LIS 628 - Data Librarianship

Vicky Steeves

Fall 2019

Contents

Welcome

Welcome to Data Librarianship! In this class, I'm hoping to impart some practical skills, an understanding of best practices, and prepare you for a variety of data librarianship positions.

On this site, you'll find the syllabus, materials for each class, and links to where you should submit assignments.

Past Materials

Here I will list out the downloadable version of each past classes' materials:

- 2018 Fall Book & 2018 Syllabus

Figure 1:

Syllabus

Download a PDF copy of this syllabus here.

Instructor Information

Vicky Steeves, Visiting Assistant Professor
Pratt Institute School of Information
144 W. 14th St., 6th Floor
New York, NY 10011-7301
Phone: 212-647-7682
Email: vsteeves@pratt.edu
Class website: https://vickysteeves.gitlab.io/lis-628-datalibrarianship

Class Information

INFO 628: Data Librarianship and Management
Fall 2019
Class Hours: Thursday, 6:30pm – 9:20pm
Office Hours: By appointment
Credits: 3
Prerequisites: None
Location: PMC 612

Bulletin Description

The world of data is seemingly a new frontier for libraries, yet in some ways, data and data sets are comparable to other print and electronic resources that librarians historically have been charged with locating, teaching, collecting, organizing, and preserving. This course asks how best we can serve the needs of a burgeoning community of data users/producers while meeting the new challenges that data present to our existing skill sets, workflows, and infrastructure. Topics will include data reference and literacy; archives and repositories;

Figure 2:

formats and standards; ethics and policy. Statistical/GIS software and research data management are also explored.

Detailed Description

Class sessions will include lectures, in-class lab activities, and student-led discussions of readings and data-related news. Practitioners in the field will serve as guest lecturers when available and appropriate. The methods, activities, and assignments in this course are designed to (a) maximize peer learning, i.e. opportunities to teach and learn from other students, (b) approximate some of the real world activities and challenges faced by librarians, and (c) get students excited about (rather than intimidated by) this growing niche of librarianship.

Course Goals

The course provides:

- An introduction to concepts and terminology related to data and data services.
- Broad overview of the nature and range of data products and producers.
- Knowledge of how to develop and provide different tiers of data services (including reference, instruction, and collections development) in a library setting.
- Understanding of ethical, social, and political issues related to the creation, use, and reuse of data.

Student Leaning Outcomes
By the end of this course, students will be able to:

- Describe forms, formats, and lifecycles of data and how these vary across disciplines.
- Practice effective strategies and appropriate sources for locating different kinds of data and statistics.
- Construct basic questions and considerations when collecting and appraising data.
- Self-sufficiently acquire technical knowledge.
- Demonstrate the ability to think critically and communicate confidently about issues related to data librarianship.

Course Materials & Schedule

While this syllabus provides a basic framework for the course, it is subject to change. All changes will be announced in class and on the course website (https://vickysteeves.gitlab.io/lis-628-datalibrarianship) and via email. Unless otherwise noted, the readings will be linked below and within each week"s module on this site. All readings will be open access, so you will be able to read them without logging into anything. Readings should be completed in advance of the week they"re assigned below. There is no required textbook.

If you see dead links (it does happen, usually with no notice), weird due dates, or other syllabus problems, please email me!

Readings	Assignments
Unit 1: Definitional/technical overview	
Week 1: Course Overview | August 29th	

Readings	Assignments
No readings	Bring in a sample of something that you consider to be data.
Week 2: Data Basics | September 5th Borgman, "The conundrum of sharing research data" (pp. 6-16 only!) Leek, The Elements of Data Analytic Style, chapters: "Tidying the data" and "Checking the data" (pp. 10-22) University of Leicester, Research Data Definitions NCSU Libraries, Defining Research Data	Facilitator: Vicky Lab 1
Week 3: Spatial, quant, qualitative, and "big" data | September 12th Force11, "Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data" Sutton et al, "A Gentle Introduction to GIS" USC Libraries, "Quantitative Methods" Sage, "Qualitative Research: Defining and Designing" (pp. 1-17) Ellingwood, Justin, "An Introduction to Big Data Concepts and Terminology"	Facilitator: Owen Lab 2
Week 4: Data management planning | September 19th Whyte, Angus and Jonathan Tedds, "Making the case for research data management" Akers, Katherine, and Jennifer Doty. "Disciplinary differences in faculty research data management practices and perspectives" Wiener-Bronner, Danielle "Most Scientific Research Data From the 1990s Is Lost Forever" Perrier et al, "Research data management in academic institutions: A scoping review"	Facilitator: Elizabeth Lab 3
Week 5: Data manipulation & analysis | September 26th Leek, The Elements of Data Analytic Style, chapter: "Statistical modeling and inference" (pp. 34-44) Nguyen, "Using Google Refine to clean messy data" Maceli, "Introduction to Text Mining with R for Information Professionals" Logan et al., "Choosing Statistical Software" Optional: Timmer, "Changing software, hardware a nightmare for tracking scientific data" BONUS: Programming Historian Lessons	Facilitator: Amber Lab 4
Week 6: Reproducibility | October 3rd Steeves, Vicky, "Reproducibility Librarianship" Sayre, Franklin and Riegelman, Amy, "The Reproducibility Crisis and Academic Libraries" Vitale, Cynthia R.H. "Is Research Reproducibility the New Data Management for Libraries?" Dekker & Lackie, "Technical Data Skills for Reproducible Research" (pp. 93-112)	Facilitator: Elizabeth Lab 5
Week 7: Legal & regulatory environment | October 10th	

Readings	Assignments
Boyle & Jenkins, "The genius of intellectual property and the need for the public domain" (pp. 10-14) Arzberger et al., "An International framework to promote access to data" Hagedorn et al, "Creative Commons licenses and the non-commercial condition" Stodden, "The legal framework for reproducible scientific research: Licensing and copyright" Audrey Watters, "Invisible Labor and Digital Utopias"	Facilitator: Mary Work on the project check-in
Unit 2: Library services	
Week 8: Data services in libraries | NO CLASS - October 17th	
Goben, Zilinski, and Briney. "Going Beyond the Data Management Plan: Services and Partnerships" Salo, "Retooling libraries for the data challenge" Reznik-Zellen et al., "Tiers of research data support services" Emmelhain, "Data librarians in public libraries" Coates, "Building data services from the ground up"	Facilitator: Paolo Lab 6
Week 9: Data reference | October 24th	
Witt & Carlson, "Conducting a data interview" Partlo, "The pedagogical data reference interview" Carleton College, "Data, Datasets, and Statistical Resources" Smith, Conte, and Guss, "Understanding Academic Patrons" Data Needs through Virtual Reference Transcripts"	Facilitator: Owen First project check-in
Week 10: Data literacy & instruction | October 31st	
Shields, "Information literacy, statistical literacy, data literacy" Rosenblum et al., "Collaboration & co-teaching: Librarians teaching Digital Humanities in the classroom" Kellam & Peter, "Data instruction; Statistical and data literacy" Shorish, Yasmeen. "Data Information Literacy and Undergraduates: A Critical Competency" Clement, Ryan, Blau, Amy, Abbaspour, Parvaneh, and Gandour-Rood, Eli, "Team-based data management instruction at small liberal arts colleges"	Facilitator: Amber
Week 11: Data collection services | November 7th	
Hogenboom et al., "Collecting small data" Geraci, Diane, et al. Data Basics: An Introductory Text. Chapters 15 & 16, pages 151-168	Facilitator: Paolo Work on your project check-in
Unit 3: Preservation, dissemination, and sustainability	
Week 12: Data sharing, publishing, access, & preservation | November 15th	
NIH, "Frequently asked questions about the NIH Public Access Policy" NSF, "Dissemination and sharing of research results" – pick directorate to read Fienberg et al., Sharing Research Data, "Issues and recommendations" (pp. 3-32) Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. & Dorsett, K., "Changes in data sharing and data reuse practices and perceptions among scientists worldwide" DataCite website and specifically "Why cite data?" ICPSR, "Phase 3: Best Practice in Creating Research Data"; "Phase 6: Depositing Data" (pp. 21-38)	Facilitator: Paolo Final project check-in
Week 13: Data archives & repositories | November 21nd	
Wilson, "How much is enough: metadata for preserving digital data" Thiede, "Preservation in practice: A survey of New York City Digital Humanities practitioners" Kellam & Peter, "Basic sources for supporting numeric data services" (Read pp. 89-105; Skim interesting sources from pp. 106-149) Vines et al., "The availability of research data declines rapidly with article age"	Facilitator: Mary

Readings	Assignments
Week 14: NO CLASS | November 28th Nothing :)	Enjoy your time off!
Week 15: Special concerns | December 5th Cegłowski, "Deep-Fried Data" Asher & Jahnke, "Curating the ethnographic moment" [PDF] Hurley, "When Academic Neurologists Leave, Who Owns Their Research?" Moody, "Elsevier Says Downloading And Content-Mining Licensed Copies Of Research Papers 'Could Be Considered' Stealing" Shaw & Cloud, "Anonymization and microdata: Can we open up granular info without invading privacy?"	Facilitator: Joanna Work on final project in-class during lab time.
Week 16: Future/sustainability of data | December 12th Timmer, "How science funding is putting scientific data at risk" Goldstein & Ratliff, "DataSpace: a funding and operational model"	Facilitator: Joanna Final projects presentations & hand in all final project materials.

Graded Work

Work	Percent of Final Grade
Discussions & Participation	20%
Labs (5% each)	30%
Final Project Check-Ins	20%
Final Project	30%
TOTAL	**100%**

Discussions & Participation (20%)

Each week, students should be prepared to discuss and/or ask and answer questions based on the readings or in-class exploratory labs. A student's participation grade will be based on facilitating class discussion during their assigned week (including coming up with 4-6 discussion questions) and actively participating in discussions led by other students. Discussion questions should be sent to the instructor before 11:45pm the Wednesday night before their assigned class. In addition to these in-class participation activities, some classes may require you to bring in outside materials/objects to fuel discussion.

Labs (6 @ 5% = 30%)

You will have 6 labs, hands-on activities designed to help you engage with that weeks' materials. Whatever is not finished in class will be assigned as homework and must be handed in by 11:55pm EST the Wednesday before the next class. These in-class labs are designed to underscore and amplify understanding in the lecture and readings for a given week. We will use software on the computers of the classroom, however all materials used (software, data, etc.) will be given at least a day in advance for any students who want to use their own machines.

You will be able to resubmit the lowest-graded assignment before the last day of class for regrading.

Final Project (30%)

You will have all semester to work on and refine a final project, which will be presented in the final class of the semester. You have two final projects to choose from, or you can present me an original idea for a final project by 5pm on Friday, September 6th. The format and length of presentations will be determined by the size of the class and the ratio of Project 1 choices to Project 2 choices. We will also have two graded project check-ins, to ensure all projects are on track.

Project Choice 1 – Researcher Perspectives: Design and carry out a small research project of your choice (focus on a data-informed study using either quantitative, qualitative, GIS data, or mixed methods). The end-product will be a research poster (I just need a digital copy as a .pdf, no need to print it) designed for a target conference, such as ACRL, SLA, RDAP or discipline-appropriate conference. Alongside the poster, you will also need to submit:

- Data management plan (2 pages max)
- Methods statement (2 pages max)
- Analysis plan (2 pages max)
- Raw, analyzed, and publication-ready data
- Data documentation or code-book (e,g. README and code-book if you're handing in a spreadsheet)
- Any specialized analysis tools that you used to get the work done
- 1,000 - 1,500 word write-up on your study (with a bibliography or lit review section)

Get some inspiration for posters here:

- http://blogs.lse.ac.uk/impactofsocialsciences/2018/05/11/how-to-design-an-award-winning-conference-poster/
- https://guides.nyu.edu/posters
- https://researchguides.library.tufts.edu/c.php?g=344931&p=4823350

Project Choice 2 – Creating Data Services: based loosely off Dorothea Salo's Tool/Service Review project: http://files.dsalo.info/668syll2014.pdf

First, you need to benchmark other services so you know where you want to build! Pick four institutions that currently offers data services, including data reference, data collection, research data services, and instruction services around data.

- Services intended purpose and audience (e.g. patrons)
- Services fitness for purpose and audience
- Features (what problems does it solve? What gaps does it fill?)
- Limitations (what does it not do?)
- Prerequisites (what do users need to know/do before using the services?)
- Ease of use
- Future prospects (how trustworthy is the service?)

- Cost (staff, software, etc.)

This can be submitted as a spreadsheet or as a report. After you finish your peer benchmarking, you are expected to write a strategic plan (3,000 - 5,000 words) on how to build and maintain your prospective data services department, including:

- Outline of your organization's mission
- SWOT: your organization"s strengths and weaknesses, as well as opportunities and threats
- 4-5 goals aligned with mission
- Priorities, activities, objectives, strategies more in-depth
  - Each goal should have a few different objectives/strategies associated with it
- Road map and timeline
- 1 page executive summary (should be written last!) to succinctly convey the future direction, priorities, and impact.

For inspiration, you can look at the Strategic Agenda for Research Data Services from Oregon State University.

Final Project Check-ins (2 @ 10% = 20%)

The first check-in will be on October 17th, and the second check-in will be on November 14th. These check-ins are to ensure that you are progressing on-schedule for your final projects, and to provide a way to get feedback at various points in the process from both the instructor and classmates.

Project Choice 1 – Researcher Perspectives:

1st Check-in – You will submit a 2 page data management plan and a 2 page (max) methodology statement to the instructor alongside any data that you've gathered or created. The methods statement should explain what kind of research you plan to do, the question you are asking, and how you plan to evaluate potential answers to that question. For instance, explain the qualitative analysis methods you would plan to use when evaluating or documents from the web. In-class, you are expected to present your thesis/idea, any data you have or plan on using, your methods, and your data management strategy. Your classmates will provide you feedback via a form created and circulated by the instructor.

2nd Check-in – By this point, you should be about 3/4 of the way done with your final project. As such, you'll hand in the data management plan (denoting any revisions from the DMP handed in for the first check-in and this one), data (raw and analyzed), research documentation, and a 2 page (max) analysis statement. This statement should explain how you've approached analyzing the data you've gathered or created – how you may have differed from the methods you planned on, for instance. In-class, you'll present your updates to the project. Your classmates will provide you feedback via a form created and circulated by the instructor.

Project Choice 2 – Creating Data Services:

1st Check-in – You will submit your benchmarking study to the instructor ahead of class. In-class, you will present the results, including a discussion of how you chose to evaluate (e.g. on a scale of 1-4, completely qualitatively, etc.). Your classmates will provide you feedback via a form created and circulated by the instructor.

2nd Check-in – You will submit the outline of your organization's mission, the 4-5 goals aligned with that mission, and the SWOT analysis to the instructor ahead of class. In-class, you will present this as if to a steering committee of a library, where you'd want to make the case for administrative buy-in for your goals and strategic mission. Your classmates will provide you feedback via a form created and circulated by the instructor.

Project Choice 3 - Student-designed project:

If you've chosen to design your own project, please discuss options for the check-in with the instructor.

Assessment & Grading

All assignments must be completed to receive a passing grade in the course. Assignments are to be submitted to me via the course site by 11:55pm (Eastern) on the Wednesday before the next class period. Except for medical and family emergencies, assignments submitted late will receive a lower grade. Additionally, late assignments will receive a grade but no comments.

Pratt's grading scale	
Superior work	A 4.0 (96-100), A- 3.7 (90-95)
Very good work	B+ 3.3 (87-89), B 3.0 (83-86), B-2.7 (80-82)
Marginally satisfactory	C+ 2.3 (77-79), C 2.0
Failed	F 0.0 (0-69)

Portfolio

Work completed for this course may be included in your portfolio. If you have a fall deadline (November 1st), please meet with to discus scheduling of projects you might want to include from this course. For more information on each program"s portfolio requirements, please visit the program"s respective webpage:

- MS Library & Information Science: Portfolio - http://bit.ly/prattmslisportfolio
- MS Information Experience Design: Portfolio - http://bit.ly/prattmsixdportfolio
- MS Data Analytics and Visualization: Portfolio - http://bit.ly/prattmsdavportfolio
- MS Museums and Digital Culture: Portfolio - http://bit.ly/prattmsmdcportfolio

You are encouraged to meet with your adviser about including projects in your portfolio.

Policies

Attendance Policy (this course)

When you have to miss a class, please notify the instructor in advance. Also, submit a 1-page reflection on the required readings for the missed class and complete the weekly lab that was missed. Students with 3 or more absences will be asked to drop the course, per Pratt policy. For more information on Pratt Institute's Attendance Policy, please visit http://bit.ly/prattattendance.

Academic Integrity Code

Academic integrity at Pratt means using your own and original ideas in creating academic work. It also means that if you use the ideas or influence of others in your work, you must acknowledge them. For more information on Pratt's Academic Integrity Standards, please visit http://bit.ly/prattacademicintegrity.

Students with Disabilities

Pratt Institute is committed to the full inclusion of all students. If you are a student with a disability and require accommodations, please contact the Learning/Access Center (L/AC) at LAC@pratt.edu to schedule an appointment to discuss these accommodations. Students with disabilities who have already registered with the L/AC are encouraged to speak to the professor about accommodations they may need to produce an accessible learning environment.

Course Communication

Email is the best way to get in touch with me. Generally speaking, it is my policy to respond to email within 24 hours Monday - Friday. Note: I will not answer questions, via email or ptherwise, about assignments after 5pm on the day *before* they are due.

Chapter 1

Course Overview

In today's class, we'll go over:

- Who I am, my teaching philosophy
- Who y'all are
- The class structure, policies, & the syllabus

But first! Please fill out the initial self-assessment (my eyes only): https://forms.gle/LnpVzZ9NXNWVmLcg7 & this fun survey which I'll use later on in class: https://forms.gle/3gju8HC1zUCDyEUz6

1.1 Introducing me

This is me! I'm Vicky Steeves and this is your co-instructor, Little Boss (LB for short):

I'm the Librarian for Research Data Management and Reproducibility at New York University, and I'm a Visiting Assistant Professor here at Pratt. My pronouns are she/her. Some basic facts about my professional background:

- BS in Computer Science & Information Technology from Simmons College
- MLIS from Simmons College
- National Digital Stewardship Residency, 2014-2015 at the American Museum of Natural History
  - My project was to gain a broad overview of the extent and status of AMNH digital assets pertaining to Science. To do so I developed a structured interview guide designed to measure and describe scientific digital assets resulting in a metric to predict ongoing data curation needs.
- Current position at NYU, 2015 - now
  - A dual appointment between NYU's Division of Libraries and Center for Data Science. I support students, faculty, staff, and researchers in creating well-managed, high quality, and reproducible research.
  - My research centers on integrating reproducible practices into the research workflow, advocating openness in all facets of research (manuscripts, code, data, analysis tools, etc.), and building/contributing to open infrastructure. I have a great passion for community-owned, open source tools and infrastructure for reusable, reproducible, well-documented, and responsible sharing of research. I think a lot about the corporate capture of the scholarly record, and how my work in data management and reproducibility can either contribute to or disrupt that.
- I am also the co-founder of the LIS Scholarship Archive, an open repository for library and information science scholarship.

I am a fairly laid-back instructor. I believe in balancing knowledge of best practices and theory with hands-on exercises to amplify understanding and build skill-sets in learners. I want to collaborate with learners on building a community within the classroom and in the field writ large.

If you're interested in learning more about my background, ethics, interests, I just chatted with Thomas Padilla about it here: https://acrl.ala.org/dh/2018/04/04/repro/

1.2 Introduce y'all

Please share with the class:

- Your name
- Preferred pronouns (if comfortable doing so)
- What you want to do after you graduate

Next round:

- The best breakfast you've ever had in your life
- If an extra-terrestrial civilization demanded an updated version of the Golden Voyager Record, what files would you upload?
- 1 recommendation of a restaurant, movie, book, article, yoga position, album, broadway show, etc.

1.3 Introducing the class

I want to give everyone lots of opportunities for hands-on application work of the best practices we'll discuss. As such, most classes will look like this:

- 00:00 - 00:45: Student-led discussion of readings
- 00:45 - 01:45: Lecture
- 01:45 - 02:00: Break
- 02:00 - 03:00: Exploratory lab/activity

Now, let's look at the syllabus!

1.4 Homework

1. Bring in a piece of data that's *not digital*. We'll discuss this at the beginning of next class.
2. Send me two weeks when you'd like to faciliate discussion. If everyone asks for the same weeks, I'll have to make some decisions for folks.
3. Do the readings & be prepared to discuss them.

Chapter 2

Data basics

Welcome to our first routine class! We are going to be going over the basics of data, especially to make sure we are on the same page with some vocabularies.

Agenda for today's class:

- 6:30 - 7:00 let's see your not-digital data that you brought in!
- 7:00 - 7:45: discussion
- 7:45 - 8:05: break
- 8:05 - 9:00: lecture
- 9:00 - 9:20: lab/activities

2.1 Discussion Questions

Before we dive right into the discussion questions I've come up with, what were everyone's overall impressions of the readings? What surprised you/didn't surprise you? Anything you *really* agree or disagree with?

1. What is data? **BONUS**: Is the word 'data' plural or singular? How does 'dataset' come into play here?
2. What is missing from the definitions of data you read over the last week?
3. Do you think libraries, archives, and museums are required to collect *all* the data from folks within their institution? Including all the raw, processed, analysis-ready data?
4. What do you think are the top 3 things that should be kept alongside data to make it useful?
5. What data do you think should be shared with the public, in general terms?
6. What is the difference between research data and research records?

2.2 Lecture: what about data?

So, the first thing when unpacking what it means to be a data librarian, is understanding the concept of `data` – how it evolves over time, how different domains of research or internal services conceptualize it, and why it's in the purview of libraries and librarians to be involved.

QUESTION 1: What types of data do you think data librarians are most likely to encounter on a day-to-day basis?

Let's start at the very beginning (a very good place to start). A few of you mentioned that you took some data related courses, such as:

- Information technology

Figure 2.1:

- Metadata Design
- Data Analysis
- Information visualization classes of various levels
- Digital Humanities

QUESTION 2: What data did you work with in these classes? Did you think of the materials you worked with as data before I asked you to list related classes?

And I'm sure the bevy of answers I got to this question illustrate the point – that data, in many respects, is in the eye of the beholder. One person's "just a lab notebook" is another person's "rich unstructured data". One person's "primary source documents" are another person's "dataset" or "corpus".

I find a lot of the time that when I'm discussing this question of "what is data" with researchers (students, faculty, independent researchers, etc.), they are often holding a treasure trove of data without understanding how valuable the materials are for others, or for the scholarly record!

The point I'm driving at, is what one calls *data* depends largely on discipline/field of study and methodologies. Jargon switching is an essential part of the job of a data librarian (and I'd argue, librarianship as a whole, but this isn't intro to librarianship, it's intro to *data* librarianship). What I call 'jargon switching' is actually known as code switching to folks who study linguistics – a really great illustration of the concept. Code switching is the process of shifting from one linguistic code (a language or dialect) to another, depending on the social context or conversational setting. It happens when a speaker alternates between two or more languages, or language varieties, in the context of a single conversation. It's typically used to align speakers with others in a specific situation (e.g. defining oneself as a member of a group or community) and to announce specific identities, create certain meanings, and facilitate particular interpersonal relationships.

This is really a skill to cultivate as a data librarian, and it starts right at the definition of data. Let's discuss this definition of research data from the reading:

> Research data, unlike other types of informtaion, is collected, observed, or created, for purposes of analysis to produce original research results.

What about this one:

> Data that are descriptive of the research object, or are the object itself.

Or this very general definition:

> Any information you use in your research

QUESTION 3: Which definition do you like best, and why? What types of research methods or data types might each definition exclude?

It's worth mentioning that when discussing data with researchers across domains, sometimes the word 'data' doesn't even come up – it's offensive to some, especially in the humanities where their work can be deeply interpersonal, and researchers are deeply embedded in the communities they study. These are *real people*, not data points. And so when I'm speaking to humanities scholars, I tend to use words like 'materials' and 'your work' instead of 'data' or 'corpora'.

So now let's examine a table of data types, taken and adapted from our readings –

Types of Data	
Documents & spreadsheetsLaboratory notebooks, field notebooks, diariesQuestionnaires, transcripts, codebooksAudiotapes, videotapesPhotographs, filmsProtein or genetic sequencesSpectraAdministrative dataStandard operating procedures and protocols	Slides, artifacts, specimens, samplesSource codeMetadataDatabase & database contentModels, algorithms, scriptsContents of an application (input, output, logfiles for analysis software, simulation software, schemas)Methodologies and workflows

QUESTION 4: What's missing from this table? Do you see any biases in the data types listed here?

Borgman (2011) also gives us a few different categories of data to think about:

- *Observational data*, such as weather measurements, surveys
- *Computational data*, such as models, simulations
- *Experimental data*, such as chemical reactions in a lab
- *Records of government, business, public/private life*, such as archival records, open government data, law cases

QUESTION 5: Based on your readings and life experience so far, do these categories generally fit? What nuances are we missing in these broad strokes?

In my opinion, one thing we miss when we have the large general cateogories of data is that data is created within a *variety of situations* – for example, completely unrelated to research or as a by product of research. A researchers' letters (emails now) are rich data for historians, future researchers in the same area, and for geneological research. Imagine someone taking your emails now and preserving them as valuable data. I personally would be apalled by that notion because I am real salty in my emails, but it's basically the foundation of whole fields studying the history of scholarship.

Here's an example from a guest blog post I wrote for the NMNH field book project, during my time as an NDSR at the American Museum of Natural History:

> While the scope of my project is in the digital realm, I am constantly shown the value of field books and older scientific texts through conversations with science staff. All the scientists at the AMNH are as passionate about our historic collections of field notebooks as they are about their own field notes [...] During my interviews, many scientists have expressed to me that the most important data from their work are actually their field notes–the majority of which are still done with good old fashioned pencil and paper.

I go onto describe an encounter with a curator in the mammology department at the AMNH, in their historical fieldbook collection:

> As I flipped through some of the newer, less vulnerable books he told me he often comes into this section of the archives to examine old accounts of expeditions, which tend to include species descriptions, and descriptions of environments that have changed drastically in the intervening years. He told me sometimes he visited these books as frequently as once a day because the information within these hundred year old volumes is so helpful to his research.

These fieldbooks were meant as a record of a research expedition – not as research material itself. Yet, it has become a trove of data for researchers both within and outside the AMNH.

All this underscores the main point…**data can be a LOT of stuff** and a good portion of a patron-facing job will be convincing them that their materials, including what they consider emphemera, is likely important and should be well-documented and preserved.

However, we have to think critically. The assumptions behind individual work can deeply influence the way data is created, gathered, procured, or otherwise generated. Borgman (2011) discusses two opposite ends of the spectrum (her spectrum, I'd add) at length:

- *Exploratory investigations*: pursue specific questions, usually about a specific phenomemon
  - Examples: interviews, biological research collecting water samples from the same beach to look at bacteria,
- *Observational investigations*: systematically capturing the same set of observations over long periods of time to propose a new theory – interpretation of natural phenomena
  - Examples: large-scale surveys, climate modeling, sky surveys/telescopes

When working with data, understanding the origins, assumptions, and methods involved in its creation will help frame how you (and your patrons!) can use or collect (in the library sense) it. Some good questions to ask are:

- What are the potential sources of bias in this data?
- What is the method of data collection wtihin the research study?
- What is the strongest argument for using this data?
- What is the strongest argument against using this data?

Once you understand how the data was created, for what purpose, with what biases, under which methodologies and frameworks, maybe you can work with it or assist others in working with it. This wouldn't be a class about data without at least one lifecycle diagram! But, it's true that data looks different depending on the stage in the lifecycle:

Each step in the lifecycle has it's own set of questions:

Planning for data	Processing data	Analyzing data	Preserving data	Publishing data
how will we manage this data?what data sources will we use to get the data?what format will the data be in?how will we collect this data?	how will we check, validate, or clean the data?how will we describe that process?how will we describe the data?	how will we interpret data?what research outputs will be produced?what format will they be in?how will we ready this data for publication?	what is the best archival format for our type of data?What needs to be preserved alongside our data to make it useful to others?What type of metadata and documentation do we submit with it?	which repository or archive is the right one for our data?how will we make sure our data is indexed widely?how can we get credit for sharing our data?

Planning for data	Processing data	Analyzing data	Preserving data	Publishing data
STORAGE & BACKUP	—>	—>	—>	—>
METADATA & DESCRIPTION	—>	—>	—>	—>

QUESTION 7: How is this workflow model flawed?

No matter which stage the data is at in the lifecycle, it needs the following at a minimum:

- metadata
- documentation/description/codebook
- tools used to create/modify/analyze

QUESTION 8: What metadata or descriptions might I need for data from a set of 10 interviews?

Leek (2015) has a very quanitative point of view, but I think it's worth going over his definition of a dataset anyway, which he posits as having:

1. Raw data - the data as you got it, read-only – "If you did any manipulation of the data at all, it is not the raw form of the data"
2. Tidy data - cleaned data ready for analysis
3. Codebook - describes each variable and the values in the tidy data – also contains information about the study design and choices you made
4. Recipe - explicit steps of how you get from raw to tidy – to Leek, ideally this would be a script to limit 'human error' in taking the raw data as an input and the tidy data as the output. It also has information about the software used to go from A –> B and the system you used it on (macOS, Windows, Linux).

Next week, we'll cover (including some of the reasons spreadsheets make me angsty).

2.3 Lab/Homework

Pick on of these 'Collections as Data' personas:

- Undergrad - political science major
- High scool math teacher
- Postdoc - humanities
- Data journalist

Then evalaute these four datasets to answer the question, "is this data?", from the perspective of your persona:

- Surreal art GIF
- 1984 Oscars - Winners and Nominees
- 3d Articulated Woolly Mammoth
- Heat use by household

Your evaluation should follow this template (`.odt` version: https://cloud.vickysteeves.com/index.php/s/ycQpAYJe3BbnGBt):

Name of object:

Creator:


Create Data
Process Data
Analyze Data
Preserve Data
Publish Data
Reuse Data
Research Data Lifecycle

Figure 2.2:

Link:

Date Accessed:

Briefly describe the digital object.

Why is this or is this not data?

What type of research domains or methodologies might consider this data?

What are the potential sources of bias in this data?

Is the data clean or does it require more work to make it 'tidy'? What makes it clean or messy?

Is there any accompanying material to help secondary users understand what it is? If so, please describe and link to it. If not, describe what documentation or metadata might help make it useful for others.

Please put all the answers for each digital object in one file and upload it here: https://cloud.vickysteeves.com/index.php/s/ZQBSWHt8ey3BnSG. You will be asked for the password, which will be given out in class.

Make sure that your filenames look like the following: `YYYY-MM-DD_LastName-FileName.ext` – so one might look like `2018-09-06_Steeves-DataBasics.pdf` or `2018-09-06_Steeves-DataBasics.txt`.

Chapter 3

Spatial, quant, qualitative, and "big" data

Agenda for today's class:

- 6:30 - 6:45: shall we discuss the homework assignment? What did you all think? What worked or didn't work?
- 6:45 - 7:45: discussion
- 7:45 - 8:00: break
- 8:00 - 8:45: lecture part 1 & first group quiz
- 8:45 - 9:00: break
- 9:00 - 9:20: lecture part 2 & second group quiz

3.1 Discussion Questions

Owen, our facilitator this week, prepared the following questions for in-class discussion:

- How did you feel about the readings? What were you most anxious or unsure about? Were there any particular concepts or passages that really resonated with you?
- What constitutes "big" data? What is big data's place in our day-to-day lives?
- How do big data and machine learning interact?
- What does bias look like in qualitative research? What about in quantitative research?
- Where can bias enter the study framework?
- How can people collecting data counteract and counterbalance biased findings from quant vs qual data?
- How can cultural heritage organizations use GIS to improve services for their designated communities? What data (per week 2's discussion) can these organizations ethically collect for mapping?

3.2 Lecture: data types

I want to begin the lecture with talking about FAIR data, which could be big, qualitative, quantitative, or spatial! The idea behind FAIR data standards was to come up with a "a minimal set of community-agreed guiding principles and practices, data providers and data consumers - both machine and human - could more easily discover, access, interoperate, and sensibly re-use, with proper citation, the vast quantities of information being generated by contemporary data-intensive science."

Findable: minimally contain basic machine actionable metadata that allows it to be distinguished from other Data Objects; uniquely and persistently identifiable

Figure 3.1:

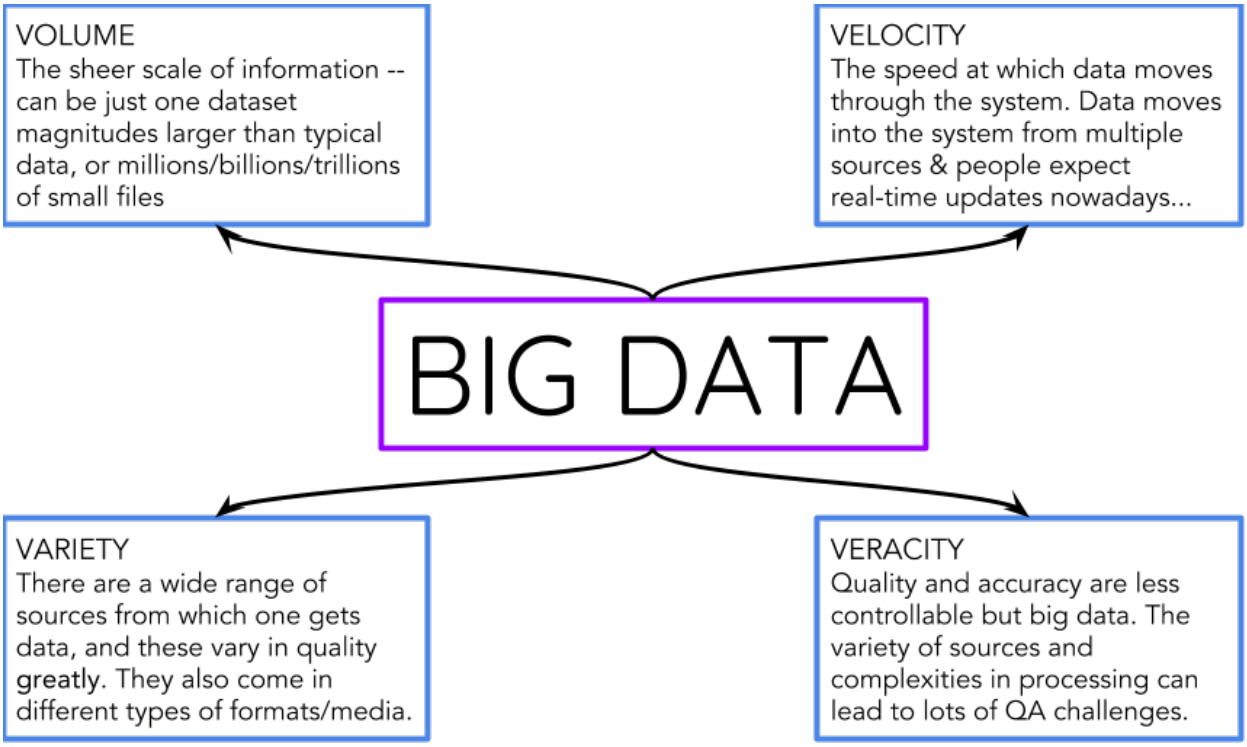


Figure 3.2:

Accessible: always obtained by machines and humans (with authorization); machines and humans alike will be able to judge the actual accessibilty of each Data Object

Interoperable: machine-actionable; formats utilize shared vocabularies and/or ontologies; syntactically parseable and semantically machine-accessible

Re-usable: compliant with FAI; (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources; rich enough metadata and provenance to enable proper citation

The point is to make data and metadata stewardship **normalized**.

3.2.1 Big Data

Big data is really hot right now and there's lots of snubbery around what constitutes "big". I typically say, if you can't open it on your local computer, then it's big. This could mean data that is hundreds of GB, data that is a TB or PB big, or millions of small files that have to be used in conjunction to make sense. Basically, big data has the following characteristics and problems:

Big data can also be defined by the processes and computing strategies/technologies used to handle the large datasets. There are significant challenges with scaling up processing and analyzing big data – imagine having a dataset that's a few TB big. You run a data cleanup program in a higher performance computing environment that takes a few days to run, only to find out at the end that it didn't work the way you hoped. It gets *real frustrating* real fast. And this is just one of many challenges in designing solutions, tools, and best practices in working with big data, or assisting patrons working with big data.

Working with big data follows similar steps to working with smaller-scale data:

- Ingesting the data

- Storing it securely
- Cleaning it up
- Analyzing it
- Visualizing the results

To do this with big data, however, you need a computer that can perform faster and at a larger scale than your desktop. This is called HPC in most places – high performance computing. It's formed with a bunch of *clusters* where people submit jobs, and wait for them to be first in the queue, then run. HPC work/clusered computing has a lot of benefits:

1. Clusters can combine all the available storage and memory they have to make storage and efficiency much better than your average laptop or desktop computer.
2. Clusters are more resilient – they are structured to prevent hard/software issues from affecting access to data and processing.
3. Clusters scale really well – a lot of people can use them at once without affecting efficiently. Yay parallel computing!

Typically you have to use the command line to access clusters. So a typical workflow might look like this:

1. SSH (securely log-in) into a cluster
2. Copy, scrape, or transfer data onto the storage layer of the cluster. These are typically distrubuted, which means that data is stored across multiple nodes to be accessed by compute resources.
3. Run some data clean-up operations via the command line, through python scripts, R scripts, bash, etc. These operations are typically about formatting the data simiarly so analysis pipelines run correctly, filtering out bad/unnecessary data, or validating data.
4. Run some analysis pipeline via the command line, through python scripts, R scripts, bash, etc. Batch processing is one method that allows researchers to analyze large data – this consists of breaking up the data into smaller pieces, scheduling each piece on a different compute node, and calculating the final result after running some job (analysis script). Folks also use real-time processing, which means data is processed and immediately fed back into the pipeline, forcing the system to react to the new information.
5. Visualize what happened during the analysis – make some graph, time series, etc. that helps others understand what happened during the analysis stage.

I think one opportunities for libraries is to start *collecting* and *serving* big data for researchers, but we can talk about this during the collection development section of the cours!

3.2.2 GIS

Geographic Information Systems (GIS) is anything that integrates, stores, edits, analyzes, shares, and displays geographic information. GIS applications are tools that allow users to create interactive queries, analyze spatial information, edit data in maps, and present the results of all these operations.

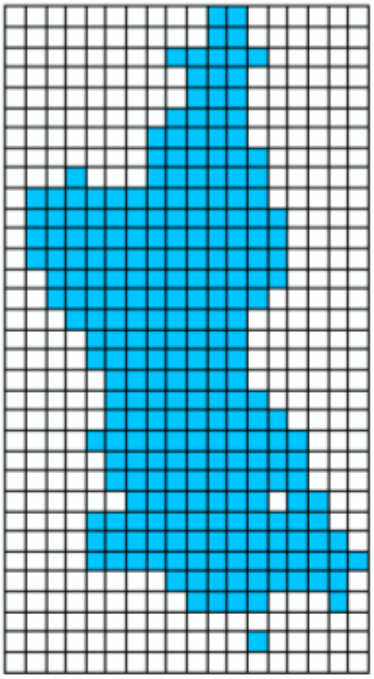
GIS applications work with a few types of data:

- Raster: data stored as grids of values. If you zoom too close to raster data, you'll get a very pixelated image.
- Vector: data stored as a seris of X,Y coordinates. These are represented in points, lines, and areas.

