Data Management

An Introduction

Leslie Barnes, Digital Scholarship Librarian and Dylanne Dearborn, Research Data Librarian

Why Should You Care?

Scholarly Reasons

- Replicability
- Findability
- Accountability

Emerging Policies

- Research ethics
- Publisher requirements
- Funding requirements

Publisher Sharing Policy

– – –

"[…] a condition of publication in a Nature journal is that **authors are required to make materials, data and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission."

nature publishing group npg

Image credit: http://www.nature.com/authors/policies/availability.html

Data Requirements - Canadian Funding Bodies

Preservation

- Social Science and Humanities Research Council (SSHRC), 2 years
- Canadian Institutes of Health Research (CIHR), 5 years

Sharing

- CIHR, must “deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database”

Management

- Data management supports other operations
- Requirements in Canada tbd…

Increasing Research Impact

- Data as an output
- Citable
- Metrics
- Facilitate innovation

Perhaps most importantly

Your own sanity...

Data and the Data Lifecycle

Research Data Lifecycle

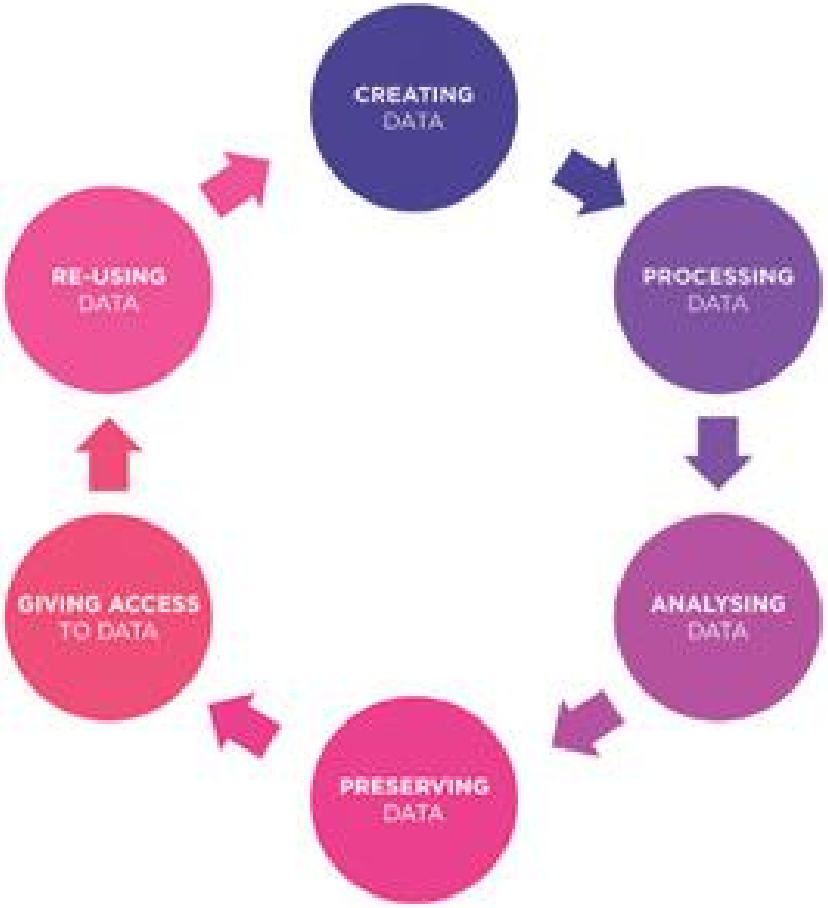
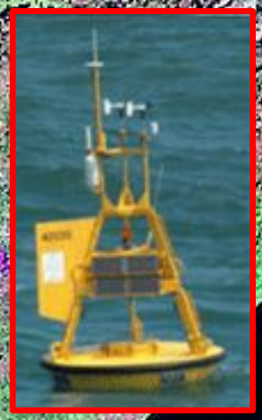
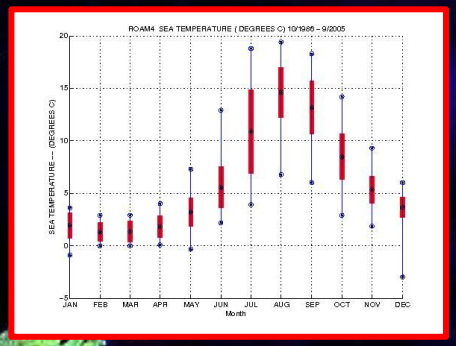


