Top 10 FAIR Data & Software Things

February 1, 2019

Sprinters:

Reid Otsuji, Stephanie Labou, Ryan Johnson, Guilherme Castelao, Bia Villas Boas, Anna-Lena Lamprecht, Carlos Martinez Ortiz, Chris Erdmann, Leyla Garcia, Mateusz Kuzak, Paula Andrea Martinez, Liz Stokes, Natasha Simons, Tom Honeyman, Chris Erdmann, Sharyn Wise, Josh Quan, Scott Peterson, Amy Neeser, Lena Karvovskaya, Otto Lange, Iza Witkowska, Jacques Flores, Fiona Bradley, Kristina Hettne, Peter Verhaar, Ben Companjen, Laurents Sesink, Fieke Schoots, Erik Schultes, Rajaram Kaliyaperumal, Erzsebet Toth-Czifra, Ricardo de Miranda Azevedo, Sanne Muurling, John Brown, Janice Chan, Lisa Federer, Douglas Joubert, Allissa Dillman, Kenneth Wilkins, Ishwar Chandramouliswaran, Vivek Navale, Susan Wright, Silvia Di Giorgio, Akinyemi Mandela Fasemore, Konrad Förstner, Till Sauerwein, Eva Seidlmayer, Ilja Zeitlin, Susannah Bacon, Chris Erdmann, Katie Hannan, Richard Ferrers, Keith Russell, Deidre Whitmore, and Tim Dennis.

Organisations:

Library Carpentry/The Carpentries, Australian Research Data Commons, Research Data Alliance Libraries for Research Data Interest Group, FOSTER Open Science, OpenAire, Research Data Alliance Europe, Data Management Training Clearinghouse, California Digital Library, Dryad, AARNet, Center for Digital Scholarship at the Leiden University, DANS, The Netherlands eScience Center, University Utrecht, UC San Diego, Dutch Techcentre for Life Sciences, EMBL, University of Technology, Sydney, UC Berkeley, University of Western Australia, Leiden University, GO FAIR, DARIAH, Maastricht University, Curtin University, NIH, NLM, NCBI, ZB MED, CSIRO, and UCLA.

TOP 10 FAIR DATA & SOFTWARE THINGS:

Table of Contents

About

The Top 10 FAIR Data & Software Global Sprint was held online over the course of two-days (29-30 November 2018), where participants from around the world were invited to develop brief guides (stand alone, self paced training materials), called "Things", that can be used by the research community to understand FAIR in different contexts but also as starting points for conversations around FAIR. The idea for "Top 10 Data Things" stems from initial work done at the Australian Research Data Commons or ARDC (formerly known as the Australian National Data Service).

The Global Sprint was organised by Library Carpentry, Australian Research Data Commons and the Research Data Alliance Libraries for Research Data Interest Group in collaboration with FOSTER Open Science, OpenAire, RDA Europe, Data Management Training Clearinghouse, California Digital Library, Dryad, AARNet, Center for Digital Scholarship at the Leiden University, and DANS. Anyone could join the Sprint and roughly 25 groups/individuals participated from The Netherlands, Germany, Australia, United States, Hungary, Norway, Italy, and Belgium. See the full list of registered Sprinters.

Sprinters worked off of a primer that was provided in advance together with an online ARDC webinar introducing FAIR and the Sprint titled, "Ready, Set, Go! Join the Top 10 FAIR Data Things Global Sprint." Groups/individuals developed their Things in Google docs which could be accessed and edited by all participants. The Sprinters also used a Zoom channel provided by ARDC, for online calls and coordination, and a Gitter channel, provided by Library Carpentry, to chat with each other throughout the two-days. In addition, participants used the Twitter hashtag #Top10FAIR to communicate with the broader community, sometimes including images of the day.

Participants greeted each other throughout the Sprint and created an overall welcoming environment. As the Sprint shifted to different timezones, it was a chance for participants to catch up. The Zoom and Gitter channels were a way for many to connect over FAIR but also discuss other topics. A number of participants did not know what to expect from a Library Carpentry/Carpentries-like event but found a welcoming environment where everyone could participate.

The Top 10 FAIR Data & Software Things repository and website hosts the work of the Sprinters and is meant to be an evolving resource. Members of the wider community can submit issues and/or pull requests to the Things to help improve them. In addition, a published version of the Things will be made available via Zenodo and the Data Management Training Clearinghouse in February 2019.

TOP 10 FAIR DATA & SOFTWARE THINGS:

Oceanography

Sprinters:

Reid Otsuji, Stephanie Labou, Ryan Johnson, Guilherme Castelao, Bia Villas Boas (UC San Diego)

Table of contents

Findability:

Accessibility:

Interoperability:

Reusability:

Description:

Oceanographic data encompasses a wide variety of data formats, file sizes, and states of data completeness. Data of interest may be available from public repositories, collected on an individual basis, or some combination of these, and each type has its own set of challenges. This “10 Things” guide introduces 10 topics relevant to making oceanographic data FAIR: findable, accessible, interoperable, and reusable.

Audience:

- Library staff and programmers who provide research support
- Oceanographers
- Oceanography data stewards
- Researchers, scholars and students in Oceanography

Goal:

The goal of this lesson is to introduce oceanographers to FAIR data practices in their research workflow through 10 guided activities.

Things

Thing 1: Data repositories

There are numerous data repositories for finding oceanographic data. Many of these are from official “data centers” and generally have well-organized and well-documented datasets available for free and public use.

- NSF / Earth Cube
- CLIVAR - CCHDO
- CLIVAR - Hawaii ADCP
- CLIVAR - JODC ADCP Data
- NOAA - NODC
- NOAA - NCDC
- NOAA - NGDC
- NSIDC

- CDIAC
- BCODMO
- GEOTRACES
- R2R
- SAMOS
- ARGO Data
- NASA - PO.DAAC
- World Ocean Database (WOD)
- Spray Underwater Glider

At some point, you may want or need to deposit your own data into a data repository, so that others may find and build upon your work. Many funding agencies now require data collected or created with the grant funds to be shared with the broader community. For instance, the National Science Foundation (NSF) Division of Ocean Sciences (OCE) mandates sharing of data as well as metadata files and any derived data products. Finding the "right" repository for your data can be overwhelming, but there are resources available to help pick the best location for your data. For instance, OCE has a list of approved repositories in which to submit final data products.

Activity 1:

- Go to re3data.org and search for a data repository related to your research subject area. How many results did you get? Which of these repositories looks most relevant to your research area? Is it easy to find a dataset in those repositories that covered the California coast (or any other region of your choice) during the last year?

Activity 2:

- What is the next journal you would like to publish in? (Alternatively: what is a top journal in your field?) Can you find the data submission requirements for this journal?

Thing 2: Metadata

High quality metadata (information about the data, such as creator, keywords, units, flags, etc.) significantly improves data discovery. While metadata is most often for the data itself, metadata can also include information about machines/instruments used, such as make, model, and manufacturer, as well as process metadata, which would include details about any cleaning/analysis steps and scripts used to create data products.

Using controlled vocabularies in metadata allows for serendipitous discovery in user searches. Additionally, using a metadata schema to mark up a dataset can make your data findable to the world.

Activity 1:

- Using schema.org markup, look at the metadata elements pertaining to scholarly articles: https://schema.org/ScholarlyArticle. Imagine you have an article you have hosted on your personal website, and you would like to add markup so that it could be more readily indexed by Google Dataset Search. What metadata elements would be most important to include? (This resource will help you: https://developers.google.com/search/docs/data-types/dataset)

Activity 2:

- OpenRefine example for making data FAIR.

Read this walkthrough of how to "FAIRify" a dataset using the data cleaning tool OpenRefine: https://docs.google.com/document/d/1hQ0KBnMpQq93-HQnVa1AR5v4esk6BRlG6NvnnzJuAPQ/edit#heading=h.v3puannmxh4u

Discussion:

- If you had *thousands* of keywords in a dataset you wanted to associate with a controlled vocabulary relevant to your field, what would be your approach? What tools do you think would best automate this task?

Thing 3: Permanent identifiers

Permanent identifiers (PIDs) are a necessary step for keeping track of data. Web links can break, or "rot", and tracking down data based on a general description can be extremely challenging. A permanent identifier like a digital object identifier (DOI) is a unique ID assigned to a dataset to ensure that properly managed data does not get lost or misidentified. Additionally, a DOI makes it easier to cite and track the impact of datasets, much like cited journal articles.

Identifiers exist for researchers as well: OCRID is essentially a DOI for an individual researcher. This ensures that if you have a common name, change your name, change your affiliation, or otherwise change your author information, you still get credit for your own and maintain a full, identifiable list of your scientific contributions.

Activity 1:

Go to re3data.org and search for a data repository related to your research subject area. From the repository you choose, pick a dataset. Does it have a DOI? What is? Who is the creator of that dataset? What is the ORCID of the author?

Activity 2:

You've been given this DOI: **10.6075/J03N21KQ**

- What would you do to find the dataset this DOI references?
- Using the above approach, you just identified, what is associated with this DOI? Who was the creator of this dataset? When was that published? Who funded that research?

Activity 3:

- Go to the ORCID website and create an ORCID if you do not have one already. Can you identify the creator associated with the DOI on the activity 1?

Discussion:

- What would be a positive benefit for having a personal persistent ID such as ORCID? Are there any drawbacks or concerns?

Thing 4: Citations

Citing data properly is equally as important as citing journal articles and other papers. In general, a data citation should include: author/creator, date of publication, title of dataset, publisher/organization (for instance, NOAA), and unique identifier (preferably DOI).

Activity 1:

- Read through this overview of citing data from DataONE. This has information application to any data citations, as well as guidelines specific to DataONE.
- Think of the last dataset you worked with. Is it something you collected, or was it from a public repository? How would you cite this data?
- Websites/data repositories will often provide the text of preferred citation, but you may have to search for it. How would you cite the World Ocean Database? How would you cite data from the Multibeam Bathymetry Database?

Discussion

Long-term data stewardship is an important factor for keeping data open and accessible for the long term.

- After completing the last activity, discuss how Open is data in the discipline? Are there long-term considerations and protocols for the data that is produced?

Tip: Resources that can help make your data more open and accessible or to protect your data

- Open Science Framework
- Figshare
- Oceanographic data centers

Thing 5: Data formats

Oceanographic data can include everything from maps and images to high dimensional numeric data. Some data are saved as common, near-universal formats (such as csv files), while others require specialized knowledge and software to open properly (e.g., netCDF). Explore the Intrinsic characteristics of the dataset that influence the choice of the format, such as a time series versus a regular 3-D grid of temperature varying on time; robust ways to connect the data with metadata; size factors, binary versus ASCII file; and think about why a format to store/archive data is not necessarily the best way to distribute data.

Discussion 1:

- what are the most common data formats used in your field? What level of technical/domain knowledge is required to open, edit, and interactive with these data types?

Discussion 2:

- What are the advantages and disadvantages of storing in plain ASCII, like a CSV file versus a binary, like netCDF? Does the characteristics of the data influence that decision, i.e. the preferred format for a time series would be different than a numerical model output, or a gene sequence?

Thing 6: Data organization and management

Good data organization is the foundation of your research project. Data often has a longer lifespan than the project it is originally associated with and may be reused for follow-up projects or by other researchers. Data is critical to solving research questions, but lots of data are lost or poorly managed. Poor data organization can directly impact the project or future reuse.

Activity 1:

Considerations for basic data organization and management

Group Discussion 1:

- Is your data file structure something that a new lab member could easily learn, or are datasets organized in a more haphazard fashion?
- Do you have any documentation associated describing how to navigate your data structures?

Group Discussion 2:

- Talk about where/how you are currently storing data you are working with. Would another lab member be able to access all your data if needed?

Activity 2:

Identifying vulnerabilities

- **Scenario 1:** Your entire office/lab building burns down overnight. No one is harmed, because no one was there, but all electronics in the building perish beyond hope of repair. The next morning, can you access any of your data?
- **Scenario 2:** The cloud server you use (everything from Google Drive to GitHub) crashes. Can you still access your most up to date data?

Discussion 1:

- From either of the two scenarios, can your data survive a disaster? What are some of the things that you think you are doing incorrectly to prevent data loss?

Discussion 2:

- Think about a time when you had or potentially had a data disaster - how could the disaster have been avoided? What, if anything, have you changed about your data storage and workflow as a result?

The Data Management Plan (DMP)

Some research institutions and research funders now require a Data Management Plan (DMP) for new research projects. Let's talk about the importance of a DMP and what should a DMP cover. Think about it you would you be able to create a DMP?

What is a DMP?

A Data Management Plan (DMP) documents how data will be managed, stored and shared during and after a research project. Some research funders are now requesting that researchers submit a DMP as part of their project proposal.

Activity 1:

- Start by watching The DMPTool: A Brief Overview 90 second video to see what the DMPTool can do for researhers and data managers.
- Next, review this short introduction to Data Management Plans.
- Now browse through some public DMPs from the DMPTool, choose one or two of the DMPs related to oceanography and read them to see the type of information they capture.

Activity 2:

There are many Data Management Plan (DMP) templates in the DMPTool.

- Choose one DMP funder template you would potentially use for a grant proposal in the DMPTool. Spend 5-10 minutes starting to complete the template, based on a research project you have been involved with in the past.

Discussion:

- You will have noticed that DMPs can be very short, or extremely long and complex. What do you think are the two or three pieces of information essential to include in every DMP and why?
- After completing the second activity, what are strengths and weaknesses of your chosen template?

Thing 7: Re-usable data

There are two aspects to reusability: reusable data, and reusable derived data/process products.

Reusable data

Reusable data is the result of successful implementation of the other “Things” discussed so far. Reusable data (1) has a license which specifies reuse scenarios, (2) is in a domain-suitable format and an “open” format when possible, and (3) is associated with extensive metadata consistent with community and domain standards.