Let's look at these data types in a bit more detail! **Vector** data is a way to represent *features* in a GIS environment, or something you can see in the environment. Some examples of a features include libraries, trees, houses, museums, etc. Vector features have attributes, typically non-GIS data that describes it. Attributes can be numerical or not – whatever is most useful as a descriptor. Vector attributes are typically stored in tables where each row represents one instance or *record*, with one record per feature in the vector layer. Columns are called 'fields', and fields define the properties of each attribute, such as height, color, etc.

A vector feature is represented using geometry, specifically connected vertices (X,Y or X,Y,Z coordinates). If there's only one vertex, it's a *point*. If there's two, it's a *line*. If there's four or more where the first vertex

Vector representation of lake polygon features. Note the detail maintained in the shoreline that defines the boundary between water and upland.



Raster representation of the same lake. With a raster data format, cells are used to encode geographic data. The entire area of each cell is assigned to a single category and boundary details are lost.

Figure 3.3:

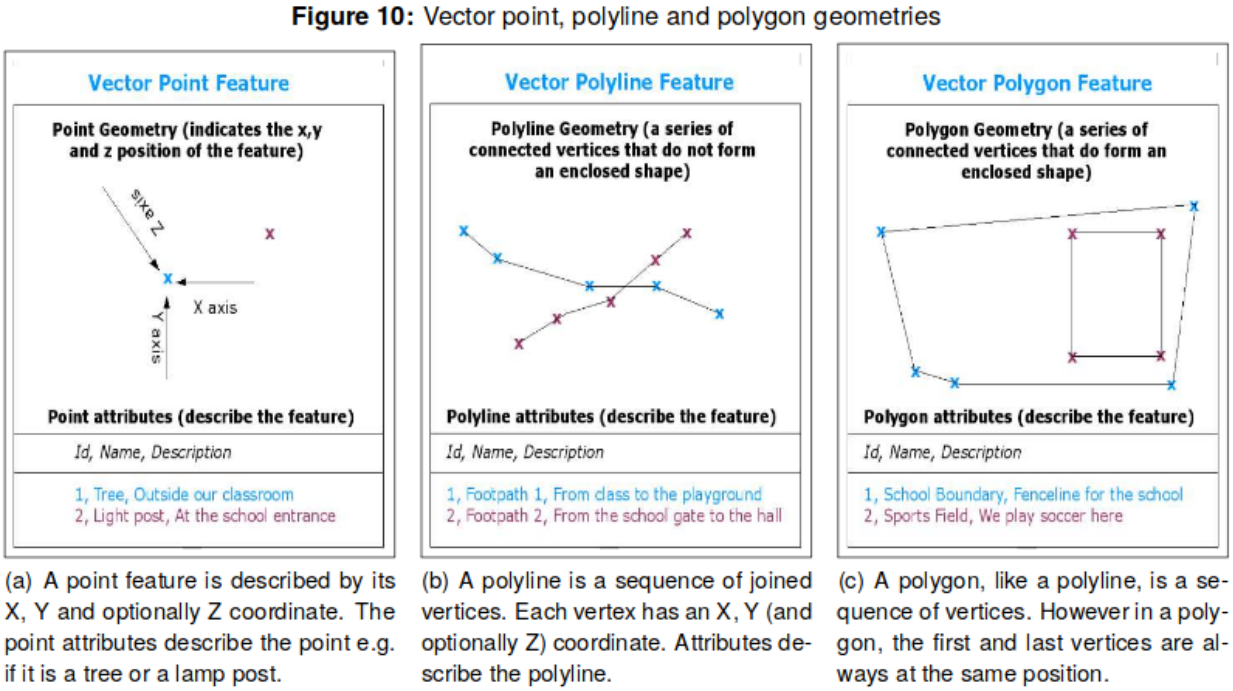


Figure 3.4: From our GIS reading this week

is the same as the last, it's a *polygon*.

Vector data can be stored in a database, or as files. The most common format is a *shapefile*, which is a group of three or more files, typically `.shp` (geometry of vector features), `.dbf` (attributes of vector features), and `.shx` (index that helps GIS program find features quickly) files.

Raster data is a grid of regularly sized pixels. The size of those pixels determines the spatial resolution. Raster data is very good for showing continually varying information, and can provide a lot of interesting contexual data when layered with vector data.

3.2.3 Qualitative

The choices you make as a researcher working with text, images, sounds, networks/relationships, are determined by theoretical framework and *that* is qualitative research – e.g. grounded theory, or phenomenology, to inform choices to code from voices of participants vs. inductive coded theory.

In general, qualitative researchers attempt to describe and interpret human behavior based primarily on the words of selected individuals ('informants', 'respondents', or 'participants' typically) and/or through the interpretation of their material culture or occupied space.

The advantage of using qualitative methods is that they generate rich, detailed data that leave individuals' perspectives intact and provide multiple contexts for understanding the phenomenon under study. Qualitative methods are used by a wide range of fields, such as anthropology, education, nursing, psychology, sociology, and marketing. Qualitative data has a similarly wide range: observations, interviews, documents, audiovisual materials, and more.

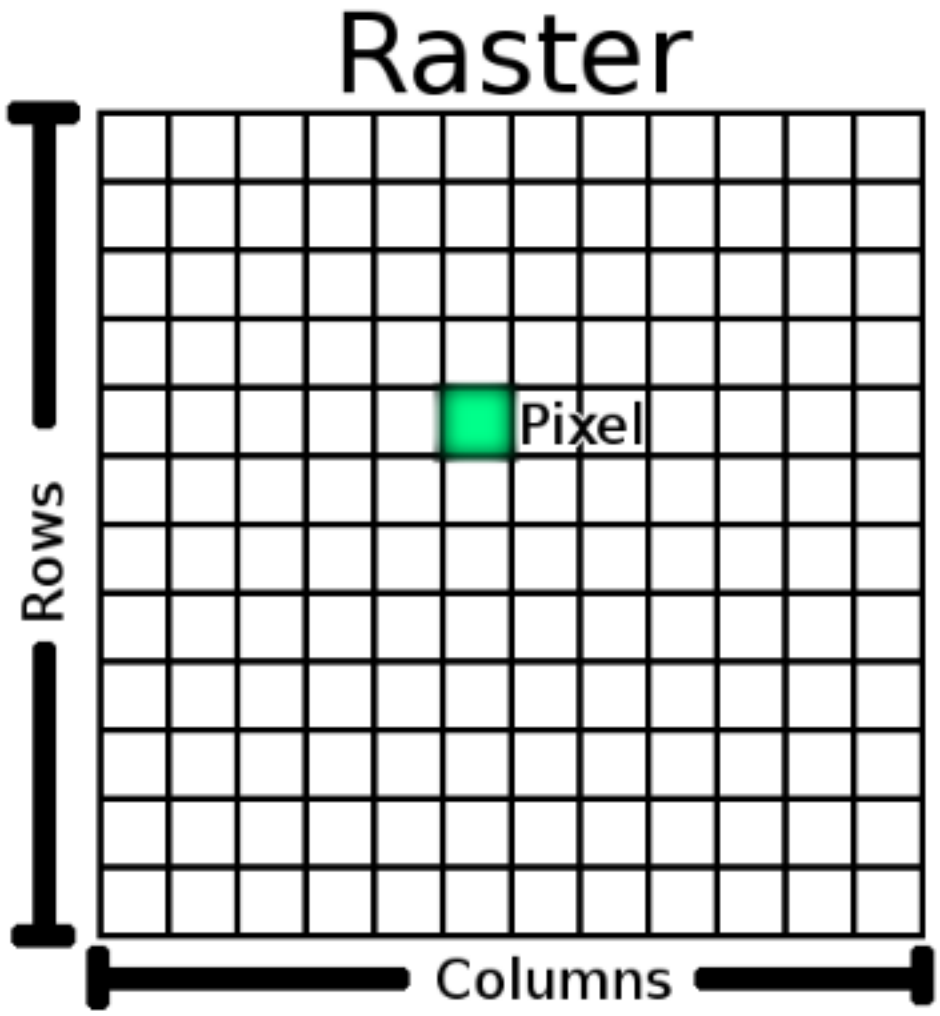


Figure 3.5:

Strengths	Challenges
small sample sizes - can get more in-depth	Usually small sample sizes - hard to generalize
researchers are often imbedded in the cultures and experiences of others	cultural embeddedness increases the opportunit
a holistic view of the phenomena under investigation	Pretty impossible to replicate results
Obtain a more realistic view of the lived world	Arriving at different conclusions based on the s
Create a descriptive capability based on primary and unstructured data	Research using human subjects increases the ch
Respond quickly to changes that occur while conducting the study	Data gathering and analysis is often time consu

Table 1. Strengths and challenges of conducting qualitative research.

Qualitative Data Collection Methods

Qualitative researchers utilize many methods of data collection throughout their study, such as:

- Direct Observation - Learn about behaviors and interactions in natural settings
- Oral History - a series of in-depth interviews conducted with one or multiple participants over an extended period of time
- Focus Groups - Generate unique insights into shared experiences and social norms
- In-depth interviews - Explore individual experiences and perceptions in rich detail. Provides a record of a subject's unique perspective or experience as it relates to a particular issue
- Autoethnography - self-reflection and writing to explore anecdotal and personal experience and connect this autobiographical story to wider cultural, political, and social meanings and understandings
- Document review - a systematic procedure for analyzing and interpreting data generated from documents; in qualitative research, document analysis is often used to corroborate findings from other data sources such as surveys, interviews, etc.

Triangulation is a process of verification that increases validity by incorporating three different viewpoints and methods. Triangulated techniques are helpful for cross-checking, or for seeking out varying perspectives on complex issues and events. You can triangulate your study with different techniques in qualitative research. For instance, you might do in-depth interviews, document review, and direct observation! If you want to take a mixed-methods approach, you could do in-depth interviews, document review, and summary statistics!

Study Design

There are five, not necessarily sequential, components in qualitative research designs. How they are presented depends upon the research philosophy and theoretical framework of the study, the methods chosen, and the general assumptions underpinning the study:

Goals: Describe the central research problem being addressed but avoid describing any anticipated outcomes. Questions to ask yourself are: What issues do you want to clarify, and what practices and policies do you want it to influence?

Conceptual Framework: Questions to ask yourself are - What do you think is going on with the issues, settings, or people you plan to study? What theories, beliefs, and prior research findings will guide or inform your research, and what literature, preliminary studies, and personal experiences will you draw upon for understanding the people or issues you are studying?

Research Questions: Usually there is a research problem that frames your qualitative study and that influences your decision about what methods to use, but qualitative designs generally lack an accompanying hypothesis or set of assumptions because the findings are emergent and unpredictable. More specific research questions are usually the result of an iterative design process rather than the starting point.

Methods: *Structured approaches* to applying a method or methods to your study help to ensure that there is comparability of data across sources and researchers and they can be useful in answering questions that deal with differences between phenomena and the explanation for these differences. An *unstructured approach* allows the researcher to focus on the particular phenomena studied. This facilitates an understanding of the processes that led to specific outcomes, trading generalizability and comparability for internal validity and contextual and evaluative understanding.

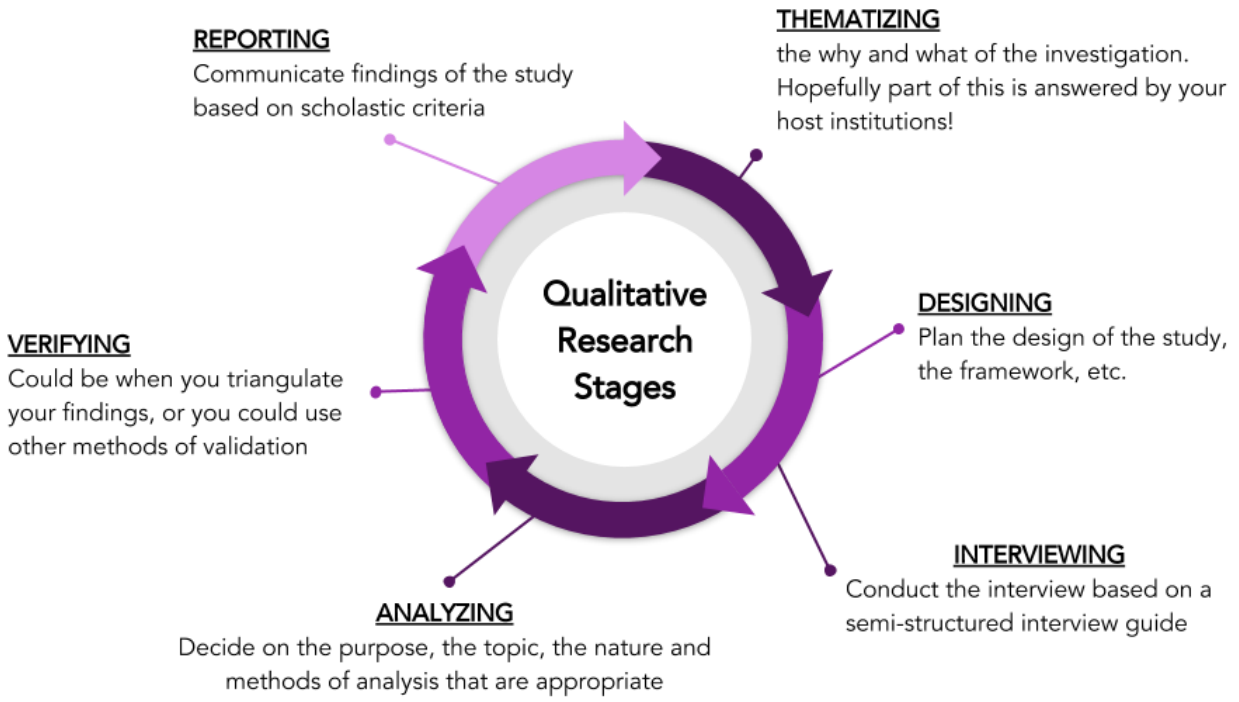


Figure 3.6:

Validity: Qualitative researchers must attempt to rule out most threats to validity after the research has begun by relying on evidence collected during the research process itself. This is in order to effectively argue that any alternative explanations for a phenomenon are implausible. Questions to ask yourself are: How might your results and conclusions be wrong? What are the plausible alternative interpretations and validity threats to these, and how will you deal with these? How can the data that you have, or that you could potentially collect, support or challenge your ideas about what's going on? Why should we believe your results?

Qualitative studies can fall into a category of design, such as:

1. Naturalistic – studying real-world situations as they unfold naturally; nonmanipulative and noncontrolling; the researcher is open to whatever emerges [i.e., there is a lack of predetermined constraints on findings].
2. Emergent – acceptance of adapting inquiry as understanding deepens and/or situations change; the researcher avoids rigid designs that eliminate responding to opportunities to pursue new paths of discovery as they emerge.
3. Purposeful – cases for study (e.g., people, organizations, communities, cultures, events, critical incidences) are selected because they are "information rich" and illuminative. That is, they offer useful manifestations of the phenomenon of interest; sampling is aimed at insight about the phenomenon, not empirical generalization derived from a sample and applied to a population.

3.2.4 Quantitative Data

Quantitative research focuses on gathering numerical data and generalizing it to explain a particular phenomenon. The goal of quantitative studies is to determine the relationship between an independent variable and a dependent/outcome variable. Quantitative research designs typically fall into two categories:

- Descriptive: subjects are measured once; establishes associations between variables

- Experimental: subjects are measured before and after a treatment; establishes causality

In quantitative research, data is usually gathered through structured research instruments and results are based on large sample sizes. These are *theoretically* very replicable, though we'll get to that in week 6. All aspects of the study are designed before data is collecting to avoid HARKing – Hypothesizing After Results are Known. HARKing is defined as presenting a post hoc hypothesis as if it were an a priori hypothesis – basically, making the data fit the hypothesis instead of the other way around. It's seen as very inappropriate. Quantitative research attempts to generalize concepts widely as well as predict future causal relationships and future results.

The most common types of quantitative data you'll likely create, or be asked to find by patrons, are:

- Microdata: information at the level of individual respondents (e.g. survey respondents)
- Aggregated data: data combined from several measurements
- Statistics: an aggregated description of datasets – they interpret or summarize datasets. The output of some kind of analysis and usually made to be read by humans (and machines)

Quantitative research follows a similar research pattern to what we've seen above:

1. Identify your research question – what is the problem or phenomenon being investigated?
2. Identify the conceptual framework under which you'll do your work. What's your hypothesis?
3. Design your methods accordingly:
    1. How will you sample the population/phenomenon you plan on studying?
    2. How will you collect relevant information about it? How will you test that data to make sure it's accurate?
4. Data analysis – what mathematical techniques, software, or instruments are you using to test your claims?
5. Interpret the results you found during analysis – did the data refute or affirm your hypothesis? What trends did you see? What are the key implications of your work? What were the limitations and biases?
6. Visualize that ish! It helps present the numerical data in a succinct and precise manner when done right

Strengths	Limitations
Larger sample sizes means more generalizable results	Missing contextual data such as attitude, motivation, etc.
Supposedly, more objective and accurate	Very inflexible & rigid
Supposedly, easily replicable	Chance for structural bias
Summarize vast amounts of information & compare across categories	Carried out in artificial environment, so higher possibility of "lab results" instead of "real-world results"

3.3 Lab/Homework

3.3.1 In-class

I am going to split you into two groups to do two in-class quizzes! Then we'll compare and contrast. These aren't graded, but are to help underscore parts of the lesson.

GIS group quiz: https://forms.gle/VHYPNLeLeWTxToZr8

Quantitative data quiz: https://forms.gle/FmMbgmkcUW8wDV4V8

3.3.2 Outside class

Pick a type of data to work with – either quanitative, qualitative, or GIS. "Big" was left out because I do not have the compute resources readily available to give you to do any labs with large datasets – sorry!

Hand in the assignments here: https://cloud.vickysteeves.com/index.php/s/iRLiyFyEDjZar7f. The password will be given in class.

GIS

Let's continue with comparing and contrasting the variety of GIS types.

1. Examine these four datasets and their metadata:
    1. https://geo.nyu.edu/catalog/nyu-2451-34171
    2. https://geo.nyu.edu/catalog/nyu-2451-34502
    3. https://geo.nyu.edu/catalog/nyu-2451-34495
    4. https://geo.nyu.edu/catalog/stanford-jt342yd7485
2. Either use the link to import the datasets directly into CARTO with the direct link via the catalog page, or download the datasets and upload them into QGIS or your GIS program of choice.
3. Answer the following questions about the process of playing with this data:
    1. Based on the data provided, can you tell me which school district has the most subway stops? And which school districts have the most public libraries nearby? BONUS: make a deritive map to show me your answer in addition to the written part.
    2. How can you use different types of GIS data to help you come to different conclusions?
    3. When might you use points over polygons, and why?
    4. When would you use raster data over vector data, and why?
    5. A researcher comes in with a query – they want to overlay a map of NYC from 1857 with a current map of NYC. How would you advise them? What GIS data types would you recommend they use to see the differences in the cityscape?
4. BONUS: Is the data FAIR?

Hand-in the answers to these questions along with any deritives of the files you made while answering the above questions (maps, subsets of the data, etc.).

Qualitative

Let's do a deeper dive into some open qualitative data.

1. Examine these two collections of data and the associated metadata:
    1. https://data.qdr.syr.edu/dataset.xhtml?persistentId=doi:10.5064/F6CN723S
    2. https://data.qdr.syr.edu/dataset.xhtml?persistentId=doi:10.5064/F6PN93H4
2. Based on your in-depth examination of the data and metadata, please answer the following questions for each data collection:
    1. Why are some datasets in the collection private and some public?
    2. What are some of the ethical considerations you might mull over when requesting to access data from this collection?
    3. What are some of the ethical considerations you might mull over when *granting access* to data to fellow researchers?
    4. Describe the datasets in relation to the idea that qualitative research is about immersion in a given community; that the information gathered is a *gift*.
    5. A researcher comes to you with interview audio and transcriptions they collected via fieldwork among an at-risk community in Turkey. They want to know if they can release a redacted version of the transcripts to the general public. What do you advise?
3. BONUS: Find an example of non-textual qualitative data and hand it in with the answers to the above.
4. BONUS 2: Is the data FAIR?

Quantitative

1. Download and examine these four datasets and their metadata:

    1. https://www.openicpsr.org/openicpsr/project/100094/version/V1/view
    2. https://www.kaggle.com/neuromusic/avocado-prices
    3. https://www.openicpsr.org/openicpsr/project/100117/version/V1/view
    4. https://osf.io/zafr6/
2. Based on your in-depth examination of the data and metadata, please answer the following questions for each dataset:
    1. What type of quantitative data is represented in the datset?
    2. Do you understand how the data was gathered or generated from the information provided?
    3. Could you use the dataset as-is without much cleaning? If no, why not? What needs to be cleaned?
    4. What are the limitations of the types of questions you might ask of this data?
    5. A researcher comes to you asking for a dataset that describes the demographic information of a given neighborhood where there has been a lot of real estate development happening. What type of quantitative data might you advise they use?
3. BONUS: Find some quantitative data that isn't microdata, statistics, or aggregate data and hand it in with the answers to the above.
4. BONUS 2: Is the data FAIR?

Chapter 4

Research data management

Agenda for today's class:

- 6:30 - 6:45: check-in about the homework
- 6:45 - 7:45: discussion
- 7:45 - 8:00: break
- 8:00 - 9:00: lecture
- 9:00 - 9:20: lab on the OSF

4.1 Discussion Questions

Elizabeth, our facilitator this week, prepared the following questions for in-class discussion:

1. Before these readings, what would you have said were the biggest benefits of well-managed research data?
2. After the readings what benefits would you add to that list?
3. In the article "Disciplinary Differences" there are figures showing how data management varies across the disciplines surveyed in their research (p.8-12, figures 1-4). Did any of the discrepancies surprise you?
    a. What might account for these differences and are they something that needs to be rectified?
4. What are the potential consequences of the loss of research data from the 90s?
    a. On a scale of "ehh" to armageddon how worried are you about the "digital dark age" of research data?
5. Imagine there's been a research paper written on exactly what you want to study, but you can only access either the paper or the data. Which would you choose? What might change your answer?

4.2 Guest Lecturer

Today I have the virtual pleasure to introduce Nick Wolf as the guest lecturer for this class section. Nick works with me at NYU as a fellow Data Management Librarian and is also an affiliated faculty member in Glucksman Ireland House. He completed his PhD in history at the University of Wisconsin-Madison and is a specialist in the cultural and linguistic history of eighteenth- and nineteenth-century Ireland. In addition

Figure 4.1:

Current Biology 24, 94–97, January 6, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.11.014

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,[1,2,*] Arianne Y.K. Albert,[3] Rose L. Andrew,[1] Florence Débarre,[1,4] Dan G. Bock,[1] Michelle T. Franklin,[1,5] Kimberly J. Gilbert,[1] Jean-Sébastien Moore,[1,6] Sébastien Renaut,[1] and Diana J. Rennison[1]
[1]Biodiversity Research Centre, University of British Columbia,

sets (23%) were co
down of the data t
We used logistic
tionships between
that at least one e-

Report

and indeed many studies have found that authors are often unable or unwilling to share their data [8–11]. However, there are no systematic estimates of how the availability of research data changes with time since publication. We therefore requested data sets from a relatively homogenous set of 516 articles published between 2 and 22 years ago, and found that availability of the data was strongly affected by article age. For papers where the authors gave the status of their data, the odds of a data set being extant fell by 17% per year. In addition, the odds that we could find a working e-mail address for the first, last, or corresponding author fell by 7% per year. Our results reinforce the notion that, in the long term, research data cannot be reliably preserved by individual researchers, and further demonstrate the urgent need for policies mandating data sharing via public archives.

Figure 4.2:

to his research, he serves as the assistant editor of the journal Éire-Ireland, the web editor for the American Conference for Irish Studies, and as co-chair of the collections committee for the HathiTrust digital library.

You can follow Nick on Twitter @nicholasmwolf and follow his publications/work on his ORCID page - 0000-0001-5512-6151.

4.3 Vicky's Lecture

First, let's start with a video: https://youtu.be/N2zK3sAtr-4

What are the problems RDM tries to solve?

- Data becoming *useless* because of:
  - Little or no documentation (metadata, provenance, etc.) about it
  - It's in a file format no longer supported by modern computers
- Data becoming *lost* because:
  - It's stored on an old and unstable storage medium (e.g. zip drive)
  - No one wrote down where it is or it wasn't backed up…people actually just lose it
  - Folks keep it in boxes under desks and then retire or die and no one can understand it (loop back to the first problem)

So now let's look at some real-world examples of the repurcussions of a lack of data management:

This study published in *Current Biology* evaluated 516 studies published between 1991 – 2011, and found that the odds of the data (when it was reported, e.g. "email me for it") being extant fell by 17% per year. The authors report:

> The major cause of the reduced data availability for older papers was the rapid increase in the proportion of data sets reported as either lost or on inaccessible storage media.

Additionally, the authors found that the odds of finding a working email address for any of the paper authors fell about 7% per year. There are lots of RDM issues wrapped up in this paper: safe storage, unique indentifiers for researchers, data documentation, and data publishing.

One of the more famous data management horror stories is about a paper called *Growth in a Time of Debt* (it even has it's own Wikipedia page!). In the study, the authors ('RR') argued that when "gross external debt reaches 60 percent of GDP", a country's annual growth declined by two percent, and "for levels of

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

In t… g pop-
histor… s? Are
search… a man-
public… main… pirical,
growt… torical
mal" … central
tries … hen M.
of GD… 009b).
wise; … ifficult
perce… f pub-
betwe… es, and
simila… arkets.
econo… untries
find … er, the
debt … vations
mies … , insti-
excep… range-
trast, …
debt l… xternal
Ou… uments
Public… arkets,
recen… more
the epicenter countries. This should not be sur- severe threshold for total gross external debt
prising, given the experience of earlier severe (public and private)—which is almost exclu-

Figure 4.3:

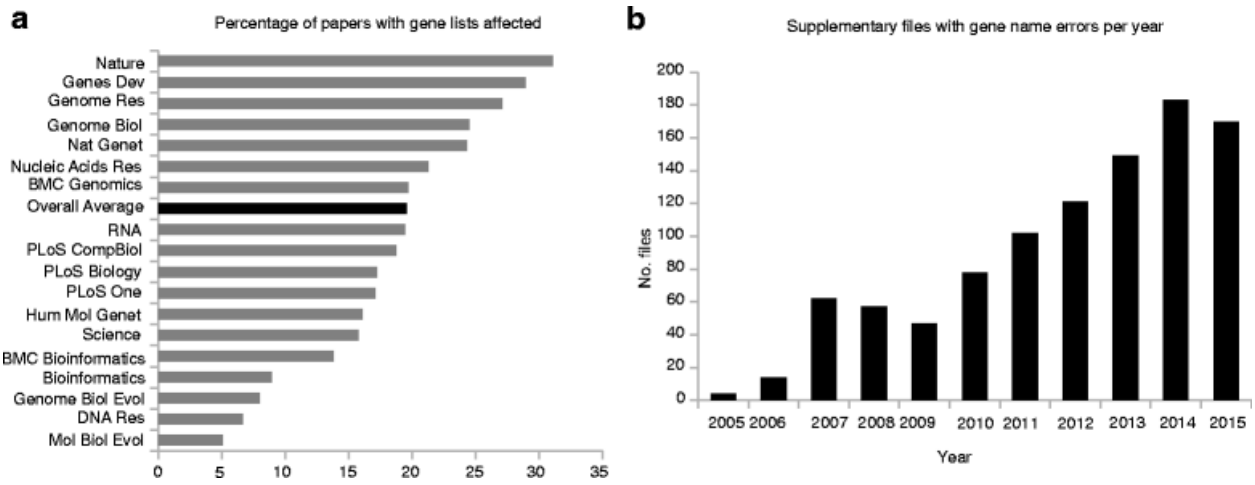


Figure 4.4:

external debt in excess of 90 percent" GDP growth was "roughly cut in half." This paper influenced policies such as *Paul Ryan's Budget* (also known as the GOP's *Path to Prosperity* budget), as well as the work of economic councils in the UK and the EU.

However – a group (HAP) led by a grad student at the University of Massachusetts at Amherst, got a copy of RR's data from the authors themselves (as they data was not published) and found that there were serious flaws in the Excel spreadsheet (coding errors, exclusion of available data from formulas, unconventional summary stats). After correcting for those errors in the spreadsheet, HAP found:

> When properly calculated, the average real GDP growth rate for countries carrying a public-debt-to-GDP ratio of over 90 percent is actually 2.2 percent, not -0.1 percent as published in Reinhart and Rogoff. [The] combination of the collapse of the empirical result that high public debt is inevitably associated with greatly reduced GDP growth and the weakness of the theoretical mechanism under current conditions, … render the Reinhart and Rogoff point close to irrelevant for current public policy debate.

Here's one more study, just to underscore how spreadsheets can deceive: Gene name errors are widespread in the scientific literature. In this paper, authors found that using Microsoft Excel for data input and analysis was extremely problematic for researchers in their field:

> The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

They also found in many cases, they couldn't reverse engineer or restore the actual gene names the authors intended to include. This effected over 1,000 datasets across over 900 papers. If we could restore the actual gene names, how many of the papers would draw the same conclusions?

So, what do we do to help limit/eliminate these problems??

DATA MANAGEMENT

Research data management is the process of managing the way data is collected, processed, analyzed, preserved, and published for greater reuse by *the community* and the *original researcher*. It's about making research materials findable, organized, documented, and safe, while also making the research process as efficient as possible.

Here's a high level look at some of the topics encompassed in research data management:

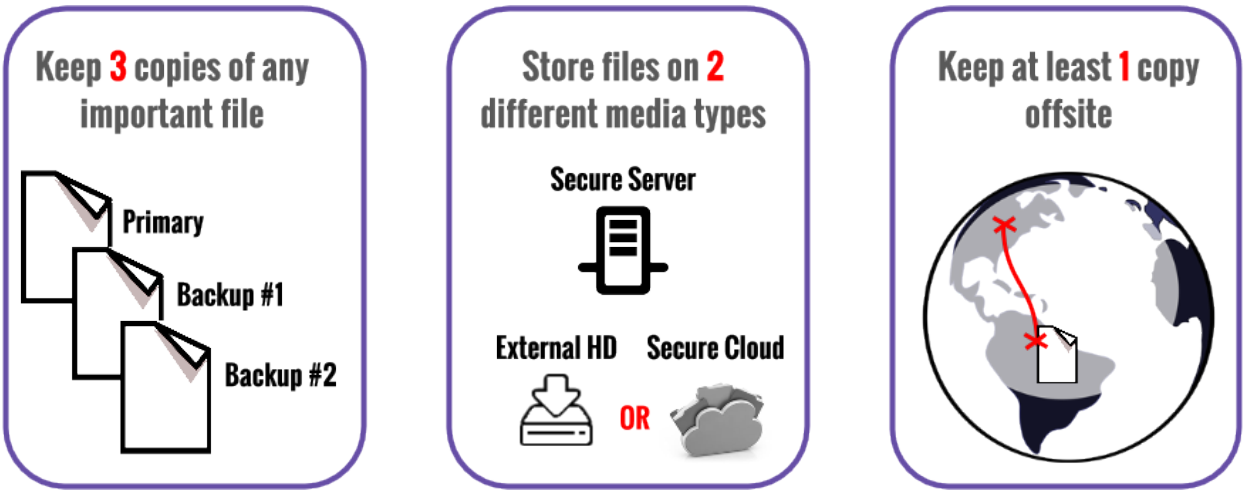


Figure 4.5: 3-2-1 Rule

Data Type	Group Roles	Data Storage	Data Archiving
format of data to be generated	who is primarily responsible for carrying out RDM? Set group norms	where will you store your data and how will you backup your data?	how will you preserve and make your data available to others?

Let's go through these and what they entail in more detail. The bottom of the RDM pyramid is **storage**. You could have the most efficient and sensical plan and execution for research data management, but if your materials only ever exist on a USB drive…it's not good.

To keep data safe, it is recommended that folks follow the 3-2-1 Rule which dictates you keep 3 copies of your data in various locations:

1. Original copy (on laptop, desktop, etc.)
2. External hard drive in a different physical location from the original (e.g., saved to external hard drive that is periodically updated)
    1. The lifetime of an external hard drive is 3-5 years!
3. Secure cloud service

Individual institutiona typically has some storage to offer folks, but a lot don't. It comes in many forms: Google Drive, Box, OneDrive, or something more centered to research such as access to cloud platforms (GCP, Amazon, Azure), mountable storage (we just put this into production at NYU – Research Workspace), or large storage infrastructure in High Performance Computing clusters.

You may have noticed that I use Nextcloud for you to submit your homework and when I need to distrubte files. This how for me this satisfies the 321 rule:

1. Access on all devices (computers, browser, phone, tablet)
2. Sync data between local copies (on all my computers) and on the server where I have it installed, which is on a different continent.
3. Run my external hard drive over the `Nextcloud` folder on my home laptop whenever there are changes (like when you upload your homework!s)

Google Drive, Box, and OneDrive all have desktop applications where folks can do the same thing. When downloaded, it appears just like a `My Documents` folder, only it's connected to your account on whatever service. Then it operates like a two-way door: if you change something on your laptop, it'll be changed on Google Drive in-browser (for instance). If you change something in the browser (or your collaborators do), it'll be changed on your laptop.

Once your files are stored correctly, the next low-hanging fruit is to make sure you **name your files** in a concise and efficient way. Actually, Kristin Briney (Data Services Librarian at University of Wisconsin-Milwaukee) just published a paper about this, narrowing in the problem of how people represent dates in file names and data broadly: The Problem with Dates: Applying ISO 8601 to Research Data Management.

File naming, when done in a well-organized fashion, can contribute to project documentation, workflow organization, and sharing. Moreover, certain choices in file naming are essential to accessing and sharing files across a computing systems. Some best practices include:

- Prefix your files with the date created – use ISO 8601 or Kristen Briney will haunt you (just kidding). This looks like YYYY-MM-DD
- Avoid special characters like `&`, `%`, `$`, `#`, `@`, and `*`. Just use letters and numbers.
- Do not make file identity dependent on capitalization unless implementing camel case (e.g. `fileName.xml`).
- Never use spaces in filenames – many systems and software will not recognize them or will give errors unless such filenames are treated specially. Use an underscore `_` or a dash `-` instead of a space.
- Use short file names. For your sake and the sake of systems that'll fail if you give it like a 50 character file name.

It's the difference between `VS_IMG%Archive2&3 Jan 2016.tiff` vs `2018-01-04_VS-Archive2-3.tiff`. One big note: you don't want *all* the medata about your files in the file name. Most metadata for researchers is unstructured (if there's any) – in the form of READMEs or field/lab notes!

There are a few solutions to help people bulk rename files to something consistent and in line with the best practices we discussed:

- Mac:
  - In the Finder, select and highlight the files you want to change using Shift or Command. Right click/Control + click on the selected files and choose "Rename X Items". Select one of the rename options: swap out text, add a set text, or apply a custom sequential format.
  - NameChanger
- Windows:
  - In the File Manager, select and highlight the files you want to change using Shift or Control. Right click on the selected files and choose "Rename". One of the file names will become active for editing. Enter a systematic file name and press enter. All files will be renamed using the chosen file name and numbered sequentially (1) (2), etc.
  - Rename It!

If folks don't try to put all the metadata in the their file name, they *definitely* try to put metadata in the file path.

Avoiding this tangled nest of files and folders is why file naming is so important, but ALSO why documenting file locations is important. You can do this easily by adding a READme file in your folders, which can tell you, link you to, and describe, all the files in a particular folder.

Projects often develop over the course of many years, and usually involve periodic work interrupted by spans of inactivity. To ensure that naming conventions are understood months or years after they are initially conceived, include a README file or some kind of file manifest (in a plain text or other sustainable format) in your directory that explains the contents of files and the naming system developed. It's really useful when, let's say, you are trying to go back to some data from 6 months ago and you forget what it is, or why you named a variable XYZ way, or you forget when it was collected. READMEs and documentation help you make your work more useful for you down the line.

These could look like a text file with the file name and a few sentences about the way the file was generated, the variables (if any), and some basic metadata (who made it, when, where, etc.). Others use an HTML file so they see it in-browser and can have links that go the files! Like this template:

```
<html>
<head>
```

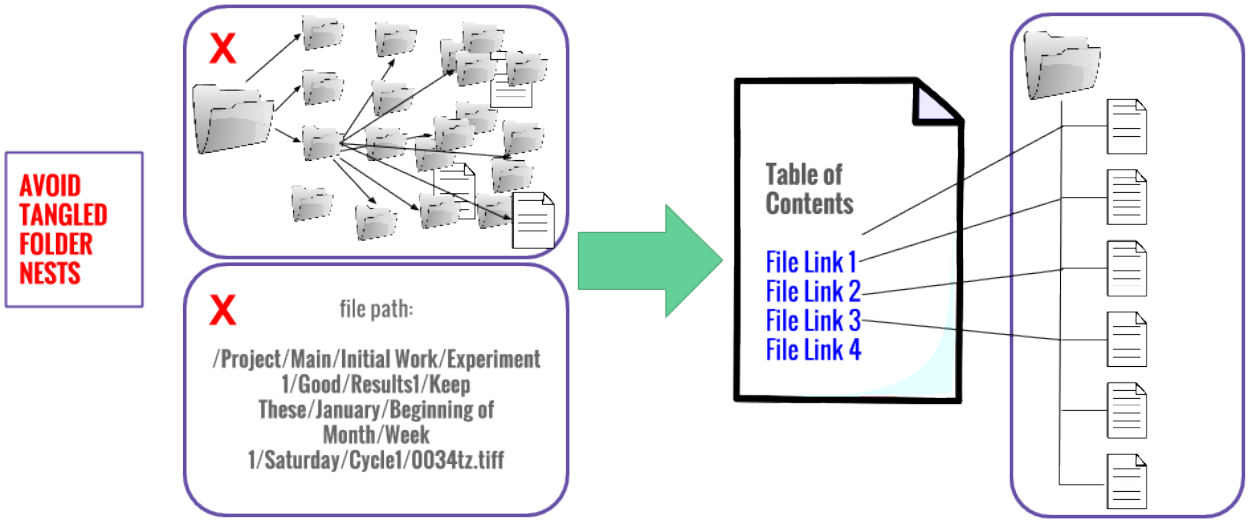


Figure 4.6:

```
<title>My Annotated File Directory</title>
</head>
<body>
<h1>My Annotated File Directory</h1>
<h2><a href="#">File Name</a></h2>
<p>This is an annotation</p>
</body>
</html>
```

Repositories also put out what they're expecting from researchers in terms of READMEs, such as Dryad: https://datadryad.org/pages/readme. Dryad asks that every data file has their own README, and then collections of data have a group README. Let's take a look!

There are lots of structured metadata that researchers generate, depending on the domain. Metadata supports:

- Discovery (i.e. the findability and citability) of data
- Re-use of other researchers of data in pursuit of knowledge
- Long-term preservation of resources (hopefully by professionals)

At the risk of generalizing, metadata in research data is typically the `HEAD` of a file coming off a machine (e.g. off a gene sequencer or telescope) or hand-made to be included in a gallery or portfolio (e.g. Omeka site for a humanities project). Let's look at some domain specific metadata: https://en.wikipedia.org/wiki/Metadata_standard#Available_metadata_standards

So, you've documented your well-named files. The next thing to look out for is the *type* of files you're generating. The gold standard for data and other research output is nearly the same for digital preservation: open, well-documented, and software agnostic.