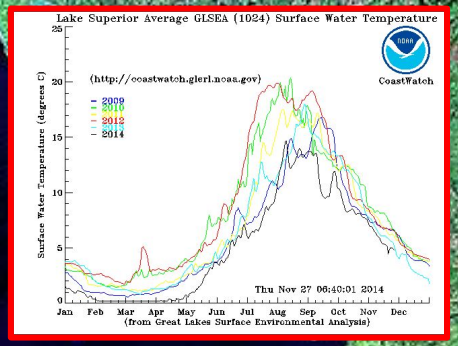
Image credits: http://www.cla.ca/feliciter/2014/2/mobile/data2.jpg

Data moving through the lifecycle



#YY	MM	DD	hh	mm	WDIR	WSPD	GST	WVHT	DPD	APD	MWD	PRES	ATMP	WTMP	DEWP	VIS	TIDE
#yr	mo	dy	hr	mn	degT	m/s	m/s	m	sec	sec	degT	hPa	degC	degC	degC	mi	ft
2013	01	01	00	00	315	13.7	15.7	99.00	99.00	99.00	999	1018.3	-9.0	999.0	-12.4	99.0	99.00
2013	01	01	01	00	310	12.6	14.4	99.00	99.00	99.00	999	1018.6	-8.4	999.0	-15.8	99.0	99.00
2013	01	01	02	00	328	12.3	14.0	99.00	99.00	99.00	999	1018.9	-8.3	999.0	-11.9	99.0	99.00
2013	01	01	03	00	315	11.0	13.2	99.00	99.00	99.00	999	1019.0	-8.4	999.0	-15.3	99.0	99.00
2013	01	01	04	00	328	11.0	13.3	99.00	99.00	99.00	999	1019.1	-8.7	999.0	-15.3	99.0	99.00
2013	01	01	05	00	319	11.8	13.1	99.00	99.00	99.00	999	1019.4	-9.0	999.0	-12.7	99.0	99.00
2013	01	01	06	00	327	9.8	11.6	99.00	99.00	99.00	999	1019.4	-9.2	999.0	-13.9	99.0	99.00
2013	01	01	07	00	325	9.5	11.4	99.00	99.00	99.00	999	1019.5	-9.1	999.0	-14.5	99.0	99.00
2013	01	01	08	00	316	7.4	9.2	99.00	99.00	99.00	999	1019.6	-8.6	999.0	-13.0	99.0	99.00
2013	01	01	09	00	317	9.3	13.3	99.00	99.00	99.00	999	1019.7	-8.2	999.0	-13.3	99.0	99.00
2013	01	01	10	00	299	11.6	13.0	99.00	99.00	99.00	999	1019.6	-8.0	999.0	-12.6	99.0	99.00
2013	01	01	11	00	284	9.7	12.6	99.00	99.00	99.00	999	1019.7	-8.1	999.0	-12.0	99.0	99.00
2013	01	01	12	00	294	9.2	10.4	99.00	99.00	99.00	999	1019.9	-8.0	999.0	-12.7	99.0	99.00
2013	01	01	13	00	301	10.5	11.9	99.00	99.00	99.00	999	1019.8	-8.0	999.0	-12.6	99.0	99.00
2013	01	01	14	00	285	11.8	13.6	99.00	99.00	99.00	999	1019.4	-7.9	999.0	-13.2	99.0	99.00
2013	01	01	15	00	278	10.5	12.6	99.00	99.00	99.00	999	1019.2	-7.8	999.0	-12.5	99.0	99.00
2013	01	01	16	00	285	13.0	15.4	99.00	99.00	99.00	999	1018.5	-7.1	999.0	-12.5	99.0	99.00
2013	01	01	17	00	268	14.5	15.9	99.00	99.00	99.00	999	1017.5	-6.6	999.0	-12.9	99.0	99.00
2013	01	01	18	00	263	14.3	16.1	99.00	99.00	99.00	999	1016.6	-6.7	999.0	-11.3	99.0	99.00
2013	01	01	19	00	253	14.8	16.6	99.00	99.00	99.00	999	1015.6	-6.8	999.0	-10.5	99.0	99.00
2013	01	01	20	00	261	15.0	19.6	99.00	99.00	99.00	999	1014.4	-6.5	999.0	-10.2	99.0	99.00
2013	01	01	21	00	270	16.8	20.0	99.00	99.00	99.00	999	1014.2	-6.5	999.0	-11.0	99.0	99.00
2013	01	01	22	00	263	14.4	18.2	99.00	99.00	99.00	999	1014.8	-6.5	999.0	-10.8	99.0	99.00
2013	01	01	23	00	259	15.5	17.1	99.00	99.00	99.00	999	1014.7	-6.0	999.0	-10.0	99.0	99.00
2013	01	02	00	00	261	15.5	16.9	99.00	99.00	99.00	999	1014.6	-5.9	999.0	-9.6	99.0	99.00
2013	01	02	01	00	254	14.4	17.3	99.00	99.00	99.00	999	1014.1	-5.6	999.0	-9.4	99.0	99.00