Process/derived data products

What is often overlooked in terms of reusability are the products created to automate research steps. Whether it's using the command line, Python, R, or some other programming platform, automation scripts in and of themselves are a useful output that can be reused. For example, data cleaning scripts can be reapplied to datasets that are continually updated, rather than starting from scratch each time. Modeling scripts can be re-used and adapted as parameters are updated. Additionally, these research automation products make any data-related decision you made explicit: if future data users have questions about exclusions, aggregations, or derivations, the methodology used is transparent in these products.

Discussion 1:

- How many people have made public or shared part of their research automaton pipeline? If you haven't shared this, what prevented you from sharing?

Discussion 2:

- Are there instances where your own research would have been improved if you had access to other people's process products?

Thing 8: Tools of the trade

When working with your data, there are a selection of proprietary and open source tools available to conduct your research analysis.

Why open source tools?

Open source tools are software tools developed, in which the source code is openly available and published for use and/or modification by any one free of charge. There are many advantages to using open source tools:

- Low software costs
- Low hardware costs
- Wide community development support
- Interoperable with other open source software
- No vendor control
- Open source licensing

Caution: be selective with the tools you use
There are additional benefits you may hear about using open sources tools which are:

- Higher quality software
- Greater security
- Frequent updates

Keep in mind, in an ideal world these three ideas are what we all wish for, however not every open source tool satisfies these benefits. When selecting an open source tool, choose a package with a large community of users and developers that proves to have long-term support.

Things to consider when using open source tools

Benefits:

- Open source tools often have active development community. Quality for end users is usually higher because the community are users of the software being developed. In turn, open source costs for development are cheaper.

- With a larger community of development, security problems and vulnerabilities are discovered and fixed quickly. Another major advantage of open source is the possibility to verify exactly which procedures are being applied, avoiding the use of "black-boxes" and allowing for a thorough inspection of the methods.

Issues:

- Open sources tools are only as good as the community that supports it. Unlike commercial software there is no official technical support. Additionally, not all open source licenses are permissive.
- Training time can be significant.

If Open source tools are not an option and commercial software is necessary for your project, there are benefits and issues to consider when using proprietary or commercial software tools.

Benefits:

- This type of software often comes with official technical support such as a customer service phone number or email.

Issues:

- Proprietary or commercial tools are often quite expensive at the individual level.
- Universities may have campus-wide licenses, but if you move institutions, you may find yourself without the software you had been using.

Discussion:

- Think about the tools you use for conducting data clean up, analysis, and for creating visualizations and reports for publications. What were the deciding factors for selecting the applications you used or are using for your project?

Thing 9: Reproducibility

Can you or others reproduce your work?

Reproducibility increases impacts credibility and reuse.

Read through the following best practices to make your work reproducible.

Best practices:

Making your project reproducible from the start of the project is ideal.

- Documenting each step of your research - from collecting or accessing data, to data wrangling and cleaning, to analysis - is the equivalent of creating a roadmap that other researchers can follow. Being explicit about your decisions to exclude certain values or adjust certain model parameters, and including your rationale for each step, help eliminate the guesswork in trying to reproduce your results.
- Consider open source tools. This allows anyone to reproduce research more easily, and helps with identifying who has the right license for the software used.

This is useful not only for anyone else who wants to test your analysis - often the primary beneficiary is you!

Research often takes months, if not years, to complete a certain project, so by starting with reproducibility in mind from the beginning, you can often save yourself time and energy later on.

Discussion:

Think about a project you have completed or are currently working on.

- What are some of the best practices you have adopted to make your research reproducible for others?
- Were there any pain points that you encounter or are dealing with now?
- Is there something you can do about it now?
- What are the most relevant "Things" previously mentioned in this document that you could use to make your research more reproducible?

Thing 10: APIs and applications (apps)

APIs (Application Programming Interfaces) allow programmatic access to many databases and tools. They can directly access or query existing data, without the need to download entire datasets, which can be very large.

Certain software platforms, such as R and Python, often have packages available to facilitate access to large, frequently used database APIs. For instance, the R package "rnoaa" can access and import various NOAA data sources directly from the R console. You can think of it as using an API from the comfort of a tool you're already familiar with. This not only saves time and computer memory, but also ensures that as databases are updated, so are your results: re-running your code automatically pulls in new data (unless you have specified a more restricted date range).

Activity:

On the ERDDAP server for Spray Underwater Glider data, select temperature data for the line 90 (https://spraydata.ucsd.edu/erddap/tabledap/binnedCUGN90.html).

- Restrict it to measurements at 25 m or shallower.
- Choose the format of your preference, and instead of submit the request, generate an URL.
- Copy and paste the generated URL in your browser.

Discussion:

- Think about the last online data source you accessed. Is there an API for this data source? Is there a way to access this data from within your preferred analysis software?

TOP 10 FAIR DATA & SOFTWARE THINGS:

Research Software

Sprinters

Anna-Lena Lamprecht, Carlos Martinez Ortiz, Chris Erdmann, Leyla Garcia, Mateusz Kuzak, Paula Andrea Martinez

Description:

The FAIR data principles are widely known and applied today. What the FAIR principles mean for (scientific) software is an ongoing discussion. However, there are some things on which there is already agreement that they will make software (more) FAIR. In this document, we go for some 'low hanging fruit' and describe 10 easy FAIR software things that you can do. To limit the scope, "software" here refers to scripts and packages in languages like R and Python, but not to other kinds of software frequently used in research, such as web-services, web platforms like myexperiment.org or big clinical software suites like OpenClinica.

A poster summarizing these 10 FAIR software things is also available.

Audience:

- Researchers who develop software
- Research Software Engineers

Goals:

Translate FAIR principles to applicable actions for scientific software

What is FAIR for software

In the context of this document, we use the following simple definition of FAIR for software:

Findable
Software with sufficiently rich metadata and unique persistent identifier

Accessible
Software metadata is in machine and human readable format. Software and metadata is deposited in trusted community approved repository.

Interoperable
Software uses community accepted standards and platforms, making it possible for users to run the software.

Reusable
Software has clear licence and documentation

Things

Findability

Thing 1: Create a description of your software

The name alone does not tell people much about your software. In order for other people to find out if they can use it for their purpose, they need to know what it does. A good description of your software will also help other people to find it.

Activity:
Think of minimum set of information (metadata) which will help others find your software. This can include short descriptive text and meaningful keywords.

Codemeta is a set of keywords used to describe software and way to structure them in machine readable way. For examples of Codemeta used in software packages see:

- https://github.com/NLeSC/boatswain/blob/master/codemeta.json
- https://github.com/datacite/maremma

Edam is an example of an ontology that provides terminology that can be used to describe bioinformatics software.

Take the 4OSS lesson episode about metadata and registries and walk through the exercise.

This example: http://r-pkgs.had.co.nz/description.html#description

Thing 2: Register your software in a software registry

People search for research software using search engines like Google. Registering your software in a dedicated registry will make it findable by search engines, because the registries take care about search engine optimization etc. The registries will usually ask you to provide descriptions (metadata) as above.

Activity:
Think of the registries most used in your domain? Do you know about any? How and where do you usually find software? What kind of keywords do you use when searching?

Here are some examples of research software registries:
* bio.tools * Research Software Directory (check if your institution hosts one) * rOpenSci Project * Zenodo

4OSS lesson episode about metadata and registries

Thing 3: Get and use a unique and persistent identifier for your software

It will help others find and access the particular version of your software. Unique means that the identifier will point on and only version and location of your software. Persistent means that it will pointing to the same version and location for long, specified amount of time. For example, Zenodo provides you with a DOI (Digital Object Identifier) that will be resolvable for at least the next 20 years. Recent initiatives, such as Software Heritage, propose to associate a permalinks as intrinsic SHA1 identifier to software (see example through the id: swh:1:dir:005bc6218c7a0a9ede654f9a177058adcde98a50 / permalinks: https://archive.softwareheritage.org/swh:1:dir:005bc6218c7a0a9ede654f9a177058adcde98a50/)

Activity:
If you have registered your software in a registry, chances are good that they provide a unique and persistent identifier. If not, obtain an identifier from another organization. If you have multiple identifiers, choose one that you use as your main identifier. Make sure you use it consistently when referring to your software, e.g. on your own website, code repository or in publications.

Making your code citable with Zenodo

Accessibility

Thing 4: Make sure that people can download your software

In order for anyone to use your software, they need to be able to download an executable version along with documentation. For interpreted languages like Python and R, the code is also the executable version. For compiled languages like Java and C, the executable version is a binary file, and the code might not be accessible. Downloading the software and documentation is possible, for instance, from a project website, a git repository or from a software registry.

Activity:
Using the identifier as your starting point, ask a colleague to try to get your software (binary/script). Can he/she download it? Does he/she also have access to the documentation? Is there anything preventing him/her from getting to it? Is it hosted on a reliable platform (long term persistent, such as Zenodo, PyPI, CRAN)?

Interoperability

Thing 5: Explain the functionality of your software

Your software performs one or more operations that take an input and transform it into the output. To help people use your software, provide a clear and concise description of the operations along with the corresponding input and output data types. For example, the wc (word count) command line tool takes a text as input, counts the number of words in it and gives the number of words as output. The ClustalW tool takes a set of (gene or protein) sequences as input, aligns them and returns a multiple sequence alignment as output.

Activity:
List all operations that your software provides, and describe them along with corresponding input and output data types. If possible, use terms from a domain ontology like EDAM.

Thing 6: Use standard (community agreed) formats for inputs and outputs

In order for people to use your software, they need to know how to feed data to it -- standard formats are easy ways to exchange data between different pieces of software. By sticking to standards, it is possible to use the output from another piece of software as an input to your software (or the other way around). For example, FASTA is a format for representing molecular sequences (DNA, RNA, protein, …) that most sequence analysis tools can handle. NetCDF is a standard file format used sharing of array-oriented scientific data.

Activity:
What are the relevant standards in your field? Which are the groups/organizations that are responsible for standards in your field? Is there a place where you can find the relevant standards and a detailed description? What other tools use these standards? If possible, use such standard formats as input/output of your software and state which you are using. (Avoid to define your own standards! http://imgs.xkcd.com/comics/standards.png)

Reusability

Thing 7: Document your software

Your software should include sufficient documentation: instructions on how to install, run and use your software. All dependencies of your software should be clearly stated. Provide sufficient examples on how to execute the different operations your software offers, ideally along with example data. Write the Docs page explains and gives examples of good documentation.

Activity:
Ask a colleague to look at your software's documentation. Is he/she able to install your software? Can he/she run it? Can he/she produce the expected results?

Thing 8: Give your software a license

A license tells your (potential) users what they are allowed to do with your software (and what not to do), and can protect your intellectual property. Without a license people may spend time trying to figure out if they are allowed to use your software -- make things easy for them. Therefore, it is important that you choose a software license that meets your intentions. Choose a license website provides a simple guide for picking the right license for your software.

Activity:
* Follow the 4OSS lesson to learn more about licenses and their implications. * Read 4OSS paper

Thing 9: State how to cite your software

You want to get credit for your work. By providing the citation guideline you will help users of your software to cite your work properly. There is no single right way to do it. Software Sustainability Institute website provides more information and discussion on this topic in a blog post How to cite and describe software.

Activity:
Read "Software citation principles" paper. Read documentation of Citation File Format and create CFF file for your software.

Thing 10: Follow best practices for software development

Reusability benefits from good quality of software. There are a number of actions you can take to improve the quality of your software: make your code modular, have code level documentation, provide tests, follow code standards, use version control, etc. There are several guidelines which you can use to guide you in the process such as the eScience Center Guide, the best practices and the good enough practices.

Activity:
Familiarize yourself with the guides provided above. Have a look at your software and create a list of actions which you could follow to improve the quality of your software. Ideally, follow these practices from the very beginning.

TOP 10 FAIR DATA & SOFTWARE THINGS:

Research Libraries

Sprinters:

Liz Stokes, Natasha Simons, Tom Honeyman (Australian Research Data Commons), Chris Erdmann,(Library Carpentry/The Carpentries/California Digital Library), Sharyn Wise (University of Technology, Sydney), Josh Quan, Scott Peterson, Amy Neeser (UC Berkeley)

Description:

To translate FAIR principles into useable concepts for research-facing support staff (e.g. librarians).

Audience:

- Library staff who provide research support
- Those who want to know more about FAIR and how it could be applied to libraries

Goals:

- Translating FAIR speak to library speak (What is it? Why do I need to know? What do I tell researchers?)
- Identifying ways to improve the 'FAIRness' of your library
- Understanding that FAIR data helps us be better stewards of our own resources

Things

Thing 1: Why should librarians care about FAIR?

There's a lot of hype about the FAIR Data Principles. But why should librarians care? For starters, libraries have a strong tradition in describing resources, providing access and building collections, and providing support for the long-term stewardship of digital resources. Building on their specific knowledge and expertise, librarians should feel confident with making research data FAIR. So how can you and your library get started with the FAIR principles?

Activity:

1. Read LIBER's *Implementing FAIR Principles: the role of Libraries* at https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf (5 minute read)

Consider:
* Where is your library at in regard to the section on ‘getting started with FAIR’? * Where are you at in your own understanding of the FAIR Data Principles?

Thing 2: How FAIR are your data?

The FAIR Principles are easily understood in theory but more challenging when applied in practice. In this exercise, you will be using the Australian Research Data Commons (ARDC) Data self-assessment tool to assess the 'FAIRness' of one of your library’s datasets.

Activity:
1. Select a metadata record from your library’s collection (e.g. your institutional repository) that describes a published dataset. 2. Open the ARDC FAIR Data Assessment tool and run your chosen dataset against the tool to assess its ‘FAIRness’.

Consider:
* How FAIR was your chosen dataset? * How easy was it to apply the FAIR criteria to your dataset? * What things need to happen in order to improve the ‘FAIRness’ of your chosen dataset?

Want more?
Try your hand at other tools like the CSIRO 5 star data rating tool and the DANS FAIR data assessment tool.

Thing 3: Do you teach FAIR to your researchers?

How FAIR aware are your researchers? Does your library incorporate FAIR into researcher training?

Activity:
Go to existing data management/data sharing training you provide to Graduates, Higher Degree Researchers (HDRs) or other researchers. For example, review the Duke Graduate School’s Responsible Conduct of Research topics page. Review how well the 15 FAIR Principles are covered in this training and adjust accordingly.