Ideally, file types for a project should be standard, non-proprietary, and open source. If these features are not possible, at the very least file format selection should be made with the suggested preferences of a digital archive in mind, or with an eye to the format with the greatest stability, longest period of usage, most widespread community, and most organized governing body for standards.

Analysis software often relies on proprietary file formats that are subject to obsolescence as new versions are created or tools lose relevance. Where possible, export data files to stable formats for long-term preservation, or convert proprietary files into equivalent standardized files that will be able to represent that data for

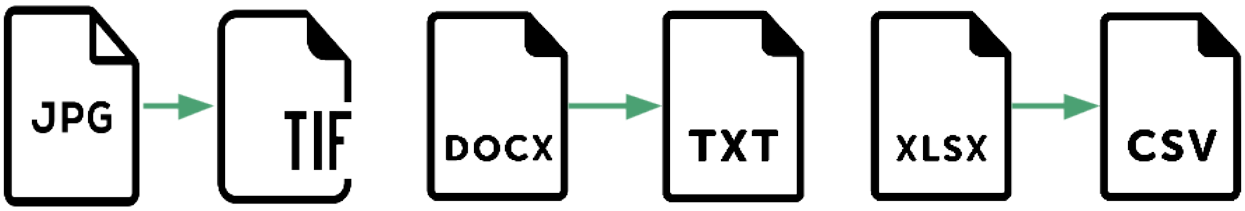


Figure 4.7: Some file format transformations

preservation purposes. In the case of spreadsheets or tabular data, this means that you should keep a master copy of your data as a `csv` file rather than an Excel file.

The LIBRARY Library of Congress (y'all saw the new logo right?) have a few good sites of interest for file format selection:

- Recommendations of file formats: https://www.loc.gov/preservation/resources/rfs/data.html
- Descriptions of file foramts (not only those they recommend): https://www.loc.gov/preservation/digital/formats/content/dataset.shtml

So we've discussed some strategies for your data, let's talk about that email failure from the *Current Biology* article. This is caused by changing insitutions, changing names, or other circumstances that lead to defunct addresses (anyone have their own email server here?). One solution that has arose around this problem is ORCID – the Open Researcher & Contributor ID. It's a free, persistent identifier for researchers (think DOI). It allows you to link all your publications to you uniquely – so no matter if your name changes, your institutional affiliation changes, or whatever, your ORCID stays with you. In fact, many journals are asking for an ORCID upon submission of materials (authority control!).

> ORCID is a non-profit organization supported by a global community of organizational members, including research organizations, publishers, funders, professional associations, and other stakeholders in the research ecosystem.

The other great thing is that ORCID data to do a lot of cool things! Like this: https://profiles.impactstory.org/u/0000-0003-4298-168X

I also wanted to point you all to one great project management tool that can help solve a lot of organizational problems, individually or when working in a group. It's called the Open Science Framework, OSF for short. The OSF is "a free and open source project management tool that connects researchers to the tools they are already using to make management easier through the research cycle." It's used across disciplines, so don't let the name get in the way of trying it out!The OSF is the work of a non-profit, the Center for Open Science, dedicated to supporting open scholarship using open-source tools.

I like the OSF particularly because it:

- Allows you to oreserve your current workflow, file types, and standard operating procedures
- Enhances and makes more efficient those workflows by better documentation, sharing, and discovery of research materials
- Gathers together group permissions, various storage options, bibliographic management, and publication in one place

Some key features include:

- Wiki: document your lab procedures, standards, etc.
- Collaborators: add collaborators of all levels, on different parts of your project
- Components: sub-projects to organize your research
- Version Control: upload files of the same name & OSF will track your versions!
- Add-Ons: use OSF to bring together tools you use
- Registrations: when you have an unchanging version of your project, register it & get a DOI!

Figure 4.8: https://twitter.com/PaulMinda1/status/1005132484093210624

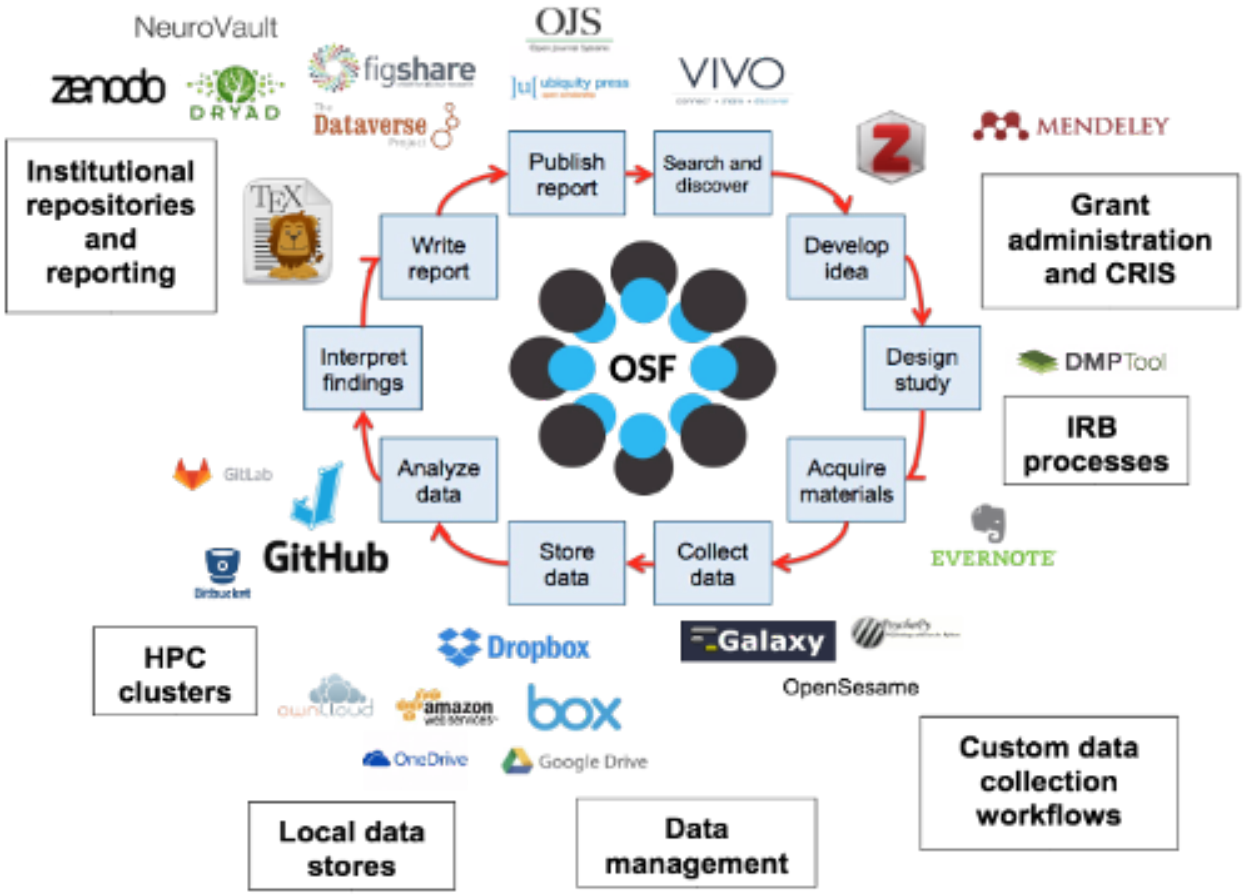


Figure 4.9: OSF Connections

- Connection to publication services: I can do all my work in the OSF and then in a few clicks, have my work published in a discipline-specific repository (like the LIS Scholarship Archive and others!)

Data Management Plans for Grants

A data management plan (DMP) is a document that describes how you will collect, organise, manage, store, secure, backup, preserve, and share your data. DMPs started out in the public funding agencies and have been adopted widely across private funders as well. A DMP is a critical part of grant applications – more and more, researchers' grants are being sent back because of a poor or sparse DMP.

The particular requirements of a DMP will vary among funding agencies, so it is best to always consult the agency's resources for their specific needs. However there are a few common attributes to all data management plans, including:

- An overview of the formats and types of data to be produced.
- Research methodology (data collection, processing, and analyzing).
- Roles & responsibilities in regards to data collection, description, processing, analyzing, and disseminating.
- Standards you will use to describe your data (metadata).
- Storage and backup procedures.
- Long-term archiving and preservation plan.
- Access policies and provisions for secondary uses.
- Security measures taken to protect data and/or participant confidentiality

For instance, NSF's data management plan guidelines include the following 5 criteria:

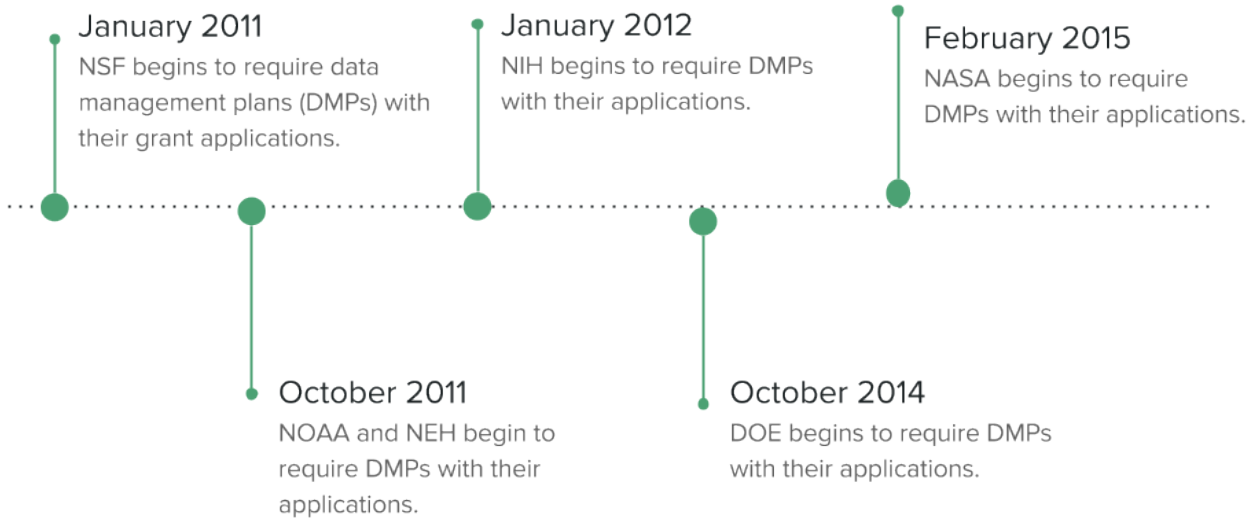


Figure 4.10:

- The *types of data*, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project
- The *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 re-use*, re-distribution, and the production of derivatives
- *Plans for archiving data*, samples, and other research products, and for preservation of access to them

There is one tool specific to the US that is helpful in getting started with writing DMPs. DMPTool Online is a web-based tool to help you build your data management plan. It provides step-by-step guidance and information specific to many US granting agencies and their directorates.

4.4 Lab/Homework

Please submit your homework here: https://cloud.vickysteeves.com/index.php/s/7JyjPJMndpdFW6S. The password will be given out in class. If you have more than one file to submit, please to it as a `.zip` file.

4.4.1 In-class

Let's do some demos of the OSF together!

Getting Started

To really use the OSF to the fullest potential, we have to first make sure that our relevant, useful accounts are linked in our profile (for easy linking in a project). So let's get started!!

Setting Up Profile

Please sign into the OSF if you haven't already: https://osf.io/signin. After you log in, you should see your name in the top right hand corner of the OSF landing page. If you don't have an account, take a few minutes to make one.

Click your name, then click 'Settings' in the drop down. This brings you first to your profile information page, where you can fill in any social links you want (like your ORCID, Twitter, ImpactStory, Academia.edu,

personal website, etc.), change your name to what you'd prefer on your citations, and any information about where you work/where you studied (if you want).

Next, let's go to the `Account Settings` tab. This is where you can set primary and secondary emails, export all your account data, and change your password. I would recommend setting one institutional address and one personal account. This way, folks reading your publications with your institutional address can find you easier, and also if you leave your institution you can quickly change the primary email on your OSF account to your personal one.

The next tab, `Configure Add-on Accounts` is where we can authorize the accounts we'd want to link into the Open Science Framework.

Setting Up a Project

Since we're already logged into the OSF, just click "My Projects" on the top bar. This will bring you to a list of all your OSF projects. Click the bright green "Create Project" and fill in the form.

If you click the "More" button, you'll see some extra options we can have at the start of our project. First is a brief description of the project – you can always fill that in later if you want to leave it blank for now. The other is a template – this is when you have an existing OSF project that you like the structure of (for instance: a lab template, another similar research project, etc.) and want to duplicate. It duplicates the *structure* of the project, not the files. You **can't** go back and add this in automatically later (you could do it manually, however).

When you click 'Create', you'll have the option of staying on the landing page or going to your newly created project. Let's go to the project!

When you get there, don't be afraid of how blank it looks. This is the most common thing I hear when first getting into the OSF – that they have a hard time knowing where to start bceause it's so blank. I am going to walk you through the best first steps in filling in your project.

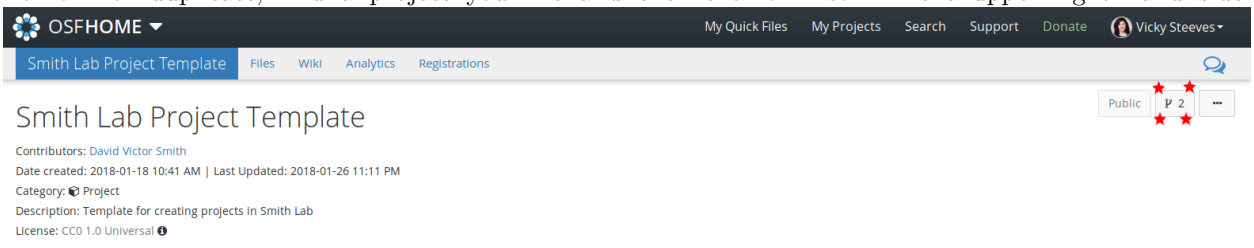
All OSF projects start private, and you choose to make them public when you want to. **A forewarning** that all the project's activities are logged in the activity log, and there's no way to delete it.

Forking & Duplicating

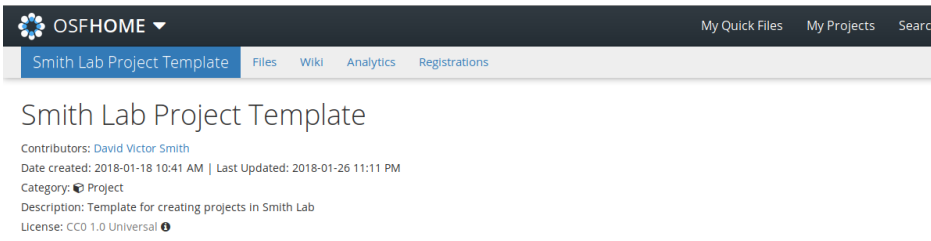
If you see a project you like and want to duplicate it's structure, you have two options:

- *Forking* a project creates a copy of an existing project and its components. The fork always points back to the original project. However, there are no merge requests in the OSF. Your fork is for you only.
- *Duplicating* does the same thing as a fork, but it doesn't have the link back to the original project.

To fork or duplicate, find a project you like and click the fork icon in the upper right hand side:



Select which type of copy you'd like to make:



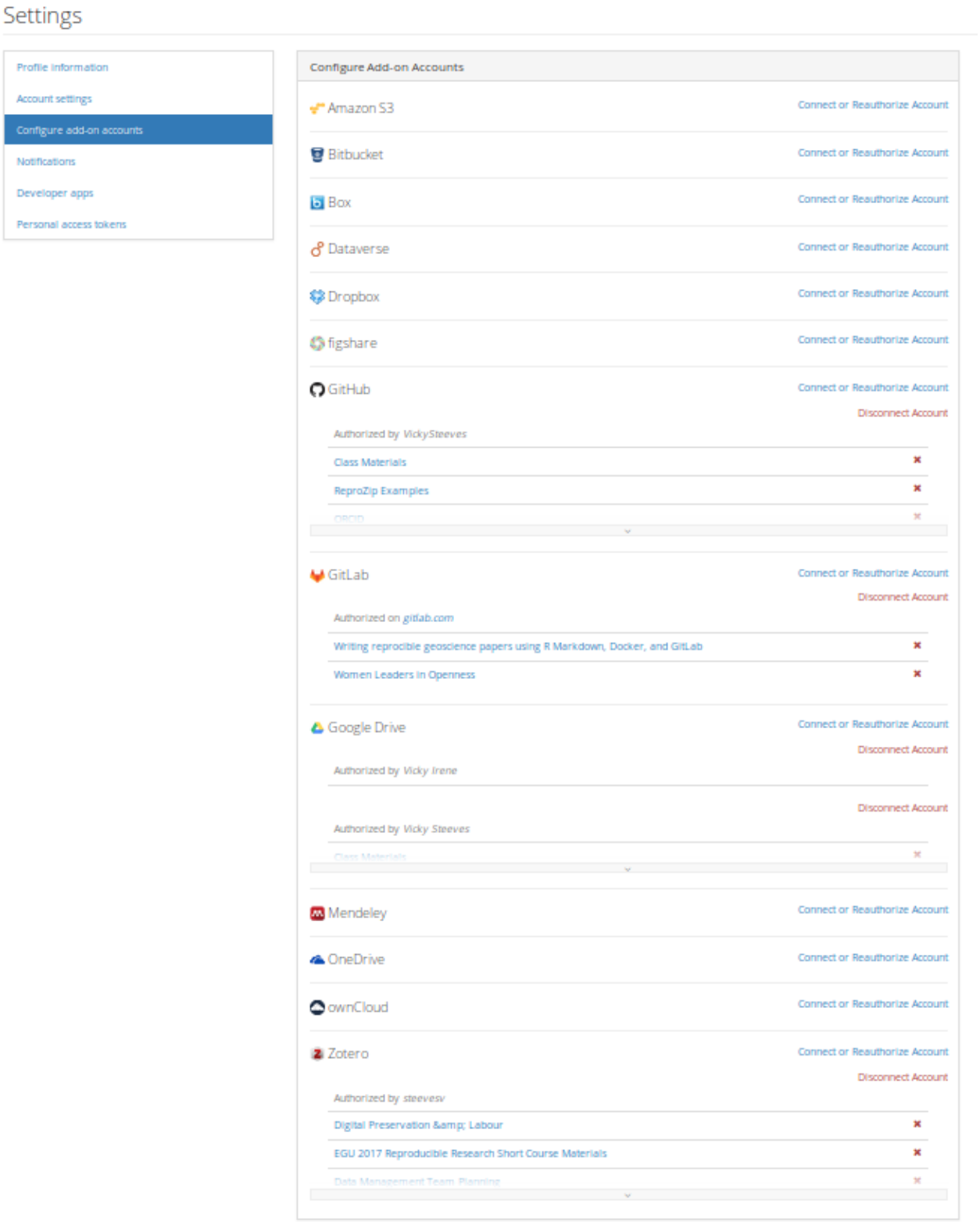


Figure 4.11: Viewing all available add-ons

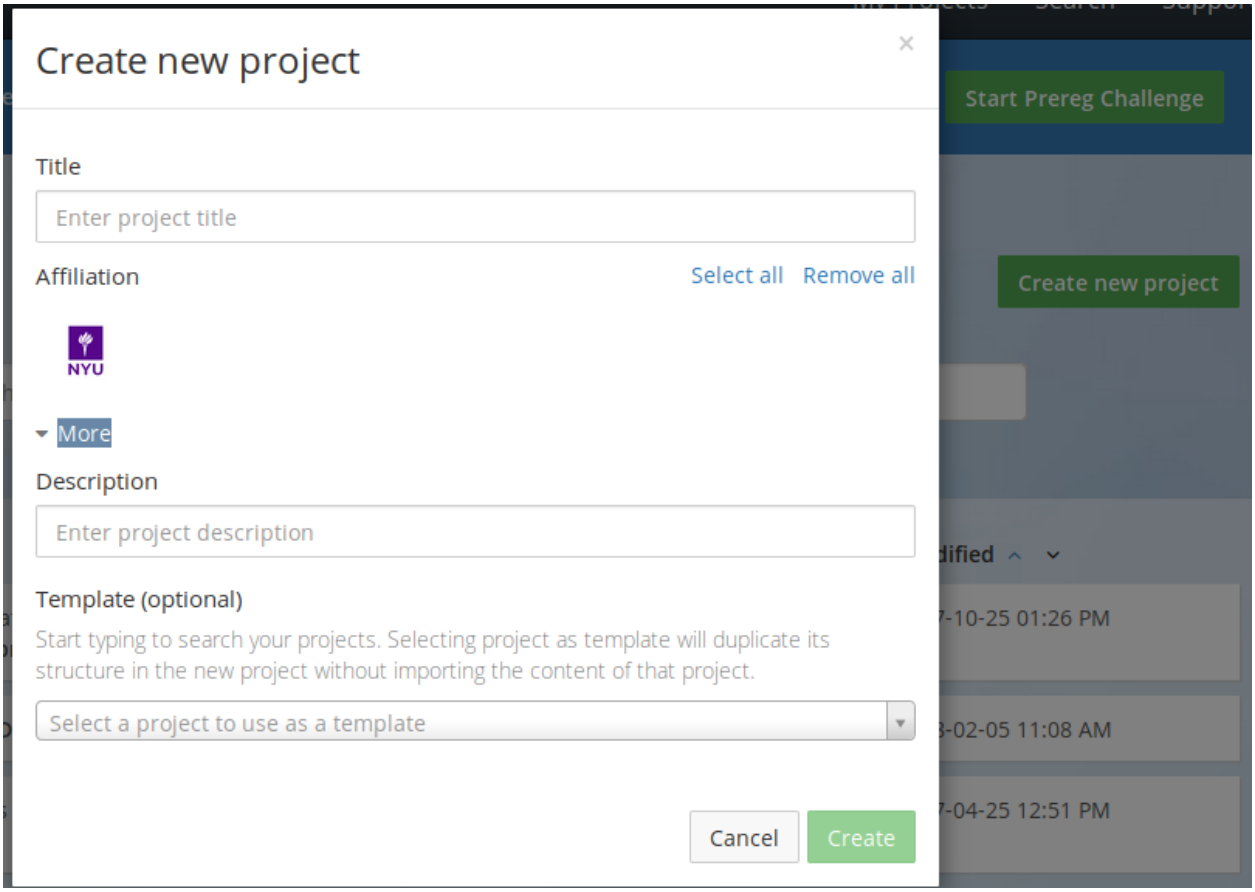


Figure 4.12: Creating a new OSF project

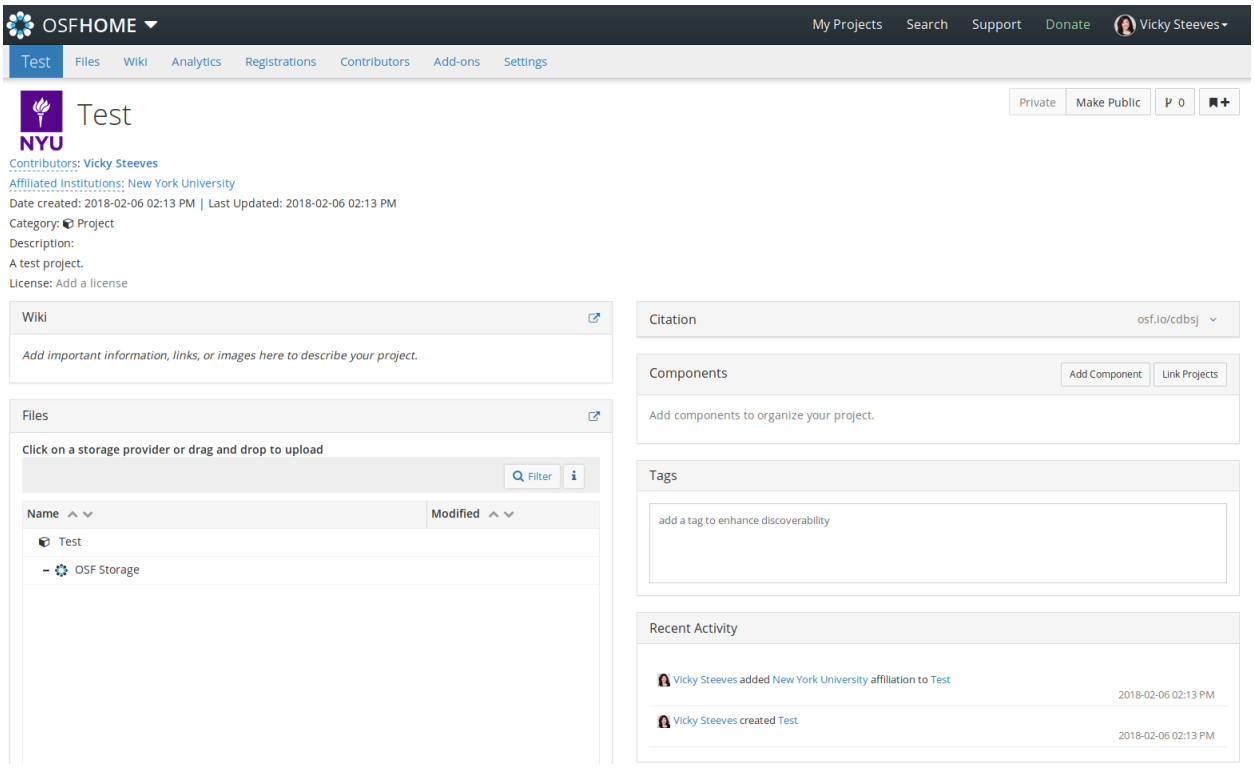


Figure 4.13:

When you try to make a fork or duplicate a project, a popup will appear asking you to confirm whether you want to create a new project using this project as a template. Any add-ons connected to the project will not be copied into the template.

You then have the project under your account, and you can do whatever you want with it! But no matter what, you'll still need to add collaborators, addons, write in the wiki, and the other skills we'll cover today.

Let's Add Collaborators

You can use the OSF by yourself, but it's really great when you can collaborate with others. So, the OSF has a few ways that you can let others see your workand collaborate with you.

The first is giving someone permissions on the project! To add a collaborator, click 'Collaborators' on the navigation bar of the *project* (not the OSF) and click the "add" button next to the title of the page.

You can then search for your partner, and add them to the project. If you have a collaborator who isn't on the OSF, you can still add them as a co-author or collaborator on the

You have three options for permissions: Read (view-only), Read + Write (view and add content, but can't delete anything), and Administrator (full permissions).

You can also share projects via a view-only link, including an option to anonymize contributors for anonymous peer review.

Adding Add-ons

As I mentioned at the start of the session, one of the best things about the OSF is the ability to link in files from lots of locations associated with one project. This saves folks having to dig through their `Shared with me` links on Google Drive, or searching through their repositories on GitHub. Everything associated with one project can be viewed and discovered in one place!

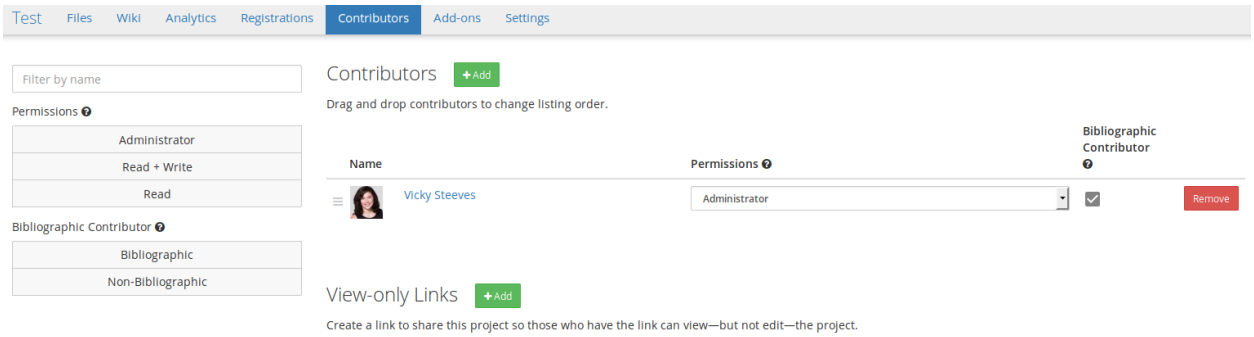


Figure 4.14: Contributors page on OSF project

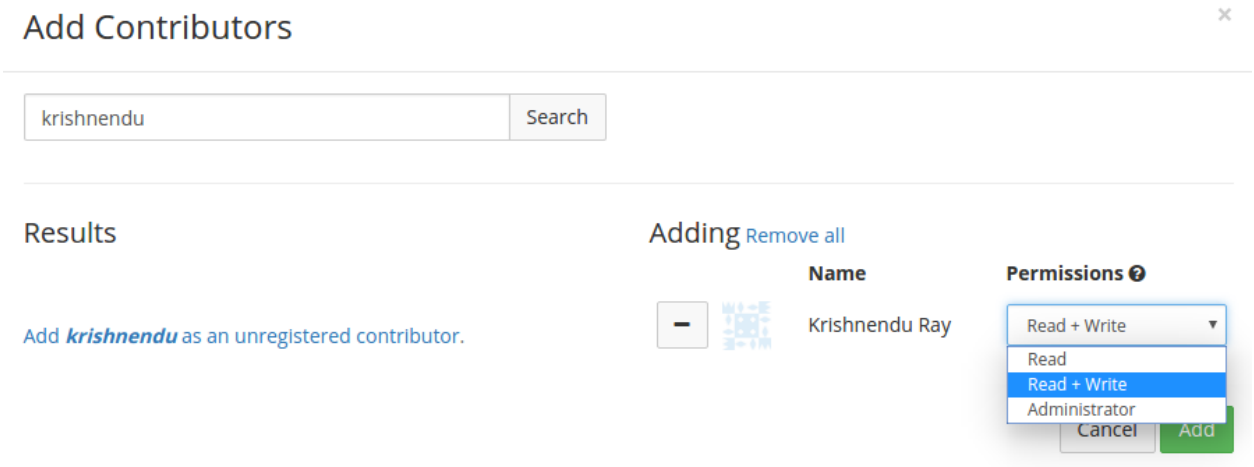


Figure 4.15: Finding our collaborators on the OSF

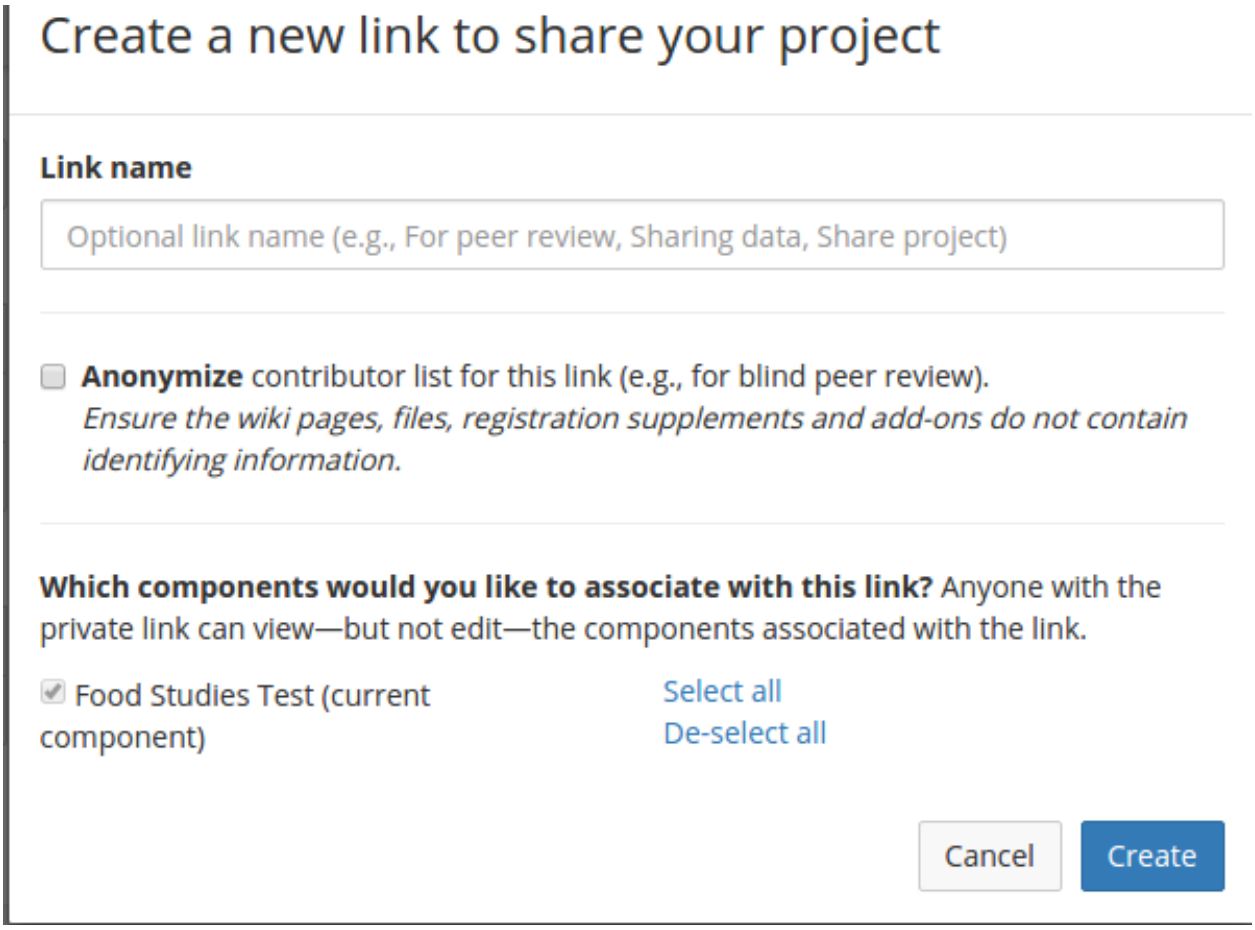

Create a new link to share your project
Link name
Optional link name (e.g., For peer review, Sharing data, Share project)
Anonymize contributor list for this link (e.g., for blind peer review).
Ensure the wiki pages, files, registration supplements and add-ons do not contain identifying information.
Which components would you like to associate with this link? Anyone with the private link can view—but not edit—the components associated with the link.
Food Studies Test (current component)
Select all
De-select all
Cancel
Create

Figure 4.16:

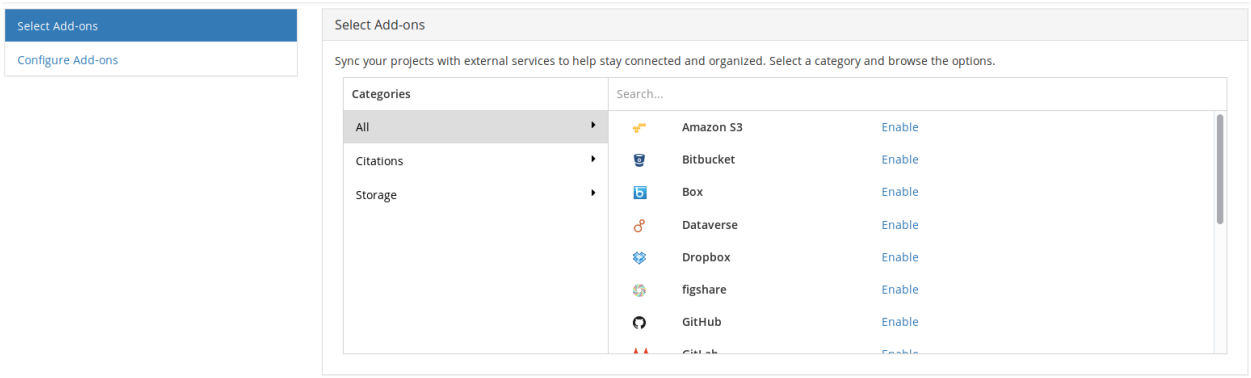


Figure 4.17:

Collaborators can independently add add-ons to projects as well. So, to get started, click `Add-on` in the project navigation. You'll see first a place where you can decide what add-ons you'd like to add to the project.

When you add an add-on to a project, it'll take you through what each action on the OSF does in relation to the add-on. For instance, for the Google Drive add-on:

Make sure you understand and are comfortable with these, and then click authorize! Then we can configure the accounts that we've added to the project. That's the next tab on this page, but you should see it right underneath the initial place where we selected what add-ons we'd like.

If it's the same add-on you authorized in your profile, you should see `Import account from profile`. You'll then be prompted to choose an account, though most times it's only one. For instance, I linked two Google Drive accounts in my OSF profile – my NYU account and my personal account (yes, Irene is my middle name…). I have to choose which one I'd like to associate with this particular project.

The page will reload, and in `Configure Add-ons` you should be able to see radio buttons which you'll use to select the folder, repository, or citation group you wish to import into your project space.

Add-ons work like a two-way door: if you change something in your add-on, those changes will be reflected on the OSF. If you change files on the OSF (only available for some add-ons), those changes will be reflected back in the add-on.

Another quick note about add-ons: files are versioned very well in OSF storage, but versioning on the add-ons is up to them. For instance, Dropbox only keeps versions for 30 days unless you have a paid account. Exposing those files in OSF storage doesn't change that.

Working with Your Files

You can have your files in add-ons, but you also have free and unlimited storage using the OSF Storage. The per-file limit is 5GB. All the files, both via add-ons and OSF storage, show up in the `Files` tab in the project navigation.

In the Files space, you can:

- See all files from OSF storage and any configured addons
- Drag and drop files into any kind of storage and also between storage types, OSF or otherwise
- Rename files
- Download all the files as a `zip` file
- Create folders in OSF storage for better organization

When you click on a file in OSF, it renders right in-browser (even 3d images!). On the bottom right is the "tag" field, where you can enter tags to allow others to find your files easier.

Google Drive Add-on Terms

Function	Status
Permissions	Making an OSF project public or private is independent of Google Drive privacy. The OSF does not alter the permissions of linked Google Drive folders.
View / download file versions	Google Drive files and their versions can be viewed/downloaded via OSF.
Add / update files	Adding/updating files in the project via OSF will be reflected in Google Drive.
Delete files	Files deleted via OSF will be deleted in Google Drive.
Logs	The OSF keeps track of changes you make to your Google Drive content through the OSF, but not for changes made using Google Drive directly.
Forking	Forking a project or component does not copy Google Drive authorization unless the user forking the project is the same user who authorized the Google Drive add-on in the source project being forked.
Registering	Google Drive content will be registered, but version history will not be copied to the registration.