Raw **Processed** **Analyzed** **Finalized/Published**

Lake Superior - Temperature Data

Image credits: http://commons.wikimedia.org/wiki/File:Lake_Superior_NASA.jpg, http://www.ndbc.noaa.gov/tour/virtr1.shtml, http://www.ndbc.noaa.gov/view_climplot.php?station=roam4&meas=st, http://www.ndbc.noaa.gov/view_text_file.php?filename=stdm4h2013.txt.gz&dir=data/historical/stdmet/, http://coastwatch.glerl.noaa.gov/statistic/avg-sst.php?lk=s&yr=0

Planning

What is a Data Management Plan (DMP)?

- A short document to state your RDM plans
- Supports management throughout the data lifecycle
- Helps to anticipate RDM challenges in your research project

What is in a DMP

- What types of data will be created?
- Who will own, have access to, and be responsible for managing these data?
- What equipment and methods will be used to capture, process and document the data?
- Where will data be stored during and after research?
- Will the data be shared?

Sample DMP

2 Data and Metadata Standards

Data will be shared in matlab MAT file format and/or as netCDF files. Data quality will be in accord with published uncertainty ranges for each instrument and within error bars for standard processing techniques. These PIs have experience with this mix of data types from previous collaborative efforts. Data responsibilities include:

PI	Responsibility
A. Thurnherr	LADCP-CTD analysis.
L. St. Laurent and E. Shroyer	HRP microstructure analysis
S. Jachec	Ongoing model output prediction
J. Moum, J. Nash	χpod microstructure data
M. Alford, J. Nash, J. MacKinnon	Mooring data

3 Data access and sharing

All field data collected under this program will be made available as per NSF guidelines within 2 years of collection via published manuscripts, publicly available final reports to NSF, and data archiving with NODC. Recognizing that any individual PI server may become unavailable over time, data will be made available by PI website locations and also by specific request to any colleague. Models codes to be employed are all public domain. Published peer-reviewed manuscripts will document the simulations and forcing sufficiently.

