The common sense approach to managing research data
How to avoid crying over messy datasets: 7 TIPS

Pauline Lam
Scholarly Communications Librarian | Hong Kong Baptist University Library
30 April 2018
“Data Intensive science… must move from data to information to knowledge.”

*The Fourth Paradigm: Data-Intensive Scientific Discovery*
Tony Hey, Microsoft

“Organising data in a presentable and useful way was the most often stated reason for not sharing data.”

*Practical Challenges for Researchers in Data Sharing white paper*
Springer Nature

“It’s not about technology, it’s about understanding the value to society.”

*Scottish Information Commissioner*
Rosemary Agnew, Scottish Information Commissioner

“60% of our time is spent on data clean up.”

*Data Studio*
HK Science & Technology Park

“It’s not just about open or close, it is about control.”

*Is Figshare open?*
Mark Hehnel, Founder of Figshare
Global efforts in RDM

Data initiatives around the world

RDA
Research Data Alliance

DDC
Digital Curation Centre

HKU
ANDS
Hong Kong University Libraries

ICPSR
Inter-University Consortium for Political & Social Research

RDA

ICPSR

DCC

ICPSR

RDA

DDC

HKU

ANDS

Research Data Services

Australian National Data Service
What is RDM?

Name the problem so that you can solve it!
Data Lifecycle
American Geophysical Union
Data Services | AGU
Data Lifecycle

Australian National Data Service
ANDS

Research Data Lifecycle

- Research data planning & design
- Data collection
- Data processing
- Data study & analysis
- Data publishing & access
- Data reuse (impact)
- Data preservation

Experiment / project:
- Dataset metadata
- Appraise, select, dispose of - data
- File formats, naming conventions
- Research output - research data

Long-term data storage & data security

Short-term data storage, data security & file sharing

Data curation
Data sharing
Access rules, e.g. open access, mediated access
Metadata services
Discovery services
Data citations
Impact metrics

- Literature review
- RDMP
- Collaboration tools
- Citation management tools
- Ethics & compliance
- IP (Copyright) & licensing

Where to publish data
Funders’ & publishers’ data access policies
Author rights
Data licensing
Persistent identifiers - ORCIDs, DOIs

Planning is Investing

A DMP helps you:
- think through objectives
- meet funding requirements
- follow through changes
- facilitate sharing
- ensure data longevity

“the PI should assess data archive potential and opportunities for data sharing. Due additional weight will be given to an application where the applicants are willing to make research data available to others.”

Research Grants Council (RGC) revised General Research Fund (GRF) Explanatory Notes for 2016/17 applications
SHERPA Juliet:
- funder policies database
- single focal point
- searchable
- open access publications

Planning is Investing

SHERPA Juliet is a searchable database and single focal point of up-to-date information concerning funders’ policies and their requirements on open access, publication and data archiving. Read more...

Open access services from Jisc
Services to support open access

SHERPA Services
Helping authors and institutions make informed and confident decisions in open access publication and compliance

Managing open access costs
A guide from Jisc

v2.sherpa.ac.uk/juliet
Planning is Investing

HKBU Template & Tools

HKBU DMP Template

digital.lib.hkbu.edu.hk/digital/MDP_template%20FINAL.pdf

dmponline.dcc.ac.uk

dmptool.org
Manage your metadata too

So that your data can be:
- shared
- accessed
- interpreted
- reused

Lab notebook  Data dictionary  Readme txt files

A Data Dictionary can:
- clarify meaning
- build relationships to other data
- trace origin
- facilitate usage
- specify format

Survey Details

ACS 2006 (1-Year Estimates)

The American Community Survey (ACS) is a new nationwide survey designed to provide communities a fresh look at how they are changing. It is a critical element in the Census Bureau’s reengineered decennial census program. The ACS collects population and housing information every year instead of every ten years, which provides more up-to-date information throughout the decade about the U.S. population at the local community level. The ACS collects information such as age, race, income, commute time to work, home value, veteran status, and other important data, and is comparable to Summary File 3 in Census 2000.

