Data management tools and practices

Anna Dabrowski

LIBRARIES
TEXAS A&M UNIVERSITY

Goals

Facilitate a conversation about data management;

Provide you with recommendations for good practices;

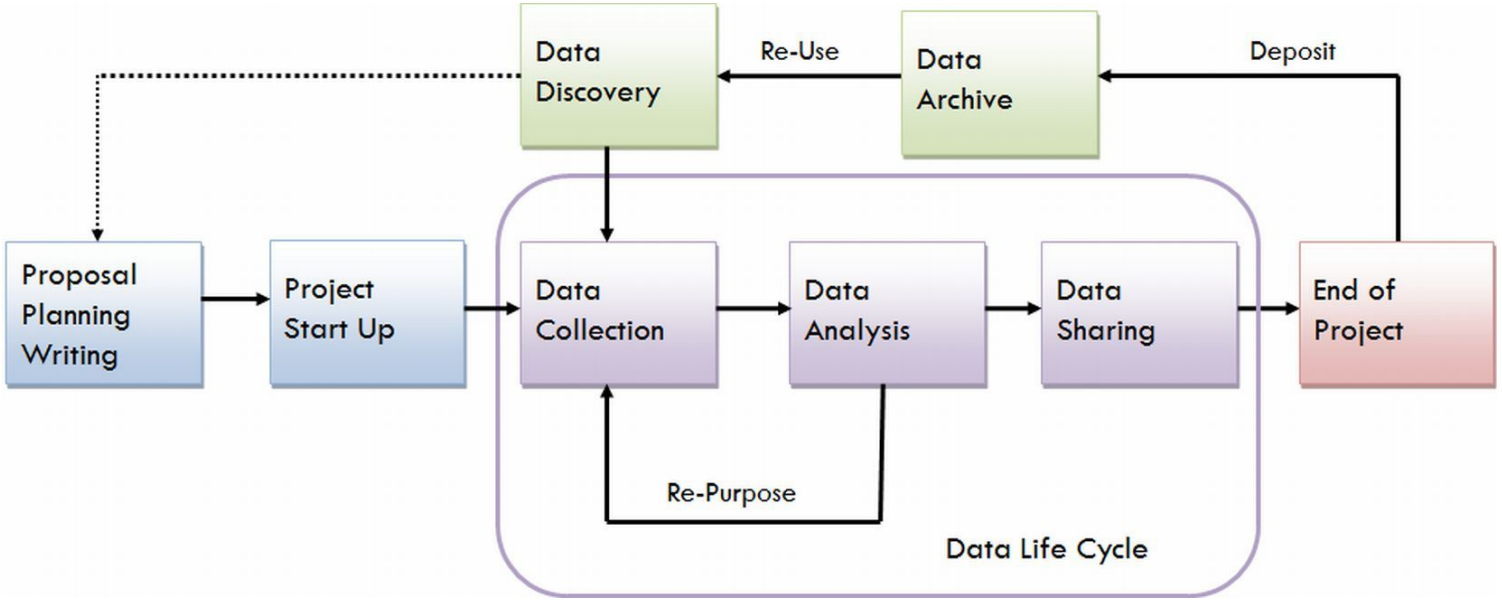
Introduce tools that may benefit you individually and as a lab.

Outline

2-hour discussion

- The data lifecycle
- Data practices for collaborative work
  - File naming
  - Version control
  - Storage and sharing
- Thinking about the long-term and new requirements
  - Documenting data and processes
  - All about repositories
  - Data management plans

The data lifecycle



University of Virginia Libraries. Steps in the Data Life Cycle.
http://data.library.virginia.edu/data-management/lifecycle/

Data practices for collaborative work

Do you have a way of naming your files?

How do you keep track of changes and updates to files?

Where do you store your files?

File naming

Use descriptive information that will help you identify the content within files when you:

- search,
- browse,
- sort.

Especially valuable when you have loads of data, when additional context is lost or missing, and next year—when your future self forgets what you're doing now.