4 Data archiving and preservation

Aside from the LADCP-shipboard CTD profiles, there are currently no established standards for archiving or data from many of the fine- and micro-scale sensors used in the proposed work. This is a concern of the Climate Process Team on Ocean Mixing, of which many PIs are members. We propose to work with the CPT to evolve formats for data and metadata suitable for archiving both sensor and (critically) model output from the experiment. Field data will be provided to NODC upon project completion. Ultimate archival formats will be determined in consultation with NODC and with the CPT. Adequate archiving is anticipated to be an expensive, time-consuming task. All PIs have included funds for this effort in their budgets.

Image credit: http://idi.ucsd.edu/_files/DMP%20Example%20Jennifer%20MacKinnon.pdf

Sample DMP

The proposed project will include human subjects data consisting of background demographic information (including age, gender, education, and handedness), reaction times and error rates from computer-based language and working memory tasks, and structural and functional neuroimaging data from MRI scans, with the functional data collected during the performance of the computer-based tasks.

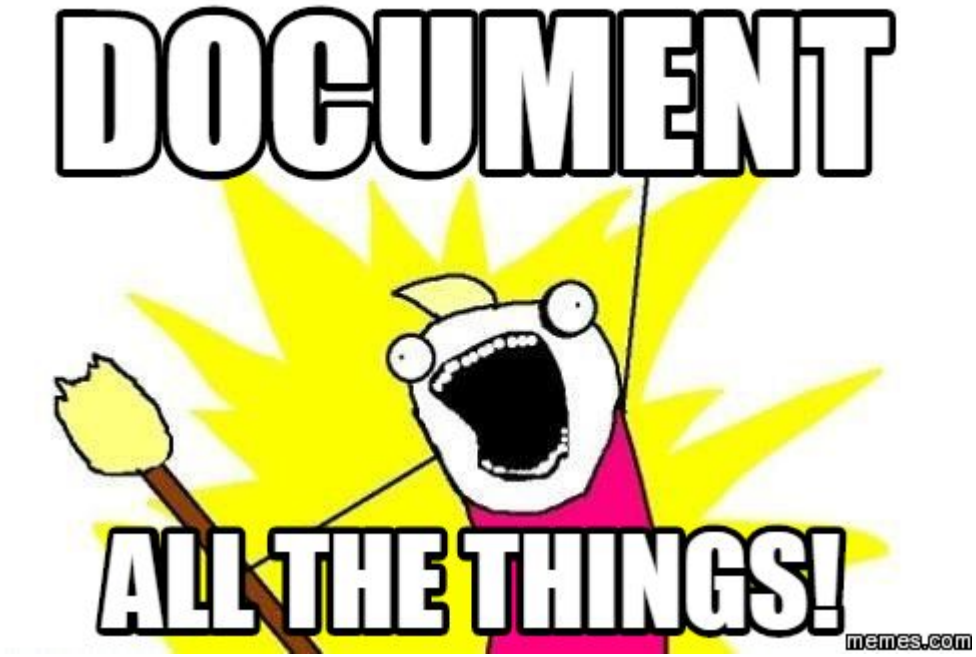
The functional neuroimaging data will be used to determine the brain regions subserving language processing and to determine their overlap with brain regions supporting working memory. The behavioral data will be used to verify that the findings from the computer-based tasks replicate previous findings regarding the conditions affecting the difficulty of sentence comprehension and the capacity of working memory. The demographic data will be needed for published reports to convey the characteristics of the subject population.

The demographic data will be collected from a questionnaire administered by the PI and student/postdoctoral researchers associated with this project, and will be entered into electronic spreadsheets. The data from the computer-based tasks will consist of tab-delimited output from the programs running the tasks. Raw structural and functional magnetic resonance data will be obtained in DICOM format and transferred on DVD or external hard drives to Rice University. DICOM files will be converted to NiFTI format for compatibility with various neuroimaging analysis packages. Analyzed neuroimaging data will be stored in the AFNI file format.