Figure 4.18:

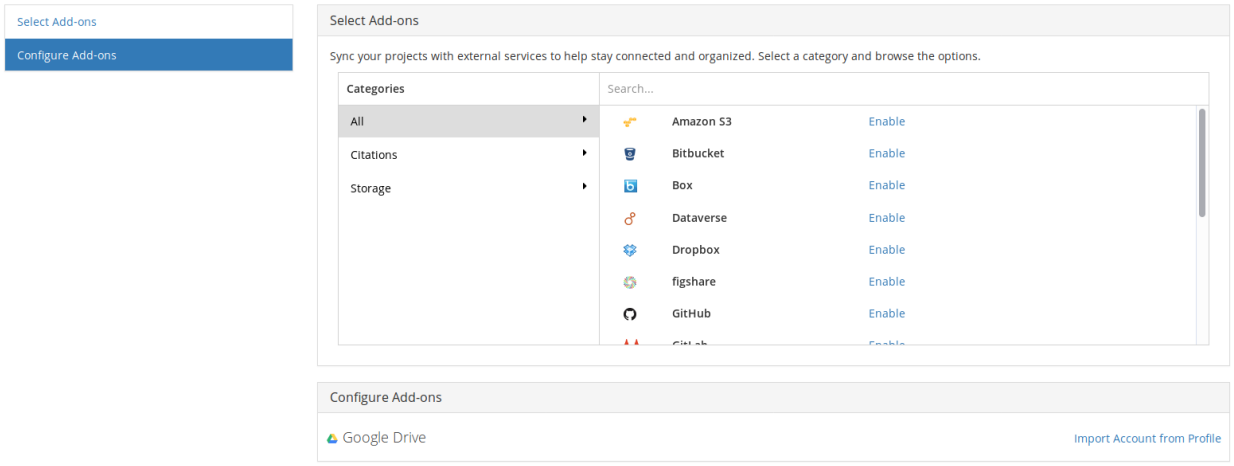


Figure 4.19:

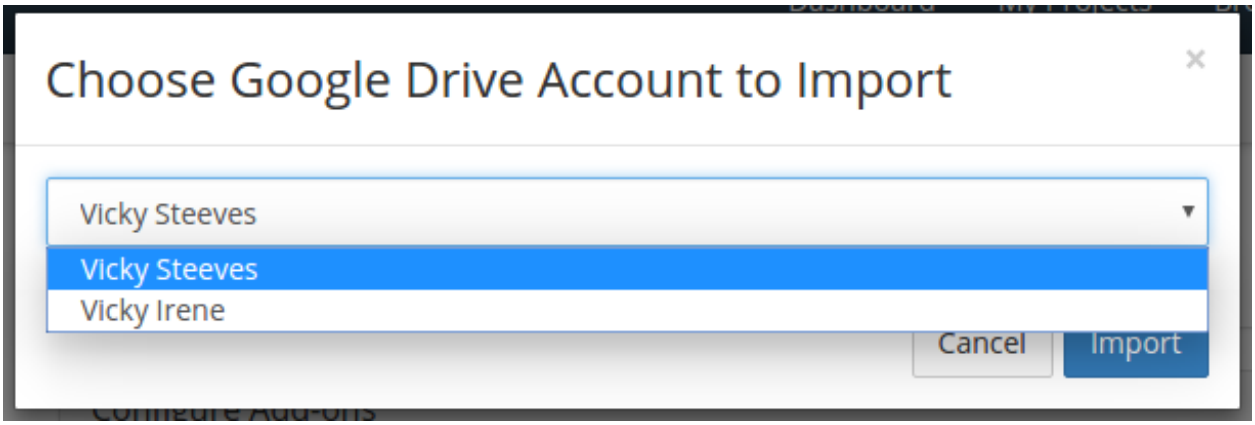


Figure 4.20:

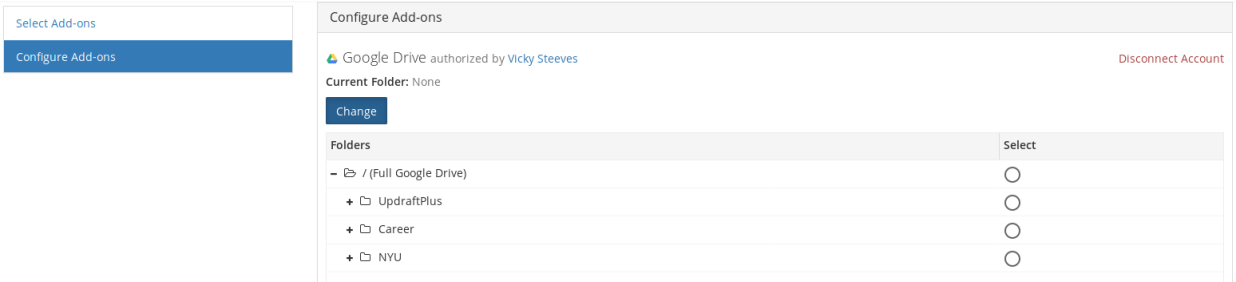


Figure 4.21:

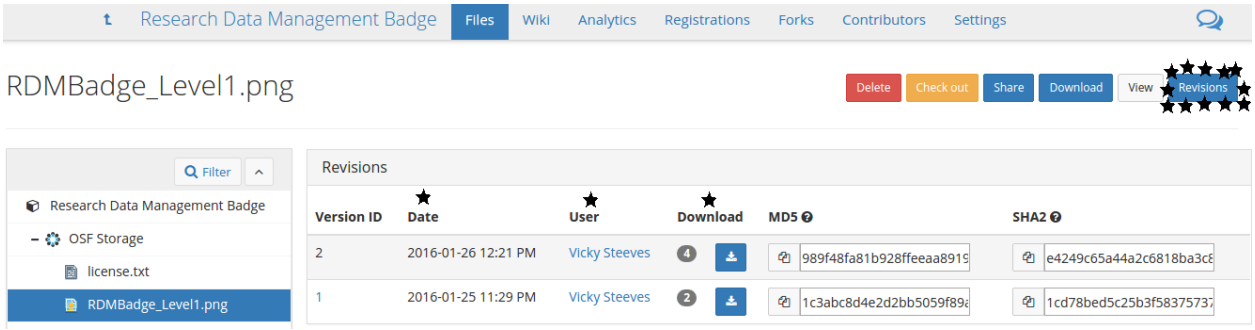


Figure 4.22:

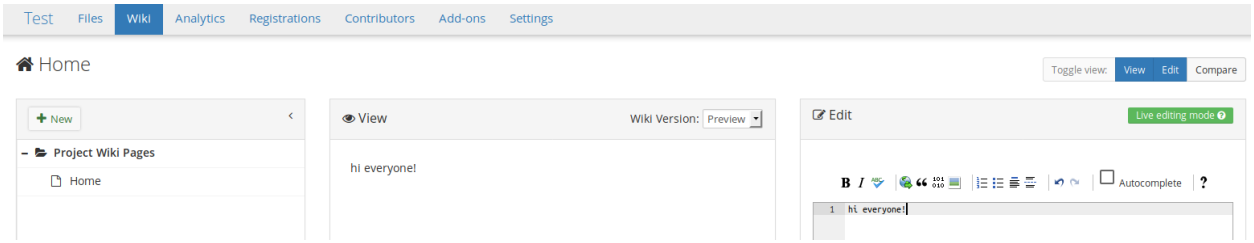


Figure 4.23:

When you click on a file in OSF storage, you can also see and download all the versions of that file that have been uploaded. The file has to be uploaded with **the same name** into OSF storage to be able to see this.

However! If your add-on has good versioning than you can view and download past versions right from the OSF as well, with that same workflow.

One other unique aspect of working with files in the OSF is that users with Read + Write or Administrator permissions can *check out* files. This means that other collaborators cannot edit the file while it is checked out. When you click on a file on the OSF (you should see it render in-browser), on the file toolbar – click the `Check out` file button from the toolbar.

Documenting with the Wiki

The OSF has a built-in wiki page that you can use to document your project. A best practice is to use the "Home" wiki page as a table of contents listing project goals, personnel, sub-components, and links to important files.

You can have multiple wiki pages to use for different purposes as well. I know some labs have a wiki page for their meeting notes, another for their standard operating procedure, another to document the variables from datasets hosted in the OSF, etc. You can use the wiki to do whatever documentation you need!

The wiki has a robust versioning – and you can compare versions side-by-side to see what's changed!

Components

Components are essentially "sub-projects" that can have their own set of collaborators, add-ons, and access controls. They can help you organize different parts of your research.A component's privacy settings, contributors, tags, wikis, add-ons, and files are separate from the parent project. If you choose, components can inherit the contributors and tags of a parent project.

In the `Components` part of your OSF project, click the `Add Component` button. You'll see some options in the screen, including:

- Title
- Add collaborators from the parent project (checkbox)

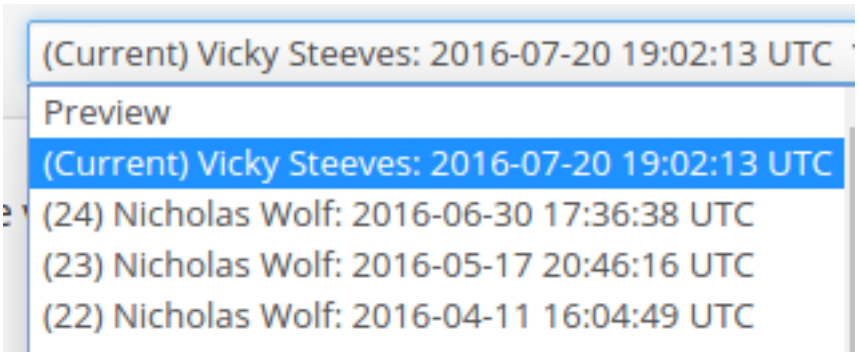


Figure 4.24:

- Add tags from the parent project (checkbox)
- Description – short blurb about the add-on
- Category – these help discovery of your work as well as helping *you* organize the different aspects of your research. You can pick one per component:
  - Analysis
  - Communication
  - Data
  - Hypothesis
  - Methods & Measures
  - Procedure
  - Project
  - Software
  - Other
  - Uncategorized

You can reorder components within a project by dragging and dropping them into the correct order from the parent project's "Overview" page. Too.

Linking projects

Linked projects are found in the "Components" section of a project "Overview" page. Links have a link symbol in front of them to distinguish them from components.

To add a link to another project from within your project, click the `Link Projects` button in the `Components` section.The `Link other OSF projects` popup will appear, and you'll see all your projcets listed first by default.

To find the project you want to link to, you can either enter a project title or project GUID into the search box. The GUID is the last part of a link in the OSF, those last 5 characters. You can either the `Search all projects` if the project you want to link to is not yours, or the `Search my projects button` if you want to link to one of your own projects. `Search all projects` takes longer to load since there are *a lot* of public projects on the OSF.

After you get to the project you want to link to, click the green `+` icon next to the project to which you want to link. Then, click `Done`. The linked project should show up in the `Components` tab of your project now! If you want to ever remove that link, click the `X` icon next to the linked project's title.

Sharing Projects

Everything (files, subcomponents, wiki docs) gets a short permalink in OSF. That makes it easy to share via e-mail, Twitter, pastebins, etc. You can get a DOI for a project or a component, and include it in a "Supplementary Materials Section" of a journal article!

To complement the easy sharing, public OSF projects also have access to some analytics about when peo-

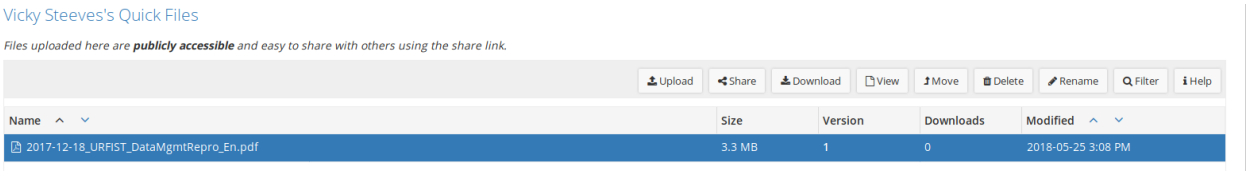
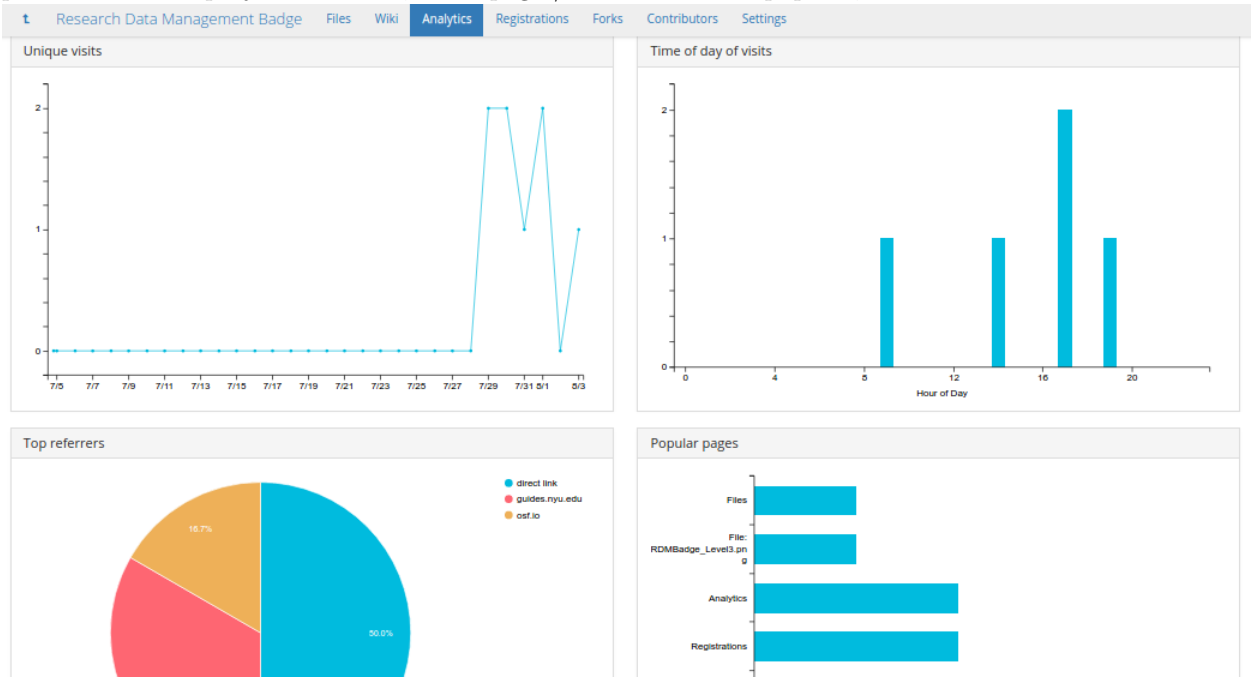


Figure 4.25:

ple accessed their projects the most, which pages/files are the most popular, and where folks are referred from:



You can also share a snapshot of your OSF project, frozen in time, via a *registration*. You can have multiple registrations of a project, all getting unique DOIs, which is useful if your project evolves over time. If you publish multiple papers on the same project, but the data/methods evolve, you can register it at each point of publication so that the paper always links back to the relevant files, not just the most recent.

When you register a project, it essentially means all the files, wiki, etc. are frozen into a separate space. The registration is a read-only copy of you project at a given point in time. It always links back to the editable, updateable project.

Quick files

If you have a file you want to share and you don't need all the functionality of the OSF (if it's too bloated to just share one file, for instance!), you can juse the Quickfiles. In the navbar of the OSF, click `My Quickfiles`. There you can add a file or download all the quickfiles as a `zip` file. You can version quickfiles the same way as other files in OSF storage – just upload a file with the same name and the versions will be tracked in the OSF.

OSF for Meetings

OSF Meetings is a way to provide conference organizers with a branded landing page, simple submission process, and simple search and discovery for their conference. Researchers who share materials can add supplementary data, code, preprints, or other material alongside their poster or slides.

- Easily collect presentation materials in advance of your meeting
- Allow attendees to discover interesting content prior to arriving

- Presenters can make revisions to their materials and update them without the need to track down the latest powerpoint file or pdf. Presenters can even present directly from their OSF Meeting project
- All meeting materials are archived and available for future reference.
- It's FREE

Read & learn more: https://cos.io/our-products/osf-meetings/

4.4.2 Outside class

Hands-on option

A researcher has come to you with the following 3 problems:

1. They have 1,000 images from an archive, all named `IMG_001.png`, `IMG_002.png`, etc. What do you advise they do, and what tool would be most helpful in solving their problem? Create a maximum 30 second screencast showing them your solution using this [extremely cute] test dataset: https://cloud.vickysteeves.com/index.php/s/KyjXEzorrEMmHg4.
2. They have their data (not confidential) strewn throughout Dropbox, Google Drive, and Amazon S3. They also have some code in GitHub, and their collaborator uses GitLab. They need some help thinking through how to stay organized when the materials they're working with are so distributed. What do you advise? Either create the basic template with a tool you advise, or write out/illustrate an organization plan that they can draw from moving forward.
3. They have no documentation about their files – only about their methods, meetings, administrative work, etc. What do you advise to help them get started with basic file (data) documentation, and what tool (if any) would be most helpful? Create a README file for the test dataset from question 1 to show them what it might look like. Feel free to use from my README maker scripts and edit the resulting file from there: Mac, Windows.

BONUS: Upload all your materials to the OSF and add me as a collaborator to the project. Add your written answers to the wiki of the OSF project too. (In addition to submitting your files via the upload link)

Reading/writing option

1. Read through these public data management plans:
    1. IMLS: https://dmptool.org/plans/22755/export.pdf
    2. DOJ: https://dmptool.org/plans/31122/export.pdf
    3. NIH: https://dmptool.org/plans/17259/export.pdf
    4. NSF: https://dmptool.org/plans/20638/export.pdf
2. Answer the following question about each DMP:
    1. Give your overall impressions of the DMP, in relation to the best practices we discussed in class.
    2. Do the researchers give sufficient detail so you have some idea of how they'll manage their data? What's missing?
    3. Are the researchers using open file formats? Is there enough information for you to find out?
    4. Do you believe the data resulting from this study will be readily discoverable to other researchers? Why/why not?
    5. Do the researchers discuss metadata? If not, what documentation are they planning on writing?

BONUS: Submit two DMPs marked up with comments and highlights showing the advice you'd give to each researcher/research team.

Chapter 5

Data manipulation & analysis

Agenda for today's class:

- 6:30 - 7:30: discussion
- 7:30 - 7:45: break
- 8:00 - 8:45: lecture part 1 & first lab
- 8:45 - 9:00: break
- 9:00 - 9:20: lecture part 2 & second lab

5.1 Discussion questions

Amber, our facilitator this week, prepared the following questions for in-class discussion:

1. Maceli states: "Text mining can be a highly useful tool in the beginnings of research exploration, allowing the textual data to suggest themes and concepts to the researcher during analysis." Do you believe that a technique like this could allow for a forced or more bias correlation between datasets, particularly as Marceli finishes this thought by suggesting that text mining helps to frame questions and analysis approaches?
2. We have seen in class before an example of misguided correlations, such as the NYT article. Can you think of any other recent example where inference of a specific population has become skewed in the media?
3. Do you think there are major differences in the benefits or challenges between word clouds and cluster dendrograms?
4. When choosing statistical software do you think that (other than the reasons already listed such as cost, hardware, ease of use, and graphics capability) academic siloed disciplines change this decision?

- if so, why?
- what are some of the benefits and challenges this presents?

5.2 Lecture: working with data

Most of the time on any given research project is going to be cleaning and preparing data for analysis and visualization. Some have hypothesized data cleaning and tidying takes 80% of the time on a project. If only this was just once, too – the process of data preparation gets repeated as new data are collected, analyses refined, and new visualizations desired.

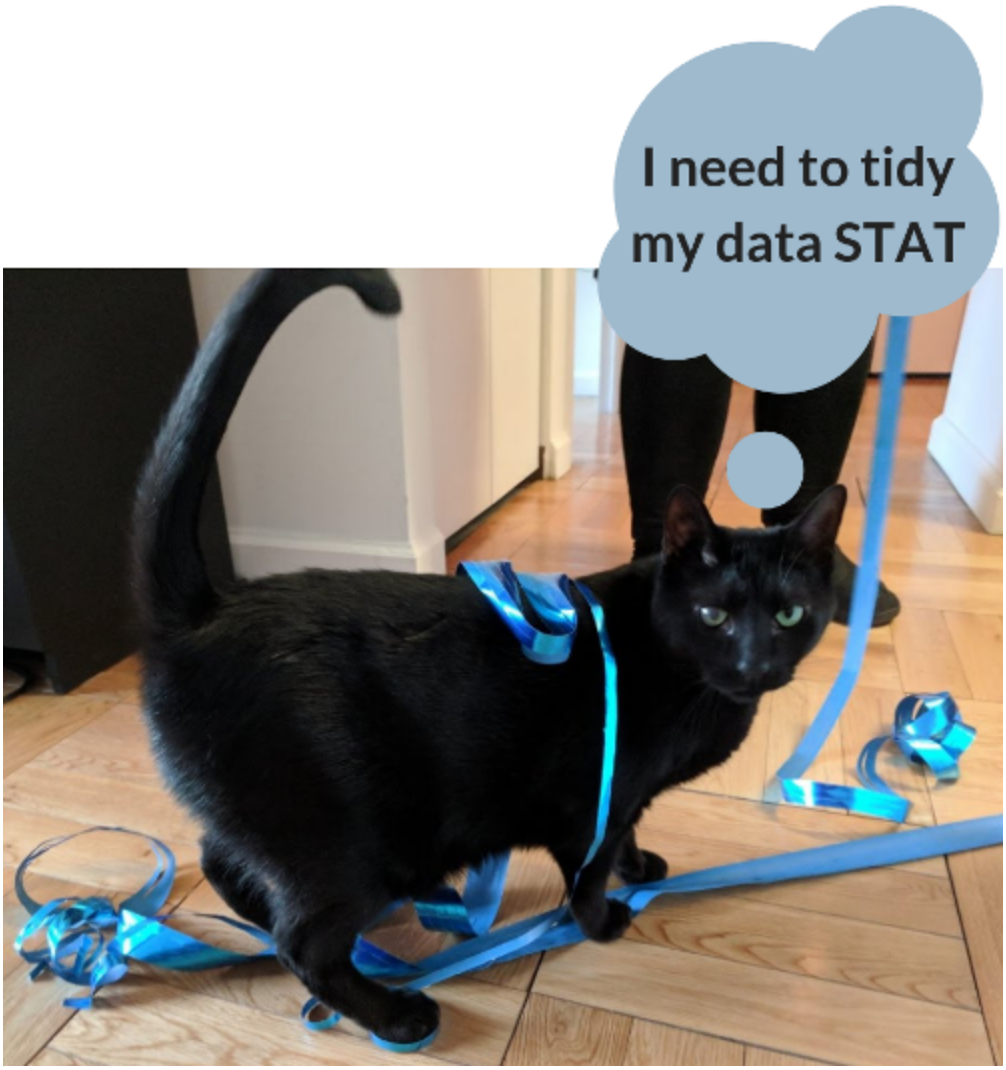


Figure 5.1:

Table 5.1: "Using OpenRefine by Ruben Verborgh and Max De Wilde, September 2013"

Dates	Price	State
2015-10-14	$1000	ID
10/14/2015	1000	I.D.
10/14/15	1000	US-ID
Oct 14, 2015	1000 dollars	idaho
Wed, Oct 14th	US$1000	idaho,
42291	$1K	Idhaho

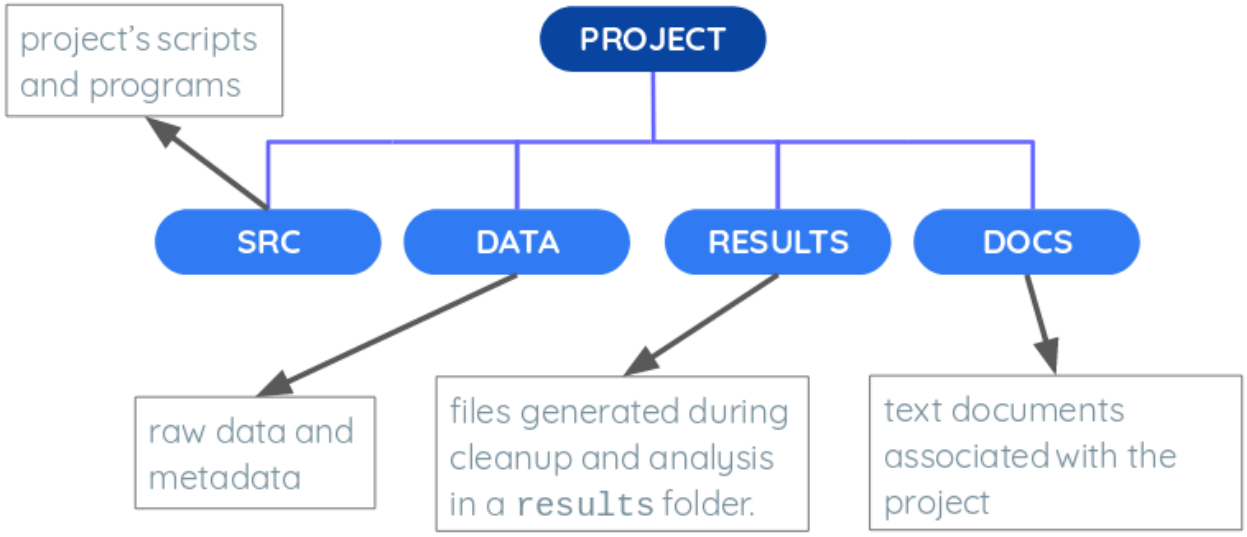


Figure 5.2: A good general outline for project structure.

For qualitative data, this could be digitizing hand-written notes, OCR-ing images, or transcribing audio recordings. For quantitative data, it means ensuring each column has a consistent type. For GIS data, it's making sure layers are computed or generated with accuracy or that your geocodes are represented in a `csv` file correctly.

So let's get into how we can make our data ready for analysis!

5.2.1 Cleaning data

"Messy data" is basically data that's full of weird inconsistencies, either because of human error or poorly designed systems (like Excel! Remember our two real-world examples from last week…). These inconsistencies make it neigh on impossible to do any meaningful searching, statistics, or other analysis!

Sometimes we can't help the way that we receive data, either from a machine, from a public repository, or carefully hand-crafted surveys. So today we'll learn some best practices for prepping data for analysis!

The first thing: **raw data should be read-only**. This means that you should *never* change the original data. If you make changes, this should be become a new file. This can be helped by the way you structure your project folder! Data management!

This is a generally good outline:

Having this structure to begin with allows you to visually and computationally structure your project for maximum efficiency:

- Put each project in its own directory, which is named after the project.
- Put text documents associated with the project in the `doc` folder.
- Put raw data and metadata in the `data` folder, and files generated during cleanup and analysis in a `results` folder.
- Put source for the project's scripts and programs in the `src` folder.
- Name all files to reflect their content or function, with NO special characters (!@#$%^*) or spaces! Use underscores or dashes, A-Z, and numbers!

Data cleaning involves making sure the records in a dataset are complete – for instance, if you want to survey people between 20 – 60 years old, part of cleaning is making sure there are no rows with ages under 20 or over 60. Or, if you ask people for their height and you notice half the respondents give their height in meters and the others in feet/inches, this is something you'll have to *clean* in your data.

When data is clean, it should be:

- *Consistent*: you don't want data that contradicts each other – such as two tables with the same survey respondent in each, but with different mailing addresses. You want to ideally have the right **one** mailing address in both tables.
- *Uniform*: age is expressed as whole numbers (15, 22, 47), addresses are expressed as lat and long, weight is in KG and not pounds
- *Valid*: if you survey people between the ages of 20 - 60, you shouldn't see anyone with their age as 18 or 61 in your dataset
- (Possibly) *Unique*: make sure that if you can't have a repeating value in one column (like SSN or Pratt ID number), then it doesn't actually repeat (you'd be surprised). You'll probably be deduping a lot.
- *Complete*: you have recorded all possible knowns and made provisions for the unknowns (e.g. designating a null value as NaN or 999).

Some common ways to clean data include:

- *Parsing* data to look for errors like "HoWARD STREET", "Howard Street", "Howard St", and making them consistent.
- *Transforming* data – such as the values of weight that are expressed in pounds into kg.
- *Deduping* data – finding duplicate values or rows that should be merged, and merging them (the most common thing in libraries I'd wager)
- *Statistical methods* like we saw in the Leek reading – this is actually using some data analysis methods to find errors in the data to clean!

Some of this just isn't possible, but we try to get as close as we can!

> There aren't these hard and fast rules for qualitative data preparation. It's very much field and method/framework dependent. Some qualitative researchers leave out "um", "ah", "well……" when they transcribe interviews. Others leave them in as a valuable part of the interview – looking at the sentiment and nature of the participant. In the case of OCR-ing images or PDFs into machine-readable text, it's about checking the accuracy of the scan – are the long 'F' looking characters that used to be 's' transformed into the modern characater? Do you want them to be? These questions around data tranformations are embedded in the framework under which qualitative researchers operate, whereas for quantitative data, the rules apply across domains.

You might also hear the phrase 'tidy' data. This is a facet of cleaning that is slightly different, and relates closely to tabular data. Hadley Wickham, who coined the phrase, defines the five most common problems that necessitate tidying in this paper called *Tidy Data*:

1) *Column headers are values, not variable names*; a table with colums 'blue', 'brown', or 'green' meant to represent eye color, and the rows contain either 0 (no) or 1 (yes). This should really be one column, "eye color", with the color as the value.

2) *Multiple variables are stored in one column*; a table with age and sex combined in one column (e.g. m014, m1524). This should be two columns: sex and age.

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, `$75-100k`, `$100-150k` and `>150k`, have been omitted

Figure 5.3:

3) *Variables are stored in both rows and columns*; a table with the column "measurement" that contains data like `n_dogs`, `n_cats` with a column 'value' next to it depicting the number of cats and dogs per person. This should really be at least two columns: n_dogs, n_cats, with numeric values showing the number of dogs and cats per person.

4) *Multiple types of observational units are stored in the same table*; a table that contains someone's age and height alongside their pet's name, type of pet, and pet's age. While it's cute, these should be two tables: one for people and one for pets.

5) *A single observational unit stored in multiple tables*; a new table of an individual's medical history for each year of their life. This should really be one table with each row representing a year of someone's life.

Here's an example of a table that represents some messy data (from Dr. Wickham's paper):

And here's the tidy version (also from Dr. Wickham's paper):

At this point, let's go to the first in-class lab!

5.2.2 Data analysis

Now that we've done some hands-on preparation of data, let's talk about analysis! Data analysis can include:

Data processing: may include selecting a subset of clean data for analysis or manipulating data for usability (or plot-ability) *Visualization*: To give you an accurate picture of your data quickly, the best way to do that is to visualize it! This can also makes it easier to see patterns and help spot errors you may have missed in your earlier cleaning

There are *many*, *many*, ways to analyze data. In keeping with the way we've structured previous lessons, I'll talk broadly in terms of qualitative, quantitative, and GIS.

We'll discuss the type of qualitative data analysis that involves coding research materials to find themes. There are other methods of qualitative analysis, where researchers do *no* coding for themes, but likely you'll find folks using thematic analysis – I would wager it will be the most common type of qualitative analysis you'll come across. This is essentially the process of coding, when the qualitative researcher reads, re-reads, codes, and re-codes until the grouping becomes higher and higher level, less redundant, and clear. When

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 6: The first ten rows of the tidied Pew survey dataset on income and religion. The `column` has been renamed to `income`, and `value` to `freq`.

Figure 5.4:

coding is complete, the researcher may prepare reports via a mix of: summarizing the prevalence of codes, discussing similarities and differences in related codes across distinct original sources/contexts, or comparing the relationship between one or more codes.

The 'code' is the word or short phrase suggesting how the associated data segments inform the research objectives (e.g. time management, health issues, etc.). Creating qualitative 'codes' can be be done in a variety of ways:

- *Preset or «a priori» categories*: When researchers identify categories in advance then search data for these topics (may come from theory, lit, or document review).
- *Emergent or grounded categories*: When researchers work with the data and inductively create categories. This is an iterative process. One builds categories until no new themes emerge. As you do more interviews/code more information, you may group/split categories until you reach your final categories. Begins by exploring, then confirming findings, guided by analytical principles rather than rules.
- *Combined approach*

This type of coding is done by going through all of the text and labeling words, phrases, and sections of text (either using words or symbols) that relate to your research questions of interest. After the data is coded you can sort and examine the data by code to look for patterns.

Researchers can code by hand, by printing out their materials and highlighting the paper, or by using qualitative data analysis software. Qualitative data analysis software are not the most accessbible to the everyday researcher. The market is dominated by a few really powerful but super expensive and inaccessible softwares.

Software	OS	Cost	Analyses	Tools
NVIVO	Windows, Mac, Browser	$1,399	Text, video, audio, pictures, survey, webpages, social networks	Coding
ATLAS.ti	Windows, Mac	$1,290	Text, video, audio, pictures, survey, webpages, social networks	Coding
MAXQDA	Windows, Mac	$635	Text, pictures, audio, video, webpages, social networks	Coding
Dedoose	Browser-based	$15/mo	Text, audio, video, survey	Coding
Taguette	All OS + Browser	FREE	Text, webpages, ebooks	Coding

Quantitative data analysis is a way to understand behavior by using mathematical and statistical modeling. Quantitative research projects usually fall into one of these categories:

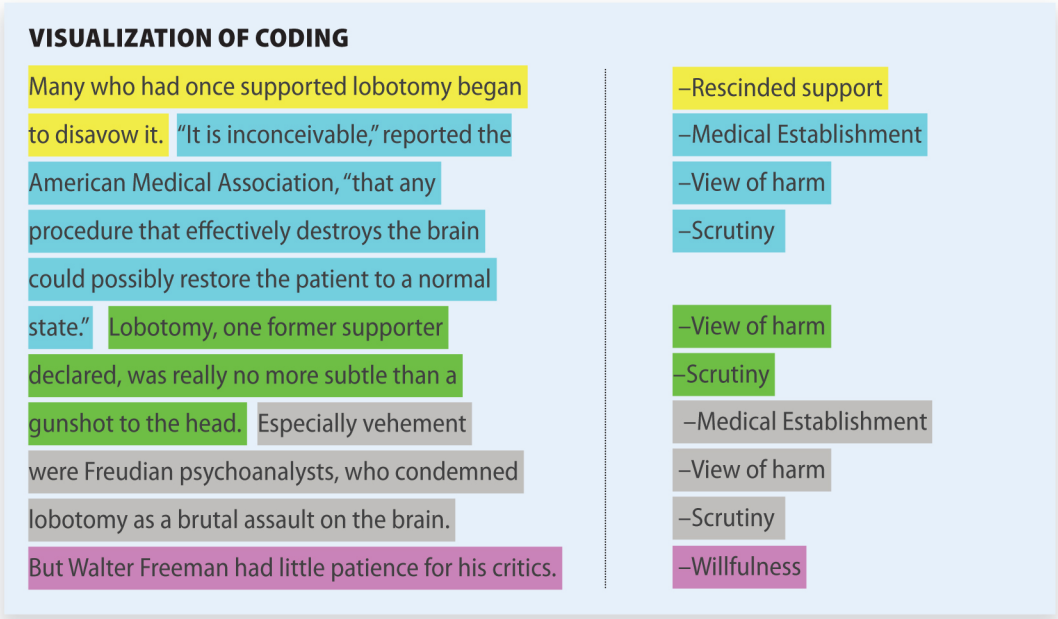


Figure 5.5: Image from: https://www.meetingsnet.com/cme-design/7-faqs-about-qualitative-research-and-cmes

1. Descriptive: subjects are measured once; establishes associations between variables
2. Experimental: subjects are measured before and after a treatment; establishes causality

The data gathered from quatitative analysis is usually expressed in tables – what we call `tabular` data. Researchers with quantitative data typically use statistics to try to measure causality or associations:

Inferential statistics are used to analyze experimental data, to try to deduce probability distribution (the probabilities of occurrence of different possible outcomes in an experiment). I won't focus on this, as more than likely you all will be working with materials more ready for descriptive stats!

Descriptive statistics are used to analyze observational or descriptive data. You might often see various summary statistics, which is used to communicate:

- a measure of location, or central tendency, such as the arithmetic mean
- the dispersion of a variable, such as the range (min, max, quartiles)
- a measure of statistical dispersion like the standard deviation
- a measure of the shape of the distribution like skewness
- if more than one variable is measured, a measure of statistical dependence such as a correlation coefficient

Descriptive statistics aims to summarize a sample rather than use the data to learn about the population the data represents (inferential). You'd use this for survey data, for example. You might see these descriptive statistics about one variable depicted as a histogram or stem-and-leaf plot (brings me back to 8th grade!). When you have more than variable, you might see descriptive statistics represented as scatterplots or cross-tabulation.

For GIS, analyzing data a lot of the time is visualizing it. GIS can be used to depict two and three dimensional characteristics – such as rainfall in the desert. This can be mapped with contour lines that indicate the amounts of rainfall over a given area. A two-dimensional contour map created from the surface

Table 3: Funder/Journal Requirements vs Willingness to Share

		Q2: Do any of your funders or journal publishers require that you share your data upon publishing an associated paper?			
		Yes	**No**	**Unsure**	**All**
Q5: Have you shared, would you share, or are you required to share any of your data with other researchers?	**Yes**	50	14	6	70
	No	0	3	0	3
	Unsure	2	1	3	6
	All	52	18	9	79

Figure 5.6: Let's look at some of my cross-tabs: https://osf.io/preprints/lissa/q36uv/ && source file: https://osf.io/y6nmh/

modeling of rainfall point measurements may be overlaid and analyzed with any other map covering the same area.

GIS also involves a lot of modeling:

- Topological: examines relationships between geometric entities traditionally include adjacency (what adjoins what), containment (what encloses what), and proximity (how close something is to something else)
- Geometric: linear networks of objects that can be used to represent interconnected features, composed of edges connected at junctions.
- Hydrological: analysis of variables such as slope, aspect and watershed or catchment area
- Cartographic: process where several thematic layers of the same area are produced, processed, and analyzed. Operations on map layers can be combined into algorithms, and eventually into simulation or optimization models.

There is also a type of analysis called geostatistics – a branch of statistics that deals with spatial data with a continuous index. It provides methods to model spatial correlation, and predict values at arbitrary locations.

No matter what type of analysis you are doing, there should also be a plan in place to formally or informally document the workflows used to create results. This includes documenting:

- Data provenance
- Parameters used in the analysis (e.g. what qualitative framework, what)
- Connections betweeen analyses via inputs and outputs

One great way to do this and have a mostly-reproducible and well-managed pipelines for data preparation and analysis is **literate programming**. Donald Knuth first defined literate programming as a script/notebook/computational document that contains an explanation of the program logic in a natural language (e.g. Enlish, Mandarin), interspersed with snippets of macros and source code, which can be compiled and rerun.

