Research Data Management (at HBS)

Agenda

• Brief introductions
• Why research data management
• Introduction to our narrative
• RDM lifecycle:
  • Planning
  • Data Acquisition
  • Storage & Analysis
  • Data Sharing & Archiving
• Closing remarks

Goals

• Become familiar with the RDM lifecycles
• Understand the details and requirements at each stage
• Know what resources and services are available and recommended at HBS, Harvard, and beyond
• Engage in utilizing best practices throughout your work at HBS and in your future careers
Why research data management?
Intro: Discussion Questions

1. Who can describe to me what Research Data Management is all about?
2. So what? Why should I care? Why should you care?
3. What’s in it for the faculty member? For the School?
4. Who might be partners in this endeavor?
5. Is anyone legally obligated?
Data come in many ways...

**Thousands of Social Media Posts fabricated by Chinese Government for Strategic Distraction**

![Time Series of 43,757 Known 50c Social Media Posts with Qualitative Summaries of the Content of Volume Bursts](image)

*How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument (King, Pan, Roberts, 2017)*

**Songs from 60 human societies exhibits form-function relations detectable by listeners around the world**

![Form and Function in Human Song](image)

*Form and Function in Human Song (Mehr, Singh, York, Glowacki, Krasnow, 2018) [https://www.naturalhistoryofsong.org/](https://www.naturalhistoryofsong.org/)*

"**Research data management** concerns the organization of data, from its entry to the research cycle through the dissemination and archiving of valuable results. It aims to **ensure reliable verification** of results, and permits **new and innovative research** built on existing information."

Why should you care?

**Benefits** for yourself, researchers, and science:

- Your future self will thank you!
  - Ever have to reproduce an analysis that you did a few months ago or even a few years ago?
- Facilitate and ensure seamless team transitions
- Conduct analysis effectively when collaborating with others
- Check and verify research results
- Support FAIR principles
  - Findable, Accessible, Interoperable, & Reusable

**Compliance**, to minimize risks:

- Be **compliant with University and School policies**
- Be **compliant with research funding organizations** that require a data management plan and data accessibility
- Be **compliant with journals** that require to submit your data accompanying the article
Your role as a Researcher...

• Responsibility as a proxy for faculty researchers
• Dual role of minding of administrative responsibilities and conducting the faculty research program
• Provides a reference framework for how to conduct research in line with best practices
• Practices aligned with career- and established research professionals (such as RCS personnel)

You may have to exercise some, many, or all of the following recommendations...
The Old Data Lifecycle

Data Roadmap, based on Briney, 2015. Data Management for Researchers
The New Data Lifecycle...

_Data as a Product of Research_

Data Roadmap, based on Briney, 2015. Data Management for Researchers
Research Data Management Lifecycle

A reference guide with information and resources to help you manage your research data

DATA LIFECYCLE

Planning Data Management
- How can I best manage my data throughout the lifecycle of my research to save time and money in the future?
  - Data Management Plans (DMPs)
  - DMP requirements and tools
  - What research objects should be tracked and documented

Data Acquisition and Collection
- How can I acquire data in an efficient and ethical way, and how can I ensure that my data is used appropriately?
  - Data Use Agreements (DUAs)
  - Institutional Review Boards (IRB and IACUC)
  - Subscription data

Storage, Security, and Analysis
- What are my options for effectively organizing, storing, securing, computing, and analyzing my research data?
  - Data security
  - Computing, research methods, data science, and viz support
  - Electronic Lab Notebooks

Dissemination and Preservation
- Why is it worthwhile to share my data? What do funders and journals require? Can I get help with data curation?
  - Data repositories
  - Open Access
  - Data citation, FAIR principles
  - Data disposal

Source: https://researchdatamanagement.harvard.edu/

* In mid-July, https://research.harvard.edu will supercede the above site
Research Data Management is a Collaborative Effort: *You need to do your part*
Introduction to our Narrative

You have just come on board as an RA for Professor Smith. Professor Smith would like your help expanding on a previous research study she worked on five years ago with a different RA. In this study, Professor Smith surveyed CEOs of fast-casual restaurants (e.g., Chipotle; Shake Shack) to learn about their behaviors. The data and code for this study were saved on the HBS research computing grid, and you have been granted access to the files.