Since these data will be from human subjects, approval for human subjects research will be obtained through the Rice Institutional Review Board. As detailed in the human subjects section of the proposal, because of confidentiality issues, each subject will be assigned an arbitrary code,

Image credit: http://comp.rice.edu/WorkArea/DownloadAsset.aspx?id=518

Staying Organized


DOCUMENT
ALL THE THINGS!
memes.com

No Seriously…

- Who is involved
- What everyone’s role is
- **WHAT YOU HAVE** (keep an inventory)
- Where it is
- What has been done to it
- Keep a copy *with* your data

But...

“I’ll remember!”

But...

“I’ll remember!”

No, you won’t

But...

“I’ll remember!”

No, you won’t.

“It’s self-documenting!”

But...

“I’ll remember!”

No, you won’t.

“It’s self-documenting!”

No, it isn’t.

Documentation continued

- Use version control whenever possible
- If you've run scripts, retain them
  - Indicate what datasets are related to what scripts
- If you've used specialized software, keep a copy
  - Indicate where it is!

Description

- Retain context for your materials
- Explain any codes you've used in your data
- Explain the methods for creation, analysis, etc.
- Describe who created the data
- Who has rights to the data
- Anything else you might find relevant

General Organization

- Create and *use* a directory system
  - Consider what a meaningful principle of organization would be for you
- Create and *use* a naming convention
  - Try to embed human-readable information in file names
  - Don’t use special characters in filenames
  - Think about how you might like to sort files

Retention and Preservation

Retention: can I throw it away?

- You can't keep *everything*
- Be willing to get rid of stuff that you don't need
- Ask yourself:
  - Can this be quickly replicated?
  - Is this important?
  - Is this meaningful or intermediate?
  - What use do I have for this in the future?
  - Will other people need this in the future?
  - **Would you pay for this to be stored?**

Should I keep it?

- Consider REB or funder regulations and policies
  - Sometimes require you to retain data
- Consider publisher policies
- Consider your own future uses
- Consider other future uses

Preservation: Meaningful and Useful Over Time

- Description is key: need to understand the context of your data
- Non-proprietary file types
- If you used proprietary software, keep it with the data or at least document a version

Metadata

Repositories may ask for metadata in a particular standard

- Check with any possible repositories about their requirements
- If a standard is unfamiliar: ask a librarian!
- Lots of domain-specific standards, but they generally ask for descriptive information that provides necessary context

Sharing

Consider Sharing Your Data

- You benefit
- Others benefit
- Discipline benefit
- How to mitigate concerns

Sharing Requirements

- You may be required to share your data to:
  - Meet funding requirements
  - Meet publisher requirements

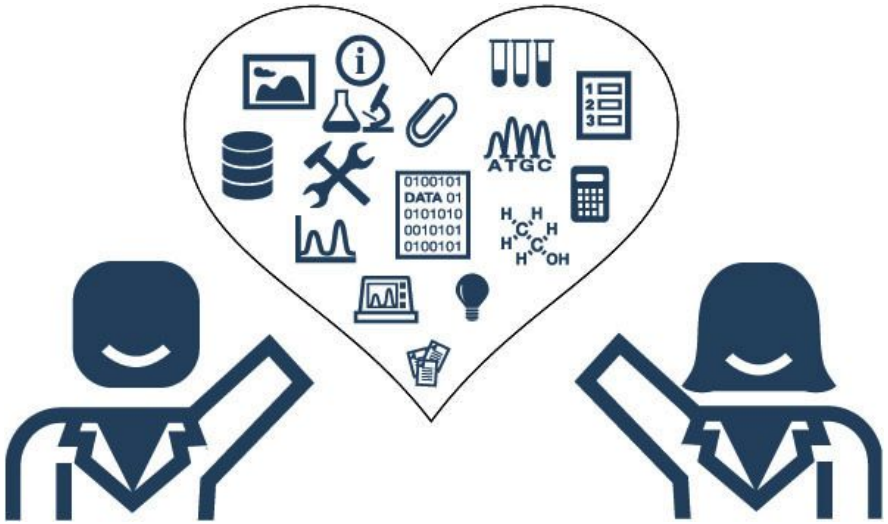


Image credit: https://blog.labfolder.com/data-sharing/

Data as a Publication

- Data can also be a <u>research output</u>
  - put data in a repository
  - publish a data paper