File naming: Useful components

Component	Use cases	Tip
Name or acronym	Creator, project, team, named data	Relevant and simple
Sequential #	Run of experiment, version number	Use leading zeros
Date and time	Creation, range of experiment	YYYY-MM-DD (ISO 8601)
Identifier	Subject, project, grant	Relevant and simple
Research condition	Instrument, temperature, model	Relevant and simple
Type or keyword	Denote type of content in a file	Use a standard list
Extension	Denote file format	Use them

File naming: Putting components together

Consistently use the same components, in the same order.

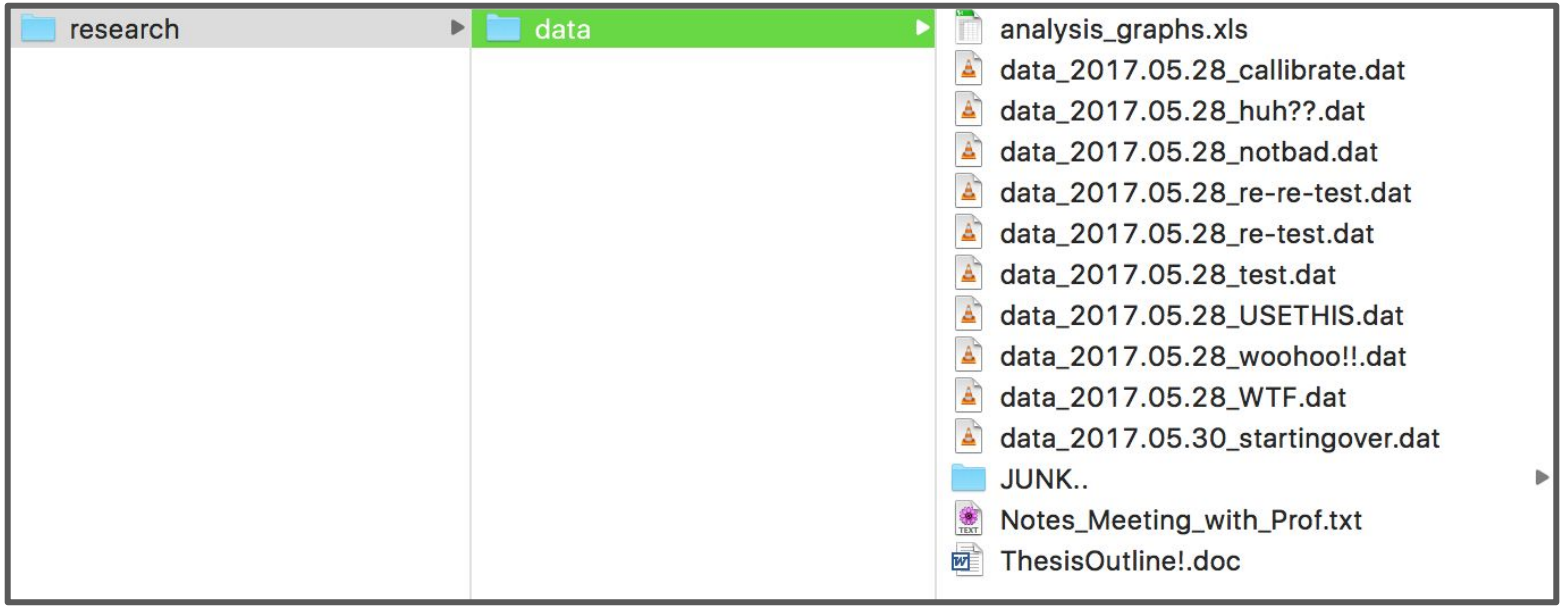
Use less than 32 characters, enough to uniquely identify files.

Avoid special characters like & * % $ £] { ! @).

Separate components with a -d-a-s-h- or _u_n_d_e_r_s_c_o_r_e_.

Include file extensions, after a .p.e.r.i.o.d.

File naming: This is not a naming scheme



Version control

Ensure what you do is transparent and reversible.

Track your progress.

Improve workflow reproducibility.

Reduce work when reconstructing how and why changes were made.

Version control: Quick and dirty

Make copies of files as you work, and adjust the file names sequentially.

> “-v[##]” as a component of file names: “-v00” , “-v01”, “-v02”.