Recently, Professor Smith received a grant from the federal government to build upon this research by exploring what CEO-reported behaviors from five years ago are related to current company financial records. To explore this, she will ask you to obtain company financial data compiled by a firm called FinanceCorp. These data include confidential company financial data, and are therefore considered extremely sensitive. The FinanceCorp data will be merged with the data Professor Smith collected five years ago from company CEOs.

In addition to you, Professor Smith’s team also includes another professor from the University of Southern California. As a result, all the files you receive and create for the project will need to be easily accessible to individuals outside of HBS.

Finally, note that as a condition of receiving the grant to explore this topic area, Professor Smith has agreed to de-identify the data and make it available for public use at the conclusion of the study.
Intro to Research Data Management
Planning for Data Management

How can I best manage my data throughout the lifecycle of my research to save time and money in the future?

Goals:
- Learn about Planning for Data Management
- Utilize DMP Checklist for work
Planning: Discussion Questions

1. We talk about a Data Management Checklist. What is this in lieu of?
2. What is the purpose of this checklist?
3. What research artifacts is it targeted towards?
4. Who should be using this? And for how long?
The checklist will help you define:
  o how the data will be created
  o how it will be documented
  o who will be able to access it
  o **where it will be stored**
  o who will back it up
  o **whether and how it will be shared & preserved**

Planning is not simply naming files and folders so that only your research team understand their content. Instead...
  • putting standards and guidelines into action
  • documenting detailed metadata for better data discovery and illumination
  • ensuring the value and accessibility of your research long after your project is complete.

The checklist can inform a Data Management Plan (DMP), which is often required by funding agencies or philanthropic funders
  • E.g. NSF, NASA, Gates Foundation, Sloan Foundation
Use the Checklist for routine work

**Research Data Management Checklist**

This document serves as a reference checklist to keep track of the elements that make up good research data management (RDM) in the lifecycle. Remember that it is not linear and you may find yourself jumping around this lifecycle throughout your project.

### Planning and Data Acquisition

<table>
<thead>
<tr>
<th>Activity</th>
<th>Questions</th>
</tr>
</thead>
</table>
| Project description (background/rationale) | 1. What research question(s) are you addressing?  
2. What study methods and design will you use? |
| Data description and collection | 1. What instruments/questionnaires will you use?  
2. What format will your data be in (audio, text, database, images)?  
3. How large do you anticipate your files will be?  
4. How will you collect or acquire your data? |
| Participating researchers | 1. Who will be a part of the research study?  
2. Are all researchers HBS affiliated, or will there be individuals from outside organizations? |
| Ethical review and legal compliance | 1. Does your data include human subjects and need to go through the Institutional Review Board (IRB)?  
2. Does your data include sensitive information (e.g., birthdates, social security numbers)?  
3. Are you collecting or obtaining data from the European Union that is subject to General Data Protection Regulation (GDPR)?  
4. If you are obtaining existing data, what agreements need to be put in place (e.g., Data Use Agreement; Memorandum of Understanding; Nondisclosure Agreement) and reviewed by a Harvard lawyer? |
What should you manage?

Any data, code, documentation used throughout the research lifecycle:
• Quantitative and qualitative data
• Primary and secondary data
• Notebooks
• Codebooks
• Records and notes
• Code or software used to run analysis
• Workflows or pipelines
• Metadata or documentation describing the data (‘data dictionaries’)

Record and retain sufficient information to enable others to understand and reproduce your work (aka winning the lottery scenario)
Planning Exercise

Professor Smith has asked you to obtain company financial data compiled by a firm called FinanceCorp. These data include confidential company financial data, and are therefore considered extremely sensitive. The FinanceCorp data will be merged with the data Professor Smith collected five years ago from company CEOs.

- Using the Checklist, how will you **plan** for this study?
Planning
Planning: Recap & References

• Talk to RCS and/or KLS early in the planning process
• Use the Data Management Checklist to plan for your research
  • NB! Some funders may require a Data Management Plan
• Manage all notes, code, data, etc. to enable others to understand and reproduce your work

References:


• 'Everyone Needs a Data Management Plan', Nature 555, 286 (2018); doi: 10.1038/d41586-018-03065-z
• http://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm
Data Acquisition & Collection (& Security)

How can I acquire data in an efficient, ethical, and secure way, and how can I ensure that my data is used appropriately?

Goals:
- Know what services available
- Understand DUAs, NDAs, and IRB
- Plan for data security at all stages
How are you acquiring your data?

