4. What forms of the data will be needed for future access?

As a first step in your research, create a "Data Management Plan" that documents your practices for collecting, organizing, backing up, and storing any data you generate. This will help you think through ways of structuring your data that increase its long-term accessibility and use.

Do:

-Consider what your NULL values are and how they are represented
-Use standard data representation (e.g., (YYYYMMDD for dates)
-Use consistent capitalization

***Do Not*:**

-Use formatting to convey information
-Include units in cells along with the data value
-Place comments in cells
-Use special characters in field names
-Use blank spaces or symbols in column names

Discipline-based data repository examples:

-Social Sciences: ICPSR
-Genomics: GenBank
-Earth Sciences: NASA's Earthdata
-Archaeology: tDAR
-Oceanography: NODC
-BioSciences: Dryad

Source - Guidance Briefs: Managing Your ETD Research Files

File Formats in brief

There is no perfect file format. Each will have advantages and disadvantages depending on your research uses. Select a file format, or set of file formats, that helps you complete your research now, and that you can access again in the future. This is true both for your research outputs (what you create) and your research inputs (materials you use in the research process).

Common file types include:

-Images: jpg, gif, tiff, png, ai, svg, ...
-Video: mpeg, m2tvs, flv, dv, ...
-GIS: kml, dxf, shp, tiff, ...
-CAD: dxf, dwg, pdf, ...
-Data: csv, mdf, fp, spv, xlx, tsv, ...
-Text: txt, rtf, tvi, doc, pdf...

What will you do if you no longer can use the software you create your research files in – either because you no longer can afford the software, or the publisher goes out of business, or the latest version is not backwards compatible? Plan for these possibilities by saving your final research files in multiple formats – including a non-proprietary format.

How to select file formats:

-Use software that imports and exports data in common and non-proprietary formats
-Consult with advisors and colleagues
-Convert files from proprietary to non-proprietary formats (e.g., .doc to .txt and/or .pdf)
-Choose a format with functions that support your research needs
-Save final versions of your content in multiple formats in order to spread your risk across multiple software platforms (e.g., docx, pdf, and txt; or mp4, avi, and mpg)

When using website-based materials as evidence or references, take precautions to ensure that if the content moves, changes, or disappears, you still have evidence of its existence. Current tools to help you ensure the longevity of these materials include Robust Links, PermaCC, and Archive-It. You can also take screenshots of important digital content in order to preserve the look and feel of an object.

Many ETD programs favor pdf files. If you export your research outputs to pdf, make sure that you:

1. Embed your fonts
2. Embed (and test!) hyperlinks
3. Archive web-based resources and citations (using a tool like Robust Links, Archive-It, or PermaCC)
4. Store supplementary materials as separate files

Before you undertake any conversion, you need to identify what characteristics of your data are important to maintain during the conversion. For example, are the colors in a document or image important? Is the pagination essential? What about references? You will want to test these after your conversion is complete to ensure that you have a conversion that will meet your needs.

Additional Resources:

-List of File Formats (Wikipedia)
-Sustainability of Digital Formats (Library of Congress)
-Evaluating Your File Formats (UK National Archives)
-Reformatting Guides (US National Archives)

Source - Guidance Briefs: Managing Your ETD Research Files

Metadata in brief

Metadata describes and documents research, data, and publications. More simply, it is information that is created and stored alongside content (such as a thesis or dissertation) in order to help users find and understand that content. It can be especially useful in providing descripting context for the research files that may accompany your dissertation.

For every research file you create, you should also produce metadata describing:

-**Who** created the content
-**What** is the content
-**When** was the content created
-**Where** is it geographically
-**How** was it developed
-**Why** was it developed

What is a Metadata Standard?

Metadata standards provide a structure for consistent (predictable) information. They define the structure and categories of information (e.g., "title," "author," "date") and provide controlled vocabulary to enable interpretation across a discipline. Metadata standards foster uniformity, which permits search/retrieval systems to identify and share the content metadata describes.

ETD metadata tips:

1. Your abstract needs to include a clear description and keywords relevant to your work, including any research files that accompany your dissertation.
2. Be careful with over-reliance on spell-check functions. For example, Microsoft Office does not spell-check capital letters, which can impact chart or graph titles.
3. Create keywords that are not in your title. This will increase the discoverability of your work.
4. Define any acronyms you use (repeat them in both letters and in natural language).
5. Proofread all of your metadata, including department name and advisor name, prior to submission.