Requires you to:

- Separately document changes between versions.
- Decide how to manage storage, and when to delete old versions.

Version control: Using a system

Implement a Version Control System (VCS) to:

- Automatically track changes within files.
- Progressively document changes in a single place.

Requires investment in software:

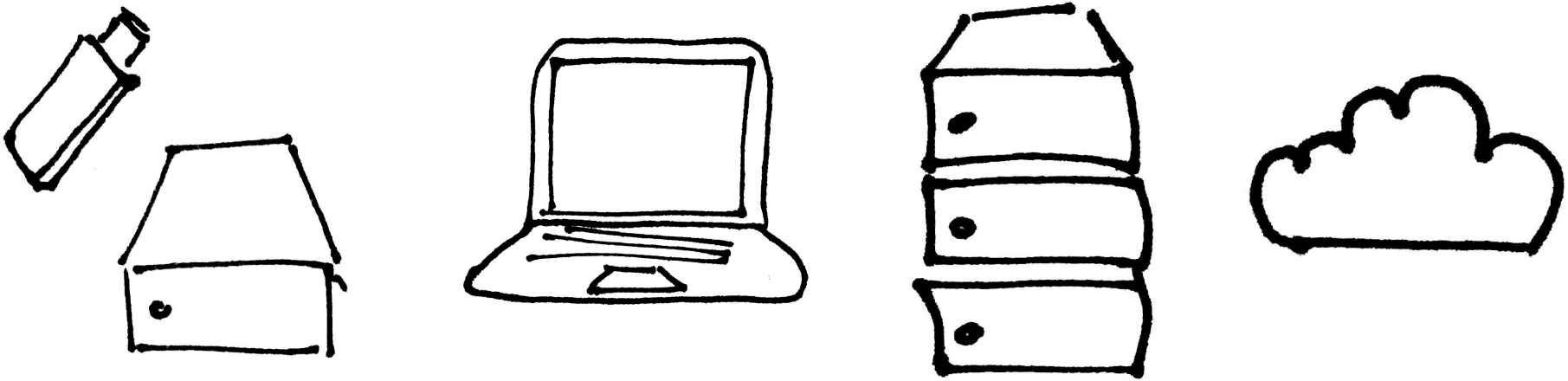
- **git** - Distributed Version Control System. Also integrates with services like Github and Bitbucket. (https://www.git-scm.com)
- **Mercurial** - Distributed Version Control System. Simpler than git. (https://www.mercurial-scm.org)
- **Subversion** - Centralized Version Control System. Older. (https://subversion.apache.org)

Demo

+

Storage and sharing

Consider the accessibility, available space, security, and integrity of storage locations.



Storage and sharing: Backup

Backup is all about redundancy for when things go wrong.

3 Copies

2 Geographic locations

1 Remote copy

Original, external copy kept locally, external copy kept remotely.

Software can help: Time Machine, Arq, Backblaze

Storage and sharing: Making a choice

FileX - TAMU-supported file transfer tool. (http://filex.tamu.edu)

Network Attached Storage - Managed by lab.

Department Server - Managed by local IT staff.

Dropbox - Cloud storage option with desktop syncing.

Amazon S3 - Large-scale cloud storage option.

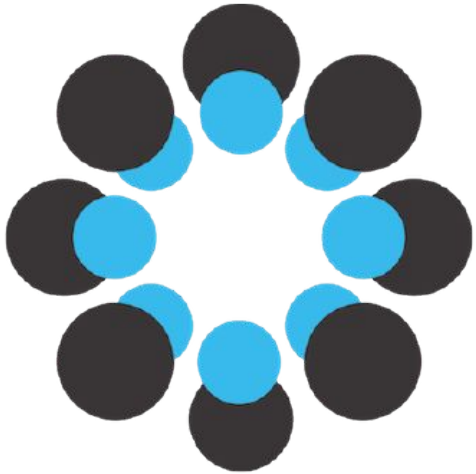
Google Drive - TAMU-supported cloud storage with Team Drives available. No sensitive data. (https://google.tamu.edu)

Syncplicity - TAMU-supported secure cloud storage with desktop syncing. (https://tamu.syncplicity.com)