Thing 4: Is FAIR built into library practice and policy?

Your library may do a great job advocating the FAIR Data Principles to researchers but how well have the Principles been incorporated into library practice and policy?

Activity:
1. Review your library or institutional policies regarding research data management and digital preservation with the FAIR Principles in mind. Consider that in most cases library policy will have been written before the advent of FAIR. Are revisions required? 2. Review

the data repository managed by your library. How well does it support FAIR Data? 3. Review your library's Data Management Planning tool. Does it have features that support the FAIR Data Principles or are changes required?

Thing 5: Are your library staff trained in FAIR?

Reusing the wide range of openly available training materials available in the FAIR Data Principles e.g. you could start here.

Activity:
* Conduct a skills and knowledge audit regarding FAIR with your library team. * Based on the audit, identify gaps in FAIR skills and knowledge. * Design a training program that can fill the identified gaps. To help build your program, read the blog post, A Carpentries based approach to teaching FAIR data and software principles.

Consider:
Reusing the wide range of openly available training materials available in the FAIR Data Principles e.g. you could start here.

Thing 6: Are digital libraries FAIR?

While the FAIR Principles are designed for data, considering their potential application in a broader context is useful. For example, think about what criteria might be applied to assess the 'FAIRness' of digital libraries. Considerations might include:
* Persistent identifiers * Open access vs. paid access * Provenance information / metadata * Author credibility * Versioning information
* License / reuse information * Usage statistics (number of times downloaded)

Activity:
1. Select one of these digital libraries (or another of your choice): * British Library * National Digital Library of India * Europeana * National Library of Australia's Trove 2. Search/browse the catalogue of items.

Consider:
* Does the library display reuse permissions/licenses on how to use the item? * Is there provenance information? * Are persistent identifiers used?

Thing 7: Does your library support FAIR metadata?

A number of FAIR principles make reference to "metadata". What is metadata, how is it relevant to FAIR and does your library support the kind of metadata specified in the FAIR Data Principles?

Activity:
1. Watch this video in which the Metadata Librarian explains metadata (3 mins) 2. Select three metadata records at random for datasets held in your library or repository collection. 3. Open the checklist produced for use at the EUDAT summer school and see if you can check off those that reference metadata against the records you selected. 4. Make a list of what metadata elements could be improved in your library records to enable better support for FAIR.

Thing 8: Does your library support FAIR identifiers?

The FAIR data principles call for open, standardised protocols for accessing data via a persistent identifier. Persistent identifiers are crucial for the findability and identification of research, researchers and for tracking impact metrics. So how well does your library support persistent identifiers?

Activity:
Find out how well your library supports ORCIDs and DOIs: * Do your library systems support the identification of researchers via an ORCID? Do you authenticate against the ORCID registry? Do you have an ORCID? * Do your library systems, such as your institutional repository, support the issuing of Digital Object Identifiers (DOIs) for research data and related materials?

Consider:
* What other types of persistent identifiers do you think your library should support? Why or why not?

Want more?
If you library supports the minting of DOIs for research data and related materials, is there more that you could do in this regard? Check out A Data Citation Roadmap for Scholarly Repositories and determine how much of the roadmap you can check off your list and how much is yet to do.

Thing 9: Does your library support FAIR protocols?

For (meta)data to be accessible it should ideally be available via a standard protocol. Think of protocols in terms of borrowing a book: there are a number of expectations that the library lays out in order to proceed. You have to identify yourself using a library card, you have to bring the book to the checkout desk, and in return you walk out of the library with a demagnetised book and receipt reminding you when you have to return the book by. Accessing the books in the library means that you must learn and abide by the rules for accessing books.

Activity:
* Familiarise yourself with APIs by completing Thing 19 of the ANDS 23 (research data) Things * Consider the APIs your library provides to enable access (meta)data for data and related materials. Are they up to scratch or are improvements required?

Thing 10: Next steps for your library in supporting FAIR

In Thing 1 you read LIBER's Implementing FAIR Principles: the role of Libraries. You considered what your library needed to do in order to better support FAIR data. In Thing 10 we will create a list of outstanding action items.

Activity:
1. Write a list of what your library is currently doing to support and promote the FAIR Data Principles. 2. Now compare this to the list in the LIBER document. Where are the gaps and what can you do to fill these? 3. Create an action plan to improve FAIR support at your library!

Consider:
* Incorporate all that you learnt and progress that you made in "doing" this Top 10 FAIR Things!

TOP 10 FAIR DATA & SOFTWARE THINGS:

Research Data Management Support

Sprinters:

Lena Karvovskaya, Otto Lange, Iza Witkowska, Jacques Flores (Research Data Management (RDM) support at Utrecht University)

Description:

This is an umbrella-like document with links to various resources. The aim of the document is to help researchers who want to share their data in a sustainable way. However, we consider the border between librarians and researchers to be a blurred one. This is because, ultimately, librarians support researchers that would like to share their data. We primarily wish to target researchers and support staff irregardless of their experience: those who have limited technical knowledge and want to achieve a very general understanding of the FAIR principles and those who are more advanced technically and want to make use of more technical resources. The resources we will link to for each of the 10 FAIR Things will often be on two levels of technicality.

Audience:

Our primary audience consists of researchers and support staff at Utrecht University. Therefore, whenever possible we will use the resources available at Utrecht University: the institutional repositories and resources provided at the RDM Support website.

Things

Thing 1: Why bother with FAIR?

Background: The advancement of science thrives on the timely sharing and accessibility of research data. Timely and sustainable sharing is only possible if there are infrastructures and services that enable it.

1. Read up on the role of libraries in implementing the FAIR Data Principles. Think about the advantages and opportunities made possible by digitalization in your research area. Think about the challenges. Have you or your colleagues ever experienced data loss? Is the falsification/fabrication of data an issue with digital data? How easy it to figure out if the data you found online is reliable? Say you found a very useful resource available

online and you want to refer to it in your work; can you be sure that it is still there several years later?

2. For more information, you can refer to this detailed explanation of FAIR principles developed by the Dutch Center for Life Sciences (DTLS).

Thing 2: Metadata

Background: Metadata are information about data. This information allows data to be findable and potentially discoverable by machines. Metadata can describe the researchers responsible for the data, when, where and why the data was collected, how the research data should be cited, etc.

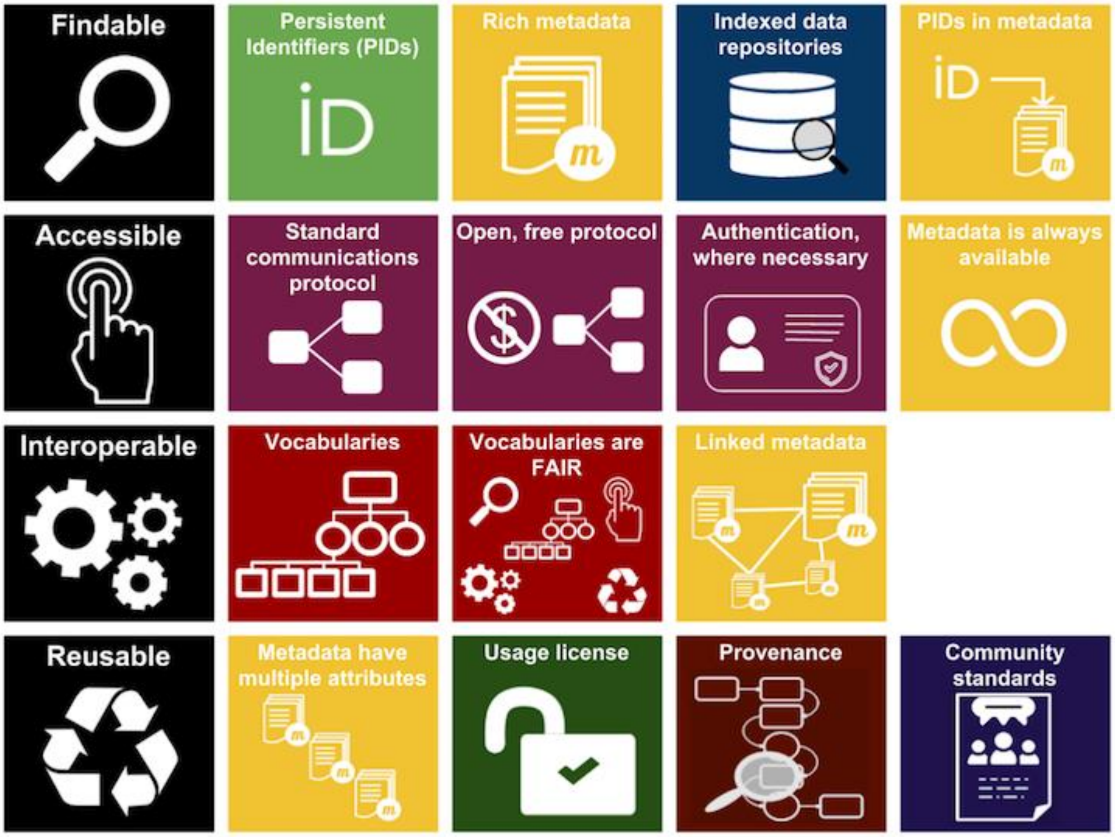
1. If you find the discussion on metadata too abstract, think about a traditional library catalogue record as a form of metadata. A library catalogue card holds information about a particular book in a library, such as author, title, subject, etc. Library cataloging, as a form of metadata, helps people find books within the library. It provides information about books that can be used in various contexts.

Now, reflect on the differences in functionality between a paper catalogue card and a digital metadata file.

1. Reflect on your own research data. If someone who is unfamiliar with your research wants to find, evaluate, understand and reuse your data, what would he/she need?
2. Watch this video about structural and descriptive metadata and reflect on the example provided in the video. If the video peaked your interest about metadata, watch a similar video on the *Ins and outs of metadata and data documentation* by Utrecht University.

Thing 3: The definition of FAIR metrics

Background: FAIR stands for Findable, Accessible, Interoperable and Re-usable.


Findable
Persistent Identifiers (PIDs)
Rich metadata
Indexed data repositories
PIDs in metadata
Accessible
Standard communications protocol
Open, free protocol
Authentication, where necessary
Metadata is always available
Interoperable
Vocabularies
Vocabularies are FAIR
Linked metadata
Reusable
Metadata have multiple attributes
Usage license
Provenance
Community standards

1. Take a look at the image above, provided by the Australian Research Data Commons (ARDC). Reflect on the images chosen for various aspects of the FAIR acronym. If we consider this video, already mentioned in Thing 2, how would you describe the photography example in terms of FAIR?
2. Go to DataCite and choose data center "Utrecht University". Select one of the published datasets and evaluate it with respect to FAIR metrics. In evaluating the dataset, you can make use of the FAIR Data self-assessment tool created by ARDC. Which difficulties do you experience while trying to do the evaluation?

Thing 4: Searchable resources and repositories

Background: To make objects findable we have to commit ourselves to at least two major points: 1) these objects have to be identifiable at a fixed place, and 2) this place should be fairly visible. When it comes to finding data this is where the role of repositories comes in.

1. Utrecht University has its own repository YODA, short for "YOur DAta". It is possible to publish a dataset in this repository so that it becomes accessible online. Try to search for one of the datasets listed on YODA in Google Data Search. Take "ChronicalItaly" as an example. Was it difficult to find the dataset? Now try to search for one of the databases stored at the Meertens Institute using Google Dataset search. Why are the results so different?

2. Take a look at the storage solutions suggested by Utrecht RDM Support. Identify searchable repositories among these solutions.

Thing 5: Persistent identifiers

Background: A persistent identifier is a permanent and unique referral to an online digital object, independent of (a change in) the actual location. An identifier should have an unlimited lifetime, even if the existence of the identified entity ceases. This aspect of an identifier is called "persistency".

1. Read about the Digital Object Identifier (DOI)) System for Research Data provided by the Australian National Data Service (ANDS).

2. Watch the video "Persistent identifiers and data citation explained" by Research Data Netherlands. Read about persistent identifiers on a very general level (awareness).

Thing 6: Documentation

1. Browse through the general overview of data documentation as provided by the Consortium of European Social Science Data Archives. Think of the principal differences between object-level documentation of quantitative and qualitative data.

Thing 7: Formats and standards

1. Take a look at data formats recommended by DANS. Which of these formats are relevant for your subject area and for your data. Do you use any of the non-preferred formats? Why?
2. Read the background information about file formats and data conversion provided by the Consortium of European Social Science Data Archives. Reflect on the difference between short-term and long-term oriented formats. Think of a particular example of changing from a short-term processing format to a long-term preservation format, relevant for your field.

Thing 8: Controlled vocabulary

Background: The use of shared terminologies strengthens communities and increases the exchange of knowledge. When the researchers refer to specific terms, they rely on common understanding of these terms within the relevant community. Controlled vocabularies are concerned with the commitment to the terms and management standards that people use.

1. Browse Controlling your Language: a Directory of Metadata Vocabularies from JISC in the UK. Reflect on possible issues that may arise if there is no agreement on the use of a controlled vocabulary within a research group.

2. Consider the following example from earth science research: *"to be able to adequately act in the case of major natural disasters such as earthquakes or tsunamis, scientists need to have knowledge of the causes of complex processes that occur in the earth's crust. To gain necessary insights, data from different research fields are combined. This is only possible if researchers from different applicable sub-disciplines 'speak the same language'"*. Choose a topic within your research interests that requires combining data from different sub-disciplines. Think about some differences in vocabularies between these sub-disciplines.

Thing 9: Use a license

Background: A license states what a user is allowed to do with your data and creates clarity and certainty for potential users.

1. Take a look at various Creative Commons licences. Which licenses put the least restrictions on data? You can make use of Creative Commons guide to figure this out.
2. Watch this video about Creative Commons licences.

Thing 10: FAIR and privacy

Background: The General Data Protection Regulation (GDPR) and its implementation in the Netherlands called Algemene Verordening Gegevensbescherming(AVG) requires parties handling data to provide clarity and transparency where personal data are concerned.