Data generated by investigator:

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment</td>
<td>A scientific procedure undertaken to make a discovery, test a hypothesis, or demonstrate a known fact</td>
</tr>
<tr>
<td>Observation</td>
<td>The action or process of observing something or someone carefully or in order to gain information</td>
</tr>
<tr>
<td>Simulations</td>
<td>The production of a computer model of something, especially for the purpose of study</td>
</tr>
<tr>
<td>Derived / compiled</td>
<td>Base data on a logical extension, modification, or collection of items</td>
</tr>
</tbody>
</table>

Data acquired from others:
- Does the data you need already exist? Do you know how and where to find it?
- Is it already licensed by Harvard or need to be acquired? Are appropriate funds available if needed?
- Does it require a Data Use Agreement (DUA)?
Assistance in acquiring your data

Services to help faculty acquire data:

<table>
<thead>
<tr>
<th>Service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baker Library subscriptions</strong></td>
<td>wide range of data available</td>
</tr>
<tr>
<td><strong>Baker Research Services</strong></td>
<td>custom discovery and delivery of data</td>
</tr>
<tr>
<td><strong>Baker Faculty Data Licensing Service</strong></td>
<td>negotiation of licenses/DUAs with vendors (for faculty acquisition or purchase)</td>
</tr>
<tr>
<td><strong>Behavioral Research Services</strong></td>
<td>supports the data collection needs of HBS faculty and doctoral students conducting a broad range of experimental and behavioral research</td>
</tr>
<tr>
<td><strong>DRFD</strong></td>
<td>also supports DUAs</td>
</tr>
<tr>
<td><strong>Research Computing Services (RCS)</strong></td>
<td>data collection via web scraping; wrangling via cleaning, matching, merging, etc.</td>
</tr>
</tbody>
</table>
Guarantee efficient & accurate collection

Consider using tools, templates, & data dictionaries when collecting data

• Increases accuracy & efficiency
• Promotes collection & preservation of metadata (source, year, etc): Consistency & reliability (where, how, what, etc)

For surveys: Qualtrics, used at HBS & Harvard

For collecting data, use *electronic notebooks*

• OneNote/O365
• Documents in HBS SharePoint/O365
• Evernote ($$)
• FileMaker Pro ($$)
• Open Science Framework (OSF)
• RSpace as a possible Harvard-wide tool

*Use case: Collecting information from Chinese texts on family migrations when one could not remove books. One needs to document book, title, year, GIS info for prefectures / provinces, etc.*

Again, keep security considerations in mind with 'synchronized' services

Data Security: An important planning step!

• A thorough understanding of the data and its custodianship will drive the RDM narrative

• All persons should have some basic understanding based on HBS IT and HU data security requirements

• Know in advance what your options are for transferring and storing your data, and how this applies to computing
  • See our website for data transfer options: https://grid.rcs.hbs.org/transferring-data
  • E.g. PII / Human Subjects data can be stored on L4 research storage as part of the compute grid environment, but cannot be accessed via Windows & Mac desktops & laptops
Data Security Plans via Data Safety

• “Data Safety” is the Harvard Portal for submitting data security plans.

• A Data Safety plan will be required for DUAs & IRB submissions

• Helps faculty research groups plan and execute good RDM practices, including:
  • What resources should be used
  • What persons should or should be involved in the data acquisition, analysis, and sharing
  • What restrictions may apply based on the data content, stewardship, or geographic location/source of the data

• Will dictate compliance with L2, L3, or L4 data security protocols

• You might be involved in helping to prepare the Data Safety plan.

• HBS IT Security (informationsecurity@hbs.edu) and RCS can help you create a data security plan compliant with data and university requirements and discuss what may be the best options for short- and long-term projects.
Data Protection Regulations & Policies

- **HIPAA** (18+ identifiers – alone or in combination datasets) Informed Consent
- **FERPA** (education information and special protections)
- **MA** data protection law (security requirements to handle private data from state residents)
- **Stem Cell** data and **Genomics** data must be published in approved repository, but also must be de-identified.
- **GDPR** (General Data Protection Regulation in Europe)

**At Harvard:**

- Data retention (7 years)
- Data security (5 security levels)
Be Familiar with Data Security!