Github - TAMU-supported collaborative coding and storage platform based on git. (https://github.tamu.edu)

Open Science Framework (OSF) - General-purpose collaborative platform with storage integration. (http://osf.io)

Demo



Open Science Framework

Data practices for collaborative work

Strasser, C., Cook, R., Michener, W., Budden, A. (2012). Primer on Data Management: What you always wanted to know. https://doi.org/10.5060/D2251G48

Git. Getting Started - About Version Control. https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

Software Carpentry. Version Control with Git. http://swcarpentry.github.io/git-novice/

Center for Open Science. OSF 101. https://youtu.be/ENz8OkUK40U

Borer, E. T., Seabloom, E. W., Jones, M. B., Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America, 90: 205–214. https://doi.org/10.1890/0012-9623-90.2.205

Thinking about the long-term and new requirements

Data that are

Findable

Accessible

Interoperable

Reusable

by others

Force 11. FAIR Data Principles.
https://www.force11.org/group/fairgroup/fairprinciples

Documenting data and processing: In writing

Data dictionaries - Give a detailed description of the data at the level of each variable within a dataset. May be contained within a README file.

README files - Describe a collection of data and how the digital files and data within them are organized.

Electronic lab notebooks (ELN) - Allow for collaborative work and sharing notes and responsibilities in a single place.

Demo

README file template

https://cornell.app.box.com/v/ReadmeTemplate

Cornell University. Guide to writing "readme" style metadata.

https://data.research.cornell.edu/content/readme

Documenting data and processing: In workflows

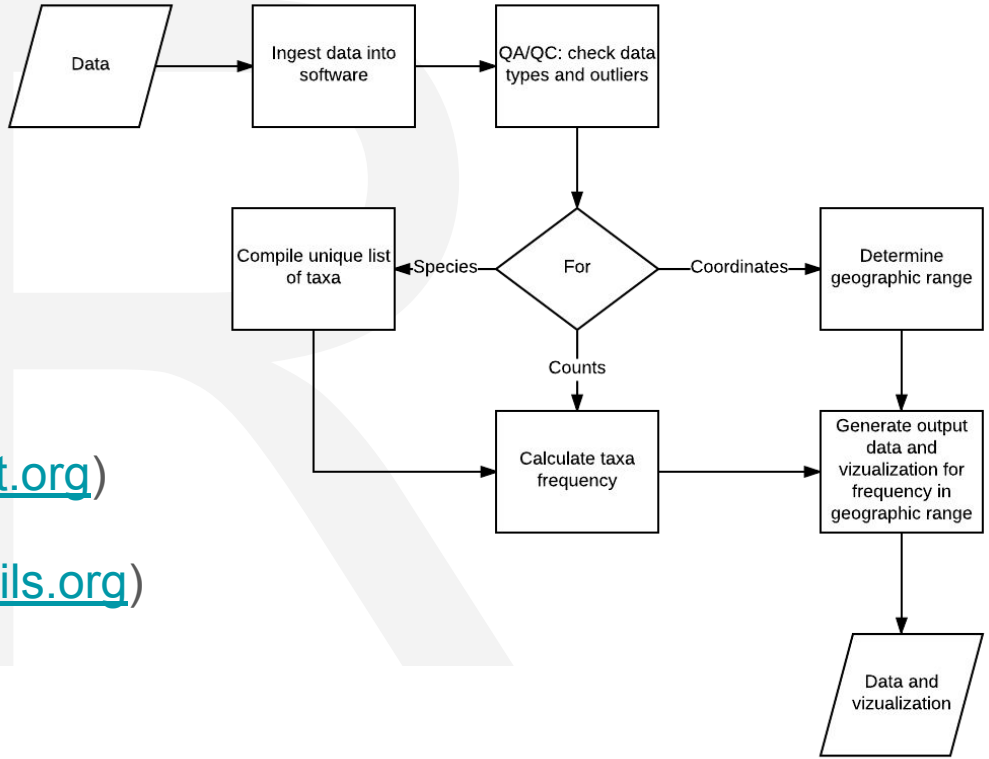
Flowcharts

Lucidchart

Scientific workflow applications

Kepler (https://kepler-project.org)

VisTrails (https://www.vistrails.org)



All about repositories

Provide a means of preserving and openly sharing (publishing) data online. May also be called *data centers*, *data archives*, or *scientific databases*.