1. Take a look at at the Handling personal data guide from the Utrecht University RDM website. Reflect on how personal data can be FAIR.

TOP 10 FAIR DATA & SOFTWARE THINGS:

International Relations

Sprinter:

Fiona Bradley, UNSW Library, and University of Western Australia (PhD Candidate)

Description:

International Relations researchers increasingly make use of and create their own datasets in the course of research, or as part of broader research projects. The funding landscape in the discipline is mixed, with some receiving significant grants subject to Open Access and Open Data compliance while others are not funded for specific outputs. Datasets have many sources, they may be derived from academic research, or increasingly, make use of large-N datasets produced by polling organisations such as YouGov, Gallup, third-party datasets produced by non-governmental organisations or NGOs that undertake human rights monitoring, or official government data. There is a wide range of licensing arrangements in place, and many different places to store this data.

What is FAIR data?

FAIR data is findable, accessible, interoperable and reusable. For more information, take a look at the Force 11 definition.

Audience:

International relations and human rights researchers

Goal:

Help researchers understand FAIR principles

Things

Thing 1: Getting started

Is there a difference between open and FAIR data? Find out more: https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/

ACTIVITY:
Are there examples in your own research where you have used or created data that may be FAIR, but may not necessarily be open? * Does the material you used or created include personal information? * Does it include culturally sensitive materials? * Does it reveal information that endangers or reveals the location of human rights defenders, whistleblowers, or other people requiring protection? * Does it involve material subject to commercial agreements?

Thing 2: Discovering data

United Nations (UN) agencies, international organisations, governments, NGOs, and researchers all produce and share data. Some data are very easy to use - they are well-described, and a comprehensive code book may be supplied. Other data may need significant clean up especially if definitions or country borders have changed over time, as they will in longitudinal datasets. A selection of the types of datasets available are linked below:

- Polity IV dataset
- World Bank Open Data
- ITU Global IT statistics
- Freedom House reports
- American Journal of Political Science (AJPS) Dataverse
- UK dataset guidelines (provides advice on using many open datasets)
- ICPSR: Inter-university Consortium for Political and Social Research

Thing 3: Data identifiers

A unique, permanent link helps make it easy to **identify** and **find** data. A Digital Object Identifier (DOI) is a widely used identifier, but not the only one available. If you are contributing a dataset to an institutional repository or discipline repository, these services may ‘mint’ a DOI for you to attach to your dataset.

Zenodo is an example of an open repository that will provide a DOI for your dataset. The AJPS Dataverse and UK Data Service, linked in **Thing 2**, both use DOIs to identify datasets.

Thing 4: Data citation

Using someone else’s dataset? Or want to make sure you are credited for use of data? The Make Data Count initiative and DataCite are developing guidelines to ensure that data citations are measured and credited to authors, in the same way as other research outputs.

Currently many researchers, NGOs, and organisations contribute data to the UN system or at national level to show progress on the UN 2030 Agenda for Sustainable Development, including the Sustainable Development goals. There are several initiatives aimed at strengthening national data including national statistical office capacity, disaggregated data, third-party data sources, and scientific evidence.

Thing 5: Data licensing

Depending on your funder, publisher, or purpose of your dataset, you may have a range of data licensing compliance requirements, or options. Creative Commons is one licensing option. The Australian Research Data Commons (formerly known as the Australian National Data Service) provides a guide with workflows for understanding how to licence your data in Australia.

ACTIVITY:
When might a Creative Commons licence not be appropriate for your data? For example: * When you are working on a contract and the contracting body does not permit it? * When you are producing data for a body with a more permissive licence or different licencing scheme in place? * When you are producing data on behalf of a body with an Open Government Data licence? (Linked example is for UK) * Are there other examples?

Thing 6: Sensitive data

Human rights researchers, scholars studying regime change in fragile and conflict states, and interviews with security officials are among the cases where data may need to be handled carefully, and be sensitive. In these cases, procedures utilised in collecting the data must remain secure, and the data may be FAIR, but not open, or require specific access protocols and mediated access. See:

- A human-rights based approach to data, UN OCHR
- Data security, UK Data Service

Thing 7: Data publishing

Data sharing policies in political science and international relations journals vary widely. See:

- Data policies of highly-ranked social science journals

ACTIVITY:
* What might some general data requirements look like for international relations? Are the Data Access, Production Transparency, and Analytic Transparency guidelines for APSR (American Political Science Review) helpful? * Or do you prefer a less defined set of criteria, such as that set out by International Organization?

Thing 8: Funder requirements

Funder requirements vary. Gary King has compiled the policies of most major social science funders (and journals, see **Thing 7**).

Thing 9: Data sharing

Your funder or publisher may set requirements for data sharing, either as 'supplementary data', or in a data repository. But, what if you aren't funded, and aren't required to provide supplementary data or comply with data publishing conditions? Make it a habit and practice to prepare and release your datasets as FAIR data when appropriate. Choose a repository, claim an identifier (**Thing 3**), and licence it appropriately (**Thing 5**). Add links to your homepage and ORCID profile. See:

- Guide to Social Science Data Preparation and Archiving

Thing 10: Learn more

The Carpentries provide training and workshops on fundamental data skills for research.

TOP 10 FAIR DATA & SOFTWARE THINGS:

Humanities: Historical Research

Sprinters:

Kristina Hettne, Peter Verhaar (Centre for Digital Scholarship at Leiden University), Ben Companjen, Laurents Sesink, Fieke Schoots (Centre for Digital Scholarship at Leiden University, reviewer), Erik Schultes (GO FAIR, reviewer), Rajaram Kaliyaperumal (Leiden Universitair Medisch Centrum, reviewer), Erzsebet Toth-Czifra (DARIAH, reviewer), Ricardo de Miranda Azevedo (Maastricht University, reviewer), Sanne Muurling (Leiden University Library, reviewer).

Description:

This document offers a concise overview of the ten topics that are most essential for scholars in the field of historical research who aim to publish their data set in accordance with the FAIR principles. In historical research, research data mostly consists of databases (spreadsheets, relational databases), text corpora, images, interviews, sound recordings or video materials.

Things

Findable

To ensure that data sets can be found, scholars need to deposit their data sets and all the associated metadata in a repository which assigns persistent identifiers.

Thing 1: Data repositories

Data repositories enable researchers to share their data sets. The following data repositories accept data sets in the field of history:

- DANS EASY
- Figshare
- Zenodo
- B2SHARE

A number of additional data repositories can be found by going to re3data.org, and by clicking on Browse > Browse by subject > History

Choosing a repository that complies with the CoreTrustSeal criteria for long term repositories is recommended. This way, the durable findability of the data is guaranteed.

ACTIVITIES:
1. Study the data set that can be found via https://doi.org/10.17026/dans-zw3-fkxb. How can the dataset be downloaded? Which formats are available?

Thing 2: Metadata

Once a certain data repository has been selected, the data set can be submitted, together with the metadata describing this data set. Metadata is commonly described as data about data. In the context of data management, it is structural information about a data set which describes characteristics such as the quality, the format and the contents. Most repositories require a minimum set of metadata, such as name of the creator, the title and the year of creation. Check what kind of metadata the repository you choose asks. Remember that the effort you put into metadata will contribute to the findability of your dataset.

Metadata are often captured using a fixed metadata schema. A schema is a set of fields which can be used to record a particular type of information. The format of the metadata is often prescribed by the data repository which will manage the data set.

ACTIVITIES:
1. Read the Digital Scholarship @ Leiden blog to learn about metadata for humans and machines 2. Log in at Zenodo.org and click on Upload > New Upload. On the web page that appears, take stock of the various metadata fields that need to be completed. Zenodo is an international repository. Different countries and institutions might have other preferred repositories, such as DANS EASY. DANS EASY list the following specific requirements for historical sciences: Historical sciences: 1) a description of the (archival) sources; 2) the selection procedure used; 3) the way in which the sources were used; and 4) which standards or classification systems (such as HISCO) were used. Read more at https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing

Thing 3: Persistent identifiers

Datasets need to be deposited in repositories that assign persistent identifiers (PIDs) to ensure that online references to publications, research data, and persons remain available in the future. A PID is a specific type of a Uniform Resource Identifier (URI), which is managed by an organisation that links a persistent identification code with the most recent Uniform Resource Locator (URL).

Academic journals mostly work with DOIs. DOIs are globally unique identifiers that provide persistent access to publications, datasets, software applications, and a wide range of other research results. DOI has been an ISO standard since 2012. A typical DOI looks as follows:

http://doi.org/10.17026/dans-x4b-uy8q. When users click on this DOI, the DOI is resolved to an actual web address.

Next to identifiers for data sets and for publications, it is also possible to create PIDs for people. Open Researcher and Contributor Identifier (ORCID) is an international system for the persistent identification of academic authors. It is a non-proprietary system, managed by an international consortium consisting of universities, national libraries, research institutes and data repositories. When your research results are associated with an ORCID, this information can be exchanged effectively across databases, across countries and across academic disciplines. You always retain full control over your own ORCID id. It is the de facto standard when submitting a research article or grant application, or depositing research data.

ACTIVITIES:
1. Watch the video "Persistent identifiers and data citation explained" by Research Data Netherlands. 2. Watch the video "What are persistent identifiers" for an example on how they are used in digital heritage. 2. If you don't have one, request an ORCID. Add all your information as completely as possible. 3. Read Alice Meadow's blog post Six Things to do now you have an ORCID iD. 4. Go to a data record and click on the DOI to see how the DOI can be resolved to current URL of the data set: http://dx.doi.org/10.17026/dans-x4b-uy8q. 5. Read "Digital Object Identifier (DOI) System for Research Data".

Accessible

Thing 4: Open data

The FAIR principles stipulate that data and metadata ought to be "retrievable by their identifier using a standardised communication protocol" (requirement A1). This requirement does not necessarily imply that the data should fully be available in open access. It principally means that there needs to be a protocol that users may follow to obtain of the data set. There can be many good reasons for limiting the access to a file. Public accessibility may be difficult because of privacy laws or copyright protection regulations, for example.

The accessibility of the data may occasionally be complicated by the fact that the data have been stored using a so-called proprietary format, i.e. a format that owned exclusively by a single company. For formats which are associated with specific software applications, it can be difficult to guarantee their long-term usability, accessibility and preservation. For this reason, the DANS EASY archive in The Netherlands works with a list 'preferred formats'.

ACTIVITIES:
1. Read the article on the website of DANS about preferred formats, and about what you can do to improve the durability of non-preferred formats. 2. Read the web page on open data on the ANDS website. 3. Consider the following three articles. To what extent can the data sets

that are mentioned in the articles be accessed? Are the data sets also in preferred formats? * https://doi.org/10.1080/0969594X.2016.1194257 * http://dx.doi.org/10.1371/journal.pone.0139563 * http://doi.org/10.1111/lang.12172 4. Look at the data set that can be found via https://doi.org/10.17026/dans-x5u-usxj. What is needed to access the data?

Interoperable

Thing 5: Data structuring and organisation

Well-structured and well-organised data can evidently be reused much more easily. This section explains how researchers can organize their data in such a way that they can be analysed effectively with data science tools. Many historians capture their data in spreadsheets. As is explained by Broman and Woo (2018), there are a number of important principles to bear in mind when you work with spreadsheets.

- It is important to be consistent. Terminology should be used invariably.
- Avoid empty cells. Use a consistent code for data which is unavailable, such as 'NA' used in R.
- Use a regular format for dates, such as YYYY-MM-DD.
- Use all cells to capture atomic data. Do not place multiple values in a single cell. Every value that you may want to use in calculations or in other analyses needs to be available separately.
- Organise the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row)
- Do not make use with colours to indicate properties of data.. Represent all data that you need as actual values in the spreadsheet.
- Do not include calculations in the raw data files.

Once you have developed a suitable data model, you are also advised to develop a data dictionary which documents the model. This document may contain the following information:

- A list of all the column names used in the data spreadsheet
- A description of the purpose and the contents of these different columns.
- If applicable, give an indication of the units of measurement.
- If applicable, describe the measures that have been taken to ensure the correctness and the consistency of the data

- Explain abbreviations or notational conventions that have been used in the data set.

ACTIVITIES:
Read Karl Broman and Kara H. Woo, "Data organization in spreadsheets".

Thing 6: Controlled vocabularies and ontologies

Tim Berners-Lee, the inventor of the Web, argued that there are five levels of open data. Creators of data can earn five stars by following the steps below.

1. Data sets can be awarded one star if it has been made public. This is clearly the case for data which have been published via an open license in a data repository.
2. In order to win a second star, the open data needs to be made available as machine-readable data. This criterion can be satisfied by providing access to an Excel Spreadsheet, for instance.
3. One disadvantage of an Excel spreadsheet is that users need proprietary software to open the data. The third star can be awarded to datasets which are captured using open formats, such as CSV or TXT.
4. A fourth star can be awarded when the entities in the data set are identified using persistent identifiers. Such PIDs have the effect that other researcher can effectively link to the data set.
5. The fifth star can be earned by linking the data to entities in other data sets via PIDs.

When researchers have published their well-structured and their well-organised data set in a data repository via a public license, as explain in things 1 to 5 above, they will have arrived at data set that can be awarded three stars, according to Berners-Lee's scheme. This section and the following section will further explain how you enhance the interoperability of their data sets even further by working with RDF and with persistent identifiers.

As a first step, it can be useful to explore whether some of the general topics that you focus on have already been assigned persistent identifiers or URIs. Many researchers and institutions have developed shared vocabularies and ontologies to standardise terminology. In many cases, the terms which have been defined have also been assigned persistent identifiers. Such shared vocabularies can make it clear that we are talking about the same thing when we exchange knowledge.

Historical research often concentrates on people, events, organisations and locations. The following ontologies and shared vocabularies concentrate on entities such as these:

- The CIDOC Conceptual Reference Model (CRM) concept search.
- Wikidata assigns identifiers to a wide range of entities, including people, locations and organisations
- The Library of Congress name authority files, e.g. http://id.loc.gov/authorities/names/n79021400.