The 2006 data products include estimates of demographic, social, and economic characteristics of people, households and housing units (both occupied and vacant) for every state in the Nation, Puerto Rico, and most areas with a population of 65,000 or more.

About three million housing unit addresses are selected annually, from across every county in the nation.

Datasets

Social Explorer Tables: ACS 2006 (1-Year Estimates)
American Community Survey Tables: 2006 (1-Year Estimates)

Documentation

socialexplorer.com/data/ACS2006/metadata
Facilitate sharing by:

- using common terms
- referencing subject-based knowledge models

Use standard terminology

TIP 3

GENERAL LINGUISTICS TAXONOMY

© 6TH OCTOBER 2014 ▸ CHRISTIAN KOHL

Short Description:
The General Linguistics Taxonomy covers all major areas of linguistics and communication science. It includes precise definitions, in-depth explanations, and synonyms and/or antonyms where applicable. Languages: English and German.

URL:
The model is not currently available externally.

Owned/Developed by:
1. Name of Owner: De Gruyter
2. Name of Developer: De Gruyter
3. Technical Contact: Bettina de Kelzner, “protected email”
4. License Contact: Not applicable

How is this KM applied?
With auto-tagging software (Lumid/TEMIS).

How is this KM used?
Formatting for data quality

Aim to be:
- logical
- methodical
- consistent
- responsive
- diligent!

Proposed folder arrangement and file naming practice for a PhD project in Earth System Science

Wageningen University & Research
Formatting for data quality

File names should be:
- descriptive
- no space, CamelCase, use _ or –
- use meaningful abbreviations
- use plain ascii text

TIP 4

TILS Document Naming Convention

Document naming for the TILS Division should follow this convention:

GDL_TILSDocNaming_V1_20090612.docx

- **Prefix** shows the document type
- The **document title** describes the content
- The **version number**
- The **date** in the format yyyymmdd

File names created from the TILS document naming convention are made up of four parts joined together with an underscore character (_). There should **not be any spaces** in the file name.
Formatting for data quality

- prevention NOT salvation
- exercise quality control
- maintain data consistencies
- data cleaning can be painful and should be last resort

**A**

Untidy Data

<table>
<thead>
<tr>
<th>species</th>
<th>habitat</th>
<th>weight</th>
<th>length</th>
<th>latitude/longitude</th>
<th>date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alligator mississippiensis</td>
<td>swamp</td>
<td>431 lb</td>
<td>4 ft 2</td>
<td>29.531,-82.184</td>
<td>Sept 15, 2015</td>
</tr>
<tr>
<td>Puma concolor</td>
<td>forest</td>
<td>125 lb</td>
<td>2.2m</td>
<td>29.125,-81.682</td>
<td>08/10/2015</td>
</tr>
<tr>
<td>Ursus americanus</td>
<td>forest</td>
<td>88 kg</td>
<td>133 cm</td>
<td>N29°7'30&quot;/W81°40'55.2&quot;</td>
<td>07-13-2015</td>
</tr>
</tbody>
</table>

**B**

Tidy Data

<table>
<thead>
<tr>
<th>species_code</th>
<th>date</th>
<th>station_code</th>
<th>weight_kg</th>
<th>length_cm</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN 551771</td>
<td>2015-09-15</td>
<td>1</td>
<td>196</td>
<td>127</td>
</tr>
<tr>
<td>TSN 55247</td>
<td>2015-08-10</td>
<td>2</td>
<td>57</td>
<td>220</td>
</tr>
<tr>
<td>TSN 180544</td>
<td>2015-07-13</td>
<td>2</td>
<td>88</td>
<td>133</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>station_code</th>
<th>habitat</th>
<th>latitude</th>
<th>longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>swamp</td>
<td>29.531</td>
<td>-82.184</td>
</tr>
<tr>
<td>2</td>
<td>forest</td>
<td>29.125</td>
<td>-81.682</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>species_code</th>
<th>class</th>
<th>genus</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN 551771</td>
<td>Reptilia</td>
<td>Alligator</td>
<td>mississippiensis</td>
</tr>
<tr>
<td>TSN 55247</td>
<td>Mammalia</td>
<td>Puma</td>
<td>concolor</td>
</tr>
<tr>
<td>TSN 180544</td>
<td>Mammalia</td>
<td>Ursus</td>
<td>americanus</td>
</tr>
</tbody>
</table>
Some preferred file formats