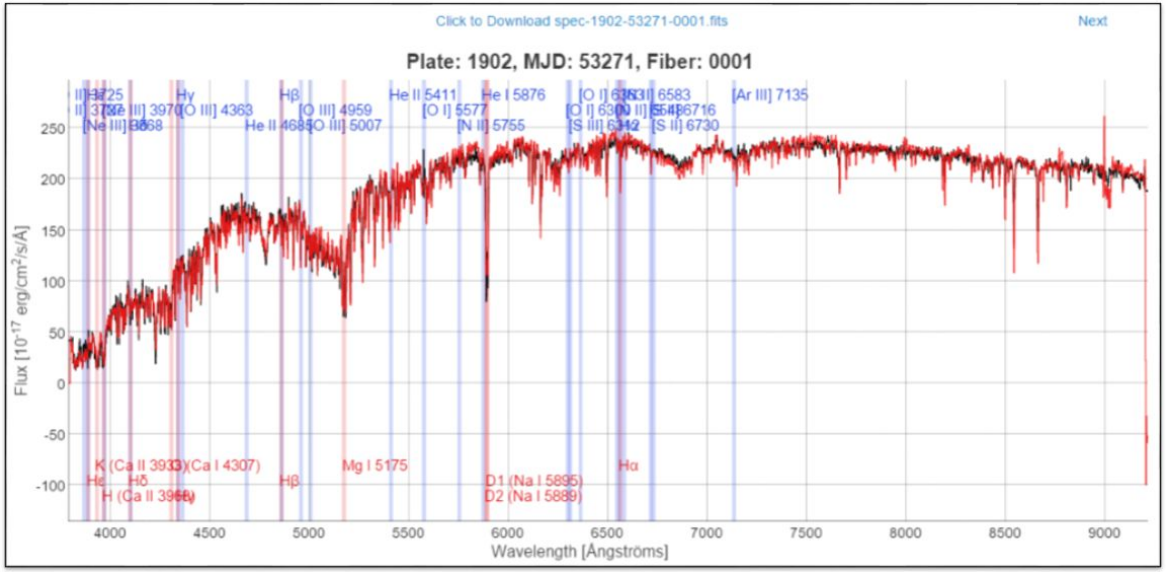
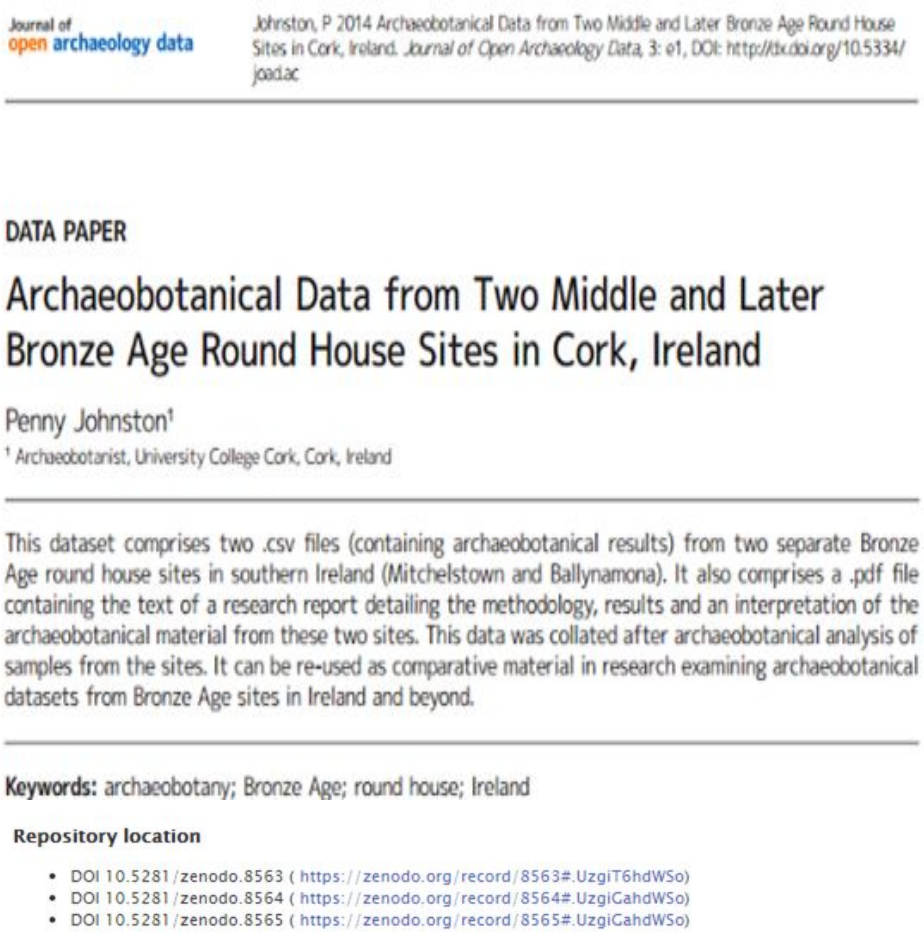