- VIAF (Virtual International Authority File (https://viaf.org/)
- Identifiers for book published in Dutch or in the Netherlands can be found via the STCN, whose contents is available as Linked Open Data.
- The UNESCO history thesaurus.
- Aspects of books can be described using terms from the Bibliographic Ontology and the FABIO ontology.
- GeoNames defined persistent identifiers to locations, e.g. https://www.geonames.org/2751773/leiden.html.
- TaDiRAh and BARTOC (Basel Register of Thesauri, Ontologies & Classifications also offer valuable overviews of the ontologies that have been developed within specific disciplines.
- One of the ways to describe the provenance of data sets is by so-called nanopublications, i.e. a set of Resource Description Framework (RDF) triples (subject-predicate-object tuples). Although you do not need nanopublications to describe provenance, Nanopublications are a way of combining argument and provenance in a single package. Nanopublications rely on the Provenance Ontology to express provenance. You can read more about them and their application in historical research in this paper by Patrick Golden and Ryan Shaw: Nanopublication beyond the sciences: the PeriodO period gazetteer

Where possible, try to use terms that have been defined in these existing ontologies in your own data set. An example where a specific vocabulary (the VOC glossary) was used to markup a dataset can be found here. The dataset is part of a project to reconstruct the domestic market for colonial goods in the Dutch Republic.

ACTIVITIES:
1. Try to find one or two terms that are relevant to your research using the resources that are mentioned above. You can aso use Swoogle to search for vocabularies related to your research. 2. Search for a term related to your research in the CIDOC Conceptual Reference Model (CRM) concept search. Were you able to find it? Tip 1: Search for “person” to get an idea of how the thesaurus works. Tip 2: All the terms used can be found in the last release of the model: http://www.cidoc-crm.org/get-last-official-release.

Thing 7: FAIR data modelling

The fourth and the fifth star in Berner Lee’s model can be awarded when the data are stored in a format in which the topics their properties and their characteristics are identified using URIs whenever possible. More concretely, it implies that you record your data using the Resource Description Framework (RDF) format. RDF, simply put, is a technology which enables you to publish the contents of a database via the web. It is based on a simple data model which assumes that all statements about resources can be reduced to a basic form, consisting of a subject, a predicate and an object. RDF assertions are also known as triples. In a FAIR data model, elements of data are organised and identified using PIDs. The same goes for the relations between these elements. The FAIR data model is a graphical view of the data that act as a metadata key to a spreadsheet but it can also be used as a guide to expose data as a linked data graph in RDF format.

Existing data sets can be converted to RDF by making use of the FAIRifier software. This application is based on OpenRefine. Other examples of tools to generate RDF are Karma and RML. In the FAIRifier, it is possible to upload a CSV file. After this, the data set can be connected to elements from existing ontologies.

ACTIVITIES:
1. Learn about the basics of RDF modeling by going through the first 15 slides of the webinar about the UNESCO Thesaurus. 2. Dig in deep by exploring the FAIRifier for a dataset you already have available in CSV.

Reusable

Thing 8: Licensing

A license describes the conditions under which your data or software is (re)usable. Picking a license can be a daunting process because of the common feeling that if you do not pick the right license something will go wrong. However keep in mind that if you do not choose a license for your data or software, it means that it cannot be used or reused. A copyright expert can help you, but to get you going you can try out the activities listed below.

ACTIVITIES:
1. Try to pick a license for a data set you are working on by using the Creative Commons license picker 2. Try to pick a license for a piece of software or code you are working on by using the choose a license picker 3. Learn more about licensing your data by reading this guide from the Digital Curation Center

If you deposit your data in a repository there will be default options available.

Thing 9: Data citation

When you have made use of someone else's data, you are strongly recommended to attribute the original creators of these data by including a proper reference. Data sets, and even software applications, can be cited in the same way as textual publications such as articles and monographs. Structured data citations can also be used to calculate metrics about the reuse of the data. Data citations, regardless of citation style, typically contain the authors, the year, the title, the publisher and a persistent identifier.

ACTIVITIES:
1. Read the ANDS guide on data citation. 2. Read the FORCE11 data citation principles. 3. Study the following data set on figshare: https://doi.org/10.6084/m9.figshare.3519755.v1. Note that there is the possibility to generate a data citation, under the link "Cite", in the citation style of your choice. 4. Consider the following publication: https://doi.org/10.1371/journal.pone.0149621. Note that the article has a "data availability" statement. 5. Explore CiteAs by typing in the figshare doi from above (10.6084/m9.figshare.3519755.v1).

Context

Thing 10: Policies

Policies for data availability can come from publishers, funders and universities. These policies are listed on the respective website, but finding these is not always straightforward. FAIRsharing is a repository for standards, databases and policies with the possibility to filter on information for a specific research domain. It started as an initiative for the life sciences but is rapidly expanding its content for other disciplines as well.

ACTIVITIES:
1. Start by going to FAIRsharing 2. Click on the blue "Policies" button at the top 3. In the left side menu under "Subjects", click on "show more" and select "Humanities". 4. Scroll down to the Taylor and Francis Data Policy 5. Which databases and standards are mentioned in this policy? 6. Go to the specific policy for the "European Review of History" journal. 7. Does it differ from the general Taylor and Francis policy? 8. Try to find the data policy for your favorite journal.

TOP 10 FAIR DATA & SOFTWARE THINGS:

Geoscience

Sprinters:

John Brown, Janice Chan, Niamh Quigley (Curtin University, Perth, Western Australia)

Audience:

Researchers

Things

Findable

Thing 1: Data sharing and discovery

Thing 6: Vocabularies for data description

Thing 7: Identifiers and linked data

Thing 10: Spatial data

Accessible

Thing 2: Long-lived data: curation & preservation

Thing 3: Data citation for access & attribution

Thing 4: DOIs and citation metrics

Interoperable

Thing 4: DOIs and citation metrics

Thing 6: Vocabularies for data description

Thing 7: Identifiers and linked data

Thing 9: Exploring APIs and Apps

Reusable

Thing 5: Licensing data for reuse

Thing 8: What are publishers & funders saying about data?

Thing 1: Data sharing and discovery

Activity 1: Data discovery

Data repositories enable others to find existing data by publishing data descriptions ("metadata") about the data they hold, much like a library catalogue describes the resources held in a library. Also, repositories often provide access to the data itself and some even provide ways for users to explore that data. Many research funding requirements reference researchers depositing their data into data repositories (which we'll discuss later in Thing 8).

Data portals or aggregators draw together research data records from a number of repositories. Because of the huge amounts of data available they sometimes focus on data from one discipline or geographic region. The EU Open Data Portal is an example that aggregates metadata records from over 30 European national data repositories and The US Government's Open Data portal data.gov aggregates from over 100 US government agencies.

1. Look at this Data.gov.au record from Geoscience Australia: Lord Howe Rise Marine Survey 2017.

- Examine the **Description** and **Additional Info** fields to see the ways that Geoscience Australia has made this record findable to other researchers. If you knew about this data portal, would you be able to easily find this dataset if it was relevant to your research?

2. Spend a few minutes exploring the Scottish Spatial Data Infrastructure Metadata Portal.

- Try browsing or searching on a topic of interest.
- Explore a record and see where it came from and if there's a way to contact the creator.
- Have a look at the map and see if you can find and add a map layer relating to fishing.

3. Look at EarthChem.

- Have a look at some of the data in EarthChem. Would it be a good place to contribute the data from your own research?

Consider: If your research appeared in the right data portal or repository, what things might result from that for yourself? What about your discipline?

Activity 2: Finding data repositories

1. Choose one of the specialised data repositories below, or find another data repository on re3data.org (perhaps one outside your particular focus area) and spend some time browsing around your chosen repository to get a feel for the data available.

- WorldClim
- Southern California Earthquake Centre
- MOPITT (Atmospheric Science Data Centre)
- International Service of Geomagnetic Indices
- Scientific Drilling Database
- Alberta Geological Survey

2. Think about how the data here differs from data you are familiar with, for example, in format, size and access method.

Consider: Could you apply a dataset from one of these repositories to your own work? Would you need to change file formats or learn a new software package?

Thing 2: Long-lived data: curation & preservation

Activity 1: Preserving born digital objects

Information sources that were commonly used in the past such as maps and handwritten observation notes and can easily survive for years, decades or even centuries. However, because most current research is done mostly on computers, it's important to remember that digital items require special care to keep them usable over time.

1. This video (2.5 min) from the US Library of Congress shows the vulnerability of "born digital" objects like research data: they are fragile; they are dependent on software and hardware; and they require active management.
2. Look at the ANDS page on file formats.

Consider: If your research was put into a time capsule and unearthed in 50 years' time, would future researchers be able to determine if your research is still useful to them? If you were allowed to update the time capsule every 5 years, what would you change to make it easier for those unearthing it?

Activity 2: Readme files

One way that researchers can ensure their data is useful in the future is to package their data with an explanation that can be opened without any software. These explanatory files mean that anyone who finds the data will know if the data is useful to them and hopefully won't have any questions for the original researcher, who may not be available or not remember. The files are usually called "readme" files in the hope that by reading the file, all the important questions will be answered.

1. Read the Guide to writing "readme" style metadata from the Cornell Research Data Management Service Group and create a readme.txt file for one of your own datasets.

Don't forget to include notes on software versions used, methodology and any special things you'd tell a colleague if you were giving them the data yourself!

Thing 3: Data citation for access & attribution

Activity 1: Citing research data

When authors cite an article they have used ideas from, they formally and publicly acknowledge the work of the earlier author. Data citation works in the same way – by citing the data created by earlier researchers they get formal and public credit for their contribution to the new work. Along with books, journals and other scholarly works, it is now possible to formally cite research datasets and even the software that was used to create or analyse the data.

1. Have a look at https://www.bgs.ac.uk/services/ngdc/citedData/catalogue/a59128b5-8e7f-4100-b0ff-87325438435b.html the Geophysical, hydraulic and mechanical properties of synthetic versus natural sandstones under variable stress conditions dataset from the British Geological Survey. If someone wanted to use this dataset for further research, would they know how to give credit to the creator of the original dataset?
2. Find a DOI of a dataset from one of the repositories you found in Thing 1 and enter it into the DOI Citation Formatter: https://citation.crosscite.org/. If you saw the citation, would you know how to go about accessing the data?
3. Read the article, "Sharing Detailed Research Data Is Associated with Increased Citation Rate" – why would it be that papers that make their data openly available get better citation counts? Would you feel more confident citing another person's work if you knew?

Consider: Data citation is a relatively new concept in the scholarly landscape and as yet, is not routinely done by researchers, or demanded by journals. What could be done to encourage routine citation of research data and software associated with research outputs?

Activity 2: Citing software

The increase in available computational power over the last 50 years has led to a massive increase in the usage of computational analysis methods in geoscience.

As such techniques become more commonplace, it's important to distinguish between the data itself, the tools used to analyse data and any discrete components within those tools. In some cases, a particular function of the software is critical to the analysis process; in other cases the critical part is an interchangeable block of code within that software package. Recognising the difference between these two is important as it changes who gets credit for their previous work and who gets left unsung.

It’s not always easy to know which to cite, but trying to give recognition for the creation of software and software components can make huge impacts on the career of a researcher, especially if they create scientific software!

1. Read https://libguides.mit.edu/c.php?g=551454&p=3900280 the How to cite software guide from the MIT Libraries.
2. Read Adding CITATION to your R package blog post.

Consider: If you wrote a package of code for a computer program to run and made it freely available to your colleagues to solve a problem in your field, would they know how they could give you credit in their work? Would they think that you would want attribution?

Thing 4: DOIs and citation metrics

DOIs are unique identifiers that enable data citation, metrics for data and related research objects, and impact metrics. Citation analysis and citation metrics are important to the academic community. Find out where data fits in the citation picture.

Activity 1: DOIs

Digital Object Identifiers (DOIs) are a type of ‘persistent identifier’. DOIs are unique identifiers that provide persistent access to published articles, datasets, software versions and a range of other research inputs and outputs. There are over 120 million Digital Object Identifiers (DOIs) in use, and in 2016 DOIs were “resolved” (clicked on) over 5 billion times!

Each DOI is unique but a typical DOI looks like this:
http://dx.doi.org/10.4225/06/577F022BA6954

1. Start by watching this short 4.5-minute video Persistent identifiers and data citation explained from Research Data Netherlands. It gives you a succinct, clear explanation of how DOIs underpin data citation.
2. Have a look at the poster Building a culture of data citation and follow the arrows to see how DOIs are attached to data sets and are used in data citation.
3. Let’s go to a Research Data Australia data record which shows how DOIs are used. Click on this DOI to ‘resolve’ the DOI and take us to the record: http://dx.doi.org/10.4225/06/577F022BA6954.
4. Click on the **Cite** icon on the upper left of the record (under the green **Access the data** tab). No matter where the DOI appears it always resolves back to its original dataset record to avoid duplication. i.e. many records, one copy.
5. DOIs can also be applied to grey literature, a term that refers to research that is either unpublished or has been published in non-commercial form, such as government reports. For example, reports like this: http://doi.org/10.4225/06/583d354b89060.

Activity 2: IGSNs

International Geo Sample Number (IGSN) are designed to provide an unambiguous globally unique persistent identifier for physical samples. It facilitates the location, identification, and citation of physical samples used in research.

Each IGSN is unique but a typical IGSN looks like this IEEVB00C3. The first five characters of the IGSN represent a name space (a unique user code) that uniquely identifies the person or institution that registers the sample. The last 4 characters of the IGSN are a random string of alphanumeric characters (0-9, A-Z).

1. Start by reading this brief introduction to IGSN.
2. Review the scope and capability of each IGSN allocation agent listed on the IGSN website and consider which allocation agent is most appropriate for your samples.
3. Have a look at an IGSN record https://app.geosamples.org/sample/igsn/IEEVB00C3 which displays what information about the sample was recorded.
4. Now have a look at how IGSNs are referenced in a dataset record http://get.iedadata.org/doi/100548.

Consider: How are you managing your physical samples? The ANDS IGSN minting service may be used by Australian researchers at no cost. Do you know of a service provider in your region?