<table>
<thead>
<tr>
<th>CLASSIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L1</strong> Information intended and released for public use.</td>
</tr>
<tr>
<td><strong>L2</strong> Information that may be shared only within the Harvard community.</td>
</tr>
<tr>
<td><strong>L3</strong> Confidential and sensitive information, intended only for those with a “business need to know.”</td>
</tr>
<tr>
<td><strong>L4</strong> High-risk information that requires strict controls.</td>
</tr>
<tr>
<td><strong>L5</strong> Extremely sensitive information requiring specific controls and isolation from the network.</td>
</tr>
</tbody>
</table>

**The University intentionally provides this information to the public.**

**The University chooses to keep this information private, but its disclosure would not cause material harm.**

**Disclosure of this information beyond intended recipients might cause material harm to individuals or the University.**

**Disclosure of this information beyond specified recipients would likely cause serious harm to individuals or the University.**

**Disclosure of this information could cause criminal liability; loss of insurability or employability; or severe social, psychological, reputational, financial, or other harm to an individual or group.**

*Nota bene!!*

Examples:
- Published research
- Course catalogs
- Published faculty and staff information
- Student directory information*  
- Basic emergency response plans (life safety)
- University-wide policies
- Harvard publications
- Press releases
- Published marketing materials
- Regulatory and legal filings
- Published annual reports
- Code contributed to Open Source
- Released patents
- Plans of public spaces

*Directory information about students who have requested FERPA blocks must be classified and handled as L3 at minimum.

Examples:
- Department policies and procedures
- Employee web/intranet portals
- Harvard training materials
- Pre-release articles
- Drafts of research papers
- Work papers
- Patent applications
- Grant applications
- Non-public building plans or layouts (excluding L3 or L4 items)
- Information about physical plant (excluding L3 or L4 items)

Examples:
- Non-directory student information
- Non-published faculty and staff information
- Information protected under FERPA, in general
- HUID tied to an individual
- Personnel records**
- Donor information (excluding L4 data points or special handling)
- Non-public legal work and litigation information
- Budget/financial transactions information
- Non-public financial statements
- Information specified as confidential by vendor contracts and NDAs
- Information specified as confidential by Data Use Agreements
- General security findings or reports (e.g. SSAE16)
- Most Harvard source code
- Non-security technical specifications/architecture schema
- Library/museum object valuations
- IRB records
- Government issued identifiers (e.g. Social Security Number, Passport number, driver’s license)
- Individually identifiable financial account information (e.g. bank account, credit or debit card numbers)
- Individually identifiable health or medical information**
- Individually identifiable research data
- Details of significant security exposures at Harvard (e.g. vulnerability assessment and penetration testing)
- Security system and architecture
- Trade secrets
- Systems managing critical Operational Technology
- Video Recordings of Research Interviews

Data Use Agreements (DUAs)

- These govern access to and treatment of data:
  1. May be required by an outside organization to your organization for use in your organization’s research, or
  2. Provided by your organization to an outside organization for use in its research.

- Also referred as: License agreement, Confidentiality Agreement, Non-disclosure agreement, Memorandum of Understanding, Memorandum of Agreement

- DUAs are required if the data will be restricted to a subset of HBS; see also Office of Vice Provost for Research policy

DRFD Research Administration & Baker Library Data Licensing Service can assist with the process of DUA review and signing

- Done in coordination with via Harvard's Office of Sponsored Programs
- No-cost DUAs: Alain Bonacossa (DRFD) or Katherine McNeill (Baker)
- Else, contact the Data Licensing Service (Katherine McNeill, Baker Library)
Research involving Human Subjects - IRB

IRB approval is required when conducting primary or secondary data collection involving human subjects or individual-level identifiable data.

**Primary data collection**
- Experiments (field, online, lab)
- Surveys, interviews, observations
- Scraping data from non-public websites

**Secondary data collection**
- Analysis of individual-level identifiable data

**Alain Bonacossa** is available to advise you on federal and state regulations and university policies that apply to research with human subjects. The team also reviews IRB applications on behalf of Harvard’s Committee on the Use of Human Subjects (CUHS).
Considerations for Sensitive Data

• Sensitive Data: human subjects, anything proprietary, or covered by DUA

• Deidentification may be tedious yet important step
  • *Highly recommend* that this be asked of data provider if possible

• Re-identification by grouping secondary data (or indirect identifiers) is very possible
  • Consider multiple approaches to permit data granularity and fidelity while preventing re-identification
  • E.g. if 1st three digits of ZIP codes + year of birth == 0.04% of individuals can be re-identified vs ZIP + birthday + sex == 87% (Sweeney et al.; 2000)