A file without metadata is like a can with no label - impossible to understand without opening it (and perhaps even then!)

Typical metadata requested about a pdf during the ETD submission process:

- Title
- Author/Creator
- Advisor
- Resource Type
- Date
- Language
- Abstract
- Subject
- Identifier
- Degree Information
- Rights information

Most ETD submission processes **do not** collect metadata about the additional files you may submit (e.g., datasets, audio or video files, image files, GIS files, CAD files, software programs, etc.). To help ensure that you and your readers will be able to understand what these additional files are and how they may be referenced, used, or built upon, you can develop a simple spreadsheet-based inventory of these items. This inventory should clearly identify how many additional files you are including, what they are, who created them, and what rights and licensing information they are governed by. Submit this inventory spreadsheet as part of your ETD package.

Source - Guidance Briefs: Managing Your ETD Research Files

Storage in brief

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*

Where and how you choose to store your research materials and writings will determine how long they survive. To mitigate against loss, make your own back-ups on a regular, formalized schedule (e.g. daily or weekly).

Threats to storage environments:

-Natural disaster
-Human error
-Human malice
-Drive failure
-Format obsolescence
-Media obsolescence
-Bit rot
-Business failure
-Software or hardware error

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.

Resources

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

Source - Guidance Briefs: Managing Your ETD Research Files

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

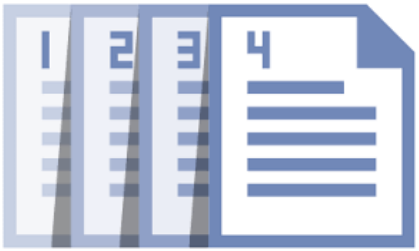
There are also software tools that can help you version your content. These tools store your content in such a way that they can remember its state from revision to revision. Usually, they also allow you to "check in" and "check out" your content, ensuring that revisions never happen simultaneously in 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/file-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
- Another helpful GitHub guide is available from 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/



Source - Guidance Briefs: Managing Your ETD Research Files

Copyright in brief

US Copyright: "that body of exclusive rights granted by law to copyright owners for protection of their work."
http://www.copyright.gov/help/faq/definitions.html

Copyright vs. Patents: "Copyright protects original works of authorship, while a patent protects inventions or discoveries.
http://www.copyright.gov/help/faq/faq-general.html

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., not having a negative impact on the copyright holder).

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/

What can copyright protect?

1. literary works
2. musical works, including any accompanying words
3. dramatic works, including any 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

Do I need a patent?

Universities often have designated offices to deal with questions arising about new inventions or innovations. These questions involve the policies of the university around ownership and IP, and understanding your own institutions' policies is a must. Examples include

-Stanford University's Office of Technology Licensing
-Columbia University Tech Ventures

Fair Use and Public Domain Resources

- Cornell University, Copyright Term/Public Domain in the United States
- CMSI Code of Best Practices in Fair Use for Scholarly Research in Communication
- CAA Code of Best Practices in Fair Use for the Visual Arts
- Columbia University Fair Use Checklist