Activity 3: Altmetrics

Data citation best practice, as discussed in Thing 3, enables citation metrics for data to be tracked and analysed. Data citations are available from the Clarivate Data Citation Index which is a commercial product.

Altmetrics is an alternative measure to help understand the influence of your work. It refers to metrics such as number of views, number of downloads, number of mentions in policy documents, social media, and social bookmarking platforms associated with any research outputs that have a DOI or other persistent identifiers. Because of their immediacy, altmetrics can be an early indicator of the impact or reach of a dataset; long before formal citation metrics can be assessed.

1. Start by looking at the altmetrics for this phylogenomics article published in Science. Note the usage statistics, including number and pattern of downloads, for this article since it was published in November 2014.
2. Now click on the “donut” or the link to ‘See More Details’ to see the wealth of information available.

3. Look also at the associated data in Dryad noting that the data has been assigned a DOI. Can you see how many times the data has been downloaded and the record viewed (scroll down to the bottom of the record)?

By way of comparison, as of early November 2018: * the same dataset had been cited once in Web of Science Data Citation Index * the article had been cited 690 times in Web of Science

Consider: Do you think altmetrics for data have value in academic settings? Why, or why not?

Thing 5: Licensing data for reuse

Understand the importance of data licensing, learn about Creative Commons and find out how enabling reuse of data can speed up research and innovation.

Activity 1: Why license research data?

Consider this scenario: You've found a dataset you are interested in. You've downloaded it. Excellent! But do you know what you can and cannot do with the data? The answer lies in data licensing. Licensing is critical to enabling data to be reused and cited.

1. Start by reading this brief introduction to licensing research data.
2. Now watch this Creative Commons Licensing introductory video or have a closer look at the Understanding CC Licences poster.
3. Check out the licence chooser from Creative Commons, which walks you through the decision of which licence is appropriate for your purpose.

Consider: If you were considering licensing a dataset on something which may have commercial value to others - what licence would you apply?

Activity 2: Data licences: unlock data for innovation

Enabling reuse of data can speed up research and innovation. Licensing is critical to enabling data reuse.

1. Start by watching this 4.30mins video in which Dr Kevin Cullen from the University of New South Wales explains their approach to licensing which aims to strengthen the University's relationship with business and industry.
2. Check out the data standards of Geoscience Australia, which refers to the Australian government policy on Public Data. Which Creative Commons licence is applied to government data by default?
3. Since November 2009, Geoscience Australia has officially adopted Creative Commons Attribution as the default licence for its website. That means thousands of products and datasets available through the website are free to be reused.

4. See the range of data products and license available at British Geological Survey.

Does your institution have policies or guidelines around data licensing?

Activity 3: Data licensing in practice

Not all research data that is shared is licensed for reuse. It should be!

1. Explore the following data repositories:
- Research Data Australia
- AuScope Geonetwork Portal
- EarthChem
2. Or review the following example records:
- Darwin Harbour marine habitats
- Mineral Occurrences - South Australia
- Whole Rock Composition Data for Garnet Pyroxenites from Arizona
3. Do all data repositories or metadata catalogues enable users to refine search by licenses? Look closely at the specific Licensing information on a small sample of those records with ‘open’ licences. How easy or difficult it is to work out if the data can or can’t be reused e.g. for commercial purposes? with international collaborators?

Consider: Assigning Open Licenses is not routine. Suggest one tip for encouraging uptake of 'open' licensing.

Thing 6: Vocabularies for data description

In addition to selecting a metadata standard or schema, whenever possible you should also use a controlled vocabulary.

Activity 1: What is controlled vocabulary?

A controlled vocabulary provides a consistent way to describe data - location, time, place name, subject. Read this short explanation of controlled vocabularies.

Controlled vocabularies significantly improve data discovery. It makes data more shareable with researchers in the same discipline because everyone is ‘talking the same language’ when searching for specific data e.g. plants, animals, medical conditions, places etc.

If you have time, have a look at Controlling your Language: a Directory of Metadata Vocabularies from JISC in the UK. Make sure you scroll down to 5. Conclusion - it’s worth a read.

Activity 2: Controlled vocabularies in action

We are going to see some controlled vocabularies in action in the Atlas of Living Australia (ALA).

1. Do a search in the ALA search engine. Type “whale” in the search box and click on search. Choose one of the records listed and click on the (red text) View record link.
2. Any metadata field where you see Supplied… tells you that the information supplied by the person who submitted the record (often a 'citizen scientist') has been changed to the controlled vocabulary being used in metadata fields e.g. Observer, Record date and Common name.
3. Have a scroll down the record and consider how many of the metadata fields probably have a controlled vocabulary in use (e.g. taxonomy, geospatial etc.).

If you have time: have a browse around the stunning level of data description and data contained in the Atlas of Living Australia.

Activity 3: Geoscience vocabularies

Explore some examples of vocabularies used in geoscience:

- American Geosciences Institute GeoRef Thesaurus
- Geological Survey of Western Australia Geoscience Thesaurus (GeMPeT)
- Geosciences Australia vocabularies register
- British Geology Society Vocabularies

Consider: Do you use controlled vocabularies to describe your data? How would you encourage other researchers to use them?

Thing 7: Identifiers and linked data

ORCID is a unique identifier for researchers. Many research data repositories record your ORCID when you submit research data for publication.

Activity 1: Check your ORCID

In your ORCID record, datasets you have published will be displayed in the Works section.

Log into ORCID now and check your details are up to date, including: * email address * biography * research keywords * other IDs such as Scopus Author ID.

If you don’t already have an ORCID you can get one, this Curtin University webpage has information on how to get the most out of your ORCID.

Activity 2: Get more from your ORCID

ORCID populates your ORCID record from many sources, one of which is peer review activities. Publishers such as the American Geophysical Union Publications now send details of peer review activities to ORCID.

- Look at your ORCID record, if you have undertaken peer review activities are they listed?
- Why do you think linking peer review activities to ORCIDs could be useful?

Activity 3: Identifiers and linked data

Because they are unique identifiers, ORCIDs can be used to link data from different datasets together. GeoLink is a network of Linked Data from multiple data repositories.

1. Go to the portal for the GeoLink demo.
2. Choose an entity e.g. Datasets, Cruises, Vessels, Instruments, Researchers and explore! The Help guide is here.

Thing 8: What are publishers & funders saying about data?

Geoscience research data is a world heritage. Researchers share the responsibility with research institutions and funders of ensuring their data is well-documented, preserved and openly available.

Many publishers have special requirements for the citation of data in publications. This can be in the form of compliance with a data policy, author guidelines or the completion of a Data Availability Statement.

Activity 1: Research data and scholarly publishing

Have a look at the Nature Data Availability Statement examples or the PLOS Data Availability policy to get an idea of what publishers expect.

COPDESS is The Coalition for Publishing Data in the Earth and Space Sciences, and they have collected links to author instructions and data policies for some geoscience journals, publishers and funders.

Activity 2: Research funders and data sharing

Activity 1 has shown us that it’s becoming more common for journals and publishers to demand your data be made available when you seek to publish. However, if your research is publicly funded it’s almost guaranteed that your grant and funding obligations with require you to make your data publicly available at the end of your project – the outputs of research funded by a population should be made available to that population.

The Australian Research Council's data management requirements states that funded researchers are expected to follow the OECD Principles and Guidelines for Access to Research Data from Public Funding. Similar principles are outlined by the UK Research and Innovation (UKRI) in their Guidance on best practice in the management of research data document.

Consider: If you were on a funding panel and were asked to assess a grant with a clear plan for making the data openly available, would you rate the future impact of that proposal better or worse than one with a poorly defined plan?

Thing 9: Exploring APIs and applications

Geosciences has many specialised services, applications and APIs which can be used to directly access and harness existing research data. Some are free, and some are subscription-based, but your research institution may have access.

Activity 1: Try an app

- The WA Geology app created by the Western Australian government, can be used in a mobile web browser and provides multiple layers of geoscience information for Western Australia.
- The British Geological Survey has created the free iGeology app to explore hundreds of British maps.

Activity 2: APIs

APIs (Application Programming Interfaces) are software services that allow you to access structured data or systems held by someone else. These are usually provided so that developers can access data held by an organisation on demand, rather than them having to hold an entire dataset (which may not be possible due to security, space requirements or if the dataset is constantly changing). Some companies charge for using their APIs, but many research-oriented organisations provide their APIs for free so that other organisations can link in to their knowledge.

- The NASA Earth Data Developer Portal provides data from the NASA Earth Science Data portal.

- The Natural History Museum API provides a range of data from their collections.

Consider: If you could systematically access and integrate the data provided from one of the sources above, can you think of a way you could enrich the outputs of your own research?

Thing 10: Spatial data

The importance of spatial data is ever increasing. Many of the societal challenges we face today such as food scarcity and economic growth are inherently linked to big spatial data. In fact, it is often said that 80% of all research data has a geographic or spatial component. It is useful then, for all of us to have an understanding of spatial data.

Activity 1: Spatial data: Maps and more

1. Start by watching this incredible, inspiring video (3.59 min) from the University of Wollongong's PetaJakarta project. It shows innovative ways of combining social media and geospatial data to save lives.
2. Now read The Application of Geographic Information Science in Earth Sciences.
3. This video combines a range of different data visualisations depicting the human impacts on our environment.
4. Geospatial data is fundamental to Australia's economic future. Check out this very short article about how GeoScience Australia is mapping the mineral potential of our continent - a world first!

Just for fun: Enter your address in the Atlas of Living Australia and see what birds and plants have been reported in your street or suburb. You may be surprised at how 'alive' your street is!

Consider: Why do you think these geospatial visualisations are so powerful?

Activity 2: Spatial data concepts

There are many types and sources of geospatial data. If you are new to the world of geospatial data, you will probably appreciate some 'busting' of the jargon of geospatial data.

1. Start by reading this Fundamentals Chapter to learn more about maps, projections, coordinate systems, datums and GIS.
2. Want more? Continue with this blog about Finding and Making Sense of Geospatial Data on the Internet which explains some basic geospatial data file formats and concepts.
3. Prefer watching? Most of these concepts are also explained in this video.
4. Read more about two important aspects of spatial data: scale and resolution.

Consider: How would you give an explanation of two new terms you have just learnt?

Activity 3: Using and visualising spatial data

Spatial data can be used in many ways, and there are many tools that you can use to manipulate and display spatial data.

You can try one of the tools below. Do one, or do them all and compare the results.

1. 13 Free GIS Software Options: Map the World in Open Source

- Browse through this site for ideas for free, open source geospatial software; the descriptions often include discipline specific advice. Download one and try your hand at mapping.

2. Spatial data visualisation with R: For those who have done the R modules in Software Carpentry - this might be a good activity to flex your R muscles! Want more? Here are some more R tutorials.
3. Create a map using Google Fusion Tables: This offers lots of features, but you need a Google account. The excellent Google Fusion tutorial uses butterfly data to show you how to import data, map the data and customise your map.

The Open Geospatial Consortium (OGC) is an international not-for-profit organization that develops open standards for the geospatial community. OGC through their dedicated global members have developed several standards to share geospatial data. Some of the most commonly use standards are:

1. Web Map Service (WMS): A standard web protocol to query and access geo-registered static map images as a web service. The outputs are images that can be displayed in a browser application.
2. Web Feature Service (WFS): A standard web protocol to query and extract geographic features of a map, these are typically attributes of a map. The latest version of WFS (3.0, Dec 2017) has created a lot of excitement in the community.
3. Web Coverage Service (WCS): Provides access to geospatial information representing phenomena that are variable over space and time, such as satellite images or aerial photos. The service delivers a raster image that can be further interpreted and processed.

Geoserver is the most popular open source reference implementation of WMS, WFS and WCS standards.

Consider: The data world is hungry for Geospatial tools and metadata and there is growing demand for people with these skills. How can these skills be encouraged in your institution?

References:

ANDS 23 (Research Data) Things https://www.ands.org.au/working-with-data/skills/23-research-data-things/all23

10 Eco Data Things https://www.ands.org.au/__data/assets/pdf_file/0003/1376121/10-Eco-Data-Things_handout.pdf

Biomedical Data Producers, Stewards, and Funders

Sprinters:

Lisa Federer (National Library of Medicine), Douglas Joubert (National Institutes of Health Library), Allissa Dillman (National Center for Biotechnology Information), Kenneth Wilkins (National Institute of Diabetes and Digestive and Kidney Diseases), Ishwar Chandramouliswaran (National Institute of Allergy and Infectious Diseases), Vivek Navale (NIH Center for Information Technology), Susan Wright (National Institute on Drug Abuse)

Audience:

- Biomedical researchers
- Data stewards
- Funding organizations

Things

Thing 1: Metadata creation and curation

Beginner activity:

1. Learn about the various types of metadata. DataOne defines metadata as "documentation about the data that describes the content, quality, condition, and other characteristics of a dataset. More importantly, metadata allows data to be discovered, accessed, and reused" - DataONE Education Module.

- Descriptive
- Technical
- Administrative
- Provenance

2. Work through the DataOne Metadata Educational Module: Lesson 7 - Metadata.
3. Explore the use of controlled vocabularies and Common Data Elements (CDE). A CDE is a "data element that is common to multiple data sets across different studies." The NIH Common Data Element (CDE) Resource Portal has identified CDEs for use in particular types of research or research domains after a formal evaluation and selection process.

- Take the NIH CDE interactive tour to learn how to use the site.

- Browse the CDEs to explore how these might be used in your discipline.

Intermediate activity:

1. Think about ways you can standardize minimal/core metadata to use across disciplines. For example, crosswalk between standards).
2. Automated metadata creation can "help improve efficiency in time and resource management within preservation systems, and alleviate the problems associated to the "metadata bottleneck".
3. Review the Digital Curation Centre (DCC) Automated Metadata Generation primer page.
4. Download the DCC Digital Curation Reference Manual and think about the ways you might be able to automate metadata creation at your organization.
5. Watch the ALCTS Session 1: Automating Descriptive Metadata Creation: Tools and Workflows webinar which examines workflows for automating the creation of descriptive metadata.