This is a great approach – one I've adopted for this course! Here are some great examples of research

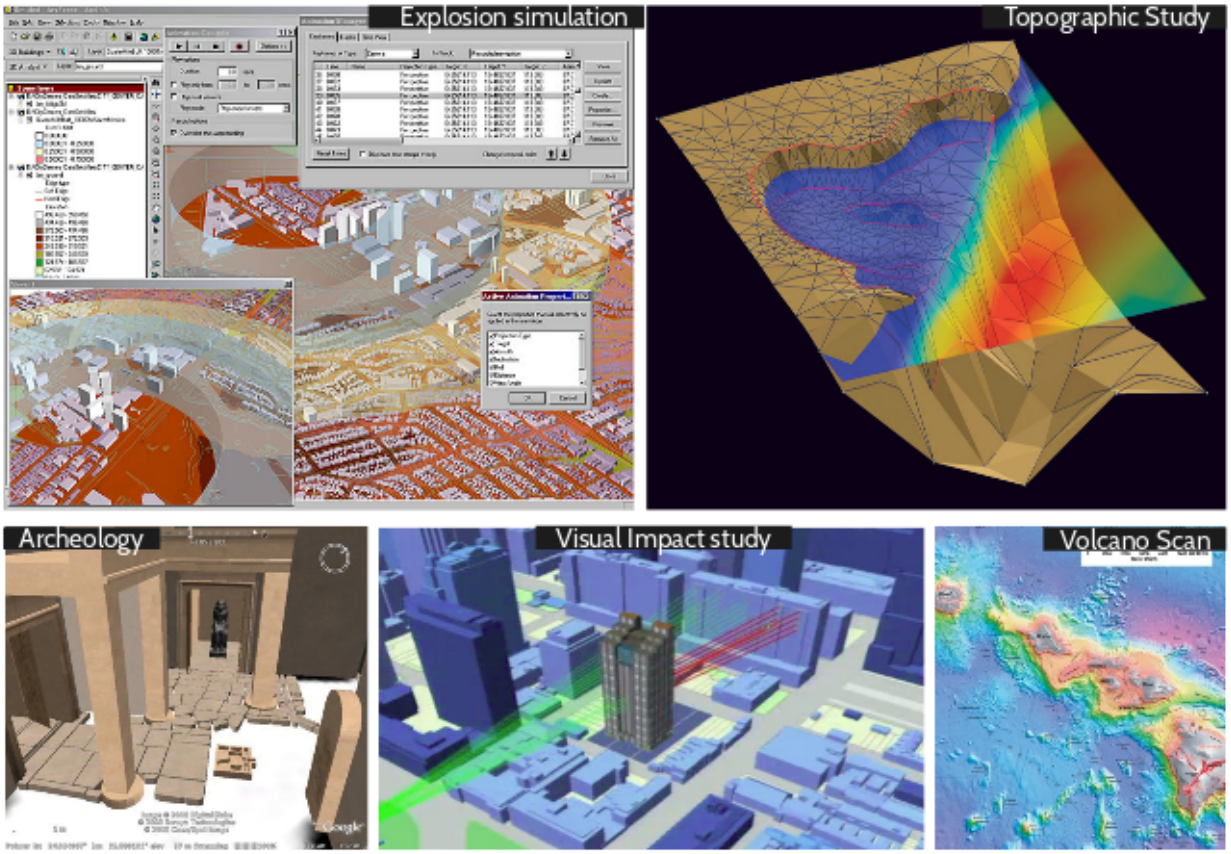


Figure 5.7:

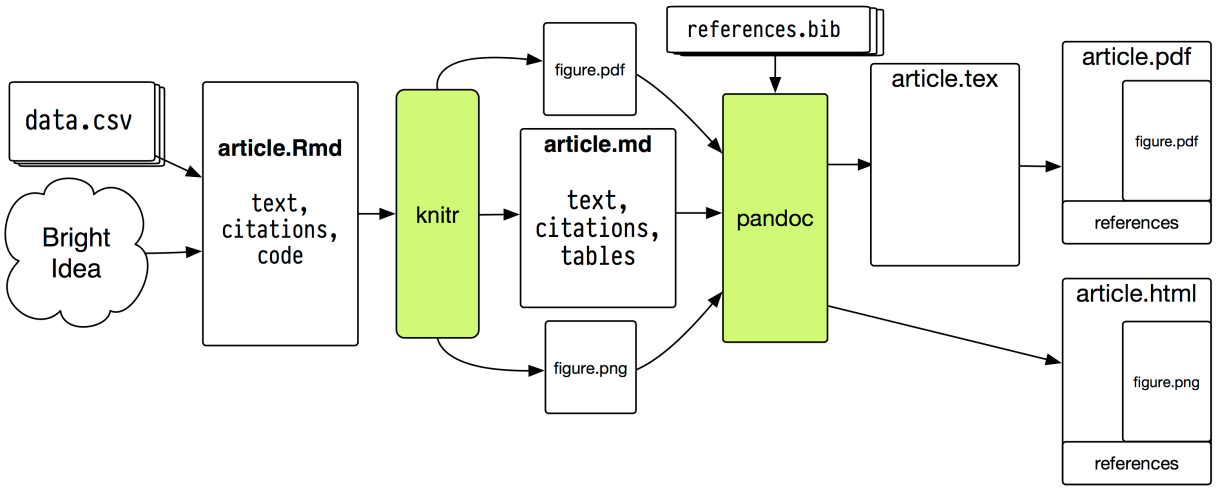


Figure 5.8:

expressed via literate programming:

- Analyzing Whale Tracks, by Roberto De Almeida
- A Reconstruction of 538 2012 Election Model, by Skipper Seabold
- Visual White Matter by Ariel Rokem
- UNCCC Data Report by Marwa Salem
- Shiny leaflet example by Matthew Leonawicz
- Reproducible research and GIScience: Computational Environment by Daniel Nuest

Now, let's do the next in-class lab!

5.3 Lab/Homework

Please submit your homework here: https://cloud.vickysteeves.com/index.php/s/b9rKQ2xm9yHoefC. The password will be given out in class. If you have more than one file to submit, please to it as a `.zip` file.

5.3.1 In-class

5.3.1.1 Data Preparation

Let's prep some data in OpenRefine! Please download the data and put it on your desktop (or somewhere you can find it): https://osf.io/39fus/

OpenRefine, formerly Google Refine and before that Freebase Gridworks, is an open source tool that allows users to load data, clean it quickly and accurately, transform it, and even geocode it. The main use of OpenRefine is data cleanup and transformation to other formats. What's more, is that all actions that were done on a dataset are stored in a project and can be replayed on another dataset!

Why Use OpenRefine?

- Simple installation
- Lots of great import formats: TSV, CSV, XML, RDF Triples, JSON, Google Sheets, Excel
- Upload from local drive or import from URL
- Many export formats: TSV, CSV, Excel, HTML table
- Works with large-ish datasets (100,000 rows). Can adjust memory allocation to accommodate larger datasets.
- Useful extensions: geoXtension, Opentree for phylogenetic trees from Open Tree of Life, and many more (listed here, scroll to 'extensions')!
- Active development community

Installation & Running To make sure your installation goes through smoothy, ensure you have the most updated Java JRE installed (get it from java.com/en/download/). You can then download and install Open Refine at openrefine.org/download.html.

In Windows, you can start the OpenRefine program by double-clicking on the `openrefine.exe` file. Java services will start automatically on your machine, and a terminal windows pops up! You don't need to do anything with this terminal – just let it run in the background. A browser window will open in your default browser to begin your OpenRefine session. On a Mac, OpenRefine can be launched from your Applications folder. If you are using Linux, you will need to navigate to your OpenRefine directory in the command line and run `./refine`.

> Note: If a new browser window does not open, then go to your favourite browser and visit the URL 127.0.0.1:3333/. Even though OpenRefine works in the browser, we are using it 100% locally.

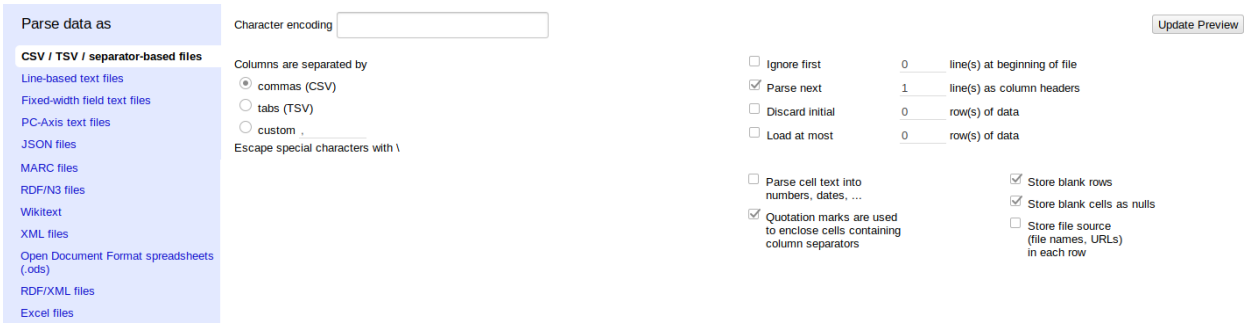


Figure 5.9: Settings when we preview our data.

Starting a Project

Once OpenRefine is launched in your browser, you'll see three options on the left sidebar: `Create Project`, `Open Project`, and `Import Project`. We are going to start a new project!

1. Click `Create Project` and select `Get data from This Computer`.
2. Click `Choose Files` and select the dataset you've downloaded for this class. Click `Open` or double-click on the filename.
3. Click `Next>>` under the browse button to upload the data into OpenRefine.

Before you 100% import the file into your project, OpenRefine lets you preview it to make sure everything is ok - set the number of headers, set the encoding (utf-8, y'all), and or pick the right file type (if CSV vs. TSV gets confused, for instance!). If this is the wrong file, click `<<Start Over` on the upper left corner of the screen. If all looks well, click 'Create Project>>. on the upper right corner of the screen.

So to start, let's:

1. Create a project in OpenRefine.
2. Import the raw data you've downloaded for the class, checking to make sure that you've configured the import correctly in the Preview.
3. Raise your hand to show you've finished!

Now that our data is in OpenRefine, let's check it out!

Faceting

Facets are a great way to see the big picture of your data – when you look at facets for a given column, it shows all unique entries with frequencies. You can use that to get a feel for how messy (or maybe not!) your data is. You can also use facets to subset rows that you want to change in bulk. One type of facet is called a 'text facet'. This groups all the identical text values in a column and lists each value with the number of records it appears in. The facet information always appears in the left hand panel in the OpenRefine interface. As well as 'Text facets' OpenRefine also supports a range of other types of facet. These include:

- Numeric facets
- Timeline facets (for dates)
- Custom facets
- Scatterplot facets

Some of the default custom facets are:

- Word facet - this breaks down text into words and counts the number of records each word appears in
- Duplicates facet - this results in a binary facet of 'true' or 'false'. Rows appear in the 'true' facet if the value in the selected column is an exact match for a value in the same column in another row
- Text length facet - creates a numeric facet based on the length (number of characters) of the text in each row for the selected column. This can be useful for spotting incorrect or unusual data in a field

where specific lengths are expected (e.g. if the values are expected to be years, any row with a text length more than 4 for that column is likely to be incorrect)
- Facet by blank - a binary facet of 'true' or 'false'. Rows appear in the 'true' facet if they have no data present in that column. This is useful when looking for rows missing key data.

  RECAP: 'Facet' groups similar values that appear in a column, and then allow you to filter the data by these values and edit values across many records at the same time.

Here we will use faceting to fix all the ridiculous characters in the `univeristy` column. 1. Scroll over to the `university` column and click 2. Click the down arrow and choose `Facet > Text facet` 3. In the left panel, you'll now see a box containing every unique value in the `university` column along with a number representing how many times that value occurs in the column.

Let's try this out some more:

1. Try sorting this facet by name and by count. Do you notice any problems with the data? What are they?
2. Hover the mouse over one of the names in the Facet list. You should see that you have an edit function available.

If you find an error you want to fix, you can do it this way! OpenRefine gives you the option to edit when you hover over a facet – you could use this to fix an error immediately, and OpenRefine will ask whether you want to make the same correction to every value it finds like that one. But OpenRefine offers even better ways to find and fix these errors, which we'll use instead. We'll learn about these when we talk about clustering.

BEFORE that, let's fix some of the data! In the `university` column you may have noticed values that look like `Lumi%C3%A8re University Lyon 2`. And there are many more of these!! This is kind of ugly and not very human **or** machine-readable. Let's fix it with something called GREL!

GREL stands for the Google Refine Expression Language, and it's a way we can automate changes. You can use GREL to query APIs, change data formats, split columns, and a whole lot more. OpenRefine lets you choose between GREL, Python or Jython (an implementation of python designed to run on the Java platform), or Clojure (dialect of the Lisp programming language.). GREL does what we need to do the most simply, so let's test GREL out on the `university` column!

Click the download arrow to the left of the `university` column header. Click `Edit column` then `Add column based on this column`. A window will pop up waiting for you to input a command! The first thing we'll do, since it's the first box, is name our new column. The GREL snippet we'll use is:

```
(value.unescape('url'))
```

We are telling GREL that for every `value` in the column, apply the `unescape` function designed for `HTML` (including `URL`s). You might notice that the weird symbols in the `university` follow a similar pattern – the % followed by some capital letters and/or numbers. This is a URL-thing.

One of my favourite things about OpenRefine is that when you are doing these advanced operations, it gives you a preview before you execute! The Preview tab in the window is the default, but you can also go to the Help pane to view **every single** command available for you – sortable, searchable, and star-able!

Now that all the names are cleaned up, we can find all the nuanced spellings and other semantic mistakes that would otherwise take forever to find. In OpenRefine, we do this with **clustering**. Clustering means "finding groups of different values that might be alternative representations of the same thing". Think back to my "NICHOLAS WOLF", "Wolf, Nicholas", and "Nicholas Wolf" example from before. Clustering is a very powerful tool for cleaning datasets which contain misspelled or mistyped entries. OpenRefine has several clustering algorithms built in. I basically click around on between each algorithm until I make sure I get all the misspellings.

Let's try it out!

1. In the `fixed_uniName` column text facet, click the Cluster button.

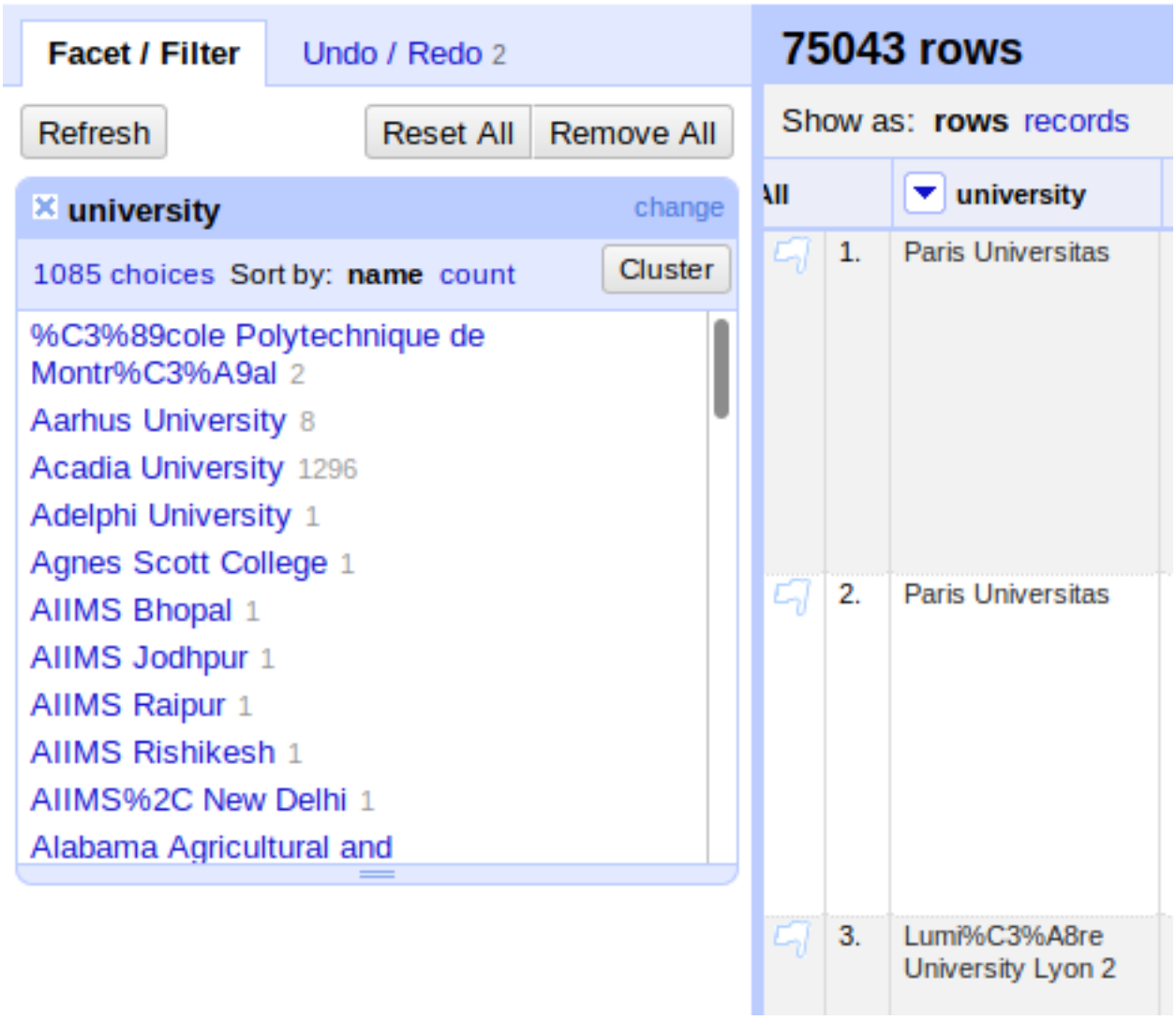


Figure 5.10: The text facet of the `university` column

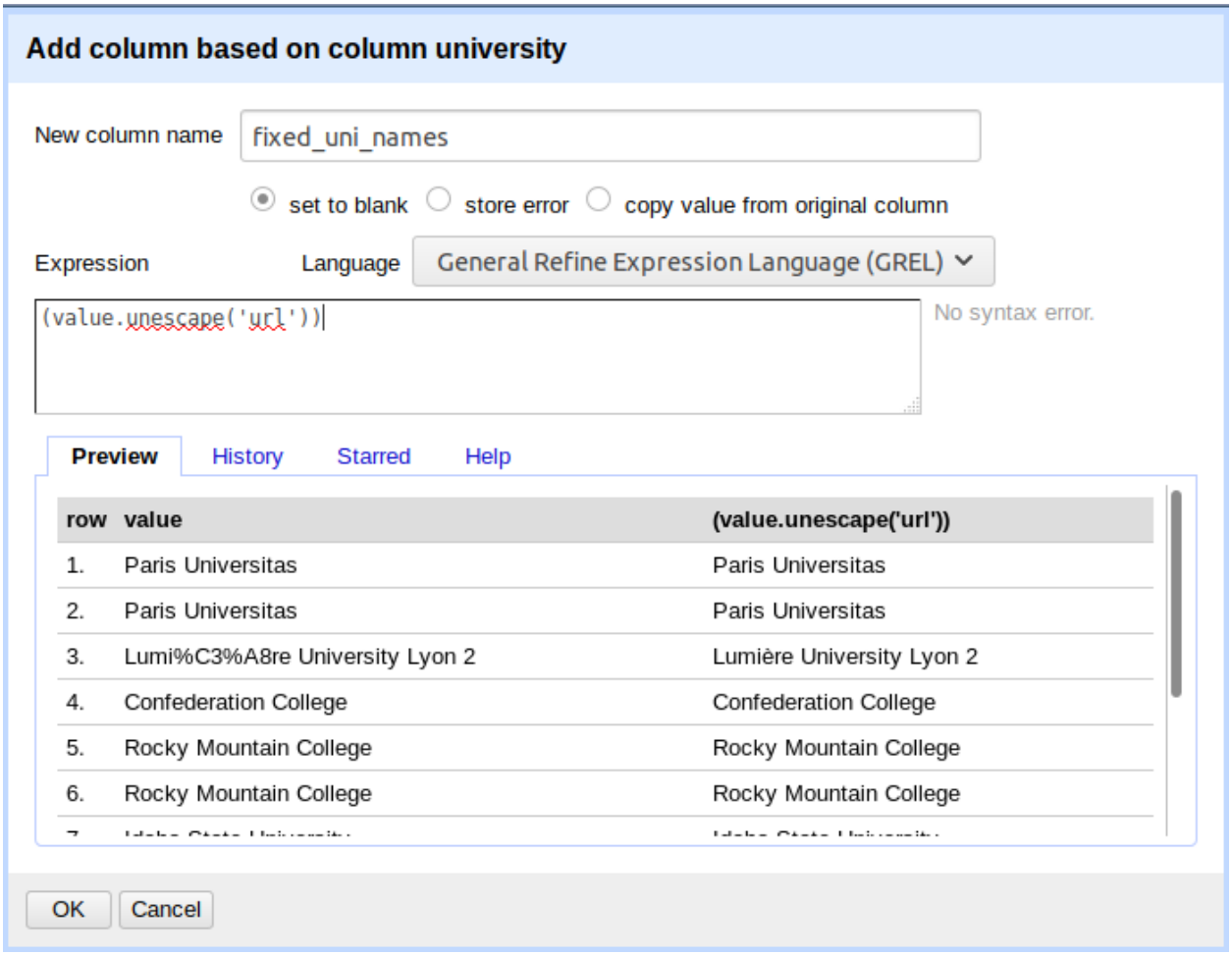


Figure 5.11: Adding a column based on a column in OpenRefine

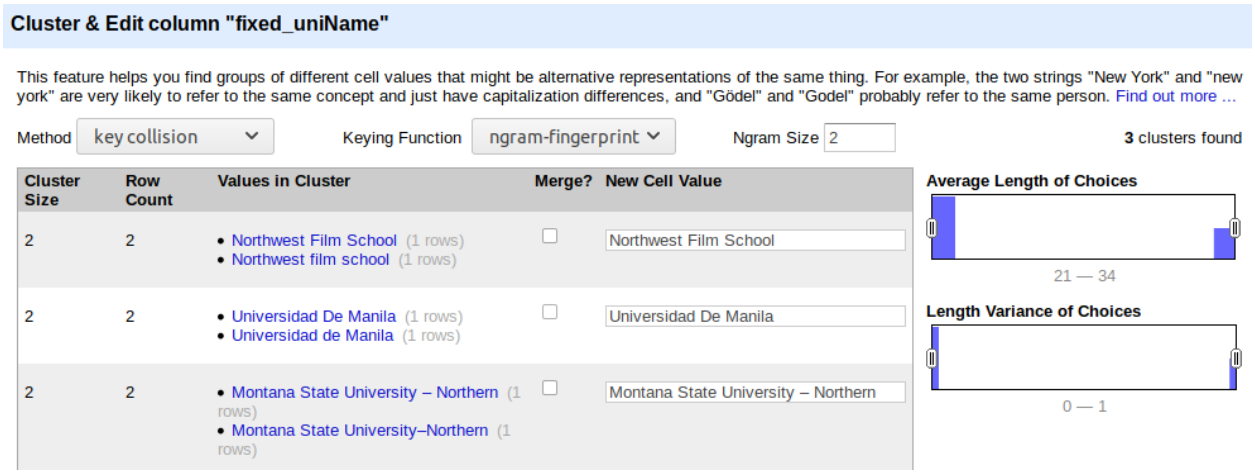


Figure 5.12: One type of clustering to find duplicates/misspellings

2. In the resulting pop-up window, you can flip between the algorithm configurations.
3. There should be at least two clusters minimum you should find.
4. Click the `Merge?` box beside each cluster, then click `Merge Selected and Recluster` to apply the corrections to the dataset.
5. Try selecting different Methods and Keying Functions again, to see what new merges are suggested.

Regular expressions?!

You can use regular expressions in GREL to powerfully repurpose and redefine your data! A regular expression, regex, is a sequence of characters that define a search pattern. There's tons of escaping to do, lots of special characters to keep watch for, and it's generally really hard to make the pattern. HOWEVER! Once we find the patterns to use (and there are also websites that can help us…), it's pretty powerful.

You may have noticed in the `established` column, there is a mix of dates and formats, such as `1963 - university status`, `1918-05-01`, and `Chartered 1984`. So, let's try to normalize all the information in this column using GREL and regular expressions!

The one thing that everything has in common is years in YYYY format. So, in the `established` column, click `Edit column`. You can run the GREL code `value.match(/.*(\d{4}).*/)[0]` to find the years in each cell, and transform the cells to only have that value!

Extra special Find & Replace

Regular expressions are only one way that we can transform the information in a cell. We can also use the `replace` and `contains` function to normalize our data!

There are a fair few things we'd like to replace. In the `endowment` column, click `Edit column`. You can run the GREL code below all to just get the

```
value.replace("USD","")
value.replace("US $","").replace("US$", "")
value.replace("US $","").replace("US$", "").replace("$","")
```

Oh no! We missed one part of the `endowment` column that still needs to be normalized - the original owner of this dataset added the word 'million' instead of adding the necessary zeroes. Let's replace that so we can have all numerical values in the `endowment` column. We need to find all the values that contain (the `contain` function!) and replace (the `replaces` function!) it with six zeroes. We can use this GREL statement in the `endowment` column > `Edit column`.

```
value.contains("million")
toNumber(value.replace("million", ""))*1000000
```

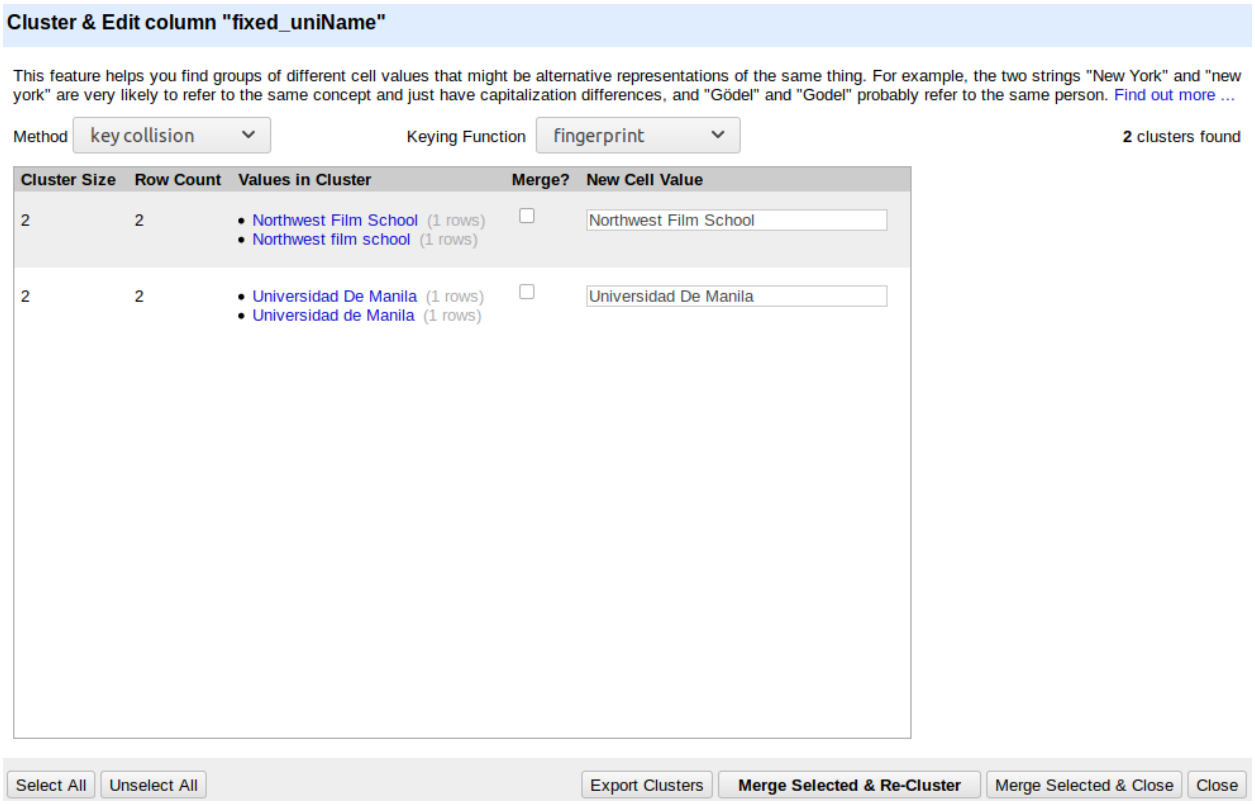


Figure 5.13: Another type of clustering to find duplicates/misspellings

Geocoding with Google's API

We can do a lot more than cleaning with OpenRefine as well! I use it a lot for geocoding – and we can do that in two lines of GREL! Let's give it a go with our dataset. Navigate to the `country` column and click `Edit column` and then `Add column by fetching URLs`. This gives us the same window we've seen many times over by now. Name the new column `geocode` and run the following GREL code:

```
http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
```

This basically takes the contents of the `country` column and queries the Google maps API with each value, and gets the resulting JSON file back in the new column. We can't do much with the JSON, though. What we really want is the lat and long coordinates! To do this, navigate to the `geocode` column in OpenRefine and click `Edit column` and then `Add column based on this column`. This gives us the same window we've seen many times over by now. Name the new column `latlong` and run the following GREL code:

```
with(value.parseJson().results[0].geometry.location, pair, pair.lat +", " + pair.lng)
```

This takes all the values of `geocode` and uses the function `parseJSON` to literally parase all the JSON. Google Maps API returns this bit of JSON from the query:

```
 "location" : { "lat" : 46.227638, "lng" : 2.213749 }
```

And we used `parseJSON` to find the lat and long in the location field of the JSON! Now, we have a column that contains the lat and long separated by a comma! This is easier to deal with – in OpenRefine, you can `Edit columun` and then `Split into several columns` and make sure you tell OpenRefine that the separator is a comma. You can then just rename the columns to `lat` and `long` respectively.

Undo some of your work

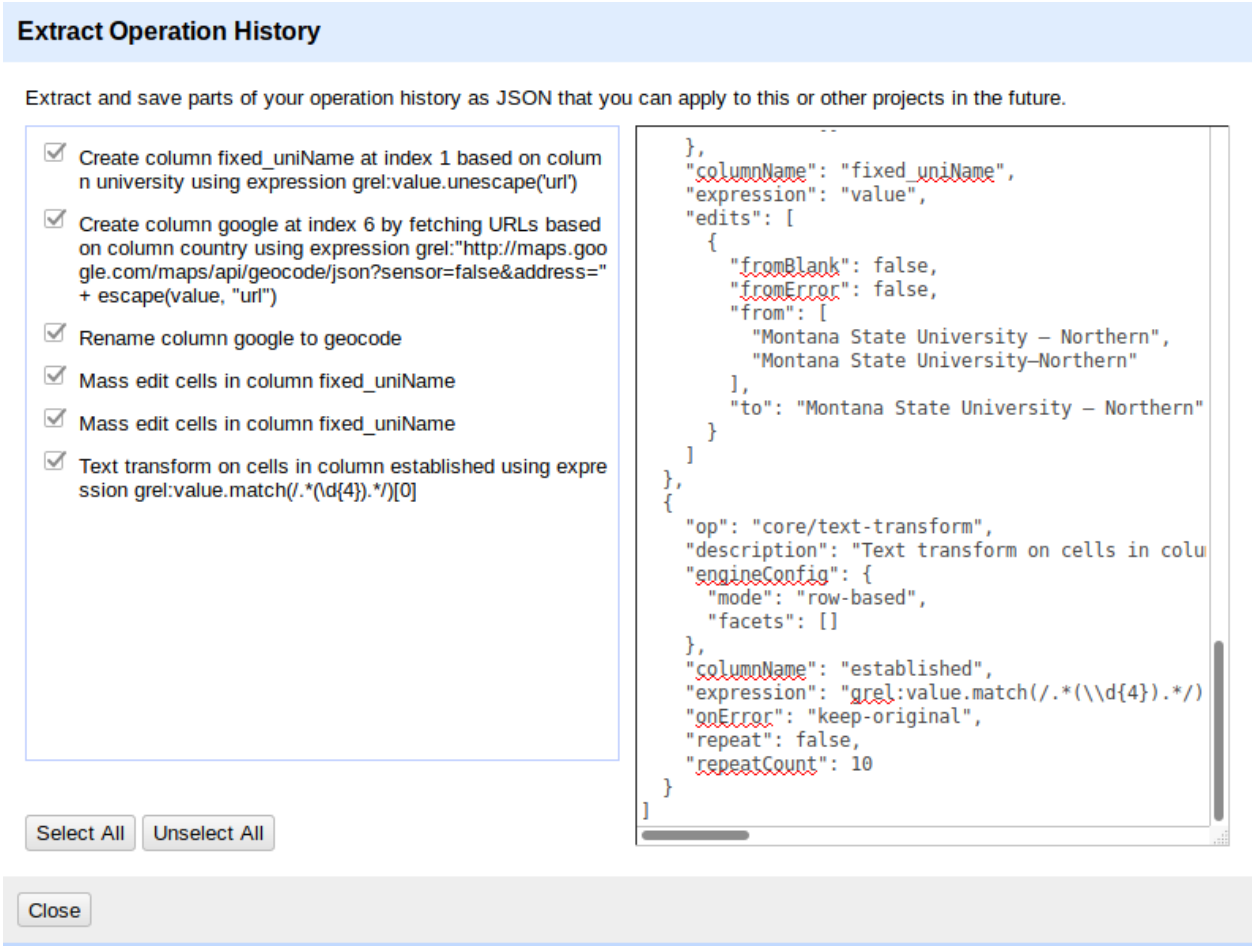


Figure 5.14: The panel that shows us our data cleaning history as JSON

1. Click where it says `Undo / Redo` on the left side of the screen. All the changes you have made so far are listed here.
2. Click on the step that you want to go back to, in this case go back several steps to before you had done any text transformation.
3. Visually confirm that those columns now contain the special characters that we had removed previously.
4. Notice that you can still click on the later steps to Redo the actions. Before moving on to the next lesson, redo all the steps in your analysis so that all of the column you modified are lacking in square brackets, spaces, and single quotes.

Documenting our work As you conduct your data cleaning and preliminary analysis, OpenRefine saves every change you make to the dataset. These changes are saved in a format known as JSON (JavaScript Object Notation). You can export this JSON script and apply it to other data files. If you had 20 files to clean, and they all had the same type of errors (e.g. genes encoded as dates, name misspellings, leading white spaces), and all files had the same column names, you could save the JSON script, open a new file to clean in OpenRefine, paste in the script and run it. This gives you a quick and reproducible way to clean all of your related data, **across operating systems**!

1. In the `Undo / Redo` section, click `Extract...`, and select the steps that you want to apply to other datasets by clicking the check boxes.

2. Copy the code from the right hand panel and paste it into a text editor (like `NotePad` on Windows or `TextEdit` on Mac). **Make sure it saves as a plain text file**. Let's practice running these steps on

a new dataset. We'll test this on an uncleaned version of the dataset we've been working with.

3. Create a new OpenRefine project and upload the uncleaned version of the dataset which you saved to your computer (in a particular folder in your git repository folder!).
4. Make sure that you name the new project something different from the one we've been working in!
5. Click the `Undo / Redo` tab > `Apply` and paste in the contents of `txt` file with the JSON code.
6. Click `Perform operations`. The dataset should go through the same data cleaning operations as the processed dataset from class!

For the sake of time, we used the same dataset, but in reality you could use that script to clean any related datasets (as long as folks use the same column headers!).

Save & Export When we save our OpenRefine project, we are saving not just the data, but also all the information about the cleaning and data transformation steps you've done. Once you've saved a project, you can open it up again and be just where you stopped before (and others can extend your existing work, and see the provenance of how your data was cleaned!).

OpenRefine by default autosaves your projects. If you close OpenRefine and open it up again, you'll see a list of your projects. You can click on any one of them to open it up again.

You can also export a project, which is quite helpful for sending to collaborators (alongside your raw data) to follow (or critique…) your data cleaning steps, since the OpenRefine project contains provenance of your work. You could even elect to share this information as a supplement to a publication!

1. Click the `Export` button in the top right and select `Export project`.
2. A `tar.gz` file will download to your default `Download` directory. The `tar.gz` extension tells you that this is a compressed file, which means that this file contains multiple files. You can extract the information
3. Look at the files that appear in this folder. What files are here? What information do you think these files contain?

You can also export just your cleaned data, rather than the entire project:

1. Click `Export` in the top right and select the file type you want to export the data in. `Comma-separated values` (`csv`) is typically the best choice.
2. That file will be exported to your default `Download` directory, from where you can share it out!

5.3.1.2 Data Analysis

Let's try some literate programming! We're going to try this in R Markdown. You can do the entire analysis pipeline in an R Markdown document: Data (pre-)processing, analysis, outputs, visualisation.

R and **RStudio** are separate downloads and installations. R is the underlying programming language, but using R alone can be daunting – we'd have to write our R files in a text editor and then run the scripts in the terminal. RStudio is a graphical integrated development environment (IDE) that makes using R much easier and more interactive. To function correctly, RStudio needs R and therefore both need to be installed on your computer.

R is an implementation of the S programming language combined with some extra secret sauce, inspired by Scheme. R is named partly after the first names of the first two R authors and partly as a play on the name of S. We'll use RStudio to write our code, navigate the files on our computer, inspect the variables we are going to create, and visualize the plots we will generate.

We are going to create a new project in RStudio and then write some R Markdown files! When a new project is created RStudio:

- Creates a project file (with an `.Rproj` extension) within the project directory. This file contains various project options and you can also double-click it to start RStudio again.

- Creates a hidden directory (named `.Rproj.user`) where project-specific temporary files (e.g. auto-saved source documents, window-state, etc.) are stored.
- Loads the project into RStudio and display its name in the Projects toolbar (which is located on the far right side of the main toolbar)

So let's:

1. Download the cleaned data: https://osf.io/qxgvn/
2. Start RStudio.
3. Under the `File` menu, click on `New project`. Then `New Directory` and make a test folder for us to work in.
4. Click on `Create project`.

You'll notice immediately that RStudio is divided into 4 "Panes": the editing window for your scripts and documents (top-left, in the default layout), your Environment/History/Git (top-right), your Files/Plots/Packages/Help/Viewer (bottom-right), and the R Console (bottom-left). The placement of these panes and their content can be customized (see menu, Tools -> Global Options -> Pane Layout).

One of the advantages of using RStudio is that all the information you need to write code is available in a single window. Additionally, with many shortcuts, autocompletion, and highlighting for the major file types you use while developing in R, RStudio will make typing easier and less error-prone.

RMarkdown

So let's get into R now. We are going to write our R code using RMarkdown. RMarkdown is an extension of Markdown. It works sort of like an executable paper – it mixes documentation & code, and not just R! You can insert code snippets from other languages (SQL, bash, Python, and more!). This allow you write documents which integrate results from your analysis. Incorporating results directly into your documents is an important step in reproducible research. Any changes that occur in either your data set or the analysis are automatically updated in your document the next time the document is created.

Five benefits of RMarkdown for your daily work:

1. You can keep an eye on text (the paper) *AND* the source code. These computational steps are essential to ensure computational reproducibility.
2. You can easily share R Markdown documents with colleagues, as supplemental material, or as the paper under review. Thanks to the package `knitr`, others can execute the document with a single click and receive, for example, HTML or PDF renderings.
3. Figures get automatically updated if you change the underlying parameters in the code. The error-prone task of exporting figures and uploading the right figure version to another platform is thus not needed anymore.
4. If you do not make any changes to document after creating the output document, you can be sure that the paper was executable at least at the time of submission. Since Markdown is a text-based format, you can also use **versioning with git**.
5. You can refer to the corresponding code lines in the methodology section making it unnecessary to use pseudo code, high-level textual descriptions, or just too many words to describe the analysis.

Creating a reproducible document in RMarkdown Ok, you should be in your `docs` folder now. To create a new R Markdown file, go to `File > New File > R Markdown`.

Some notes about R. What are known as `objects` in R are known as `variables` in many other programming languages. I am going to call them objects. Depending on the context, object and variable can have drastically different meanings. For more information see: https://cran.r-project.org/doc/manuals/r-release/R-lang.html#Objects

Create code chunks and text Let's look at some of the basics in R! Right now, this is just a narrative text written out with no special wrappers. Just text in a text box. But! I want to show you how R works. So, I need to insert a R *chunk*. In the source code pane, you might see a icon `C` with a plus sign and the

word `Insert`. If you click that, you can choose to insert a code chunk of variable types. We want R right now.