They are often divided into three categories:

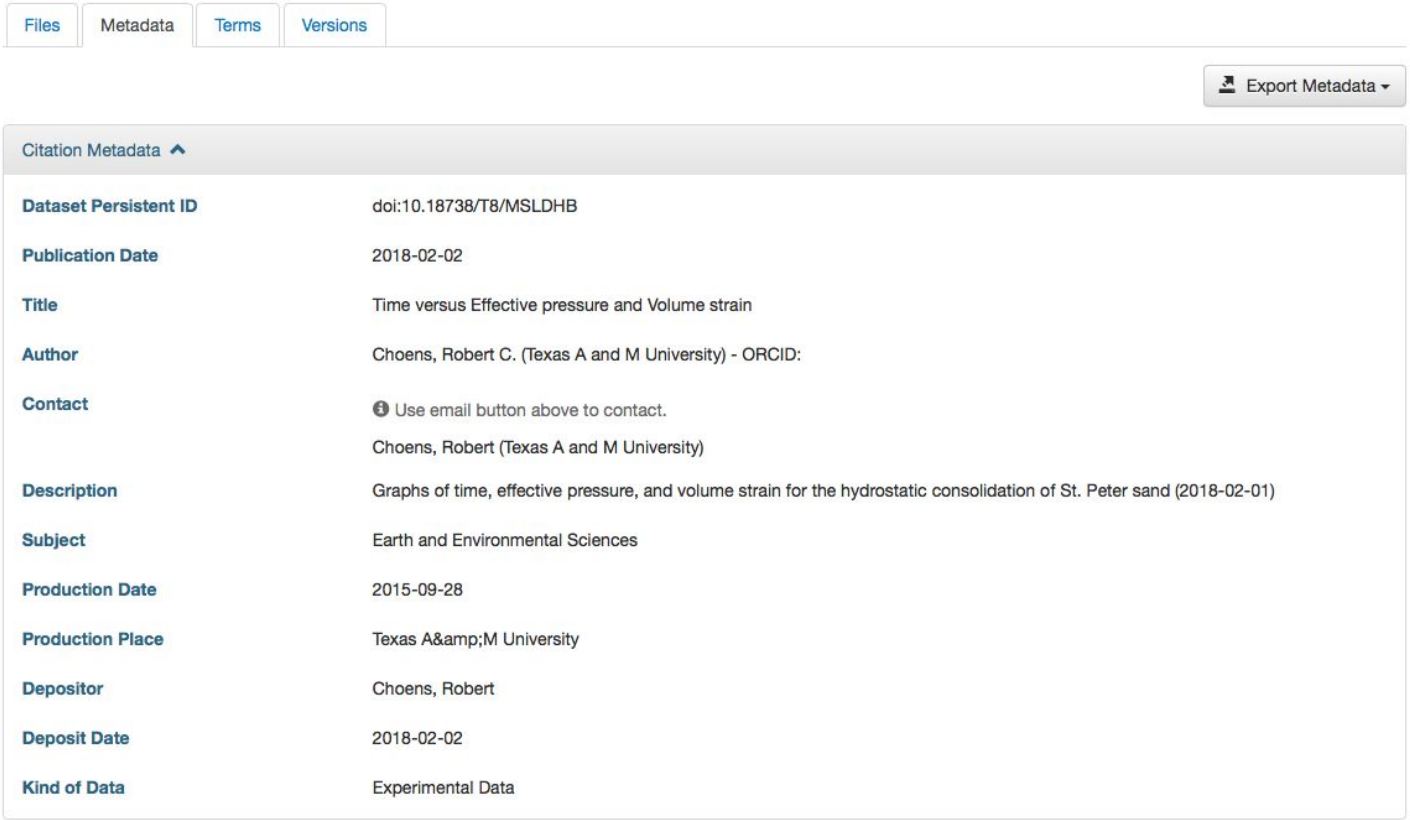
- **Institutional Repositories** (IRs) - Affiliated with a researcher's institution.
- **Disciplinary Repositories** (DRs) - Discipline-specific and often operated by a professional organization, a consortium of researchers, or a similar group.
- **General-purpose** or **Open Repositories** (ORs) - Make data available regardless of disciplinary or institutional affiliation.