Source - Guidance Briefs: Managing Your ETD Research Files

EDUCOPIA INSTITUTE

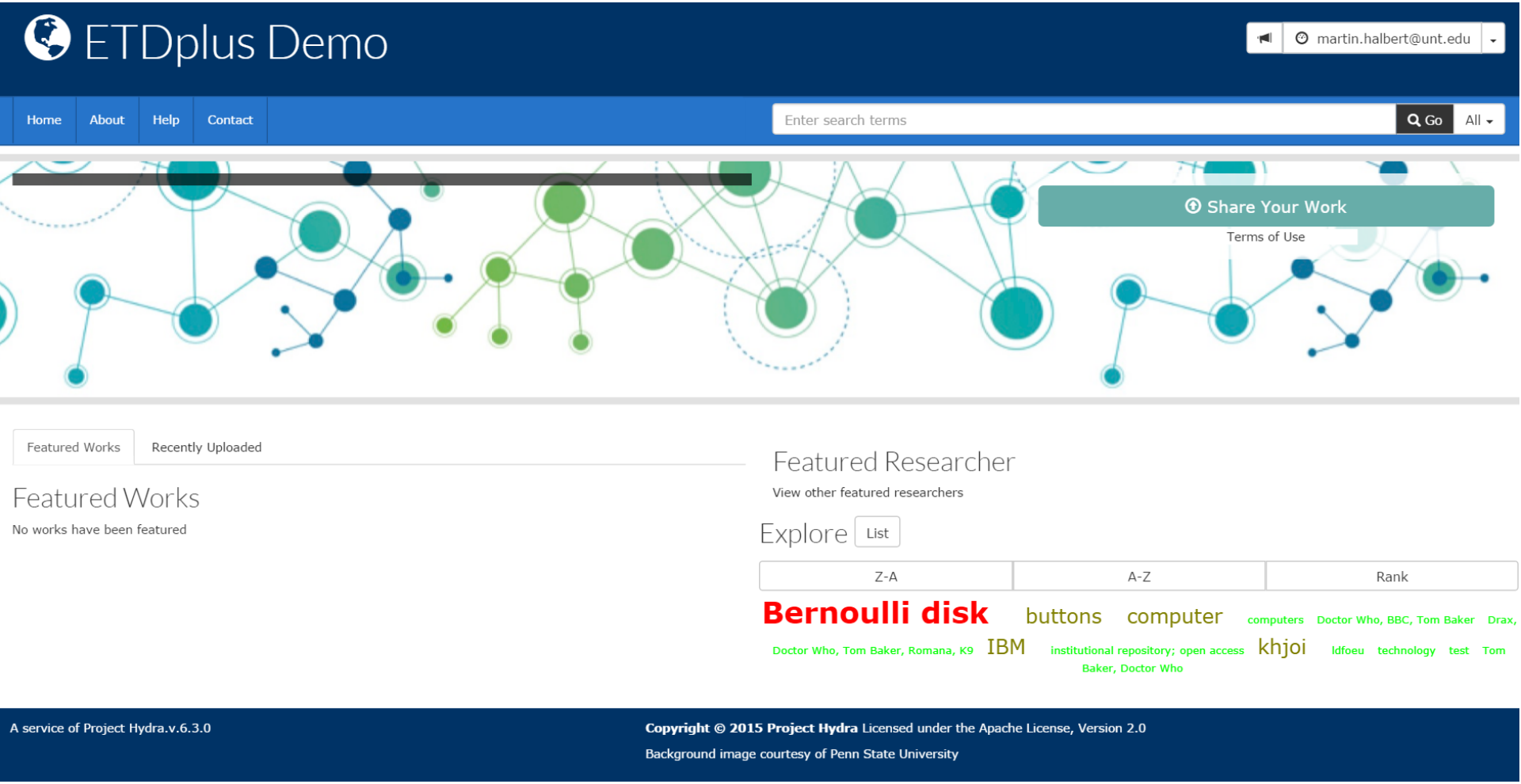
SECTION 4:

HYDRA-SUFIA REFERENCE IMPLEMENTATION TOOL FOR COMPLEX ETD MATERIALS SUBMISSION/INGESTION WORKFLOW

Software Developed in this Project: ETDplus Curation Workbench Tool

- Flexible web-based tool that is designed to assist students in preparing and packaging ETD supplementary materials for long-term preservation and access
- Built using the popular new Hydra-Sufia software combination
- Developed as open source software in the ETDplus project for adaptation and use in institutional workflows
- URL: **http://etdplusdemo.educopia.org/**

http://etdplusdemo.educopia.org/



Get a head
on your repository.

Multi-Purpose Repository Solutions
Flexible User Interfaces
Durable Digital Asset Management

About Hydra

Hydra is not just a repository software solution. Rather, we see it as having three complementary components:

- there is a vibrant, highly active community supporting the work of the project which shares an underlying philosophy behind all that it does
- there are design (and other) principles involved in constructing a successful Hydra "head" for use with compatible digital objects, and of course,
- there are the software components, the Ruby gems, that the Hydra community has constructed which are combined together to provide a local installation

Sufia: An open-source, Hydra-powered repository front-end

Learn more

Community Support

Sufia is maintained and supported by the Hydra community. Hydra is an open-source repository solution built collaboratively to address a broad range of repository needs.