Image credit: http://dr10.sdss3.org/spectrumDetail?mjd=53271&fiber=1&plateid=1902

Examples of Repositories

Example of a Data Paper



Journal of open archaeology data

Johnston, P 2014 Archaeobotanical Data from Two Middle and Later Bronze Age Round House Sites in Cork, Ireland. *Journal of Open Archaeology Data*, 3: e1, DOI: http://dx.doi.org/10.5334/joad.ac

DATA PAPER

Archaeobotanical Data from Two Middle and Later Bronze Age Round House Sites in Cork, Ireland

Penny Johnston[1]

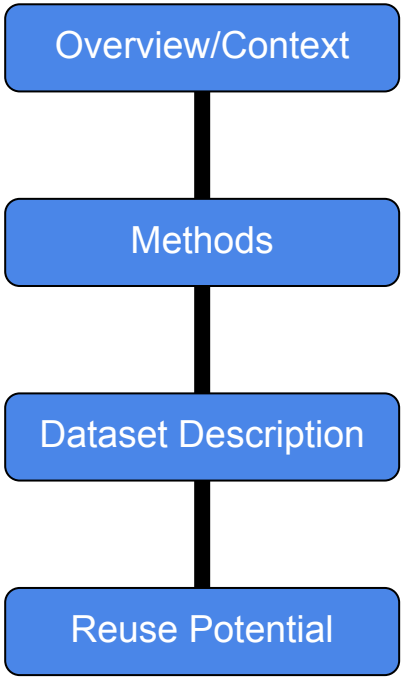
[1] Archaeobotanist, University College Cork, Cork, Ireland

This dataset comprises two .csv files (containing archaeobotanical results) from two separate Bronze Age round house sites in southern Ireland (Mitchelstown and Ballynamona). It also comprises a .pdf file containing the text of a research report detailing the methodology, results and an interpretation of the archaeobotanical material from these two sites. This data was collated after archaeobotanical analysis of samples from the sites. It can be re-used as comparative material in research examining archaeobotanical datasets from Bronze Age sites in Ireland and beyond.

Keywords: archaeobotany; Bronze Age; round house; Ireland

Repository location

- DOI 10.5281/zenodo.8563 (https://zenodo.org/record/8563#.UzgiT6hdWSo)
- DOI 10.5281/zenodo.8564 (https://zenodo.org/record/8564#.UzgiGahdWSo)
- DOI 10.5281/zenodo.8565 (https://zenodo.org/record/8565#.UzgiGahdWSo)



Questions?