- Containers: TAR, GZIP, ZIP
- Databases: XML, CSV
- Geospatial: SHP, DBF, GeoTIFF, NetCDF
- Moving images: MOV, MPEG, AVI, MXF
- Sounds: WAVE, AIFF, MP3, MXF
- Statistics: ASCII, DTA, POR, SAS, SAV
- Still images: TIFF, JPEG 2000, PDF, PNG, GIF, BMP
- Tabular data: CSV
- Text: XML, PDF/A, HTML, ASCII, UTF-8
- Web archive: WARC

Source: Stanford Libraries | Best practices for file formats
Share your data

Share your data to:
- gain impact
- enhance your reputation
- increase chance of funding
- archive your materials
- share what you can share

**EXPLAIN IT**
- contextualise your material and date
  
  *Describe the circumstances prevailing at the time of your research and the parameters within which you were working.*

- describe your research process
  
  *Help people understand your material and data in the future by explaining why you used a particular methodology, or how you analysed your data.*

- explain acronyms and jargon
  
  *Don’t assume the reader will understand specialist terms - remember they may be reading your material in several years’ time.*

- provide information (sometimes called metadata) about each file
  
  *This will help a preservation service to index your material and people to find it. Some of this might be generated automatically by the digital equipment you use.*

**STORE IT SAFELY**
- make multiple copies
  
  *Use different types of storage media and store copies in different locations.*

- use open file formats where possible
  
  *Choosing non-proprietary formats means that files are more likely to be readable in the future. Your library or preservation service should be able to advise you on suitable formats.*

- control who can access your files
  
  *Take particular care about how you handle and store sensitive information.*

- decide when to delete digital materials and date
  
  *Be selective about what you keep so that it is easier to find relevant and useful information.*

**SHARE IT**
- to gain more impact
  
  *Other researchers - in your field or in different disciplines - may want to make use of your material, review it in the future.*

- to enhance your reputation
  
  *Making research available allows you to demonstrate research excellence, increases your credibility and can lead to collaborations.*

- to increase the chance of funding
  
  *Most funding agencies respond positively to you making your material and data available to others.*

- use repositories and data centres for archiving your materials
  
  *Consider making your research openly available. Choose a repository with controlled access if this is more appropriate for your research.*

- react or embargo when necessary
  
  *Your material can still have value when personal or confidential information is removed, and most preservation services will embargo your material while you wait for publications or patents.*

---

[shard-jisc.blogspot.hk/2012/07/here-is-full-leaflet-introducing.html](http://shard-jisc.blogspot.hk/2012/07/here-is-full-leaflet-introducing.html)
**RESEARCHER DATA SHARING INSIGHTS**

- Wiley’s Researcher Data Insights Survey was launched earlier this year to understand how and why researchers make their research data publicly available. The study’s results, highlighted below, are intended to advance the global conversation about data sharing and help Wiley better meet the needs of our researchers, authors, and partners in the rapidly evolving landscape of scientific research and communications.

- The survey was deployed in March 2014 and received more than 2,250 responses from researchers around the world.

**GLOBAL DATA SHARING TRENDS**

Data sharing practices vary widely across research fields and geographic areas. Just over half of researchers report making their data publicly available, though archiving results in Repositories is not yet the norm.