Thing 2: Use of standard data models

1. Explore the OMOP Common Data Model (CDM), which allows for the systematic analysis of disparate observational databases.
2. Review one of the OMOP Community Meeting presentations and think about how this might align to the work of your organization.
3. Familiarize yourself with one of the Observational Health Data Sciences and Informatics GitHub repositories.

Thing 3: Exploring unique, persistent identifiers

Beginner activity:

Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset (GOFAIR)

1. Explore the GO FAIR F1 webpage to see examples of globally unique and persistent identifiers.
2. Learn how a Digital Object Identifier (DOI) can be used to create a unique reference to your data. Watch a video that explain what DOIs are and how they work, and how they benefit managers of digital content.
3. Read the Digital Preservation Handbook to learn about all of the elements that comprise a persistent identifier.

Intermediate Activity:

ORCID allows you to create persistent digital identifiers for authors.

1. Create an ORCID ID.
2. Link your ORCID with CrossRef and DataCite.
3. Then, go through steps included in the Getting Started with ORCID Integration guide.
4. Test the ORCID Application Programming Interface (API).
5. As a best practice, use ORCIDs from the start of data creation. For example, you can attach data creator name/ORCID to dataset as a metadata field. Include ORCIDs with datasets in repositories (e.g. in Sequence Read Archive (SRA), include the ORCID for the data creator). This allows for the tracking of your research and enables citation of your data.

Thing 4: Versioning and data "retirement"

Beginner activity:

A source-code repository is a file archive and web hosting facility where a large amount of source code, for software, web pages, and other resources, is kept, either publicly or privately. Advantages of versioning include:

1. Persistence of identifiers pointing to different/earlier versions
2. Maintaining previous versions of code, software, and data.
3. Sharing various levels of processed data (primary, secondary, or raw/clean/processed, etc.).
4. De-accessioning of data that has reached the end of its life cycle

Intermediate activity:

1. GitHub is one of the most popular options for code hosting. Explore alternative options for code hosting.
2. Work through the Library Carpentry Introduction to GitHub module.

Thing 5: Linking research objects

Beginner activity:

1. Read the following article on managing digital research objects.
2. Read the linking data CrossRef page.

Intermediate activity:

1. Using a (Github code repository or Zenodo), try to find data that goes with a published paper. Then answer some of the following questions:

- Where is the data or code stored (for example, Github repo or Zenodo)?
- Who created the objects (ORCID)?
- Was there proper documentation? License information (regarding commercial use)?

Thing 6: Human and machine readability

1. Read about the FAIR principles for making your code both human and machine readable, and the FAIR Guiding Principles article.
2. Read the following report Jointly designing a data FAIRPORT from the Lorentz Center.
3. Having code that is both human and machine readable supports:

- API access
- Allows for automatic integration of multiple datasets
- Use of standard formats widely accepted in the discipline

Thing 7: Maintain/preserve entire research environment (e.g. software)

1. Familiarize yourself with best practices for scientific computing. Read Good Enough Practices in Scientific Computing, and Top 10 Metrics for Life Science Software Good Practices to familiarize yourself with the topics of containers, software preservation, and software emulation.
2. Read more about the Long-term preservation of biomedical research data.

Thing 8: Indexing repositories to enable findability

1. re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines. Register your dataset with re3data.org
2. Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data.

- Read their Getting Started Guide to get indexed with Google.

3. Link your ORCID account to fairsharing.org, verify your email address, and create a public profile.

- Familiarize yourself with their Standards, Databases, Policies, and Collections.

Thing 9: Broad consent

Informed consent for human subjects should be broad enough to make reuse possible. See Broad Consent for Research with Biological Samples: Workshop Conclusions. Also see, Recommendations for Broad Consent Guidance from the Office for Human Research Protections.

Thing 10: Application of metrics to evaluate the FAIRness of (data) repositories

Beginner activity:

1. Explore the work of the FAIR Metrics Group. Explore their proposed FAIR Metrics.
2. Read the following paper: Evaluating FAIR-Compliance Through an Objective, Automated, Community-Governed Framework.
3. Explore the design framework for exemplar metrics for FAIRness.

Intermediate activity:

1. Explore the Make Data Count Project, where you can learn about COUNTER Code of Practice as well as the Code of Practice for Research Data Usage Metrics.
2. Learn how Zenodo and DataONE have responded to the Make Data Count recommendations.

TOP 10 FAIR DATA & SOFTWARE THINGS:

Biodiversity

Sprinters:

Silvia Di Giorgio, Akinyemi Mandela Fasemore, Konrad Förstner, Till Sauerwein, Eva Seidlmayer (ZB MED - Information Center for Life Science, Cologne, Germany), Ilja Zeitlin, Susannah Bacon, Chris Erdmann (Library Carpentry/The Carpentries,/California Digital Library)

Audience:

Researchers

Things

Findability

Thing 1: Identifiers

To make data findable, it has to be uniquely and persistently stored with an identifier.

- A digital object identifier (DOI) is a unique, case-insensitive, alphanumeric character sequence and can be very helpful for this purpose. You can reach the identified digital object by using the DOI as a URL. Just fill in the DOI in the address bar (e. g. https://doi.org/10.1109/5.771073). Also, see: ANDS Guide: Digital Object Identifier (DOI) System for Research Data.

NOTE:
The distributing DataCite-agency (i.e. issues DOIs) for Life Sciences is PUBLISSO: https://www.publisso.de/wir-fuer-sie/doi-service/

Exercise:
For easy look up, we have a list of DOIs below. Can you match the right document to the appropriate DOI? HINT: Start from here https://www.doi.org/!

1. 10.1103/PhysRev.48.73
2. 10.5962/bhl.title.28875

- *On the origin of species*
- *The Particle Problem in the General Theory of Relativity*

Which of these is not a valid DOI?

1. 10.1037/arc0000014
2. 12.1093/fMicb.2018.00257
3. 10.1101/468025 HINT: Check the prefix (before the forward slash)!

Which part indicates the publishing institution? The prefix or the suffix of a DOI?

ORCID Exercise: ORCID is a self-identifier for authors to avoid author name ambiguity. Use ORCIDs from the start of data creation, i.e. attach data creator name/ORCID to dataset as a metadata field. Include ORCIDs with datasets in repositories (e.g. in Sequence Read Archive (SRA), include the ORCID for the data creator). This allows for the tracking of data provenance (the origins, custody, and ownership of research data).

Go through the Getting Started with ORCID Integration.

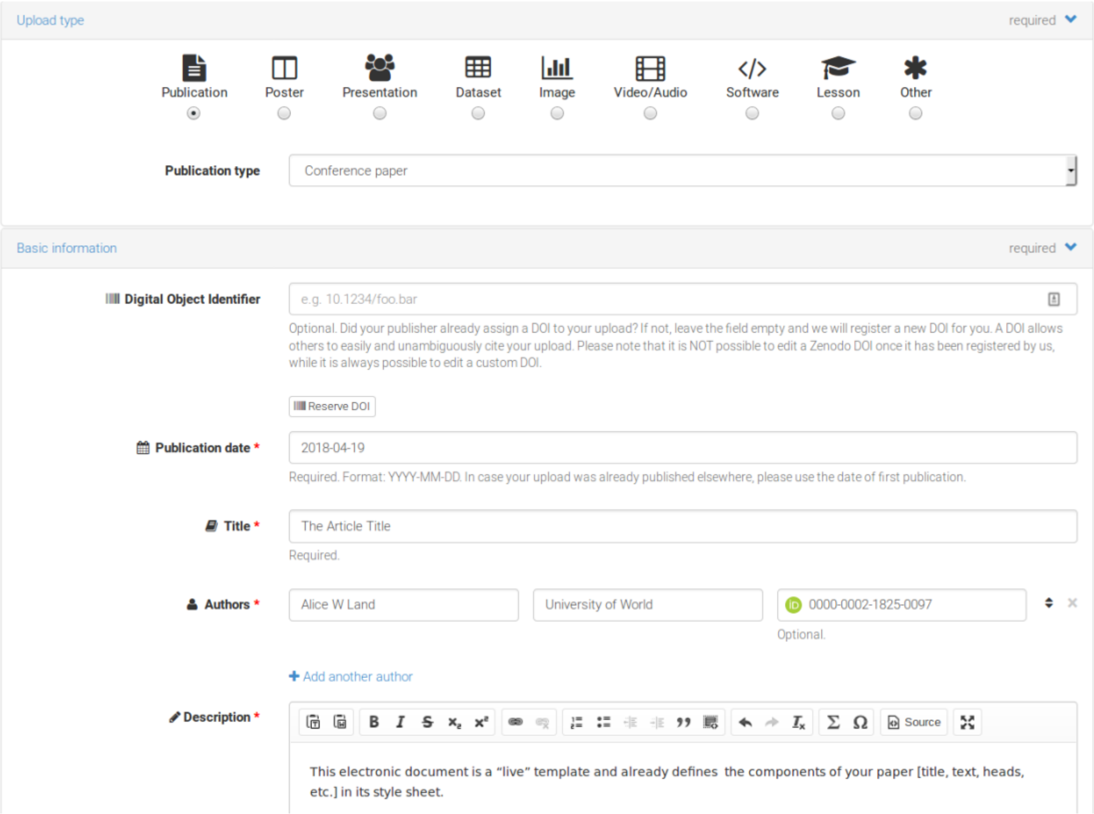
Thing 2: Citations

Zenodo, for example, is a tool that makes scientific data and publications easier to cite.
It supports various data and license types. It also supports source code from GitHub repositories.

See https://zenodo.org/

Exercise:
* Use the Zenodo Sandbox to upload an example dataset, software program, etc.
https://sandbox.zenodo.org/

Questions:
1. Which metadata fields do you have to add when uploading data and why? 2. Which fields are mandatory and which ones are not? 3. What identifiers can you use?


Upload type — required

● Publication ○ Poster ○ Presentation ○ Dataset ○ Image ○ Video/Audio ○ Software ○ Lesson ○ Other

Publication type: Conference paper

Basic information — required

Digital Object Identifier: e.g. 10.1234/foo.bar

Optional. Did your publisher already assign a DOI to your upload? If not, leave the field empty and we will register a new DOI for you. A DOI allows others to easily and unambiguously cite your upload. Please note that it is NOT possible to edit a Zenodo DOI once it has been registered by us, while it is always possible to edit a custom DOI.

Reserve DOI

Publication date *: 2018-04-19

Required. Format: YYYY-MM-DD. In case your upload was already published elsewhere, please use the date of first publication.

Title *: The Article Title

Required.

Authors *: Alice W Land | University of World | 0000-0002-1825-0097

Optional.

+ Add another author

Description *: This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet.

Uploading to Zenodo (Sandbox)

Thing 3: Wikidata

Wikidata provides a common source of open data which can be used by Wikimedia projects such as Wikipedia, and by anyone else, under a public domain license.

Exercise:
Go to Wikidata and find the publication date of the book “On the origin of species”.

- Switch over to the linked dataset of the author of the book and see his other publications.
- What did he publish in 1839?

Thing 4: Registry of Research Data Repositories (re3data)

This project aims to accelerate scientific discovery and enhance the integrity, transparency, and reproducibility of data. To enable FAIR data sharing, data need to be deposited in a repository that is taking steps to make data as open and FAIR as possible. It's not clear-cut what is FAIR at this time, there is no such thing as a FAIR stamp - although the CoreTrustSeal certification provides a good indication. Therefore, under the auspices of the Enabling FAIR Data Project, American Geophysical Union (AGU), re3data, and DataCite, these organisations have decided to develop new tools to assist researchers with finding an appropriate repository for their data:

- Browse Subject Repositories
- Repository Finder

Exercise:
1. How many entries are returned for the query specific for your research topic on re3data? 2. If you filter under "Subject", what do you find? 3. Do you think something is missing from the results? If so, suggest a repository.

Try the "browse by Subject" entry to the re3data-database since this gives a great overview on the wide landscape of research data repositories: https://www.re3data.org/browse/by-subject/

Accessibility

Thing 5: Bioschemas

bioschemas.org aims to improve data interoperability in the life sciences. It does this by encouraging people in the life sciences to use schema.org markup, so that their websites and services contain consistently structured information (metadata). This structured information then makes it easier to discover, collate and analyse distributed data.

Exercises can be found on the Bioschema website under "tutorials" and "how to".

- https://bioschemas.gitbook.io/training-portal/

Thing 6: Licenses

Knowing the appropriate licenses to use for your data can help others understand how they can use your data and can also help with improving accessibility.

- Open Source Licenses
- Data and Creative Commons licenses
- How to License Research Data

Exercise:
1. Use the Creative Commons License Tool to select the appropriate license with the following intentions; 2. Allow your work to be adapted and also allow it to be used commercially.

Thing 7: Availability via torrents

The era of Big Data is finally upon us. A prerequisite for accessibility is availability. Well established sharing protocols like torrents will ensure data are perpetually available without the constraint of time and space. Using the torrent protocol for scientific data will lead to some of the below advantages:

- Immutability
- Distribution capabilities (lower cost for distributing the data)
- No sole maintainer (we don't have to rely only on one specific maintainer because data can be cloned and maintained across the peer-networks)

The Magnet URI scheme defines the format of magnet links, a de facto standard for identifying files by their content, via cryptographic hash value rather than by their location.

Using Magnet URI scheme directly on the publication will make all the data accessible. For more information, read:

- Academic Torrents
- Magnet URI Scheme

Exercise:
1. Upload any small data set of your choice with the above link. 2. Share with a colleague a link to access it over torrent.

Interoperability:

Thing 8: ELIXIR platforms

Standardisation of life science data will ensure interoperability across different sub fields. ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe.

- ELIXIR Interoperability Platform

Exercise:
Use the ELIXIR software bio.tools to find the author of the RNA-seq python pipeline "READemption".

Thing 9: Research data management

Bio2RDF is a large network of linked data for the life sciences. The database provides interlinked life science data using semantic web technologies. To learn more about Bio2RDF, read *Bio2RDF: towards a mashup to build bioinformatics knowledge systems*.