• Just as important to promote preservation and re-use of sensitive data
  • Don’t promise to destroy your data
  • Don’t promise not to share your data
  • Do get consent to retain and share data
  • Do incorporate data-retention and -sharing clauses into IRB templates

• Many evolving techniques to safeguard privacy yet promote reuse

• Contact RCS, KLS RDP, or DRFD Research Admin if you have any questions
Acquisition/Collection: Questions

1. Name a few partners and their role in data acquisition.
2. Why use data templates?
3. What tools might one use for data collection? Why?
4. Tell me about Data Security...
5. Give me an example of L3 and L4 data
6. How are a DUA, an NDA, an IRB submission, and a Data Safety plan related? Or not?
Data Acquisition & Collection Exercise

Professor Smith has asked you to obtain company financial data compiled by a firm called FinanceCorp. These data include confidential company financial data, and are therefore considered extremely sensitive. The FinanceCorp data will be merged with the data Professor Smith collected five years ago from company CEOs.

- Should IRB be involved?
- Is a DUA needed?
- What Harvard security level might the data be?
- How will you transfer and store the data?
Data Acquisition & Collection: Recap

- Services are available to help with data acquisition, no matter if acquired by the investigator (primary) or from others (secondary).
- Use software tools to aid in efficient & accurate collection.
- However the data are acquired, be mindful of requirements related to IRB, DUAs, and HBS security levels.
- Sensitive data does require special precautions at all stages of work.

References
- [https://inside.hbs.edu/Departments/it/security/Pages/default.aspx](https://inside.hbs.edu/Departments/it/security/Pages/default.aspx)
- [https://grid.rcs.hbs.org/transferring-data](https://grid.rcs.hbs.org/transferring-data)
Data Acquisition & Collection
Storage, Security, & Analysis

What are my options for effectively organizing, storing, securing, computing, and analyzing my research data?

Goals:
- Know what resources are available
- Understand best practices
- Know where to get help
Storage & Analysis

• This is the most difficult & time-consuming RDM stage
• Likely need to perform, rinse, & repeat

So..

• Should be effortless if one has planned well...
  • 5Ps: *Proper Planning Prevents Poor Performance*
• ...and if done well 1st time around
Data Security

Security is more than where you store it – it's how you approach the care, handling, and movement of data

• This will vary depending on sensitive data level
• Is important to consider while on- and off-campus
  • Email at home?
  • Using your mobile phone or tablet
  • What about while traveling?
  • Even more important in our remote-work/pandemic status
• May be determined by Data Safety plan.
• See HBS IT Security handout for appropriate considerations
• And these other helpful websites:
  • https://huit.harvard.edu/remote
  • https://inside.hbs.edu/Departments/it/security
  • https://inside.hbs.edu/Departments/it/howto/Pages/work-remote.aspx
What resources are available for storage?

This will vary based on data sensitive data level and indicated by DUA, IRB, or Data Security plan

- Desktop/laptop: L1 - L3 data
- Research storage on compute grid: L1 - L4 data

- See HBS IT Security handout for appropriate considerations
- May often be directed by faculty or RCS member
- HBS-issued desktops / laptops storage (usually SSD)
- HBS research computing environment
  - Research storage associated with the compute cluster ('grid')
  - MariaDB database server
  - HBS and HU collaboration tools (E.g. SharePoint, OneNote)

- Strongly encourage the use *collaboration folders* on research storage for group work, and ensure that work is synchronized daily
Other storage resources?

• FASRC's compute environment*
• IQSS' compute environment
• Cloud providers: Mass OpenCloud, AWS, Azure, GCP, etc*
• 3rd party-licensed providers (e.g. Qualtrics, Zotero, etc)

• See our website for data transfer options: https://grid.rcs.hbs.org/transfer-data

*Some costs may be associated with use. Please contact RCS first
# Collaboration Tools: Data Security & Privacy

<table>
<thead>
<tr>
<th>Tool</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Email (Gmail, Yahoo, etc)</td>
<td>✓</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer Google Drive - All tools</td>
<td>✓</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumer Dropbox, Evernote</td>
<td>✓</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harvard email</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Google Drive/Docs (g.harvard)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Office365 OneDrive/SharePoint</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Microsoft Teams</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Zoom - follow security guidance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Slack (FAS first, and expanding)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Confluence/Wiki (behind HarvardKey)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Dropbox</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Network File Share</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvard Office365 &quot;HRCI&quot; SharePoint</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Accellion Kiteworks</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Level 5 data applies only to research data per IRB and requires security consulting for special handling.