All about repositories: Metadata

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

Metadata are formatted into a record according to one of many standards, and attached to a resource.

All about repositories: Metadata



All about repositories: File formats

From the data lifecycle perspective, think about:

- **Interoperability** - Usable with different software tools.
- **Preservation** - Can be opened 10 or more years later.

Features of formats that last:

1. In common usage by the research community.
2. Non-proprietary.
3. Have documented standards.
4. Uncompressed (space permitting).

All about repositories: File formats

Content	File formats
Text	PDF/A, HTML, XML, TXT
Tabular data	XML, CSV
Numbers and statistics	TXT, DTA, POR, SAS, SAV
Geospatial	SHP, DBF, GeoTIFF, NetCDF
Audio	WAVE, AIFF, MP3, MXF
Images	TIFF, JPG, JP2, PDF, PNG, GIF, BMP
Moving Images	MOV, MPEG-4, AVI, MXF
Web Archive	WARC
Containers	TAR, GZIP, ZIP

Repositories: Using them

Texas Data Repository

http://data.tdl.org/about/

Registry of Research Data Repositories

http://www.re3data.org

Nature - Recommended data repositories

https://www.nature.com/sdata/policies/repositories

Demo

Texas Data Repository

Data management plans

Often called DMPs, they describe the data management activities for a research group, researcher, or research project.

They are structured documents created in advance of collecting data, and describe practices that will be followed over the course of the data lifecycle.

Data management plans: For funders

For the purpose of research grants, DMPs are limited in scope. They describe the practices researchers intend to follow for a particular project.

In approximately two pages of text, researchers are asked to outline how they intend to ensure that data are well maintained, shared, and reusable in the future.

Find article and data sharing requirements for each federal agency with **SPARC** (http://researchsharing.sparcopen.org)

Write data management plans with the **DMPTool** (https://dmptool.org)

Find guidance and more information (https://tamu.libguides.com/research-data-management)

Example

DMP + Checklist

Data management plans: For researchers

Living documents that describe overall data management practices in detail, and are refined over time.

Allow you to:

- Identify relevant practices that improve research efficacy.
- Plan in advance to reduce later costs in time and effort.
- Easily share an overview with collaborators and stakeholders.
- Make incremental changes to meet the needs of their research community.
- Speed up the grant application process over multiple cycles.

Thinking about the long-term and new requirements

Carroll, M. W. (2015). Sharing Research Data and Intellectual Property Law: A Primer. PLoS Biology. http://dx.doi.org/10.1371/journal.pbio.1002235

Costello, M. J., Wieczorek J. (2014). Best practice for biodiversity data management and publication. Biological Conservation. 173: 68–73. http://dx.doi.org/10.1016/j.biocon.2013.10.018

DataONE. Document and store data using stable file formats. http://www.dataone.org/best-practices/document-and-store-data-using-stable-file-formats

Earth Science Information Partners (ESIP). Data Management Short Course for Scientists. http://commons.esipfed.org/datamanagementshortcourse

Michener, W. K. (2015), Ecological data sharing. Ecological Informatics, 29, 1: 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

Further references

Leonelli, S., Davey, R. P., Arnaud, E., Parry, G., Bastow, R. (2017) Data management and best practice for plant science. Nature Plants. http://dx.doi.org/10.1038/nplants.2017.86

NISO. Understanding Metadata: What is metadata and what is it for?
http://www.niso.org/publications/understanding-metadata-riley

The Gurdon Institute. Electronic Lab Notebooks - for prospective users.
https://www.gurdon.cam.ac.uk/institute-life/computing/elnguidance

Veerle Van den Eynden, V., Corti, L., Woollard, M., Bishop, L., Horton, L. (2011). Managing and Sharing Data: A Best Practice Guide for Researchers.
http://www.data-archive.ac.uk/media/2894/managingsharing.pdf