**WAYS DATA IS SHARED**

- 47% As supplementary material in a journal paper
- 53% Personal, institutional or project webpage
- 28% Institutional data repository
- 13% University or institute repository
- 19% Discipline-specific data repository
- 2% General-purpose data repository
- 7% Other

- 48% SHARE
- 52% NOT SHARE

**RESEARCHER MOTIVATIONS FOR SHARING DATA**

- 36% Transparency and quality
- 31% Institution requirements
- 21% Personal interest
- 19% Accountability
- 17% Publication
- 16% Promotion
- 15% Improving career
- 14% Research efficiency
- 13% Preserving history
- 11% Replication
- 10% Publication in a journal
- 9% Journal requirements
- 8% Transparency
- 7% Institutional reputation
- 6% Publication
- 5% Transparency
- 5% Publication
- 4% Transparency
- 4% Publication
- 3% Transparency
- 3% Publication
- 2% Transparency
- 2% Publication
- 2% Transparency
- 2% Publication
- 1% Other

**DATA SHARING TRENDS BY COUNTRY**

- **United States**: 46% SHARE, 54% NOT SHARE
- **United Kingdom**: 43% SHARE, 57% NOT SHARE
- **Japan**: 44% SHARE, 56% NOT SHARE
- **China**: 36% SHARE, 64% NOT SHARE
- **Brazil**: 52% SHARE, 48% NOT SHARE
- **Australia**: 41% SHARE, 59% NOT SHARE
- **Germany**: 55% SHARE, 45% NOT SHARE

**DATA SHARING BY DISCIPLINE**

- **Natural Sciences**: 48% SHARE, 52% NOT SHARE
- **Life Sciences**: 44% SHARE, 56% NOT SHARE
- **Social Sciences**: 28% SHARE, 72% NOT SHARE
- **Physical Sciences**: 35% SHARE, 65% NOT SHARE
- **Humanities**: 20% SHARE, 80% NOT SHARE

**REASONS WHY RESEARCHERS ARE HESITANT TO SHARE THEIR DATA**

- 43% Intellectual property or confidentiality issues
- 36% My funding/institution does not require data sharing
- 26% I am concerned that my research will be stolen
- 24% I am concerned about misappropriation or misuse of my data
- 22% I am concerned about being given proper credit or attribution
- 21% I did not know where to share my data
- 20% Insufficient time and/or resources
- 16% I did not know how to share my data
- 12% I am not sure if it is my responsibility
- 12% I did not consider the data to be shareable
- 11% Lack of funding
- 7% Other

**TIP 5**

Share your data

To share or not to share? That is the (research data) question...

*Scholarly Kitchen*
Communication Volumes in Social Relationships in Online Communication Systems

Version 3 ▼ 26.05.2016, 17:01 by Masanori Takano

These files contain the sending messages behavior. The columns of the files show row numbers, message sender’s user id (uid1), message recipient’s user id (uid2), the number of characters (cv), and date index (i). Each row shows each day data of the combination of uid1 and uid2. uid1 was determined based on the different rules of uid2.

Please cite the following paper when using the dataset.

REFERENCES
- http://www.palgrave-journals.com/articles/palcomms201714

Log in to write your comment here...
Share your data

datacite.org
Retention and disposal of research data

Why do I need to identify how long my research data needs to be retained?

TIP 6

Future-proof your data

Set up:
- archive rules
- destroy rules
- procedures to manage archived data

monash.edu/records-archives/records-management/resources/research/retention-and-disposal-of-research-data
It's not about technology

“It’s about understanding the value to society.”

Rosemary Andrew
Scottish Information Commissioner
How the Library can help...

- **Research Data Services**: [Digital Library](digital.lib.hkbu.edu.hk/digital/SCS.php)
- **RDM Library Guide**: [Library Guide](hkbu.libguides.com/rdm)
- **Data Analysis Room**: [L4 MLC Library](library.hkbu.edu.hk/forms/dar.html)
THANK YOU!