Fedora 4 Repository

Sufia is built and tested atop the latest versions of Fedora 4, an open-source, linked data-compatible, digital asset repository platform that is sustained by the repository community and stewarded by DuraSpace.

Portland Common Data Model

Sufia uses a community-developed data model — the Portland Common Data Model (PCDM) — to model and store all deposited content in a way that bolsters interoperability.

Hydra/Sufia/Fedora Software

- Sufia is an extensible, out of the box, self-deposit repository that powers sites such as Scholarsphere (https://scholarsphere.psu.edu) at Penn State and HydraDAM (https://www.openhub.net/p/hydradam) at WGBH. Until now, no Hydra application has been capable of on running on Fedora 4. Penn State and Data Curation Experts have teamed up to change that. This work has resulted in a beta version of Sufia which runs on Fedora 4. We would love for you to take a look at this work, kick the tires and let us know if you have any suggestions.
- Getting started on Sufia and Fedora 4 isn't difficult. You can create your own Sufia based application by following the directions here: https://github.com/projecthydra/sufia/blob/fedora-4/master/README.md#creating-an-application

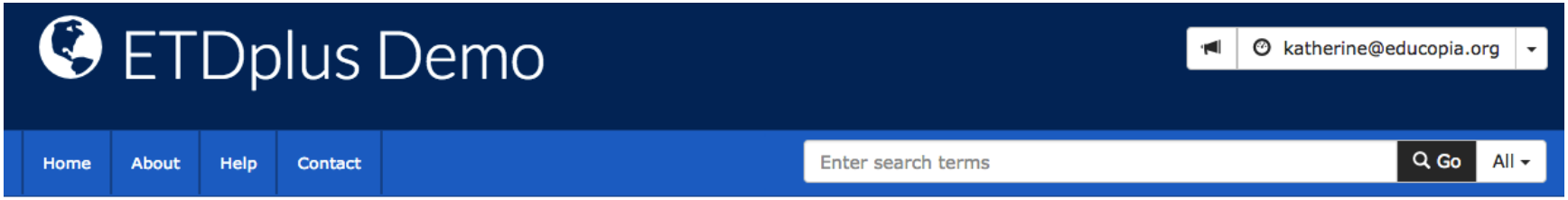
-from http://duraspace.org/articles/2415

ETDplus Curation Workbench Tool (cont.)

- Includes configurable functions and features that integrate basic preservation actions into a simple data upload and review workflow:
  - virus scans,
  - integrity checks,
  - file format identification and validation,
  - personally identifiable information scans, and
  - metadata and versioning support.

ETDplus Curation Workbench Tool (cont.)

- The Curation Workbench tool uses the BagIt specification, the emerging standard for digital content normally kept as a collection of files.
- **BagIt** is a hierarchical file packaging format designed to support transfer of arbitrary assemblies of digital content. (BagIt has been growing in popularity.)
- A "bag" consists of a "payload" (the content) and "tags" (metadata about the content).
- Packages uploaded data and metadata as Bags that users can download and ingest into an array of repository and storage environments.



Select something first

Create New Collection

Descriptions

* indicates required fields

Resource type

Capstone Project
Conference Proceeding
Dataset
Dissertation
Image

*** Title**

Jazz charts by year, 1935-1985

Creator

Katherine Skinner + Add

SECTION 5:

BRIEF THOUGHTS GOING FORWARD

Emerging Themes of ETDplus Project

- Important need exists to intervene with graduate students in the course of their work to inform them about what they need to be thinking about in terms of sustainability of their content
- These skills are also important and useful for scholars in the rest of their career
  - Storage practices
  - Metadata
  - Etc.

Emerging Perspectives Noted in ETDplus Group

- ETD programs constitute an evidence chain of intellectual products by scholars starting at the beginning of their careers
- It is important to preserve the full intellectual content of scholars in their theses and dissertations, not just a perfunctory PDF
- We must begin to adapt our ETD programs to address these newly identified gaps
  - Guidance documents
  - Better ingest software and workflows

QUESTIONS/DISCUSSION

Project Contacts:

Dr. Katherine Skinner (katherine@educopia.org)

Sam Meister (sam@educopia.org)

Q&A