This inserts a wrapper that tells R Markdown that it needs to run some R code now:

```
# blank code chunk
```

We can run code chucks by either clicking the `>` symbol in the top-right corner of the code chunk itself, or near the `Insert` menu, you'll see a `--> Run` menu as well. There, you can choose to run the current code chunk, *all* the code chunks, or even selected lines within a chunk. Ok, let's get into R and look at how to make objects. We assign values by typing out the `<-`. We use this for objects, for DataFrames – for everything that needs an assignment!

```
cool_num <- 100     # assigns the object a value and a name to call it by
(cool_num <- 100)   # but putting parenthesis around the call prints the value of `cool_num`
```

```
## [1] 100
```

```
cool_num            # and so does typing the name of the object
```

```
## [1] 100
```

If you look at the `Environment` pane, you'll see our variable `cool_num`! It's been loaded into memory. Now, we can do other stuff with it, like call it in conjunction with maths!

```
other_num <- 2.2 * cool_num # doing some arithmetic and assigning it to a new object
other_num
```

```
## [1] 220
```

One of the main reasons people use R is the bevy of `functions`. Functions are built into R and extended with R packages. Functions help automate more complicated sets of commands. We'll only use predefined functions, but you can also define your own as your R proficiency grows!

A function usually gets one or more inputs called `arguments`. Functions often (but not always) return a value.

We'll take a look at the function `sqrt()`. The input (the argument) must be a number, and the return value (in fact, the output) is the square root of that number. Executing a function ('running it') is called `calling the function`. Let's find the square root of our variable `other_num`:

```
sqrt(other_num)
```

```
## [1] 14.8324
```

Ok, let's use our own data now & process it!

Your datasets and files are ***here()***. We have some external datasets we want to to include into our RMarkdown file. What most people do is something like:

```
# source("/home/vicky/Downloads/2018-uutah-repro/anotherRscript.R")
```

This work fine. For me. On my computer. Not on yours. We want to work well with others, so we need to use something called *relatively paths*. These errors are one of the most common errors in sharing, and disturb the entire execution process of our executable paper. So let's try the following:

```
if(!requireNamespace("here"))
  install.packages("here", repos = "https://cloud.r-project.org")
library("here")
```

Alright, ***here()*** starts relative to the folder structure on my laptop. This means, when using it on your machine, it starts relative to the folder structure on your own machine.

This allows us to provide all directories relative to the top directory of the project folder, which is in this case `/test/`. Now, our scripts will work on my machine *and* on yours. Ok, let's put our *cleaned* data from OpenRefine into a data frame!

```
uni <- read.csv("results/2018-09-27_universityData.csv")
```

Now that we've read in the data we cleaned in OpenRefine, let's get rid of the duplications! We have lots of rows for the same university counting different endowments. I really only care about the *total* of endowments per school.

```
uni_money<- subset(uni, select=c("universities", "endowment")) # make a new subset
head(uni_money) # makes ure it looks right
```

```
##                universities endowment
## 1          Paris Universitas        15
## 2          Paris Universitas        15
## 3 Lumière University Lyon 2       121
## 4     Confederation College   4700000
## 5    Rocky Mountain College  16586100
## 6    Rocky Mountain College  16586100
```

```
# make the endowment column a number
uni_money$endowment <- as.numeric(as.character(uni_money$endowment))
```

```
## Warning: NAs introduced by coercion
```

```
# deduplicate the data and sum the rows so we don't lose data!
dedupe_unimoney<-ddply(uni_money,.(universities),function(x) data.frame(universities=x$universities[1],

# get the first 6 rows of our newly deduped and summed rows
head_dedupedMoney <- head(dedupe_unimoney)
head_dedupedMoney
```

```
##          universities  endowment
## 1   Aarhus University 4.5864e+10
## 2   Acadia University 5.1840e+10
## 3  Adelphi University 8.6000e+07
## 4 Agnes Scott College 2.3060e+08
## 5        AIIMS Bhopal         NA
## 6       AIIMS Jodhpur         NA
```

Now that we have some clean, relevant information to use, let's plot it! I mainly use R for plotting purposes myself. I think the range and options are vastly superior to other data viz packages out there. I can even get plots with DPIs enough for print!

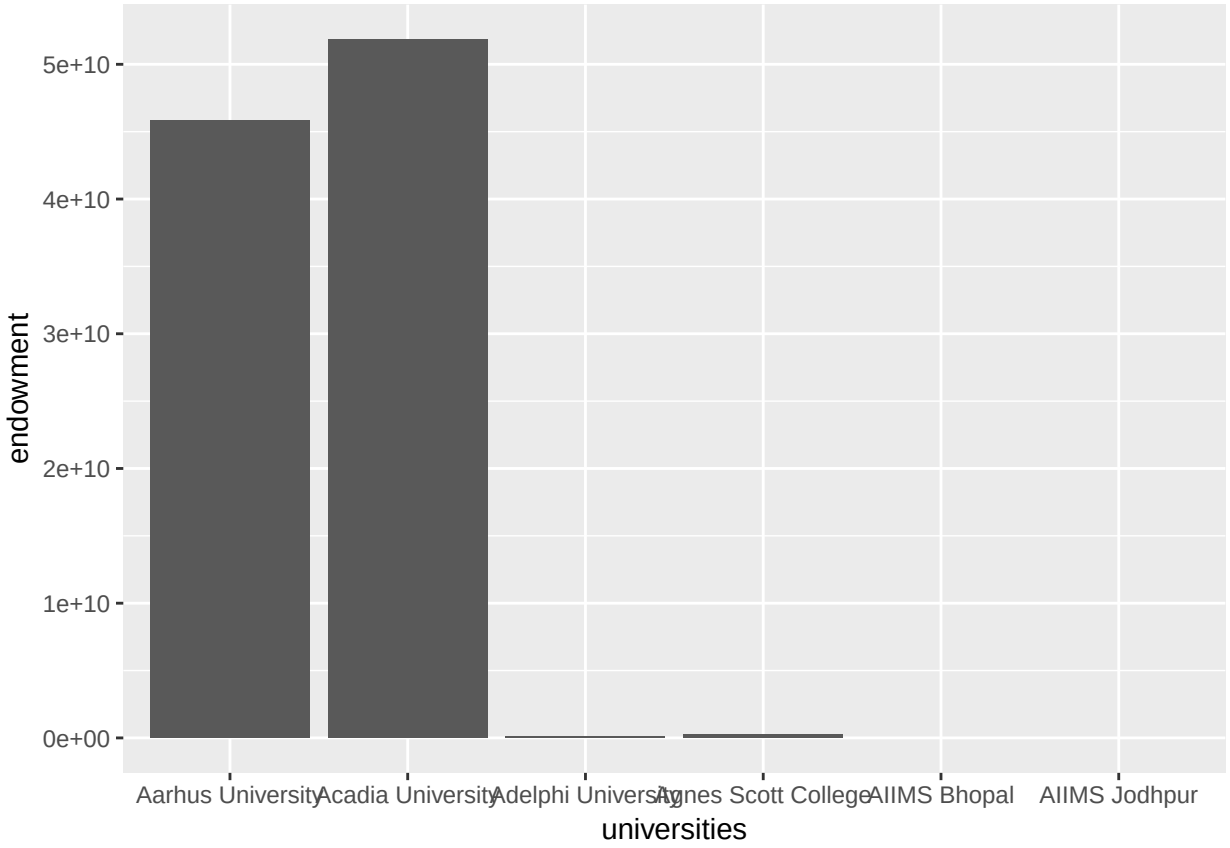
For now, let's do a basic histogram of the first few rows:

```
p <- ggplot(head_dedupedMoney, aes(x = universities, y = endowment))  + geom_histogram(stat="identity")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
p
```

```
## Warning: Removed 2 rows containing missing values (position_stack).
```



We can also do some mapping in R and make it interactive with `leaftlet`! First we need to make *another* subset, with just the relevant information – the universities and their geographic locations.

```
plot<- subset(uni, select=c("universities", "lat", "long")) # make a new subset for map
uni_plot <- unique(plot[,]) # deduplicate the data!

head(uni_plot) # makes ure it looks right
```

```
##                universities      lat        long
## 1          Paris Universitas 46.22764    2.213749
## 3  Lumière University Lyon 2 46.22764    2.213749
## 4      Confederation College 56.13037 -106.346771
## 5     Rocky Mountain College 37.09024  -95.712891
## 7     Idaho State University 37.09024  -95.712891
## 19       University of Milan       NA          NA
```

Ok, let's map!

```
m <- leaflet(uni_plot) %>%
addProviderTiles(providers$CartoDB.Positron)
m %>% setView(-72.690940, 41.651426, zoom = 8)
```


```
m %>% addMarkers(~long, ~lat, popup = ~as.character(universities), label = ~as.character(universities))
```

```
## Warning in validateCoords(lng, lat, funcName): Data contains 37 rows with
## either missing or invalid lat/lon values and will be ignored
```

Ok, we have some:

- data cleaning steps (subsetting)
- data analysis steps (summary statistics)
- narrative (our text throughout)
- plots (our map)

This way of publishing research results allows others to reuse your code and give you appropriate credit. For example, others might be interested in the same way of illustrating their data on the map, but for another region or another university dataset!

You can choose between HTML, PDF, and others, like our class book. Note: Not every template support HTML, so if you want that, choose wisely

5.3.2 Oustide class

Reading/writing

BONUS: Read this fabulous article I was going to include in the this homework: https://www.biorxiv.org/content/early/2018/07/24/344804. The reason I didn't was because *I couldn't find anything like it.*

Read either these two articles about p-hacking and HARKing:

1. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002106
2. http://journals.sagepub.com/doi/abs/10.1207/s15327957pspr0203_4

Write 700-1,000 words about the articles, giving your reaction/thoughts and also the questions outlined below:

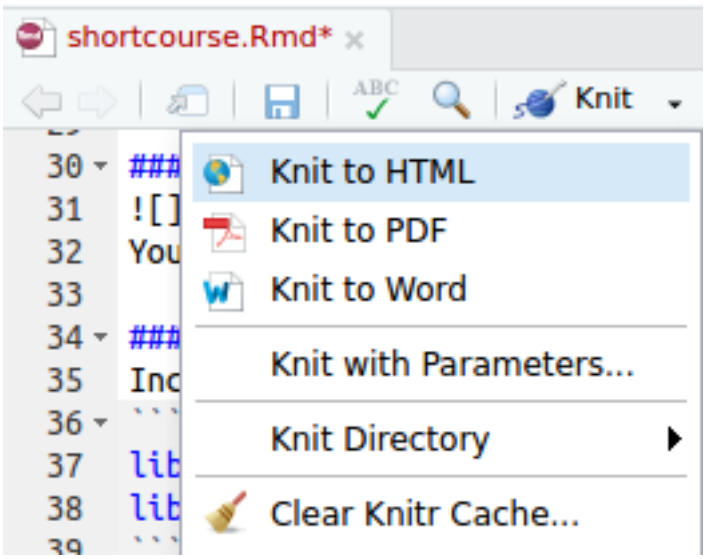


Figure 5.15: Knitting a document

- Is p-hacking really such a big deal? Don't people massage data all the time? What makes this so nefarious?
- Similarly, is HARKing really a problem or just 'exploratory' analysis?
- What can researchers do to avoid p-hacking/HARKing?
- What can journals and tenure & promotion committees do to dis-incentivize p-hacking and HARKing?

Hands-on

Pick a type of data to work with – either quanitative, qualitative, or GIS. "Big" was left out because I do not have the compute resources readily available to give you to do any labs with large datasets – sorry!

Choose one of these Programming Historian lessons to work through:

Qual

1. Fetching and Parsing Data from the Web with OpenRefine
2. Not a programming history lesson, but one good way to do qual is to practice and get comfortable with the *subjectivity* of it. Download this short (3min) audio clip: https://cloud.remram.fr/s/CF7tqH6gSx8NeWF (password is the same as the homework password) and transcribe it using oTranscribe: https://otranscribe.com/. Save the transcript as a `.txt` file. Make a copy of that and open it in Word. Go through and highlight the text for different themes that you see in your close reading. Make a key that shows which color highlight goes to which theme.

Quant

3. R Basics with Tabular Data
4. Counting Word Frequencies with Python

GIS

5. Geocoding Historical Data using QGIS
6. Using Geospatial Data to Inform Historical Research in R

Send in the raw, analyzed, and clean data, any output you were expected to make at the end of the lesson (if any – some of these lessons have you making a viz, map, or a figure), and a 250-500 word reflection on the lesson, the process, and the results. Don't spend more than ~2 hours trying to figure it out if you absolutely get stuck.

Chapter 6

Reproducibility

Agenda for today's class:

- 6:30 - 7:30: discussion
- 7:30 - 7:45: break
- 8:00 - 8:45: lecture
- 8:45 - 9:00: break
- 9:00 - 9:20: lab

6.1 Discussion questions

Elizabeth, our facilitator this week, prepared the following questions for in-class discussion:

1. Do you agree with Sayre that we have a reproducibility crisis?

- What makes it (or would make it) a crisis?

2. Can you think of a time that you have experienced the effects of poorly documented research?
3. In the spectrum of reproducibility (Reproducibility Librarianship, p.81), which do you imagine are the hardest to implement?
4. Out of the strategies for promoting reproducibility we encountered in the readings, which do you find compelling?

- What changes do they rely on? (Increased funding, R&D, communication networks, institutional changes, or a combination)

6.2 Lecture

While reproducibility has always been a part of the scientific method, research reproducibilty generally was first stressed in Western literature in the 17th century by the Irish chemist, Robert Boyle. Before research computing, reproducibility generally meant one researcher travelling to anothers' lab or workspace to try to re-do their work. Since the widespread use of born-digital research materials, pipelines, and processes, the ability to re-do others' work is just as arduous as trekking across a sea or continent. If you have ever tried to get something to work on Windows that originally worked on Mac, you might understand why that is.

While these definitions vary across disciplines, for this lecture I am using the terms reproducibility and replication like so:

Reproducibility: independent people use the same code and data to verify a claim

Figure 6.1:

Figure 6.2: My idea, made aesthetically pleasing by Andrew Rarig (NYU)

Replication: independent people use *different* code and data (collected the same way) to verify a claim

However, as with all things, reproducibility should be defined on a spectrum. The Stodden et al ICERM report (2013) outlines these five tiers for reproducibility that I particularly like:

> Reviewable Research: Sufficient detail for peer review & assessment. Replicable Research: Tools are available to duplicate the author's results using their data. Confirmable Research: Main conclusions can be attained independently without author's software. Auditable Research: Process & tools archived such that it can be defended later if necessary. **Open/Reproducible Research: Auditable research made openly available**

These can be mapped onto a pyramid, like so:

The idea is that we need the whole pyramid for research reproducibility. So the goal becomes to capture the whole pyramid – the code, data, documentation, narrative, and the computational environment. Natural scientists back in the day used to spend a lot of time documenting, with paintings and illustrations, the ecological environment in which they were doing their research. Researchers *today* use those descriptions to juxtapose the current landscape to the historical one, to draw conclusions about the changing world. Most of the time, researchers can return to these landscapes. However, this becomes much harder with computational environments – operating systems with their various configurations.

We call this **computational reproducibility**. With so much research today being completely born-digital, it's just as important to capture these computational environments for posterity – to be able to return to, access, and reproduce the research being done today.

ICERM Report Definitions	Potential Real-World Examples
Reviewable Research: Sufficient detail for peer review & assessment	the code and data are openly available

ICERM Report Definitions	Potential Real-World Examples
Replicable Research: Tools are available to duplicate the author's results using their data	the tools (software) used in the analysis are freely available for others to confirm results
Confirmable Research: Main conclusions can be attained independently without author's software	other can reach the conclusion using similar tools, not necessarily the same as the author, or on a different operating system
Auditable Research: Process & tools archived such that it can be defended later if necessary	The tools, environment, data, and code are put into a preservation-ready format
Open/Reproducible Research: Auditable research made openly available	Everything above is made available in a repository for others to examine and use

But first – **why do we care about reproducibility?**

There are many reasons to support reproducibility in research, including:

- **Build on top of previous work** – after all, research is incremental and we always rely and base ourselves in methods and techniques developed in the past.
  - Y'know. That Sir Isaac Newton quote in every reproducibility presentation ever: "If I have seen further, it is by standing on the shoulders of giants."
- **Help newcomers** – new students who want to learn the field, post-docs who might need to carry on a project.
  - Your reproducible work is their greatest teacher!
  - It's always a struggle to make things work when there's not enough detail or the right code/data to keep a project going.
- **Defeat self-deception**, which is in line with verifying the correctness of results.
  - This is not a matter of researchers not trusting each other – the issue is that even an honest person is a master of self-deception. Often, we are trying to find reasonable and acceptable outcomes to meet a certain deadline.
  - A very common fallacy in research, for instance, is to rigorously check the unexpected results but give expected results a free pass. Inviting others to reproduce your work and be the devil's advocate is a great way to verify if you are following the right path.
- Others can **re-use and extend your work** more easily! You can even find interesting collaborations and future research projects out of this.
- **YOU can re-use and extend your work more easily**! Future you is your greatest collaborator (and past you doesn't answer emails!)

Ok, so we care about reproducibility. It's so hard, though!

Yes, achieving that 100% reproducibility mark is very hard (and some say impossible). This is in part due to three major challenges:

1. **Workload & Time Challenges**: it is a time commitment to get data and code ready to share, and to share it
   - the *Incentive Problem*: reproducibility takes time, and is not always valued by the academic reward structure

   "Insufficient time is the main reason why scientists do not make their data and experiment available and reproducible." - Carol Tenopir, Beyond the PDF2 Conference

   "77% claim that they do not have time to document and clean up the code." – Victoria Stodden, Survey of the Machine Learning Community – NIPS 2010

2. **Technical Obsolescence**: technology changes affect the reproducibility
   - the *Pipeline Problem*: reproducibility requires skills that are often not included in most curricula!

> "It would require huge amount of effort to make our code work with the latest versions of these tools." – Collberg et al., Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

3. **Normative Dissonance**: espoused values do not always match behavior, for a number of reasons.

> "Scientists' views of their fields as cooperative or competitive were associated with their normative perspectives, with competitive fields showing more counternormative behavior." – Melissa S. Anderson, Brian C. Martinson, and Raymond De Vries, "Normative Dissonance in Science: Results from a National Survey of U.s. Scientists"

Some workflow-related barriers to reproducibility

We talked in week 4 about the importance of using open file formats and tools for data management and reproducibility, so I won't rehash that here, but the file format/tool problem in research (the problem being, too many tools and formats are closed and proprietary and expensive) *is* a high barrier to reproducible research. So keep it in the back of your head! You probably noticed that the Dekker & Lackie article was largely going over best practices in data management…

I want to highlight something related to workflows & reproducibility: the idea that clicking is not reproducible but learning to program is a high barrier. Dekker & Lackie highlight this in their Databrarianship chapter: "Novice researchers rarely have a chance to systematically learn about the essentials […] before they are faced with applying the necessary principles in their own research". And it often is – where in methodology sections do you see things like "played with settings in Adobe until my figure was exactly the way I want it" or "clicked into seven sub-menus in SPSS to get to this one specific feature essential for my stats".

The Yenni et. al. article from last week (it was optional) highlights these reasons for automating their workflow:

> We do this by leveraging existing tools to: 1) perform quality assurance and control; 2) import, restructure, version, and archive data; 3) rapidly publish new data in ways that ensure appropriate credit to all contributors; and 4) automate most steps in the data pipeline to reduce the time and effort required by researchers. The workflow uses two tools from software development, version control and continuous integration, to create a modern data management system that automates the pipeline.

While the authors discuss efficiency, the automation of their workflow (and thus, the reduction of click around…) is a really important step in the pursuit of reproducibility. Human error is greatly reduced (it's hard to keep track of you click in the heat of the moment, and even harder to remember after the fact), and reporting out is a lot simpler – one markdown file, one jupyter notebook, instead of a laundry list of menus to click!

Computational reproducibility

> An article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result. – Johnathan Buckheit and David Donoho, Stanford University

I always posit that good research data management enables reproducibility but does not *guarantee* it. Open file formats, good documentation, using open tools, and backing up your data is amazing and necessary, but even that isn't enough to ensure that I can rerun your work. That's computational reproducibility is about. What works on my Linux machine should give the same results when run on your macOS laptop. This phenomenon has been explored in publications such as:

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

> We investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). **Significant differences** were revealed between FreeSurfer

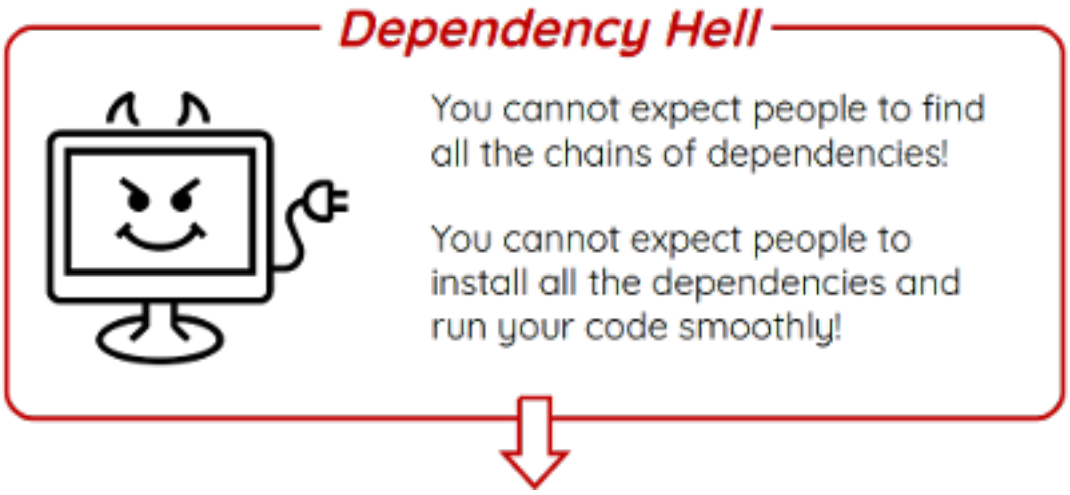


Figure 6.3:

> version **v5.0.0 and the two earlier versions**. [...] About a factor two *smaller differences* were detected between *Macintosh and Hewlett-Packard* workstations and between *OSX 10.5 and OSX 10.6*.

The challenge: environments are hard to capture

Gap: tools that can automatically capture all the dependencies in the original environment in a research compendium and automatically set them up in another environment. There are a few tools that try to address this gap in slightly different ways:

Containers: lightweight virtual operating systems you can send around to other people.

- Singularity (made for & popular in high performance computing)
  - Starting a Singularity container "swaps" out the host operating system environment for one the user controls – instantly virtualize the operating system, without having root access, and allow you to run that application in its native environment!
- Docker
  - Docker was made to "pack, ship and run any application as a lightweight container." – idea is to provide a comprehensive abstraction layer that allows developers to "containerize" or "package" any application and have it run on any infrastructure (doesn't really work on HPC though…).

The research community has been increasingly using and sharing containers (especially Docker) to try to mitigate this problem. However there are a few problems with containers:

- No idea of provenance. If I got a container and some code/data, I'd still need to know what to run first, which data is input/output, etc.
- Not trivial for new users to make *or* to use; they have a steep learning curve.
- Not sustainable; I can only use a Dockerfile with Docker, and it's not always backwards compatible. This is a big problem thinking long-term.

Packaging Systems: auto-capture of dependencies & source code used at time of running.

- ReproZip (I work on this!)
  - Open source tool that automatically captures provenance of research and packs all the necessary files, library dependencies, and variables to reproduce the results. Anyone can then unpack and reproduce the research without having to install any additional software!
- o2r
  - Give them a R workspace with an RMarkdown file, get a re-runnable paper in-browser. Uses docker to do this in the backend.

A few systems like Yale University Library's Emulation as a Service or Carnegie Mellon University's Olive Archive offer legacy base operating system access to users that could in time start to address computational reproducibility as analysis software is added to their collections, or they integrate with existing tools like Whole Tale or ReproZip.

This recent report discusses fair use in regards to software preservation (helpful for capturing proprietary environments for the scholarly record **and** sets us up for the licensing talk next week!): https://www.softwarepreservationnetwork.org/bp-fair-use/

The goal is: at the end of a project, take all the great data and project management skills and make what's called a *research compendium* or a *reproducible package* of all your work! This is a package that contains all of the things necessary to reproduce your work, taking even the computational environment into account.

> "Research compendia are an increasingly used form of publication, which packages not only the research paper's text and figures, but also all data and software for better reproducibility." - Nuest, Bottinger, & Marwick, How to read a research compendium

ReproZip, the tool I work on, was made in order to facilitate the creation of these self-contained, distributable bundles of pipelines! I thought I'd give a demo on packing and unpacking work with ReproZip to give you a better idea of how the workflow and resulting file might look like. Let's do that now! I'm using our demo virtual machine and example use cases from: https://examples.reprozip.org.

To recap:

- Good data management is necessary for reproducibility, but doesn't guarantee it
- You can work reproducibly in many ways – your data cleaning work can be somewhat reproducible, the analysis fully reproducible, or data collection not at all reproducible. It's a spectrum
- Introduce reproducible workflows in small bits, get comfortable, and expand.
- Open formats & open tools enable more reproducibility. Use them!

6.3 Lab/Homework

Please submit your homework here: https://cloud.vickysteeves.com/index.php/s/spWXRiNjBKe9rdQ. The password will be given out in class. If you have more than one file to submit, please to it as a `.zip` file.

6.3.1 In-class

We're going to explore reproducibility in the cloud with myBinder!

1. https://mybinder.org/v2/gh/betodealmeida/notebooks/master?filepath=blob%2Fmaster%2Fearth_day_data_challenge%2F
2. https://mybinder.org/v2/gh/betatim/sparql-notebooks/master?filepath=wikidata.ipynb
3. https://mybinder.org/v2/gh/ypriverol/github-paper/master
4. https://github.com/jakevdp/PythonDataScienceHandbook

They also have a neat add-on: https://addons.mozilla.org/en-US/firefox/addon/open-with-binder/

6.3.2 Outside class

Reading/Writing

Please read and respond to the following two articles:

1. How to read a research compendium
2. Enabling the Verification of Computational Results: An Empirical Evaluation of Computational Reproducibility

Write 700-1,000 words about the articles, giving your reaction/thoughts and also the questions outlined below:

- What do you think about the idea of a research compendium? Especially in the context of publishing supplementary materials?
- Why do you think there is a high barrier to computational reproducibility? What can librarians do to help that?
- How would you offer services in reproducibility to patrons? Which aspects of computational reproducibility would you focus on, if any at all?

Hands-on

Download and install Anaconda 3.6 on your computer. This includes Jupyter Notebooks and a number of relevant python packages.

1. Download this repository: https://github.com/arokem/visual-white-matter
2. Try to rerun the jupyter notebooks locally on your own computer in the right order, to get the same results.
3. Launch the myBinder instance of that repository: https://mybinder.org/v2/gh/arokem/white-matter-matters/master
4. Try to rerun the jupyter notebooks in your browser via myBinder in the right order, to get the same results.

Write 750 words maximum about the experience. Were you ever able to get the notebooks to run on your local computer? If you were reviewing Dr. Rokem's paper, what would your preference be – receiving the github link or the binder link? What are the pros/cons of trying to configure things locally vs. in-browser? Don't spend more than ~1 hour trying to get this to work locally if you are having trouble.

Chapter 7

Legal & regulatory environment

Agenda for today's class:

- 6:30 - 6:45: go over homework
- 6:45 - 7:30: discussion
- 7:45 - 8:00: break
- 8:00 - end: lecture from Megan Wacha!

7.1 Discussion Questions

Mary, our facilitator this week, prepared the following questions for in-class discussion:

1. Are intellectual property rights too broad or too narrow? Why?
    - In what ways have digital technology and the internet impacted that?
2. For academic research, from Arzberger et al.: "What reward structures might lead to better access and sharing practices?"
3. What is the role of a university librarian in either protecting, or arguing for free and open access to, the data produced by those working for the university?
4. Creative Commons (as well as other licensing resources) was developed in part to provide creators with a way to determine, on their own and independent of the state, what constitutes acceptable use of their work. In what ways has it succeeded? What still needs work?
5. Watters writes: "If digital utopia cannot imagine our existence, it makes sense it will not value our presence (or notice our absence). […] that also might mean it won't expect our resistance." What are some circumstances where resistance —where opting out of open—might be appropriate?

OPTIONAL! (But I'll answer honestly!) Have you ever violated someone's intellectual property? Assuming that many people do, should the law be changed?

7.2 Lecture

Today we are joined by Megan Wacha! Megan is the Scholarly Communications Librarian at the City University of New York's Office of Library Services. Prior to joining CUNY, they served as the Research and Instruction Librarian for the Performing Arts at Barnard College and held positions at the New York

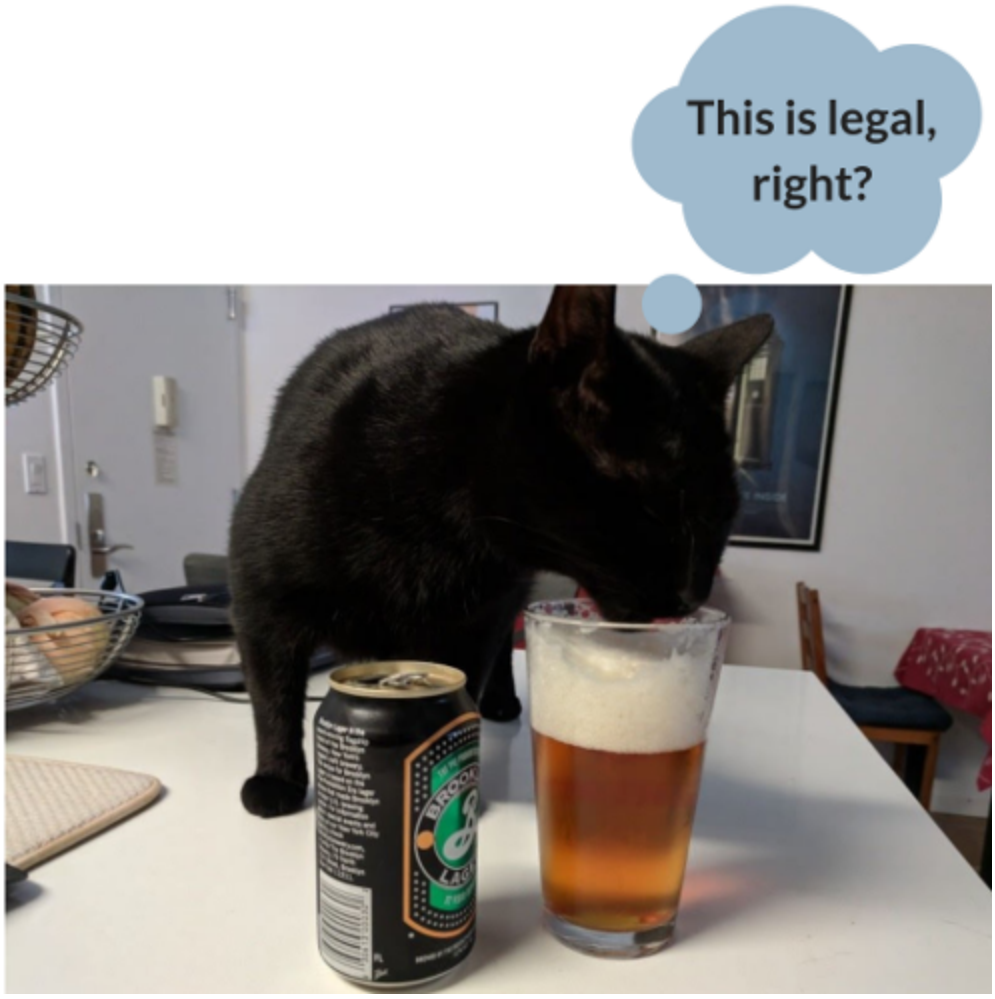


Figure 7.1:

Public Library for the Performing Arts and NYU's Fales Library & Special Collections. Megan holds an MSLIS from Pratt Institute and an MA in Performance Studies from NYU.

In addition to leading open research initiatives at CUNY, Megan is the President of Wikimedia NYC, a non profit dedicated to connecting the peoples and institutions of New York City with Wikipedia and the larger free culture movement. Megan is also on the steering committee of the LIS Scholarship Archive, a discipline-specific open source repository to enable those in library and information science and allied fields to make their work openly available.

A note from Vicky: Megan's twitter is really great so if you are on twitter you should follow them @megwacha!

7.3 Lab/Homework

No homework! Work on your first project check-in, due next week. Please submit the materials for the first check-in via this link: https://cloud.vickysteeves.com/index.php/s/CdFZAZiwFxwFz8d before 11:55pm on Thursday the 18th.

Please also send me your presentation materials (if you plan on using any) so we can minimize time setting up between presentations.

Chapter 8

Data services in libraries

NO CLASS THIS WEEK – VICKY IS OUT!

8.1 Lecture

So this is the first lecture of unit 2! Unit 1 was focused on more of the "on-the-ground" challenges of every day patrons from various domains/walks of life/areas of study. Unit 2 will focusing on how we can build and provide library services for data broadly – from collecting data and giving access, to data reference, to data

So today we are going to talk about how to build data services! While there's no one-size-fits-all, there are best practices and tips that can help get institutional and community buy-in, resource and capacity building, and collaborations/emotional labor.

We talked about the breadth of data types out there, and Dorothea Salo did a great job of underscoring that in her article that we read for this week. One part I would tease out here is that the variability of the work done with data (in addition to the variability of data itself) requires that the first thing to do before building *any* type of service is to conduct an **internal and external environmental scan**. In fact, Dorothea mentions that in the conclusion of her article & it's completely on point:

> However, unless we proceed with clear understanding of researchers and their data, as well as our own systems and habits, we will simply trip over ourselves. Research data are too important, and our role in curating them at present too insecure, to allow that to happen.

So let's talk about the ways that we can understand the way that our communities are using, creating, analyzing, storing, and generally dealing with their data! We can do that in a few key ways.

External environment scans involve analyzing opportunities and threats to an institution/service via the national environment, broader socio-economic environment, and industry environment. An example of this might be looking at the national requirements for DMPs and then searching for services that address it in similar institutions to make a case for adding services to your own institution.

> At NYU, my job came out of an external environmental scan. IT and the Libraries does a joint benchmarking endeavor once every few years that looks at things like AI/machine learning, data management, and other areas where we might expand. We compare grant requirements, other services at similar institutions (R1s), and infrastructure capabilities. They found that most R1s had data management as a service, and so my job was created!

An *internal environment scan* is a type of requirements gathering process where you look at the present capabilities/limits of your organization (infrastructure, hardware, personnel, abilities, structure, etc.) and compare the findings to what the organization will **need** in the future to achieve its strategic goals or fill gaps in service offerings (like data services!).

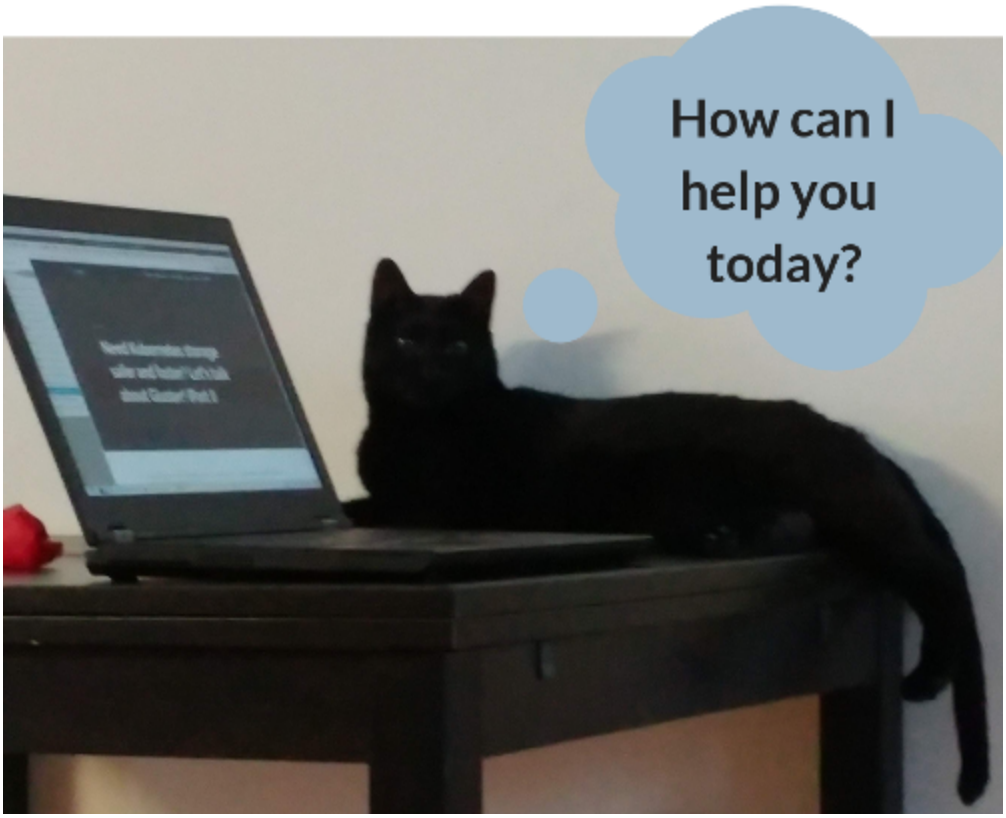


Figure 8.1:

> In my first year, this was pretty much all I did. I met with every single subject liaison to get an idea of the research done in their departments, and how I could best make services and offerings that would fill a capacity gap or be the most immediately impactful. After getting some baseline from my colleagues in the library, I did a lot of outreach to students, staff, and faculty around the university to tell them that this service exists! And we're here to help. That was **a lot** of my first year building the RDM service at NYU.

Observing the internal organizational environment includes employee interaction with other employees/staff/faculty/students/curators/etc., interaction with management, manager interaction with other managers, inter-unit collaboration, access to resources, service awareness, organizational structure, operational potential, and sometimes even doing some human-computer interaction stuff. Is the current research infrastructure usable? Watching students try to go through their giant nest of files on the lab server was a great way to find new avenues to build in teachable moments, or build out services that could help them.