* Consumer Versions not recommended for University business.
**L4 with special controls setup

See the [Data Classification Table](https://security.harvard.edu/collaboration-tools-matrix) for examples of data at levels one through four.

See also: [Know Your Data Resources, Guides](https://security.harvard.edu/collaboration-tools-matrix)
What resources are available for compute?

HBS computing environment:
- HBS-issued desktop / laptops (data-intensive work – please talk to RCS/RSS)
- HBS compute cluster (grid), FASRC Cannon*, IQSS' RCE

Cloud commercial vendors*:
- Amazon Web Services (AWS), Google Cloud, Microsoft Azure
- Please sign-up under Harvard contract (tenant)
- They provide support for secure storage & compute, BUT ensure they meet your security requirements (storage location, sufficient security)

Open-source Cloud systems (not vetted)
- OpenStack, OpenNebula, MA OpenCloud

Be thoughtful and strategic about use and efficiency
- Offload long-running work to the HBSGrid cluster or FASRC’s Cannon
- If something isn't running as expected, troubleshoot or ask for help

*some costs may be associated with use
Storage & Analysis: Questions

1. What speaks to you about Data Security?
2. What storage resources are available at HBS?
3. How might you collaborate with other when using data?
4. What compute resources do you have to use??
Storage and Analysis
You have also received the go-ahead to transfer and store the company financial data from FinanceCorp. Recall that project materials will need to be easily accessible to her collaborator from USC.

• How will you transfer the data?
• Where will you store it, taking into account any security level determination?
• How will you ensure secure file sharing with the collaborator?
Doing the work...

We organize our recommendations into the following topics (Box 1):
• Data management: saving both raw and intermediate forms, documenting all steps, creating tidy data amenable to analysis.
• Software: writing, organizing, and sharing scripts and programs used in an analysis.
• Collaboration: making it easy for existing and new collaborators to understand and contribute to a project.
• Project organization: organizing the digital artifacts of a project to ease discovery and understanding.
• Tracking changes: recording how various components of your project change over time.
• Manuscripts: writing manuscripts in a way that leaves an audit trail and minimizes manual merging of conflicts.

Good enough practices in scientific computing

Greg Wilson1☯, Jennifer Bryan2☯, Karen Cranston3☯, Justin Kitzes4☯, Lex Nederbragt5☯, Tracy K. Teal6☯
1 Software Carpentry Foundation, Austin, Texas, United States of America, 2 RStudio and Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada, 3 Department of Biology, Duke University, Durham, North Carolina, United States of America, 4 Energy and Resources Group, University of California, Berkeley, Berkeley, California, United States of America, 5 Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway, 6 Data Carpentry, Davis, California, United States of America

☯ These authors contributed equally to this work.
* g wilson@software-carpentry.org

https://doi.org/10.1371/journal.pcbi.1005510

Planning
Data Acquisition & Collection
Storage, Security, & Analysis
Dissemination & Preservation

PERSPECTIVE

We are never done doing the work...
Project Setup & Organization, part 1

1. Put each project in its own directory, which is named after the project.

2. Create folders that will separate your code and data.

3. In your data folder, ensure that your raw data are separated from any data you have processed (i.e., your clean datasets).
4. Create additional folders as needed for project. E.g., report folder for output; references folder for reference material such as survey instrument.

5. Create a "README" file that outlines basic information about the project and the folder/file structure.

6. Name files in a way that their content or function can be easily identified.

7. Use relative addressing to make the project portable.
Languages & Tools Recommendations*

Programming languages for processing and analyzing data in research:

• Most used: Python (Spyder as editor) and R (RStudio as editor)
• Others: Scala, Java, Julia

Statistical packages:

• Stata, R, SAS, & SPSS

Big Data tools:

• Spark (Hadoop), Kubernetes/containers

Data Visualization tools:

• ggplot2, Tableau, D3, Shiny, Plotly, Pandas, WorldMap (from HU's CGA)

*This list is not meant to be exhaustive!
Documenting workflows & analyses

Whatever tool you use, document all steps in your analysis and data transformations. Some tools to help with that:

• RMarkdown and RMarkdown Notebooks (used with R)
• Jupyter Notebooks: support for most languages (Python, R, Stata, MATLAB)
• Dyndoc (notebook for Stata)
• Templates with OneNote (or EverNote)
• Workflow/Pipeline Tools.
  • These help document and track process order:
  • Consider Drake (R); SnakeMake, doit or py-Make (Python); and make for other systems
  • See https://github.com/pditommaso/awesome-pipeline for a full list of options
• Be sure to update your data dictionary/codebook as you make changes to your data

and... Your future self will thank you!
Versioning and version control

• Incredibly important given the duration and lifespan of projects
• What may start out as small, test idea may grow organically into multi-person and multi-site research project

Two approaches given, for small to large...

• Manual versioning
  • In most cases, data will be organized in files under directories:
    • Use phase title, unique identifiers, and descriptive filenames
    • Prefix by date (yyyy, yyyy.mm.dd, yyyy_mmdd, yymmdd)
    • Reserve / display 3-letter file extension for file format, such as .txt, .pdf, or .csv.
  • Note all changes in a ReadMe.txt or Changes.txt document

• Use a version control system via Git or Github.com
  • Use your judgement, and talk to your faculty advisor; BUT their non-use does not prevent your use
  • Utilize Github.com web interface for external (non-HU) collaboration, and code.harvard.edu for internal-only use
  • Command-line (terminal/shell) Git or Git-GUI like GitKraken

• NB! HBS/IQSS Version Control Class offered each semester
Databases for (un)structured data

• When data contain complex relationships we recommend storing and using the data in a database for easy, fast queries:
  • SQL databases (PostgreSQL, MySQL, MariaDB)
  • Non-SQL databases (MongoDB, Cassandra)

• Side benefit: data is read-only, unless explicitly changed
• Consider versioning the data / databases also
• Update your data dictionary describing data, types, use, etc.
  • Consider storing it, as well as change information, as a part of the database

RCS provides MariaDB as a part of the RC environment
• Provision and guidance on data modeling and DB development
• Advise on best practices and performance tuning
Training Opportunities

• RCS offers tool & analysis environment training, both on-campus and in collaboration with IQSS
  • E.g. Intro to R / Python, Automating work, Version control, etc.

• Data collections & analytical methods this fall
  • Web scraping, causal inference, natural language processing

• Offered fall and spring, and announced through our newsletter and website

• (HBS RAs) Please contact RSS prior to attending
Storage & Analysis: Questions

1. How might one use Version Control?
2. Describe to me good project organization?
3. Why are workflow tools important?
Storage and Analysis
Storage, Security, & Analysis Exercise, Part 2

You have obtained access to the data and code from five years ago, and it is a bit...disorganized.

- How would you go about organizing and documenting the previous files?

The FinanceCorp data will be merged with the five year old data.

- How will you organize the new code and files?
- How will you document your processes?
- What type of version control system will you employ?
- What research software will you use?
Dissemination and Preservation

What help can I get with managing my data at the end of a project, so that it can be safeguarded for the long-term?

Why is it worthwhile to share my data?

Goals: Understand:
- Options for managing data long-term
- Resources available
Options for Safeguarding Data Long-Term

• Not sure what to do with your data at the end of a project?
  • Retention requirements
  • HBS Faculty Papers Program

• Interested in sharing, but not now?
  • Embargos

• Ready to share now
  • Data repositories
  • Baker Library Data Deposit Service for HBS Faculty
Storage, Security, & Analysis: Recap

- Many resources are available for storage and computation (desktop, laptop, HBS grid, cloud), but the storage must be appropriate for the security level.
- Important to organize code/files and document processes
- RCS offers consultations and trainings on storage and analysis.

References:
- Baker Library Research Data Program: [https://www.library.hbs.edu/Services/Research-Data-Program](https://www.library.hbs.edu/Services/Research-Data-Program)
- [http://security.harvard.edu/dct](http://security.harvard.edu/dct)
- [https://github.com/pditommaso/awesome-pipeline](https://github.com/pditommaso/awesome-pipeline)
- Github.com
- code.harvard.edu
What do I do with all the "stuff" from my project?

If you're not yet ready to think about data sharing, but still want to secure your data, Baker Library can help.
Research Data Retention

How long do I need to keep research records?

Harvard Office of the Vice Provost for Research

• "Research Records should be retained, generally, for a period of no fewer than seven (7) years after the end of a research project or activity"

What exactly is a "research record"?