Scanning Type	General Characteristics	Pros	Cons
Ad-hoc	usually in response to a crisis usually not in-depth analysis findings are more short-term	quicker turnaround of results lower commitment of resources over time	data is more superficial results are less generalization more reactive than proactive, can't really do any forward-thinking with this method always catching up, never ahead
Periodic	tied to a planning cycle (e.g. every 5 years to align with strategic planning) usually in-depth analysis findings usually last another 5 - 10 years	predictable frequency lets us plan a budget easier information is typically very timely planning is more proactive	reaction to unforeseen environmental changes may require an additional ad-hoc scan planning response to changes is more reactive than proactive
Continuous	structured, very in-depth data collection/analysis dedicated staff to doing this data collection is very comprehensive	dedicated staff means more data from more sources provides folks in power/decision-makers with more information plans can be adapted/changed/stopped more proactively	requires a lot of money/resources/institutional buy-in

A lot of the time, folks use surveys, interviews, and focus groups for internal scans, and surveys, literature/document, and calling/emailing your colleagues at other institutions for external scans. There is a great existing tool to facilitate internal scans, and some public iterations/outcomes from this tool that can serve as a great dataset for an external review (doc analysis):

- Digital Curation Profiles: "A Data Curation Profile is essentially the 'story' of a data set or collection, describing its origin and life cycle within a research project. The Profile and its associated Toolkit grew out of an inquiry into the changing environment of scholarly communication, especially the possibility of researchers providing access to data much further upstream than previously imagined: If researchers are interested in sharing or required to provide access to data sets or collections, what does that mean for the data, for researchers, and for librarians?"
  - I use the underlying questions as the basis for a lot of interviews that I did with faculty – so I didn't make a profile with it, but I used the same questions to get at the heart of the research process when conducting internal environmental scans
  - see article: "Using Data Curation Profiles to Design the Datastar Dataset Registry"

Once you have conducted your internal and external scanning, you are ready to build your service! What I tell folks who are just starting out, or maybe don't have the resources to start building a service that they can

devote all their time to (I know many librarians who part-time can maybe fit in data service activities into their schedule because of resource issues!), is to just start by reviewing data management plans. Patrons email me their DMPs, and I mark them up and send them back. Sometimes we have an accompanying phone call or meeting depending on the scale and type of study (e.g. the data needs extra security, or it's a 3-institution collaboration). I basically go over the plans and make sure they are hitting the best practices – are the formats open? Do they have a plan for backups and data sharing?

> Recently, the federal agencies have been sending back grants to PIs because of an insufficient data management plan. In fact, this happened to a number of researchers last year who I provide the service for! This is something I've noted in my annual service review (which includes other facets of environmental scanning and experiences with patrons) that goes to the higher-ups in my library.

Growing services often involve reaching out to the larger community for collaboration and general outreach about potential new service offerings (e.g. pilots, preliminary workshops, user testing, etc.). One other point that we discussed in our questions portion of class and one thing that we saw across readings was collaboration and relationship building with other units in an institution, such as:

- IT
- Institutional Review Board (review human/animal subjects research)
- Office of Sponsored Programs (grant admin folks)
- Faculty & Student Senate (if at university)
- Labs (e.g. SMaPP)
- Centers (e.g. CUSP & CDS)

Goben, Zilinski, Briney (2016) in particular outline cross-unit collaboration for data services as a good way to build services from the ground-up:

> While a skilled librarian may be able to deliver very specialized services, it is often more sustainable to have several individuals providing at least low-level data support. A group approach also enables data support to be a normalized part of the total library offerings. Additionally, partnering with other campus offices could allow for resource sharing of both personnel and budget.

While these collaborations are incredibly fruitful when done with mutual boundaries, respect, and understanding, this is a lot of emotional labor. It's very hard work to pursue and build relationships, to manage conflicting ideas. There are loads of cultural barriers to new data services in libraries. Some things I've heard from researchers at many different institutions:

- Isn't this an IT thing?
- How do you know anything about data?
- You can't possibly understand *my* data and *my* research!

These cross-unit collaborations can also help get your feet in the door of many who would share these biases.

No matter what route you take in gathering resources and staffing, as the need is continually assessed and the services grow, you can provide data services through a number of avenues within your institution. This image explains the four most common I've seen:

Coates also mentions a few key services in her article not seen in the diagram above (but could be classified in it!):

- data reference and guides
- institutional repositories (IR)
- purchasing, acquisition, and licensing data
- data citation
- metadata and standards support
- embedded librarians as data managers
- working to influence/create policies around data within institution & externally

Figure 8.2: Most common types of data services

- promoting open data

I can't say enough, that it's all about *gradual* ramping up of services in practice. Ideally, you'd have all these services, and research was being well-managed and preserved from the very start. This just is never grounded in reality. You need to do community building, gather resources for infrastructure, hire people (one person cannot do this job for a whole campus), and you'll likely try a lot of things that fail at first.

> This "Build-Measure-Learn" (*Lean Startup Method*, Eric Ries 2014) approach comes from the lean startup movement and emphasizes continuous innovation. The ideais that small, early failures produce a better product in the end. Adopting this iterative and flexible approach can help data services remain relevant in a rapidly changing research environment. – Coates (2014)

8.2 Homework

Write 250 - 500 words on one of the readings, giving your thoughts and reactions. Please submit it here: https://cloud.vickysteeves.com/index.php/s/Xttr3yxMGkxBNJb. The password follows our class convention.

Chapter 9

Data reference

Agenda for today's class:

- 6:30 - 6:45: please fill out the unit 1 self-assessment –> https://forms.gle/yBH7R64sFjcpAZRy5 and the course check-in –> https://forms.gle/nZpC7aGNP3n7hXTF8
- 6:45 - 8:00: first project check-in
- 8:00 - 8:15: break
- 8:15 - 9:00: discussion
- 9:00 - 9:20: brief lecture & activities

9.1 Discussion Questions

Owen, our facilitator this week, prepared the following questions for in-class discussion:

General discussion: How did people feel about this week's readings? Were you surprised by anything you read? Did you disagree with anything you read? Was there one reading you preferred and why?

- Of the different data-related reference interactions laid out in this week's readings, which, if any, do you presently feel able to perform? Which do you feel most and least confident about?
- How much overlap is there between the research data management support explored by Goben, Zilinski & Briney, the "data interview" as laid out by Carlson & Witt, and the data reference interview discussed by both Partlo and Smith et al?
  - Should the same librarians be involved in some or all of these patron interactions?
  - Do you see a clear flow of data through library services? Why or why not?
- In what ways can the pedagogical data reference interview be modified in virtual reference interactions?
- What web resources can libraries make readily available that will streamline these interactions?

9.2 Lecture

This week we are going to take a look at how we can adapt the reference interview to draw out patrons' needs for data, research, tools, and code. All these, to me, fall under the overarching category of 'data reference'.

I think the one thing that each article comes back to and that I'd like to just underscore for this class is *collaboration* to provide comprehensive and holistic data reference services.

Figure 9.1:

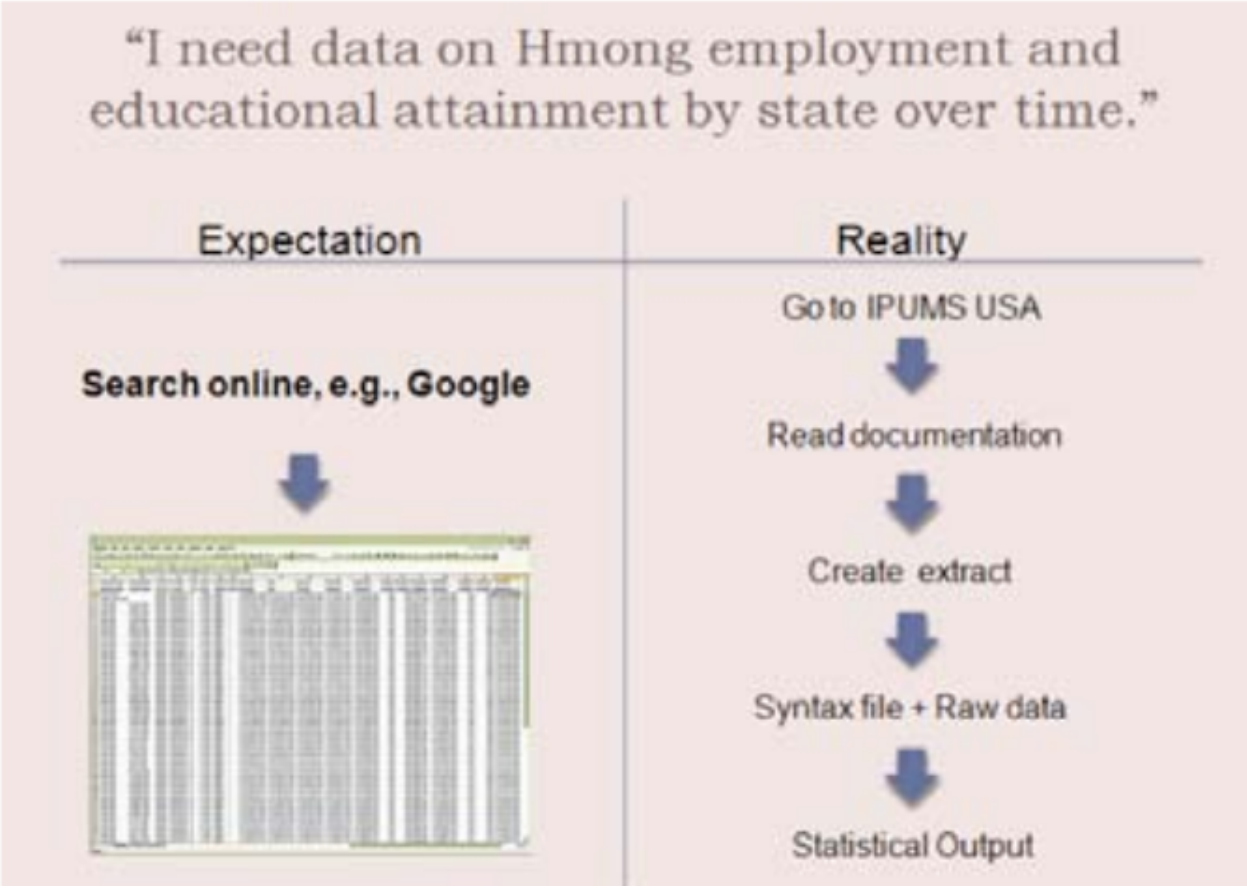


Figure 9.2:

> Data librarians on the one hand and general and subject reference librarians on the other bring distinct sets of knowledge and experience to bear on the challenges of 'assessing the user's needs,' which can be a rich point of collaboration and referral between them. - Partlo, pg. 6

I continually bring this up as a part of my sort of "lived experience" part of these lectures, but I would not be able to do my job, the job of a data librarian, as well without the relationships I've built with subject selectors and reference staff. Sometimes, I'm enough – I can find that financial information or at least point them to the LibGuide on our Virtual Business Library. Other times, I don't even understand the words in the queries coming to me to help them find financial information – and in that case, I kick it off to the business/econ librarian. Likewise, when the business/econ librarian needs help with things like API access for researchers who want to bulk download a large amount of data, they come to me to help navigate that process.

This is the ideal situation. Seamless hand-off (e.g. not forcing the patron to fill out *another* appointment form to get to the right place), offers to do joint consultations (which I do a lot! I often sit with the patron and their subject liaison to provide a full pic of information for them), and **managing expectations**. A lot of folks assume there *must* be data about *everything*. But there's not – and there's even more broken links, government whiteouts, and researchers who die without ever giving their data to an archive.

Partlo totally got it right with this image in their article:

The other big expectation is that the data will come *in exactly the right format to just start working with it* which is just….never correct. Finding data is often the easy part compared to processing & analyzing it! Often times, a consultation on finding the right dataset will end with a recommendation they see a software

specialist. For example: a request like this comes in often – "I want data about the total rainfall in NYC so I can analyze it with R" – this ends up starting with a data finding consult and ending with an R consult.

The socialization of students in particular to the data-heavy environment in which domains want to operate in is a part of our job as data librarians. In practice this means really explicitly explaining our judgements/criteria when it comes to data (a third of the most common queries at NYU, as told by Smith et al!), what goes into working with data, and getting them to sit and look with us – explaining the process instead of just delivering the data.

And this brings us to the idea of embedding pedagogy into the data reference interview. This is actually something that I see most folks do when they answer questions on finding or interacting with data, without calling it instruction or pedagogy. As a practictioner, it's really awesome to see it described and called out this way in the professional lit (part of the reason I assigned the article!). A lot of times, patrons are told to "find data" related to an assigment (students are the highest users of data services, typically) or have some vague idea of what they'd like to prove with data (never minding if the data actually *proves* what they want it to or not). The reference interview is a space to interject some critical thinking around data-informed thinking/analysis, managing expectations (really, one of the biggest parts), and socializing folks into their discipline.

Another Partlo quote comes to mind (pg. 8):

> In fact, when they come to us, they may have never encounted data 'in the wild' before [...].

And this is very much the experience of folks providing data reference – their patrons have a vague idea and they need the questions from us/someone to refine.

CHALLENGE: Find me the database of datasets

Pair up!

Where does NYU libraries keep their list of databases that contain data? Start here: https://library.nyu.edu/ and tell me when you get to it.

Do the same for Pratt: http://library.pratt.edu. And let me know when you find it.

When I do sessions on finding data, or get a query from a patron, I try to guide them with this process:

1. Form your research question, and ID the following:
   - Who (population)
   - What (subject, discipline)
   - Where (location)
   - When (snapshot, longitudinal, etc.)
2. Think about who might have collected the data:
   - nonprofit/nongovernmental organizations
   - government
   - academic researcher
   - private business/industry group
3. Look for publications, news articles, etc. that cite the dataset
4. Once you know what you want, look to see:
   - if it could logically exist (sometimes, it doesn't!)
   - does the institution subscribe to it? Most library instruction sessions go through this in a lot of detail.
   - can you request it from the creator?
   - is it openly available/licensed under a way that you can reasonably use it?

On the flip side, combining criteria from the Partlo article as well as the Carlson & Witt article, the data refence interviewer might ask some of these questions:

1. What is their research question?

    - Get the context of not only why they need the data, but their workflow, how they might plan to use, which can help inform what resources/types of data you might recommend.
2. Do they need statistics or data?
    - Are they using a particular analysis tool? That might predicate the need for some specialized format the data needs to be in.
3. Get the who (population), what (subject, topic), where (geographic constraints), when (time constraints, longitudinal data, snapshots in time, etc.), and *how much* data they need to do the analysis/work they want.
    - If they don't know, it's a good indication they are trying to explore the subject more (get a bird's eye view) or they haven't refined their research question
    - If they need 1 million pictures of images, that's a very different need and search/delivery strategy than a $<$ 1GB spreadsheet of business/financial information from a certain company.
4. What are they ultimately going to do with the data?
    - Just use it for a term project? Larger captsone?
    - Republish a derivitatve? Use it for a Kaggle contest?

I once had to make a screencapture for students as a way to provide data reference services. Some students needed historical weather data in NYC – specifically, hourly temperature and hourly precipitation data for as far back as we could find. They asked if NYU had this data. These are students from my department that I de-facto liaise with – the Center for Data Science. Part of data science work is analyzing and interpreting large swaths of open data, and a part of their socialization into the field is being able to navigate big open data sources.

I was able to point them to NOAA, which had the data they needed! But the interface was so counterinuitive that the only way I could guide them in making the choices they'd need to, virtually, was literally capturing my process. But what's more is, in my reply to them I was able to:

- give them some reasons why I looked for open data first
- impart some background info on the norms in data science
- provide them resources that matches their use case – large dataset that they'd be analyzing with a script
- show them literally the best way to grab the data they want from both this resource and others like it

I'll just end the lecture on this final quote from the Partlo piece:

> [...] the data reference interview needs to reflect the tension of providing both service and instruction. It is not sufficient for the model to include only the elements of helping a patron to determine thir need and then either getting them to the data or pointing them in the right direction.

If we have time, I thought we'd try out using Partlo's data reference worksheets with each other!

9.3 First check-in!

Just to refresh everyone, the order for this first check-in is:

1. Owen
2. Joanna
3. Elizabeth
4. Mary
5. Paolo
6. Amber

You will have a few minutes after each check-in to provide feedback via this form: https://forms.gle/kQcUUCTadtFMSceD8

Chapter 10

Data literacy & instruction

Agenda for today's class:

- 6:30 - 7:45: discussion
- 7:45 - 8:00: break
- 8:00 - 8:45: lecture
- 8:45 - 9:00: break
- 9:00 - 9:20: jeapordy!!

10.1 Discussion Questions

Amber, our facilitator this week, prepared the following questions for in-class discussion:

1. There are a few opposing views on what information literacy includes throughout the readings. Schield argues that this is due to each discipline's specific needs, while Shorish claims that information literacy does not include visual or digital literacy under its umbrella.

   "Changes in information delivery formats have led to conversations around visual literacy and digital literacy, recognizing that the traditional application of information literacy has not been able to address the nuances of these associated competencies effectively." (Shorish)

```
- Do you agree with the specific characteristics that separate each literary (statistical, digital, visu
- Do you consider the term information literacy to include all of these or is it a separate knowledge ba
```

2. Schield separates the idea of information (and data) consumer and producer and how these expectations in a research student change between undergraduate and graduate levels. He uses phrases such "ethics", "integrity", and "responsibility", for an information producer, these terms are not typically included in all literary competency definitions.

   - Do you believe they should be, or is this mostly specific to data literacy?

3. Rosenblum described 3 mostly positive experiences of collaboration in teaching DH. Can you think of any possible downsides or struggles to implementing such courses?

4. Multiple references are made to the social construct of statistical data and the misconception that students can have about the credibility of numbers, yet, Kellam encourages the use of numerical data as a primary source for young researchers. Thinking back to our undergraduate days, can you remember a time when your professors or librarians discussed numerical data or data literacy?

   - Would you have liked them to?

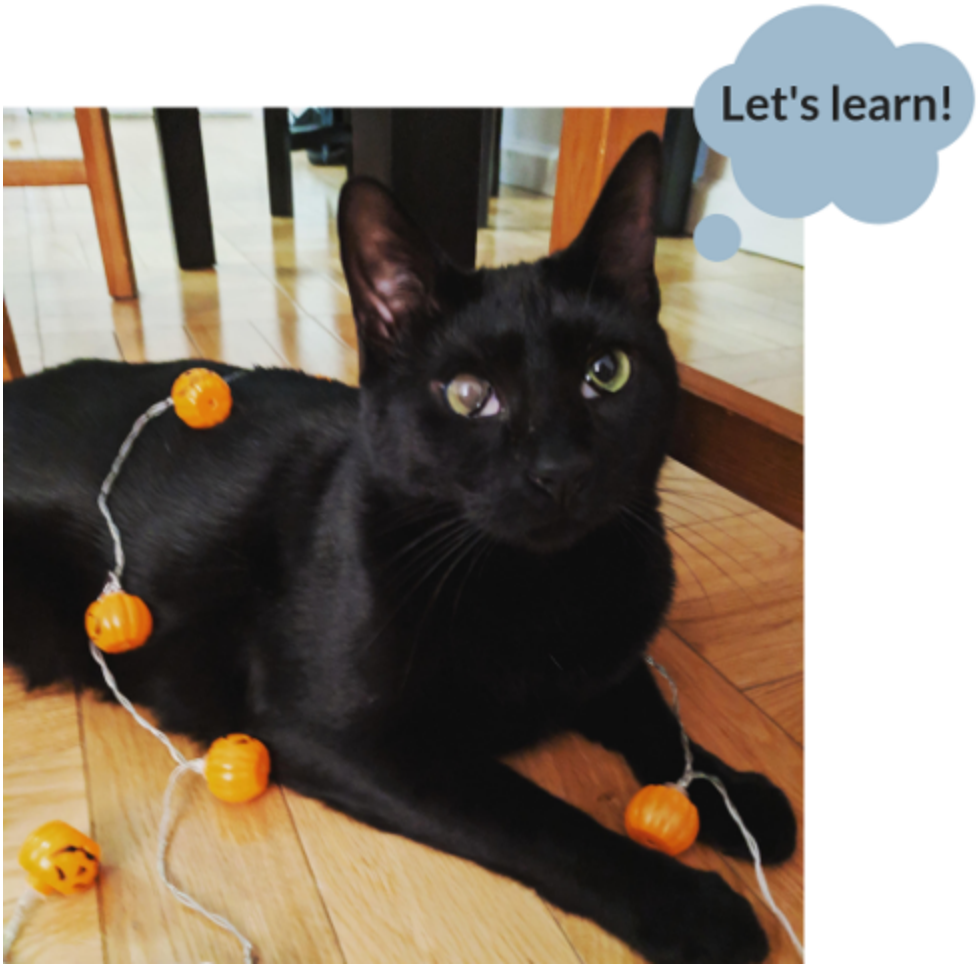


Figure 10.1:

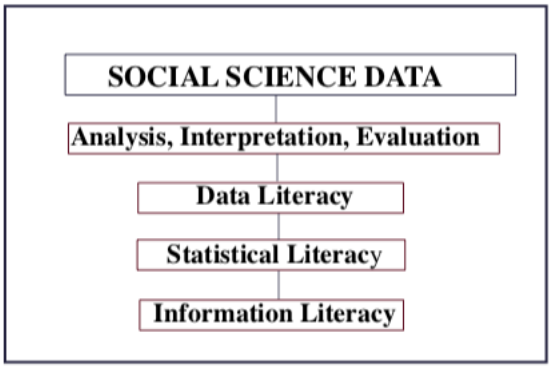


Figure 10.2: Social Science

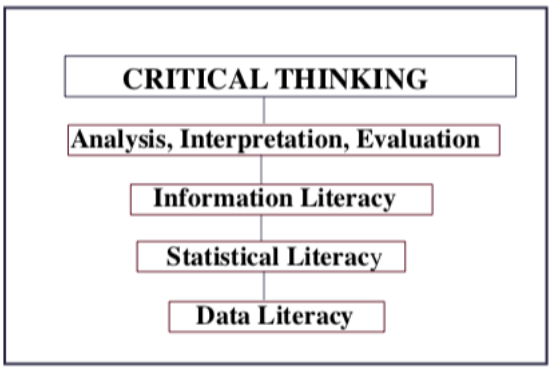


Figure 10.3: Critical Thinking

10.2 Lecture

The biggest request I get is to teach some sort of specialized lesson for different domains, departments, and communities of practice on campus. Instruction is, in my opinion, the crux of data librarian work. The other work I do is held together by my teaching momentum, shaped by what I do and observe in classrooms across campus, and can provide a foundation on which to build and/or grow services.

I am most often called in to do introductory sessions about research data management, or how to find data sources (for them to use). As a result, I've had to create, remix, destroy, and recreate entire models of my teaching because when I first started out as a librarian at NYU, I had no experience really doing heavy-duty data or RDM sessions before for researchers. I did "intro to digital preservation for data" which covered almost all the same topics, but usually for practioners of digital preservation and faculty (curators, at the AMNH!).

Information literacy is "the set of integrated abilities encompassing the reflective discovery of information, the understanding of how information is produced and valued, and the use of information in creating new knowledge and participating ethically in communities of learning." (ALA and ACRL standard) and data information literacy (DIL) adds compentencies with managing, finding, preserving, and documetning data, putting researchers in the role of both consumers and producers of data (very interesting terms to use) in a way that traditional infolit hasn't.

The challenge is always finding the right session to impart skills that are seen as more of an add-on or time-consumer than vitally important (e.g. file naming conventions). Lisa Federer wrote in a blog post for the UC3 about her work at UCLA Louise M. Darling Biomedical Library (she's now at NLM):

> [At] my institution, the researchers and students are very busy, and not likely to commit to a seven-session data literacy program. Nonetheless, it's still important that they learn how to manage, preserve, and share their data, not only because many funders now require it, but also because it's the right thing to do as a member of the scientific community. Thus, my challenge has been to design a one-off session that would be applicable across a variety of scientific (and perhaps even social science) fields.

Additionally, the DataInfoLit IMLS-funded project sought to "build infrastructure in the library community for DIL skills, to have students learn DIL skills appropriate to their disciplinary context, and to develop a robust process for librarians to articulate DIL curricula in their research communities." To that end, they tested a variety of approaches of instruction:

APPROACH	PROS	CONS
Mini-course (Cornell)	Co-teaching the course with faculty from the department increased faculty engagement Course format provided opportunities to practice application of best practices, and the ability to build on prior sessions	Time investment is substantial, both for librarian and faculty collaborator Must have buy-in from university department to offer course, and from library administration to spend librarian time teaching (Many libraries consider teaching a university course a high achievement for a librarian, so this may not be a con but should still be considered)
Online course (Minnesota)	Very scalable While initial time investment may be high, modules can be reused, repurposed, and recombined, increasing the potential impact of the training Online format provides the opportunity for students to reference the materials at the time of need, potentially resulting in improved data practices	May require assistance from an instructional designer, or someone with experience building online content Impact is increased by combining with an in-person session, due date, or some other kind of encouragement. Students tended to forget or put off completing the module until the last minute
One-shot session (Oregon)	Small group setting allows materials to be closely targeted to their specific needs, increasing student awareness of tools, resources, and best practices Because the time investment is small, increased likelihood of getting buy-in from reluctant or busy faculty and graduate students	Not very scalable if you have lots of research groups interested in such targeted training Limited time with students may mean that material covered is not retained as well as it would be if there were more opportunities for activities and repetition
Embedded librarianship (Carlson and Sapp Nelson team from Purdue)	Group setting allows materials to be closely targeted to their specific needs, increasing student awareness of tools, resources, and best practices Ongoing relationship with students and faculty provides multiple opportunities for evaluating student work and providing feedback	Not very scalable as interest increases Time investment is large, since the librarian participates in group meetings and is closely involved in development of tools and resources
Series of workshops (Bracke and Fosmire team from Purdue)	Workshops allow materials to be closely targeted to developing specific skills and best practices Clear expectations such as a checklist make it easier for students to see what is required, resulting in increased student compliance	Very specific outcomes (checklist of practices) may result in the students not realizing that the same best practices could apply in other situations Time investment is large, since librarian must develop specific checklist and accompanying instruction May not be easily scaled if other groups want such targeted instruction

Table from http://www.datainfolit.org/dilguide/.

One piece to pull out from this bevy of information is the importance of active learning in any process involving data instruction/literacy. Active learning is a:

> [..] range of teaching strategies which engage students as active participants in their learning during class time with their instructor. Typically, these strategies involve some amount of students working together during class, but may also involve individual work and/or reflection. These teaching approaches range from short, simple activities like journal writing, problem solving and

Figure 10.4:

> paired discussions, to longer, involved activities or pedagogical frameworks like case studies, role plays, and structured team-based learning. https://cei.umn.edu/active-learning

With library instruction especially, active learning is an important way to keep learners engaged and make sure that they centered in the process of instruction. I know we've talked about the Carpentries model of pedagogy in this class before, but I would pull out that organization as a very successful one in terms of bringing data information literacy to a wide array of researchers through active learning.

Their model of basically 1 instructor to every 8 learners, as well as the entirely hands-on curriculum and real-time feedback (with stickies) has really contributed to a collaborative, open model of instruction that has proven effective in imparting data literacies. The Carpentries also gives pre-workshop and post-workshop surveys to gauge effectiveness with different audiences and adjust curricula accordingly.

I want to bring together the active learning and DIL to talk about play in the classroom. Let's look at this great poster from Chealsye Bowley: https://osf.io/preprints/lissa/d5mkj/ and this great presentation from Andrew Walsh: https://osf.io/preprints/lissa/abg4r/

Walsh's presentation focuses on the theory behind play-as-instruction, positing that play/playfulness is a political act and in academia, and how play "can give permission for your learners to challenge their understanding of a topic and gain deeper understanding, helping to create a transformative learning environment as opposed to one that concentrates on the echoing of facts and basic skills." Walsh says in the abstract:

> This approach particularly matches the teaching of skills to improve information literacy, as information literacy itself is a socially constructed concept (Lloyd, 2005, Elmborg, 2006). Playful Information Literacy teaching can help critical interactions with information in a way that encourages and enables action from your learners, within and without your classroom, in a "safe" and creative environment (Francis, 2009; Gauntlett, 2011). It encourages a playful approach to formal and informal learning, important for critical social engagement with political issues (Koh, 2014) as well as increased creativity (Chang, Hsu & Chen, 2011).

Meanwhile, Bowley's poster an example and use case of playing to teach data literacy and data management

best practices. She writes:

> Teaching research data management and data literacy can be a challenge. How can one know if the information is being retained and will be applied? Using game techniques and role playing can give the presenter immediate feedback on if the information regarding data management and/or data literacy is being retained, and allow students to immediately apply the information to increase their chances of retaining and using the information when conducting research.

I find it also really effective to use play when teaching or trying to impart best practices that are…less exciting than learning R, or Python, or OpenRefine. Like file naming. Or data storage, publishing, documentation, etc. Sometimes play can be a way to make these less-than-dramatic topics engaging and interesting to learners.

So, we are going to do our own jeapordy board!

Chapter 11

Data collection services

Agenda for today's class:

- 6:30 - 6:40: do the unit 2 self-assessment: https://forms.gle/JZdT4E6iFvfM3FGk8
- 6:40 - 7:40: discussion questions
- 7:40 - 8:50: break
- 8:25 - 9:20: lecture/activities

11.1 Discussion Questions

Paolo, our facilitator for this week, has prepared the following questions for in-class discussion:

This week's readings offered some solutions, mostly best practices for issues that librarians may encounter when developing or managing data collections.

1. One of the best practices posed in "Data Basics" (Geraci et al) was to consider the quality and reputation of a vendor when deciding whether or not to purchase their services (pg 159). A sound idea in theory but is this realistic in practice when the market is dominated by a small number of vendors?
    - The University of California demonstrated a trend reversal when it chose to end its contract with Elsevier earlier this year. How can a smaller system with fewer resources and less clout replicate the UC system's decision? Is this possible?
2. "Data Basics" (Geraci et al) references a scenario in which a researcher may choose to deposit or donate data to an institutional repository. The reading says:

    It is essential when acquiring such data to ensure that the data are adequately documented, that the privacy of respondents has been protected, and that access restrictions, if any, are clear and consistent with your service and collection policies (pg 164).
3. I haven't taken the Collection Development class at Pratt. If you have, does that class discuss policy when developing collections? Is data a part of that conversation? Should it be?
4. The "Data Basics" reading references another scenario a librarian may encounter when considering data sources (Cost, pg 165): a lower price or free acquisition from a government agency may come at the cost of time and resources to convert, format, and clean data to be ready for analysis. On the other hand, a more expensive acquisition from a private vendor may come ready to be analyzed, but the data may not be preservable, or even usable in a few years. You can boil this down to an issue of time vs. money. Do you have a preference in this scenario? Why? What would change your mind?
5. The pilot program in "Collecting Small Data" (Hogenboom et al) provided a few takeaways:

Figure 11.1:

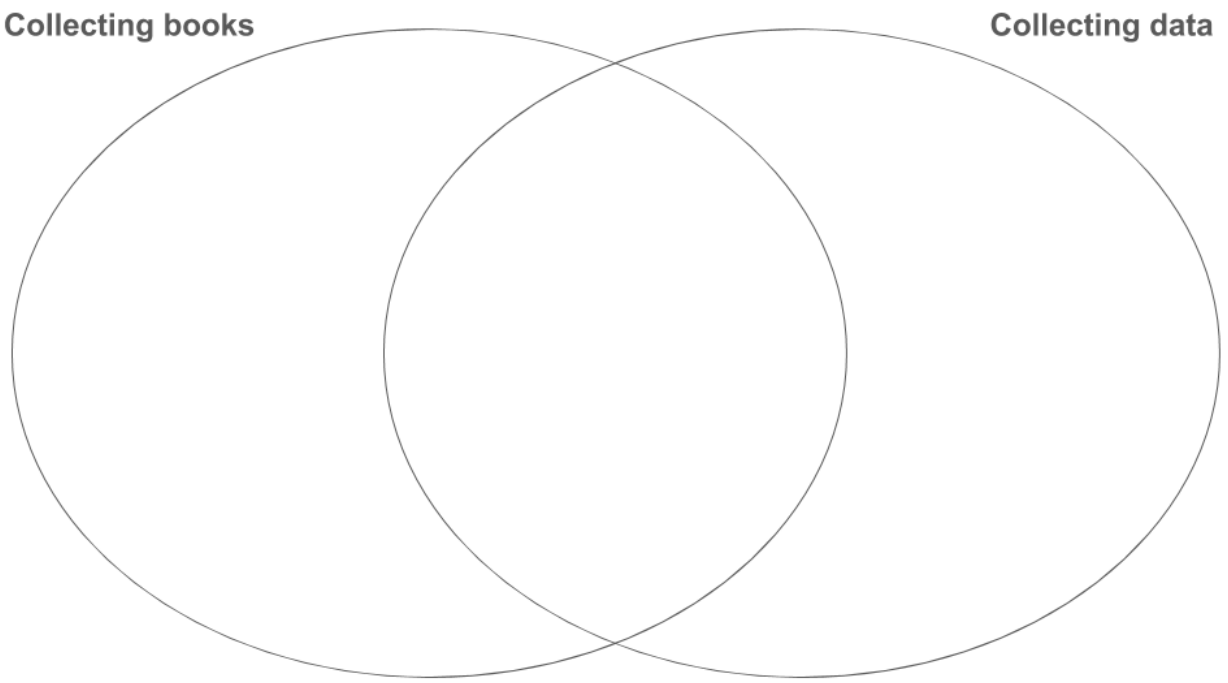


Figure 11.2:

> As previously noted, a couple of applicants requested data already in the library's collection. Another applicant requested support for processing data from a local government agency. Library personnel referred them to a service on campus that helps researchers prepare data for analysis (pg 6)

```
+ These are positives, kind of, but they confirm what we already know-the library has access to info tha
```

11.2 Guest Lecturer

Today I have the virtual pleasure to introduce Genevieve Milliken as the guest lecturer for this class section. Genevieve graduated from Pratt SLIS in 2019 with an MLIS with an advanced certificate in Digital Humanities. She is currently the Research Scientist for LIS on the "Investigating and Archiving the Scholarly Git Experience" (IASGE) at NYU (PI, Vicky Steeves (me! ha!)). She has also served as a graduate assistant for "Exploring Data Worlds at the Public Library," a research project on Youth Data Literacy designed by Dr. Leanne Bowler. She was the 2018-2019 NYARC Web-Archiving Fellow at the Frick Art Reference Library. In addition to her MLIS, she also has a Masters of Art History from George State University.

You can (and should, in my opinion!) follow Genevieve on Twitter @gen_milliken.

11.3 Vicky's Lecture

First, an activity! Let's fill in this Venn diagram and discuss the similarities/differences with collections:

OK, now let's set some terms and discuss the main differences between archives and libraries as it relates to data.

Data archives:

Matrix of Some Collection and Collection Service Alternatives

Level of service	Select	Acquire Access	Acquire data	Organize	Preserve	Service
1 (low)						x
2	x			(x)		
3	x	x				
4	x	x		x		
5	x	x		x		x
6	x	x		x	(x)	x
7 (high)	x	x	x	x	x	x

1. Provide service help for users who have found data on the Internet.
2. Select data on the Internet and provide minimum organization of your selections by presenting them on a web page, with minimum cataloging, or by relying on the organizational and finding aids providing by others (e.g., Nesstar tools, Google, etc).
3. Select commercial data services and provide access to your users through subscription, client/server software, IP recognition of their machines by service, etc.
4. Same as 3, but provide additional organization of the services you've selected such as OPAC records, web pages, special finding aids.
5. Same as 4, but add reference service to help users locate needed data, use the commercial service, and download data.
6. Select and preserve data through a membership organization such as ICPSR. Provide access through ICPSR Direct or similar facilities. Organize by adding records for ICPSR studies to your OPAC. Provide service of helping users locate, download, and use data.
7. Build a local collection of data by selecting and acquiring data. Organize by adding records to OPAC and providing specialized search services built on DDI variable level information. Provide online access through web technologies such as Nesstar. Provide reference help in identifying, locating, using data. Provide data discovery and subsetting tools.

Figure 11.3: From Geraci, D., Humphrey, C., & Jacobs, J. (2012). Data Basics: An Introductory Text.

> "The primary functions of data archives are to gather, preserve, and provide access to original research data. As a rule of thumb, the focus of collections in data archives tends to be specific in geographic scope and either general or topical in subject….Some data archives are directly affiliated with a research center engaged in the collection of original data (e.g., UC DATA at University of California, Berkeley's Survey Research Center)." From Geraci, D., Humphrey, C., & Jacobs, J. (2012). Data Basics: An Introductory Text

Data libraries:

> "The primary function of data libraries is to support established research communities interested in secondary data analysis by providing access to and assistance with data. Data libraries are most apt to be located in university libraries, computing centers, or research institutes. Sometimes departments or units within a campus or among several campuses form partnerships to create a complement of data services." From Geraci, D., Humphrey, C., & Jacobs, J. (2012). Data Basics: An Introductory Text

The main point here is that data in archives are meant to be preserved *indefinitely* and data in libraries are not considered a "copy of last report" and might be discarded if no one needs it anymore. Collection, then, for each of these institutions will vary widely.

Q for you: how do these definitions relate to the propagation of data repositories that we've seen in recent years? E.g. Zenodo and the others we've looked at in this class.

A collection service in general, but especially for data, needs to be flexible. Every research project, dataset, software, has the potential to need unique handling, and a collection policy in a library should allow for that.

In this matrix, levels of services around data collection are ranked low to high based on the responsibility taken for each of six possible activities:

So let's think about what each of those collection services might entail:

Service	Competencies	Work
Data selection: IDing user needs, locating data to fit needs, and choosing delivery methods	Not very different than the skills that liaisons or cataloger folks have already. Familiarity with methods could be useful.	A combo of identifying data and the options for accessing that data (e.g. do I buy it from a vendor and have to serve it in my own library, or can I buy a membership to something like ICPSR?).
Acquiring data: physically getting files or getting access to remotely stored data, sometimes from users themselves	Not super different either: looking at license agreements, defining 'acceptable' access to data	Depending on how users are able to access the data, it could require supporting specific software or computers
Building a local collection: getting data from community, getting data that isn't available remotely, ensuring long-term pres/access, provide subsets/make data easier to use	Can leverage existing folks like metadata/catalogers, liaisons. Need to understand data prep and distribution, including data documentation and software that may be needed to read in data	Need service to help users ID the data to meet their needs. Also need service around licensing, getting data in shape (potentially)
Organizing collection: let users quickly find & use data we've selected and collected	Cataloging! For data! Building specialized indexes/databases potentially.	Adding metadata for data sources to OPAC/catalog, potentially including metadata from outside sources (like ICPSR)

The last piece of collection service is data librarianship – which I didn't include in the table because it's this whole class ;)

While thinking through these service models, another important consideration is whether or not collection activities are *proactive* or *reactive*. A reactive collection policy would authorize acquiring a data file only when a user requires it – the collection builds in reaction to the users' specific requests. A proactive collection policy would authorize acquiring data in anticipation of future needs of users, much the way a traditional library collects books as they are published in anticipation of demand for those books.

The distinction isn't very clear between these two ideas – "just in case" or "just in time" – for instance, collecting longitudinal study releases as they come out (e.g. census data). It's also obviously resource-based. Sometimes you can't afford, either because of money or time, to be proactive, even though I think most people would want to. Data collection needs to be relevant, despite all else.

A lot of the activities of selection are evaluating data in ways similar to RDM best practices we talked about – what format is the data in, is there sufficient documentation, how do users access it, do you need the whole dataset or a subset, what's the license, etc.

Geraci, D., Humphrey, C., & Jacobs, J. in their book Data Basics: An Introductory Text outline this checklist for data collection services:

- Make sure that the data acquired meets your patrons' needs.
- Make sure the data acquired fits with your service plan.
- Make sure the data acquired fits your collection policy.
- Make sure the data acquired is technically usable in your computing environment.
- Your service plan and collection policy should include guidelines that balance access and privacy and these considerations should be important in your decision-making.
- Weigh the trade-offs you have to make among different choices.

Some Qs for You:

1. Do data collection services belong in data services, with liaisons, some other department in the library?
2. What are some of the good reasons a data librarian might choose to add data to a local collection? What are some of the bad reasons a data librarian might choose to add data to a local collection?
3. What classes at Pratt, besides LIS 628, can you think of being applicable to elements outlined in these subsections (e.g. classes that a data archivist should take)?
4. What would the #critlib say about developing and managing data collections? What would Elsevier say?
    - What kind of data would #critlib or Elsevier want you to include in a collection and why?
    - How would they want you to acquire that data and why?
    - How would they want you to make that data find-able and distribute it, and why?

So, now we know a bit about data collection and what it entails for librarians. The next step is writing a policy statement for data collections, like we do for any collection! We like these statements because:

- Outline who to ask for what (a problem in most libraries!) – who is selecting the data to add to a collection?
- Funding for the collection – personnel and data purchases – we want to be transparent!
- Provide a framework for the scope and methodology/thinking behind the collection, and the access to the collection
- Specifying limitations/boundaries on services and collections

Let's look at a real data collection policy: https://www.library.virginia.edu/services/data-purchase-program/ & https://www.library.virginia.edu/services/data-purchase-program/data-collection-development-policy/

11.4 Lab/Homework

11.4.1 In class

Group quiz!

11.4.2 Outside class

You have no out-of-class homework – work on you second check-in! Please submit the materials for the second check-in via this link: https://cloud.vickysteeves.com/index.php/s/gDji5Ha2pgw77M8 before 11:55pm on Thursday the 14th. Please also send me your presentation materials (if you plan on using any) so we can minimize time setting up between presentations.

Chapter 12

Data sharing, publishing, access, & preservation

Agenda for today's class:

- 6:30 - 8:00: second check-in presentations
- 8:00 - 8:15: break
- 8:15 - 9:00: discussion
- 9:00 - 9:20: highlights of the lecture

12.1 Second check-in!

Everyone goes 15min + 5min for questions. Order:

1. Amber
2. Mary
3. Paolo
4. Owen
5. Elizabeth
6. Joanna

12.2 Discussion Questions

Paolo, our facilitator this week, prepared the following questions for in-class discussion:

1. "Changes in Data Sharing and Data Reuse Practices.." (Dorsett, et al) presents some interesting researcher perspectives that splinter into variables such as age and geography. The report says,

   > While younger researchers are more concerned about lack of access, they are the least involved in data sharing. In fact, data-sharing behavior actually increases significantly with each older age group. So while younger researchers express a higher interest in sharing their own data, they are also more interested in requiring others to get permission to access their data.

Is this surprising to you? Why do you think this is?

2. Another variable the report cites is geography. The authors report that Asian researchers are more concerned about the lack of access to data than their western counterparts. The report continues to say that Asian and African researchers are more concerned with restricting access to data. This seems

Figure 12.1:

dissonant, but it's an echo of the perspective favored by younger researchers in the last question. Why do you think this is the case?

3. None of the researchers (measured across all demographics) expressed satisfaction regarding data management best practices, specifically searchability, long-term data storage processes, and tools for preparing documentation. About half (47.7%) reported using any metadata at all. These are items that fall squarely within our wheelhouse as librarians. With all the time and money in the world, how would you address these pressure points?

4. Several of the readings this week (and most of the semester, tbh) used words like synergy, cooperation, and collaboration across the research lifecycle to cement the best practices that the scientific community espouses. Any thoughts about how the library can be a proactive resource in this ecosystem? Outreach? What kind? More libguides?

12.3 Lecture

This week we'll talk about the ways in which folks share data. This largely depends on their field, but disciplinary research repositories have emerged for most domains, and archives for some. We didn't get a chance to full explore it this week, so maybe we can take a second here and unpack the difference between repositories and archives, both functionally and theoretically. Here's a reminder of the key differences from last week:

> The main point here is that data in archives are meant to be preserved *indefinitely* and data in libraries are not considered a "copy of last report" and might be discarded if no one needs it anymore. Collection, then, for each of these institutions will vary widely.

So, the Q for you: How do the definitions of data archive and data library relate to the propagation of data repositories that we've seen in recent years? E.g. Zenodo and the others we've looked at in this class.

So let's look at some of Tenopir et al.'s data to get a look at their findings on the disciplinary differences in data sharing (they also look at some demographic information, mostly age, which could also be interesting): https://doi.org/10.1371/journal.pone.0134826.s001

The ideal situation is that folks have been sticking to the data management best practice so that when it comes time for upload into a library/repository/archive, it's not time-intensive at all, just a literal upload.

However, we know that folks don't engage in the best practices all the time, so the first step is usually cleaning and documenting data for the deposit process. Once we have a README or codebook, or other required documentation (sometimes more is required by the repository), we put our data into an open format and find a good place (either a repository or data journal – domain-specific or journal-specific data review articles, usually peer-reviewed) to deposit that hits a few criteria:

1. Gives a DOI for the dataset, ideally versioned (like Zenodo: https://www.zenodo.org/record/1489112 && https://www.zenodo.org/record/1489121)
2. Widely indexed in places where you'd want stuff indexed (e.g. google scholar)
3. Searchable/browsable interface (some of these repositories do not meet this criteria, so it's worth mentioning even though it seems 101)
4. Data can be viewed and downloaded easily enough (no paywall or log-in required)

Ideally, anyone who discovers the data and wants to use it must go to this repository, download it, and cite their use. There are some places that aggregate information about all the repositories and make it available to help folks choose:

- https://repositoryfinder.datacite.org/
- https://www.re3data.org/
- http://oad.simmons.edu/oadwiki/Data_repositories

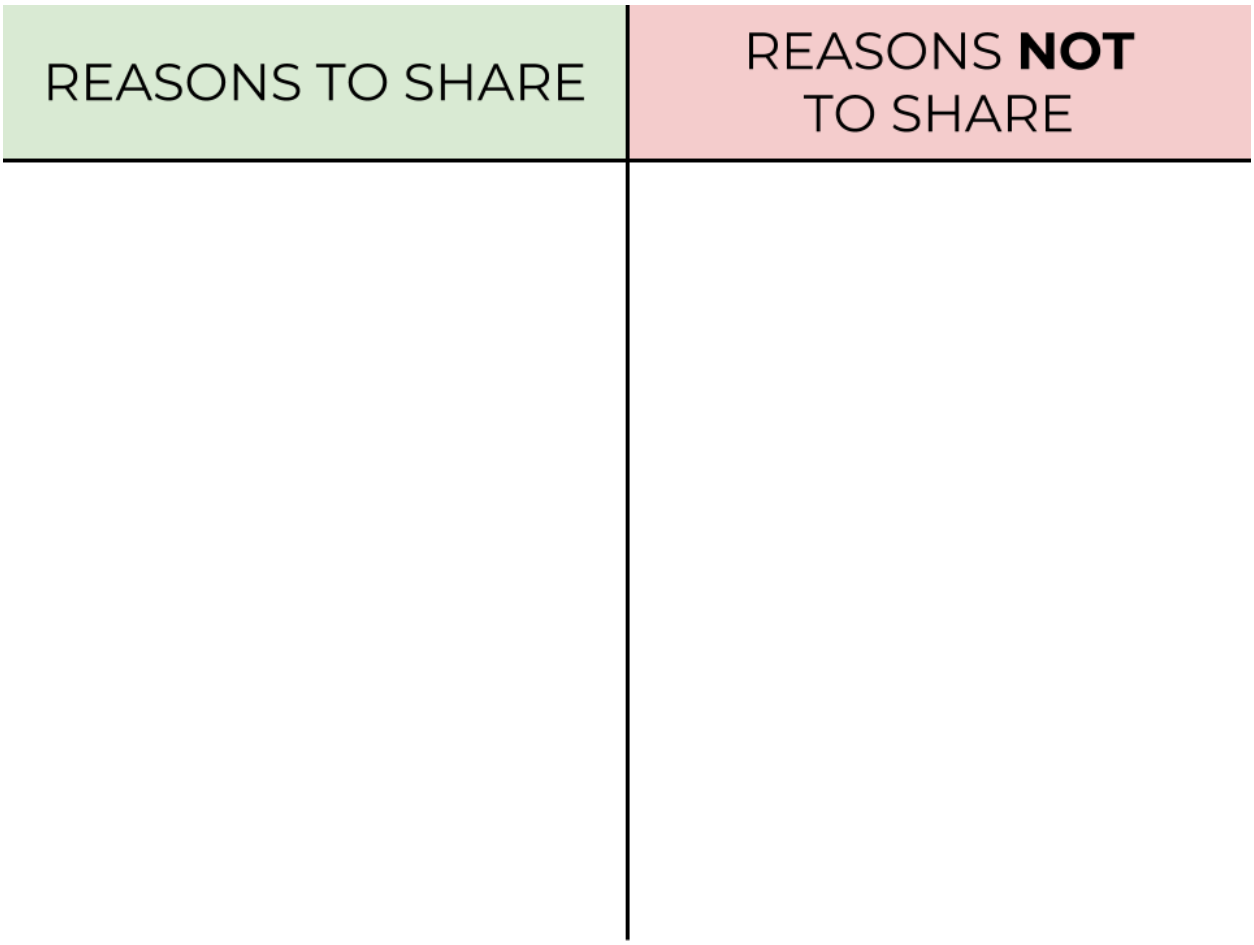


Figure 12.2:

Some general repositories folks tend to use are the Open Science Framework, DataONE node, Dryad, and Dataverse hubs.

Increasingly, folks are doing this for the proliferation of research *code* being created and shared openly too! The most popular way I've seen this work is the GitHub <> Zenodo link but there are excellent software journals gaining popularity, such as the Journal of Open Source Software (require a brief report on your software and a link to a GitHub repo) and the Journal of Open Research Software (require a manuscript and a link to any repositories), both peer-reviewed and open access. The steps are as follows:

1. Log into Zenodo using your GitHub account.
2. Zenodo will redirect you back to GitHub because permissions. Grant them.
3. Pick the public repository you want to publish.
4. Check to make sure there is now a Zenodo webhook in the repo you chose.
5. Release a version!
6. Add a brief description in Zenodo, and that version gets a DOI and a badge!

See an example here: https://www.zenodo.org/record/1488364

Let's pivot from the **how** to the **why**. How is simple compared to why :) Let's draw it out:

I thought we would watch these three films about from three different stakeholders (well, I say there's really two stakeholders but maybe y'all will have differeing opinions) about **why** folks might want to share data:

1. https://www.youtube.com/watch?v=-PcwOWiHcP8

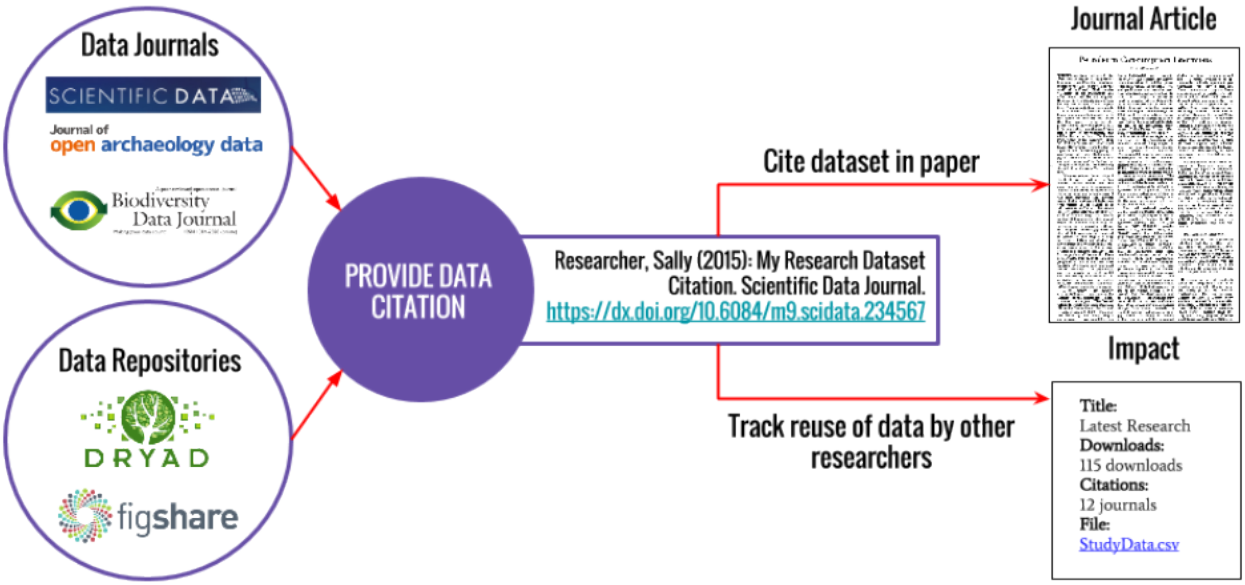


Figure 12.3:

2. https://www.youtube.com/watch?v=8qRLgQa1wT4
3. https://www.youtube.com/watch?v=94qM6tnDtcQ

Let's discuss the differences in content/emphasis between these three videos and how we might refine our answers to the graph we filled in above.

Data and code citations are the very important flip-side to sharing data. Academia has a reputational economy, and so a major incentive for sharing data is the prospect of more citations – this preprint posits a relationship between data sharing and increased citations: https://peerj.com/articles/175/.

Data and code should be cited within our work for the same reasons journal articles are cited: to give credit where credit is due (original author/producer) and to help other researchers find the material.

A data citation includes the typical components of other citations:

- Author or creator: the entity/entities responsible for creating the data
- Date of publication: the date the data was published or otherwise released to the public
- Title: the title of the dataset or a brief description of it if missing a title
- Publisher: entity responsible for hosting the data
- URL or preferably, a DOI

Some optional (but recommended) values include:

- Edition or version
- Date accessed online

These help other researchers find the exact version of the data that you used, helping to enhance the probability that your work will be reproducible.

Just like data, *software should also be cited* when you use it in the course of your work and analyses. **But what software should you cite?**

To start, you would only cite code that's not seemingly universal – e.g. don't cite Microsoft Office, but DO cite scikit-learn, or OpenRefine, or a specific R package like `ggplot2()`!

Citing code is a bit different than citing data or articles, in that there are a few anomalies:

1. *Authors vs. Contributors*: a piece of software might be created by dozens, if not thousands of contributors. Do all of them get cited? No…it's the difference between the maintainer(s) of a project (who is/are currently responsible for it), and the contributors (those who have committed code to the project, or made other contributions). You would cite the maintainers, and possibly also previous maintainers.
2. *Location?? Publisher??*: publisher name/location is similarly difficult. This could be optional, but not when software is produced solely by a specific university or software company. The geographic location is probably irrelevant, unless it's necessary for distinguishing between multiple entities with the same name.

The major citation players even have codified some styles for citing software:

- *Chicago*: Software Name. Location: Publisher/Author, Date. Link if available.
- *MLA*: Authors, Software Name, Place Written: Organization, Date Written. Link if available.
- *APA*: Author Name (first & middle name abbreviated). Title of Software/Code. [Computer software]. Location: Publisher. Link if available.

So, in summation:

- Data sharing is, barring ethical considerations for human subjects, essential for reproducibility, verifying work, and building on each's others work
- People don't share data for lots of reasons – some real, some imagined, and some as the product of fear-mongering
- Citing code and data as much as possible is a great way to normalize and incentivize sharing

Chapter 13

Week 13: Data archives & repositories

Agenda for today's class:

- 6:30 - 7:30: visit from Dan Hickey!
- 7:30 - 7:40: break
- 7:40 - 8:20: discussion
- 8:20 - 8:30: break
- 8:30 - 9:20: lecture

13.1 Guest Lecture: Dan Hickey

Today I'm so happy to have Dan Hickey (he/his) visit us tonight! Dan is the Librarian for Business & Economics at New York University. In his current role, Dan supports the academic and career success of Stern and the Economics Department through collection building (databases, journals, data, etc), research coaching, and teaching. His current research explores the information seeking behavior/needs of MBAs. Before coming to NYU, he was the Assistant Director for Business & Hospitality Research Services at Cornell's Hospitality, Labor & Management Library.

Dan has published on librarians' roles in preparing patrons for salary negotiations, and is here to talk to us today about how to prepare for *own* negotiations, including preparation best practices and workflows.

13.2 Discussion Questions

Mary, our facilitator this week, prepared the following questions for in-class discussion:

1. Wilson writes: "Digital preservation specialists, many of them with information technology rather than information management backgrounds, need to work closely with domain specialists to ensure that the aims and promise of digital preservation are realized for all concerned communities." This and Thiede's DH preservation piece call for collaboration, which has been discussed a lot in this class. With regards to digital preservation, who do you think needs to be involved? What should the librarian's role be? And what are some of the benefits and challenges of these collaborations?
2. Thiede's piece about preservation of DH projects from 2017 includes this question posed by Bethany Nowviskie: "is [digital humanities] about preservation, conservation, and recovery—or about understanding ephemerality and embracing change?" What can librarians do to better situate DH researchers before/during/after their projects for their project's preservation needs, whether that's to maintain it in perpetuity or to accept its fleeting nature?

Figure 13.1:

3. Thiede's piece also touched on national grant funding for such projects with one of the interview subjects: "When asked about the intentions of the National Endowment for the Humanities to incorporate preservation and reusability into funding requirements, the respondent expressed skepticism of the agency's conceptualization of preservation, stating that a reconsideration and reworking of the term's definition was in order." The NEH does currently include a requirement for a DMP for DH project funding applications, and some (but not all) applicants are required to complete a sustainability plan. Should institutions like the NEH specify preservation as part of their funding requirements? Why or why not? If so, how should the funding model be changed to cover preservation?
4. The Kellam & Peter piece looks at the types of institutions that may collect statistics: "government agencies, international and non-governmental organizations, researchers, and other organizations (non-profit and private)." What do you think the moral and ethical responsibilities are, if any, of those institutions to preserve that data in the long-term? If those institutions are unable or unwilling to preserve data in the long-term, to whom does or should that responsibility shift, if anyone?
5. The bottom line from the paper that tried to access data from old studies and found it was, unsurprisingly to us, very hard to do (Vines et al.): "Fortunately, one effective solution is to require that authors share it on a public archive at publication: the data will be preserved in perpetuity, and can no longer be withheld or lost by authors." We've had several discussions about the lack of incentives for authors to share data publicly. What could "requirement" for public data sharing look like, and what structures need to be in place for that to be successful?

13.3 Lecture

I'm going off slides for this one, apologies!

Chapter 14

ENJOY YOUR TIME OFF

Figure 14.1:

Chapter 15

Week 15: Special concerns

Agenda for today's class:

- 6:30 - 7:30: discussion
- 7:30 - 7:45: break
- 7:45 - 8:30: lecture
- 8:30 - end: work on final project, ask me questions!

15.1 Discussion Questions

Joanna, our facilitator this week, prepared the following questions for in-class discussion:

Key themes: ownership, anonymity, privacy, confidentiality, de-identification, ethics

1. In the Cegłowski piece, the author expresses concerns about the ethics of big data, artificial intelligence, and surveillance. He says:

   > ...the very thing the Librarian of Congress objected to in the Patriot Act (Note from Jo: in November 2019, it was extended for three more months! Thoughts?) —the intrusive surveillance—is the bread and butter of online services. Much of the valuable information is collected in ways that would never pass ethical standards in academia, and ways that even the NSA would be legally prohibited from collecting.

and

> I worry about legitimizing a culture of universal surveillance. I am very uneasy to see social scientists working with Facebook.

How would you approach the ethics of data sets collected by the researchers you work with? If the researchers you work with collect data in a way that conflicts with the values of the library (against intrusive surveillance, for example), how should the library handle it?

2. The Shaw and Cloud paper and the Asher and Jahnke paper both emphasize the virtual impossibility of complete anonymity in research data. Shaw and Cloud explain that differentially private database mechanisms will eventually make data widely available and sufficiently accurate for data analysis, but currently this technique is expensive and technically difficult. In reference to ethnographic data, Asher and Jahnke explain that:

   > Given technological advances, removing all conceivably identifying details might still be insufficient to guarantee that transcripts are rendered anonymous. In fact, anonymizing transcripts may soon no longer be technically possible.

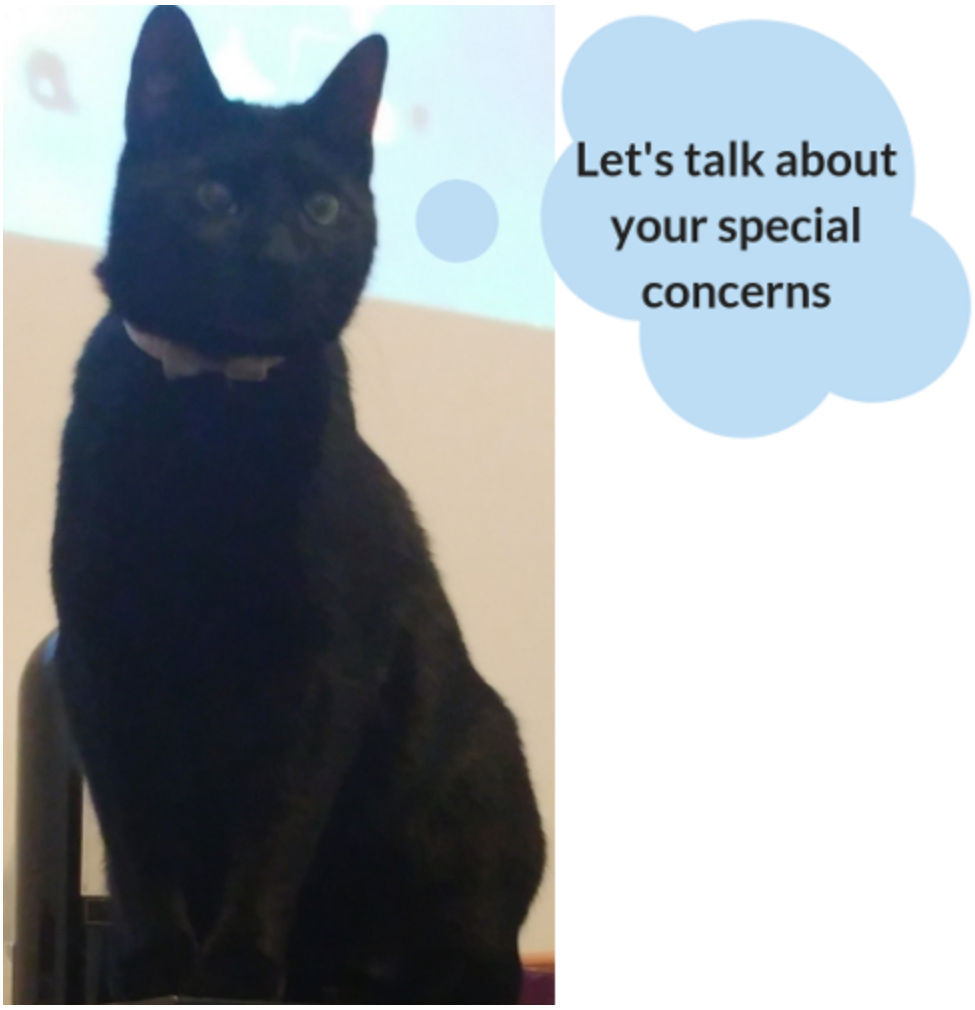


Figure 15.1:

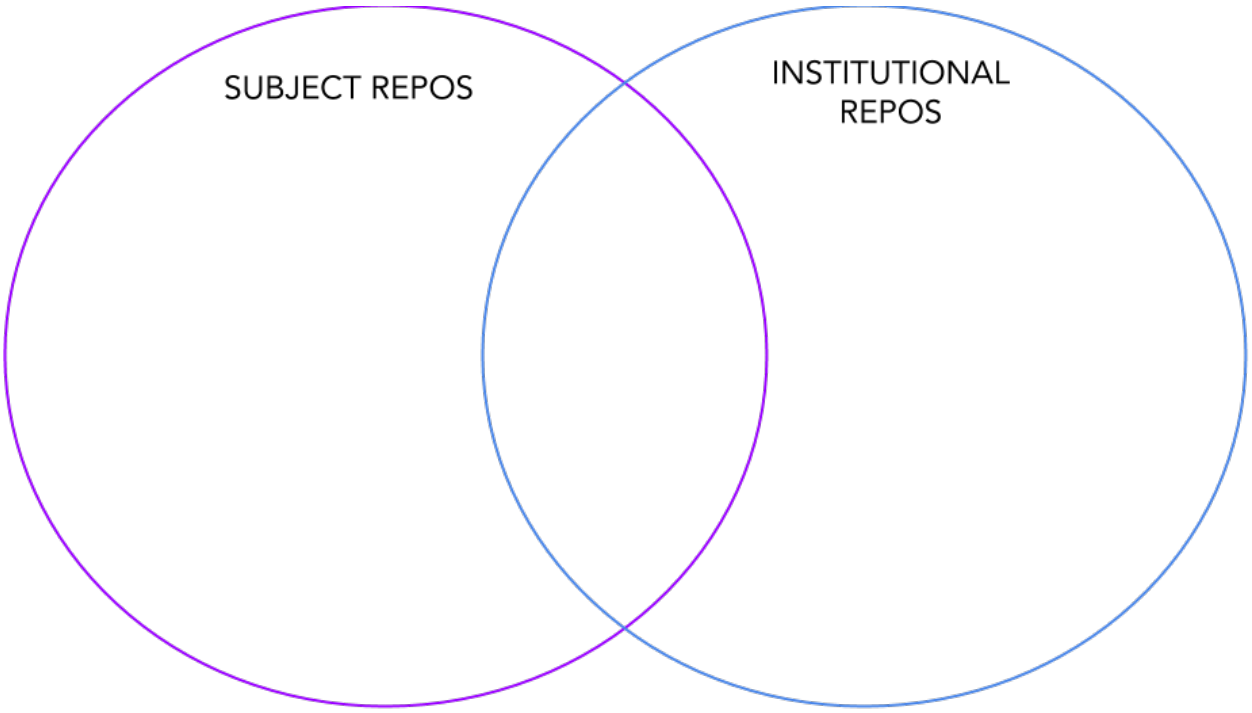


Figure 15.2:

What do these claims make you think about making all data open and accessible? What can librarians do to mediate the risk of insufficiently anonymized data, if anything?

`+ Asher and Jahnke also emphasize the importance of contextual details in qualitative research:`

> As is often the case with qualitative research, removing these contextual details would diminish, if not destroy, the usefulness of these data to future researchers—the ostensible audience for these materials. Unfortunately, each piece of contextual detail retained about a research participant to aid in data interpretation simultaneously creates more potential for identification.

`+ How might librarians and researchers balance the need for confidentiality with the need for contextual`

3. How should we differentiate between "useful" data that should be kept for future use and data that should be kept private and not collected? Does it matter who does the collecting and their intentions with the data?

4. Asher and Jahnke assert that "…a researcher's ethical obligation to his or her research participants does not end once he or she has transferred responsibility for their preservation." Do you agree with this? What obligations do researchers have to their participants? What is an original researcher's obligation when another researcher (whether individual or affiliated with a research or corporate institution) reuses the data? Do librarians and archivists have any obligations to research participants?

5. In reference to the Hurley piece, should data belong to researchers (primary investigators) or institutions (or someone else)? What steps could be taken by librarians to ease the transition from one institution to another that they discuss in this article?

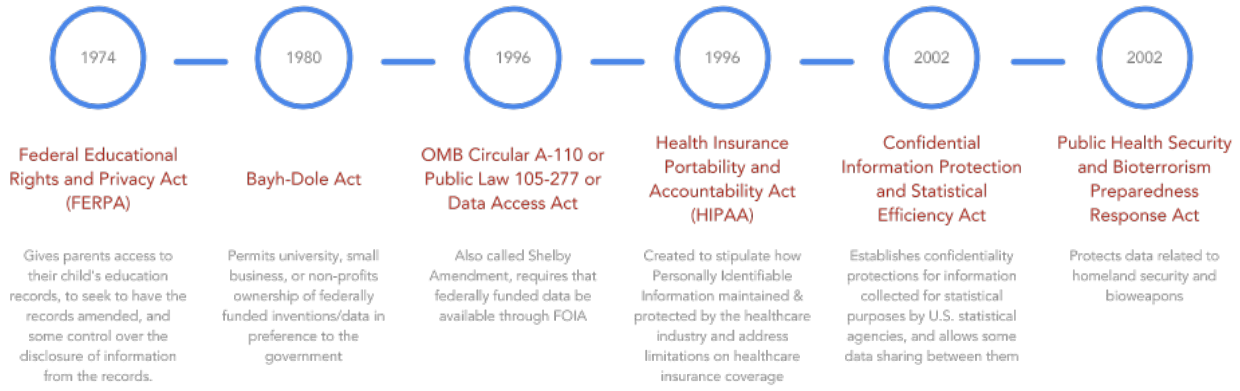


Figure 15.3:

15.2 Pre-lecture activity

15.3 Lecture

Let's take a historical look at some of the national-level responses to data sharing and how this might impact our work and the way that we discuss data sharing and management with our patrons.

Worth noting too, that "anonymized" data is not the same as "deidentified" data, legally speaking. Deidentification is specifically mentioned in HIPAA as the process of striping a list of 18 key identifiers. However, we have to think about the process of "anonymization" – removing values in a dataset so it can't be reconnected to specific people. You might not have SSN or DOB, but if you have gender, zip code, and occupation, then there's a chance you could re-identify people from deidentified data.

Another interesting point around security and identification is the fact that the Public Health Security act is why you can't get data about NYC infrastructure as well.

What do y'all think about the restrictions placed on data sharing?

The early 2000s also brought a lot of changes specifically for those who's research was federally funded:

A lot of the federal memos and laws prompted the creation of data services and data managment librarian positions. The OSTP memo in particular is cited by many data librarians as the push that created their position at their institutions. The idea was there was someone on campus who is paying attention to all the data and research related news and laws, and could help the patron population be proactive about compliance (which means teaching good data management practices, research infrastructure, etc. all the facets of data librarianship that we've discussed in this class). Now, librarians are acknowledged in the NSF-ENG's new guidelines for DMPs: https://nsf.gov/eng/general/ENG_DMP_Policy.pdf!

Librarians and archivists can also harness their expertise for social justice and community involvement RE: government data. When Trump was elected, the Data Refuge project was born because of the fear that federally hosted environmental data would be deleted from the government databases and repositories (which it has):

> Data Refuge is a community-driven, collaborative project to preserve public climate and environmental data. When we document the many ways diverse communities use data, we can also advocate for future data. Data Refuge is also an initiative committed to identifying, assessing, prioritizing, securing, and distributing reliable copies of federal climate and environmental data so that it remains available to researchers. Data collected as part of the #DataRefuge initiative will be stored in multiple, trusted locations to help ensure continued accessibility.

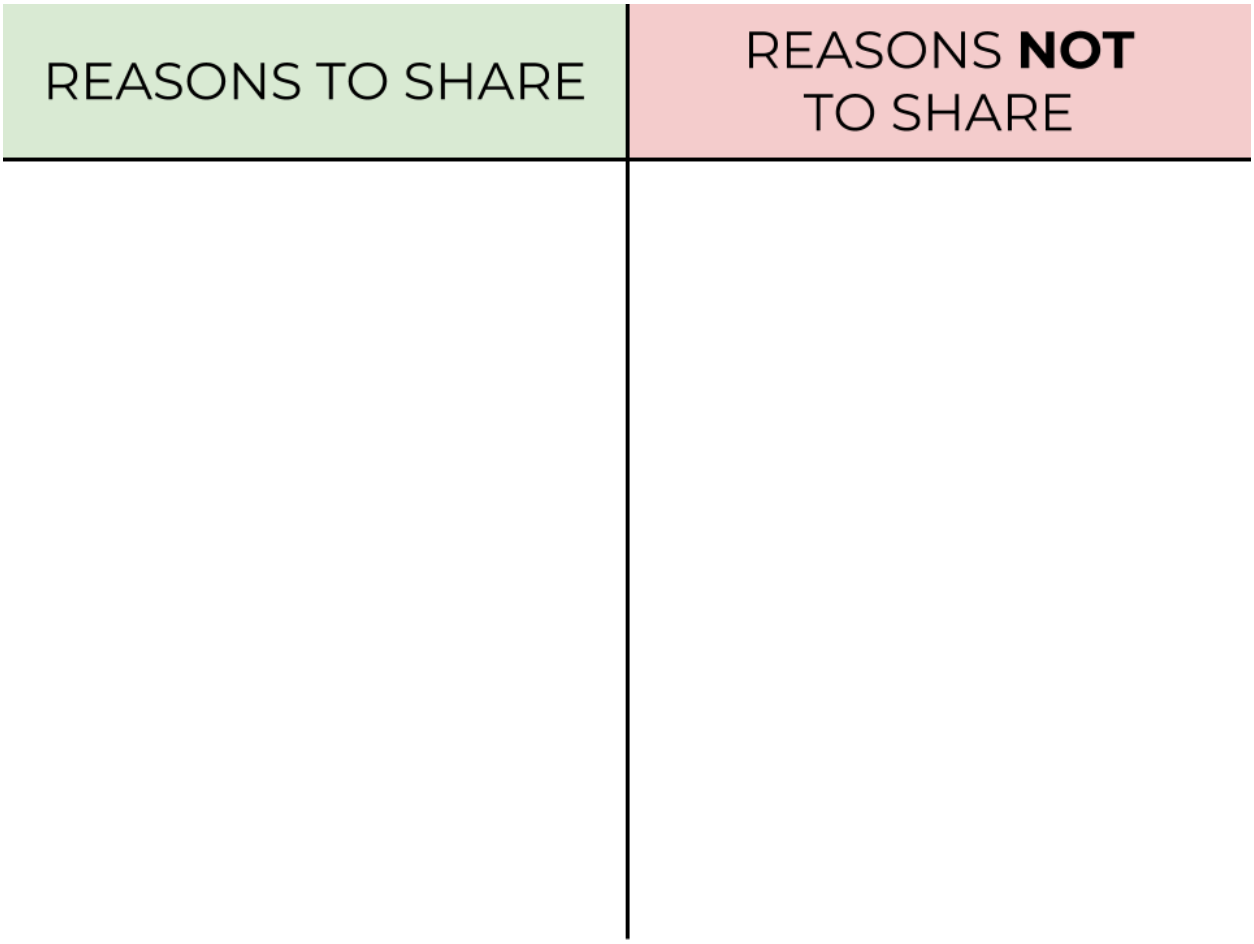


Figure 15.4:

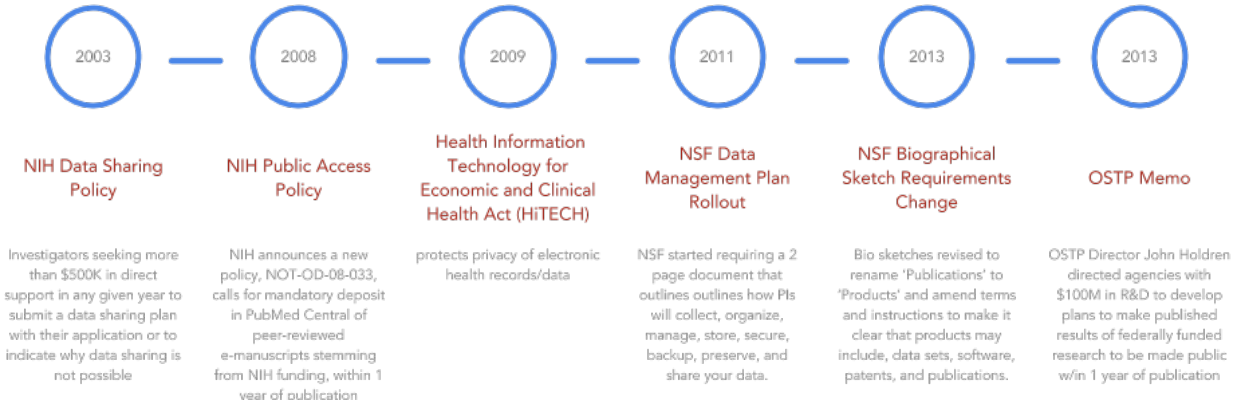


Figure 15.5:

The idea of being involved in both "guerilla archiving" events and also going through officialy channels for influence have varying levels of success. One example of that 'official channel' is the ACRL (Association of College and Research Libraries) reponse in this letter to the NIH's draft Data Management and Sharing Policy, providing some advice and voicing some concerns:

> Instead, ACRL encourages NIH to include case reports,other medical records, or data containing PII in the definition of scientific data and clearly note that researchers should share them in accordance with federal policy and other best practices (e.g., HIPAA, restricted sharing, aggregation to a level that will reduce the possibility of disclosure).

And this great piece of advice:

> ACRL also requests that NIH reconsider the exclusion of laboratory notebooks, as their exclusion is in tension with Section V, Part 1.2 of the Proposed Provisions, which states that the DMP must: 'Describe any other information that is anticipated to be shared along with the scientific data, such as relevant associated data, and any other information necessary to interpret the data (e.g., study protocols and data collection instruments)." Laboratory notebooks include recorded information that is 'necessary to interpret the data.' NIH should consider requiring that the Data Management Plan address how laboratory notebooks will be managed and how the information contained within them will be shared.

Taking a turn out of government involvement let's talk about text and data mining. Chris Hartgerink's use case, trying to access and download materials from his library in bulk, is not new. There are lots of researchers who ask me weekly if they can mine the materials that the library subscribes too. And the answer most often is: sorry, you can't. Because exactly what happened to Chris, outlined in our reading for this week, is still the answer that many publishers give us.

- What are some of the reasons why publishers would want to discourage text/data mining for institutions' subscribed materials?

15.4 Lab/Homework

You'll have time in class to work on the final project! No homework except the final project materials!

Please submit the materials for the final project via this link: https://cloud.vickysteeves.com/index.php/s/ZYwcWAyr5pYxNiA before 11:55pm on Thursday the 12th.

Chapter 16

Week 16: Future/sustainability of data

Agenda for today's class:

- 6:30 - 6:40: do the final class self-eval: https://forms.gle/tCp97SYxVC2rXH2h7
- 6:40 - 7:30: discussion
- 7:30 - end: final project poster session & final class celebration!

16.1 Discussion Questions

Joanna, our facilitator this week, prepared the following questions for in-class discussion:

1. In the Goldstein and Ratliff piece, the authors explain that their DataSpace model "does provide for storage for an indefinite period as well as basic internet-based sharing, but it would not, for example, fund the duplication and transfer of data into another repository, or the conversion of data from one format into another."
    - Is there anything else that is missing from this funding model that should be included?
    - If the model does not fund the conversion of data from one format to another, how should obsolete data formats be handled?
2. As we have discussed throughout the semester, "it costs money to accept and maintain both data and materials and, on many levels, people are faced with a choice between funding more science and saving the science we've already done."
    - When faced with the choice between funding more science and saving the science that has already been done, what should people or institutions prioritize?
    - Alternatively, how can scientific funding choices be made differently so that we can both preserve the old and create the new?
3. Timmer contends that "In the US and most of Europe, government science budgets are likely to be flat in the best of circumstances for the next few years, so coming up with a dedicated budget for community resources will mean removing it from elsewhere."
    - What are your thoughts? Should the budget for scientific research be taken from elsewhere? If so, what do you propose be cut to increase the scientific research budget? If not, why?
    - Most importantly, how can we fund scientific research without "[supporting] a system of haves and have-nots, or [limiting] access from developing world researchers"?

Figure 16.1:

16.2 Resubmitting Homeworks

As I mentioned in my email, you have until the Wednesday December 18 at 12pm to submit any re-do's of your assignments that you want regraded. Final grades for Fall 2019 due online by end of that day, so I can't accept anything past that.

Please submit your mulligans here, password will be given in class: https://cloud.vickysteeves.com/index.php/s/eXZAn6WHLr7sRpD