• Transcripts of interviews
• Photographs
• Videos/ audio files
• Data (qualitative and quantitative)
• Agreements with research subjects
• Project proposals
Long-term Storage

• For HBS Faculty, Baker Library offers secure storage through the Faculty Papers Program.
  • [https://www.library.hbs.edu/Services/Faculty-Papers-Program](https://www.library.hbs.edu/Services/Faculty-Papers-Program)

• If keeping locally, follow security protocols (for example, a locked cabinet for paper records; secure/restricted network space which is routinely backed up)
Archiving Data

**HBS Archives** includes research papers of faculty that trace the innovations in business education pioneered at the Business School.

- We collect for the HBS Archives research data with long term historical value/importance to the School.
- We can consult with HBS faculty members and help add their research to the HBS Archives collection. We can place an "embargo" on access to meet privacy or other concerns, but also ensuring this data is available for future researchers.

**Other Archives**
- Inter-university Consortium for Political and Social Research (ICPSR)

Comparison of Output with Hours of Sleep, ca. 1930. Western Electric Company Hawthorne Studies Collection.
Why share?

• Increases your reputation and the visibility of your research
• Increases your impact and informs new research
• Because your data can be cited:
  • Informs you on how your data is being used
  • Allows you to measure your greater impact
• Maximizes transparency, accountability and scrutiny of research findings
• To ensure data can be re-used, work to make it FAIR:
  • Findable
  • Accessible
  • Interoperable
  • Reusable
Data Deposit Service for HBS Faculty

Wherever HBS faculty want to share their data, Baker Library staff will do the deposit, including:

• Help researchers select the best repository for their needs
• Help researchers consider what data can be shared
• Advise on preparing data and documentation
• Liaise with the data repository
• Deposit the files
• Create repository metadata

https://www.library.hbs.edu/Services/Data-Deposit-Service
Data Repositories for HBS

HBS Dataverse

• Repository run by Harvard for sharing your data
• Enables immediate sharing of data and associated documentation
• Widely discoverable and citable with a DOI
• Online analysis features for selected formats

Other data repositories, e.g.,:

• ICPSR: full-service data archive; expert in managing and providing restricted access to sensitive data
• Journals may specify place of publication
Harvard Business School Dataverse is HBS' online repository for faculty and doctoral student research data. HBS Dataverse is managed by HBS Knowledge & Library Services. For more information, please visit the KLS Research Data Program website.
Dissemination & Preservation Recap

• Think early and often about how you can keep your data organized and well-documented while you're doing your research, to save a ton of time at the end.

• You don't have to share your data to properly keep it.

• When you're thinking about what your faculty can do with their data long-term, think of Baker Library.

• We can help your faculty find a solution that meets their unique needs--contact us for a consultation anytime as you're doing your research.

• More info -

• [https://www.library.hbs.edu/Services/Safeguard-Your-Data-Long-term](https://www.library.hbs.edu/Services/Safeguard-Your-Data-Long-term)
Dissemination & Preservation
Research Data Services for HBS Researchers

Departments across the Harvard Business School working together to help you be most effective at the various stages of the research data lifecycle.

Groups

- **Baker Library: Baker Research Services**
  - Data needs analyst
  - Data discovery
  - Secondary data retrieval
  - Text mining/analysis

- **Baker Library: HBS Archives**
  - Data management planning

- **Baker Library: Research Data Program**
  - Data management planning
  - Buying and licensing data
  - Research Datasets Tool
  
- **DRFD: Research Computing Services**
  - Research planning
  - Data management planning
  - Data needs analysis
  
- **DRFD: Behavioral Research**
  - Research planning
  - Data management planning
  - Data needs analysis

- **DRFD: Research Administration**
  - Research planning

- **Information Technology**
  - Storage

Contact Info

- **HBS Working Knowledge**
  - Alex Caracuzzo, acaracuzzo@hbs.edu
  - 617.495.2492

- **HBS Archives**
  - Rachel Wilse
  - rwilse@hbs.edu
  - 617.495.6822

- **Data deposit and publishing services**
  - Katherine McNell
  - kmcnell@hbs.edu
  - 617.495.5310

- **Research Information System**
  - rinfohelp@hbs.edu
  - 617.495.6600

Harvard Business School
Baker Library | Division of Research and Faculty Development | Information Technology
In summary...

• Plan your trajectory using the DMP Checklist
• Think carefully about data collection, including security & legal documents
• There are many options for storage, security, and analysis – choose thoughtfully & wisely
• Data dissemination and preservation are important considerations throughout all phases of your work

• Please reach out to RCS, KLS/BRS, & RSS if you have any questions. And please reach out sooner.
• We wish you success in your research