- http://bio2rdf.org/

The German Federation for Biological Data (GFBio) is the authoritative, national contact point for issues concerning the management and standardisation of biological and environmental research data during the entire data life cycle (from acquisition to archiving and data publication). GFBio mediates expertises and services between the GFBio data centers and the scientific community, covering all areas of research data management.

- https://www.gfbio.org/

Thing 10: Machine-readability

Make the data accessible via an API, in a structured data format that can be automatically read and processed by a computer. See the Open Data Handbook Glossary - Machine readable.

Exercise - Crossref:

1. Pick the DOI of a publication of your choice.
2. Open a Web browser and add the URL.
3. https://api.crossref.org/works/DOI <= replace DOI with the DOI of the publication.

Example: https://api.crossref.org/works/10.1371/journal.pcbi.1004668

Exercise - DataCite:

1. Pick the DOI of a dataset in Zenodo.
2. Open https://api.datacite.org/works/DOI <= replace DOI with the DOI of the Zenodo entry.

Example: https://api.datacite.org/works/10.5281/zenodo.1574835

Reusability

Thing 11: Digitalization

If the methods to record complex experiments are prone to error, so that reproducible results cannot be guaranteed, how can you ever be sure you’re dealing with real insights and not random information? The electronic lab notebook provides the missing infrastructure for data recording, retrieval and integrity. An electronic lab notebook must be able to create, import, store and retrieve all important data types in digital format. For more information, read:

- Kanza, Samantha et al. “Electronic lab notebooks: can they replace paper?” Journal of cheminformatics vol. 9,1 31. 24 May. 2017, doi:10.1186/s13321-017-0221-3

- Electronic Lab Notebook Matrix

Exercise:
Explore the demo lab notebook at https://demo.elabftw.net/experiments.php

Thing 12: Containers

In a scientific field, most of the time we have to deal with large amounts of data that have to be processed before publication. One important aspect of the reproducibility challenge is ensuring computational analysis can be reproduced, even in different environments. For more information, read:

- Grüning, Björn, et al. "Practical computational reproducibility in the life sciences." Cell systems 6.6 (2018): 631-635.

Exercise:

Learn Docker & Containers using Interactive Browser-Based Scenarios:
https://www.katacoda.com/courses/docker

Thing 13: Blockchain for Life Science

Blockchain technology has the potential to be a technical solution to the current reproducibility crisis in science, and could "reduce waste and make more research results true". See:

- Mapping the blockchain for science landscape
- Blockchain for science and knowledge creation: A technical fix to the reproducibility crisis?

Living document example:
See Blockchain for Open Science – the living document:
https://www.blockchainforscience.com/2017/02/23/blockchain-for-open-science-the-living-document/

Supplementary information:

Research Data Infrastructure for the Life Sciences (NFDI4Life)
NFDI4Life brings together research communities across the life sciences domain in the context of the planned National Research Data Infrastructure (NFDI). As a response to the increasing scientific and societal demand for data and data analysis, NFDI4Life brings together scientific communities and research data infrastructures broadly covering the life sciences with particular focus on the subdomains biology, medicine (with veterinary

medicine), epidemiology, nutrition, agricultural and environmental science as well as biodiversity research.

- https://www.nfdi4life.de/

Carpentries Community

The carpentries develops and teaches workshops on the fundamental data skills needed to conduct research.

- https://carpentries.org/

Go-FAIR-Initiative

GO FAIR follows a bottom-up open implementation strategy for the European Open Science Cloud (EOSC) as part of a broader global Internet of FAIR Data & Services.

- https://eosc-portal.eu/
- https://www.go-fair.org/

FAIRDOM

FAIRDOM supports researchers, students, trainers, funders and publishers to make their data, operating procedures and models, Findable, Accessible, Interoperable and Reusable (FAIR).

- https://fair-dom.org/about-fairdom/

TOP 10 FAIR DATA & SOFTWARE THINGS:

Australian Government Data/Collections

Sprinters:

Katie Hannan, Data Librarian (CSIRO), Richard Ferrers, Research Data Analyst (ARDC), Keith Russell, Manager Engagements (ARDC)

FAIR data

See ARDC image summarising what FAIR means; see also Force 11 definition.

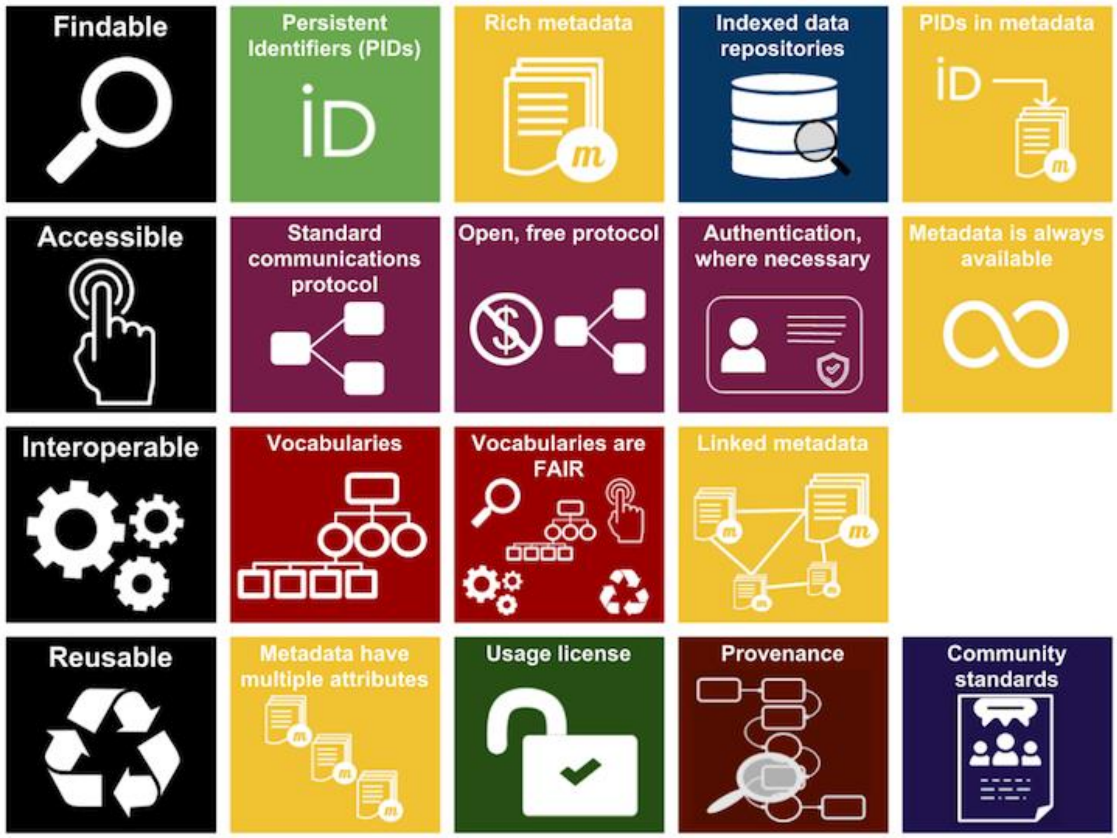


Figure 1; FAIR in a nutshell. Image: ARDC 2018 - CC-BY 4.0.

Description:

Governments have a mandate to make non-sensitive data open. For example, the Australian Government Public Data Policy Statement says “Australian Government entities will … make non-sensitive data open by default…make high value data available for use by the public, industry and academia… ensure non-sensitive publicly funded research data is made open for use and reuse… to extend the value of public data for the benefit of the Australian public." FAIR data is a way to extend the value of data. The largest 20 nations, the G20, agreed to make Open Data Principles a priority at the 2015 meeting in Turkey, saying "Transparency… Global transformation, facilitated by technology, fuelled by data and information.. Open data is at the center of this global shift." (p.2).

Audience:

Government data custodians

Goal:

Help government data custodians to understand FAIR data principles

NB: Nomenclature and data:

Where "data" is used here, we also mean collections such as Cultural Collections, historical collections, documents, artefacts and other valuable collections.

Table of contents

Things

Thing 1: Why is data important?

Read G20, Australian and States policies on Open Data

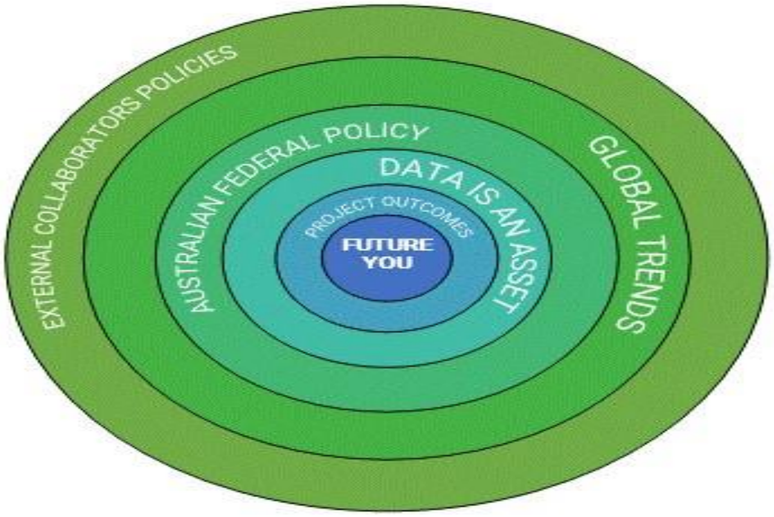


Figure 2; Data sharing drivers

Source: Katie Hannan, 2018, CC-BY.

Beginner activity:

International

G20: Open Government Forum; G20 Turkey 2015. “Transparency... Global transformation, facilitated by technology, fuelled by data and information.. Open data is at the center of this global shift.” (p.2) Read and consider G20 Open Data Principles.

Familiarise yourself with your State or Territories Data Policy. See links in Appendix 1.

Australia

* Public data policy statement Office of the Australian Information Commissioner: Principles on open public sector information * Principle 1: Open access to information — a default position “information held by Australian Government agencies is a valuable national resource. If there is no legal need to protect the information it should be open to public access.” * Principle 3: Effective information governance “ensuring agency compliance with legislative and policy requirements on information management and publication” * Principle 4: Robust information asset management * Principle 5: Discoverable and useable information “ensure that information published online is in an open and standards-based format and is machine-readable” “attach high quality metadata to information so that it can be easily located and linked to similar information using standard web search applications” * Principle 6: Clear reuse rights “The economic and social value of public sector information is enhanced when it is made available for reuse on open licensing terms.”

See Appendix 1 for a list of Australian State Open Data Policies.

Intermediate activity:

The following legislation may apply to the management of government data:

- Archives Act 1983 - https://www.legislation.gov.au/Details/C2016C00772
- Freedom of Information - http://my.csiro.au/Support-Services/Legal/FOI.aspx
- Privacy - http://my.csiro.au/Support-Services/Legal/Privacy-Law.aspx
- Australian Government intellectual property rules - https://www.communications.gov.au/policy/policy-listing/australian-government-intellectual-property-rules
- Records Disposal Authority - An agency-specific records authority may have advice that you need to follow. Find your agency here - http://www.naa.gov.au/information-management/records-authorities/types-of-records-authorities/Agency-RA/index.aspx

- New Australian Government Sharing and Release Legislation (open for public comment, shows where legislation is going): https://www.pmc.gov.au/sites/default/files/publications/australian-government-data-sharing-release-legislation_issues-paper.docx

Advanced activity:

If your organisation doesn’t have a policy on open data, who are the key stakeholders that you would need to work with to prepare an open data policy?

What main headings would you need to include as part of your data policy?

Thing 2: Open data vs FAIR data

Read https://www.go-fair.org/faq/ask-question-difference-fair-data-open-data/ Can you think of examples of data you deal with that cannot be made Open but can be made FAIR? List some advantages in making this data FAIR.

Does the current wording in the policy for Open Data encourage making the data FAIR? Where do you see gaps?

See slide 14 here https://www.slideshare.net/sjDCC/open-fair-data-and-rdm

Beginner activity:

See how Geoscience Australia implement the FAIR data principles in their work. Geoscience Australia describe themselves as “the nation's trusted advisor on the geology and geography of Australia” (GA 2018).

Advanced activity:

How FAIR is your data? - https://www.ands-nectar-rds.org.au/fair-tool Suggest using this now, and then finishing off the modules, making some changes to a data collection and then testing again using the FAIR data tool.

Thing 3: Data discovery

- What’s a data repository?
- What’s a data portal?
- Where to find data?
- Where to store data?
- Data.gov.au (and search.data.gov.au!) - find - this is an aggregator See https://data.gov.au/dataset/list-of-australian-government-data-portals for a list of Australian Government Data Portals (current as of March 2017). Some other data portals appear on https://data.gov.au/harvest.
- CSIRO DAP - find/store
- National Map - find
- Re3data.org - registry of research repositories (etc)

International government data portals:

- United Kingdom - https://data.gov.uk/
- New Zealand - https://www.data.govt.nz/
- Canada - https://open.canada.ca/en/open-data
- United States of America - https://www.data.gov/
- India - https://data.gov.in/
- Finland - https://vm.fi/en/opendata
- Singapore - https://data.gov.sg

Thing 4: Describing your data or collection

- Including a description of data. What should go in a description?
- What makes a good description? See ANDS Content Providers Guide on descriptions -> Best Practice -> Writing good descriptions
- Write the description for a reader who has a general familiarity with a research area but is not a specialist—this will make data more accessible for cross-disciplinary use.
- Don't use specialist acronyms or obscure jargon.
- Don't assume a reader has specialist knowledge.

Some reusable content here - https://ecu.au.libguides.com/10-marine-science-rdm-things/Thing6

Beginner activity:

Read a data description on data.gov.au eg Arts Victoria, ABC or Research data Australia Eg National Archive of Australia, Australian Antarctic Data Centre, CSIRO (Commonwealth Scientific and Industrial Research Org), Geoscience Australia.

Reflection: Could you understand the description? Can you think of someone for whom this data or collection would be useful? Was it clear where to go next to access the data, or to ask for more information about this data or collection? What else would you like to know about this data/collection?

Activity: Post your questions or responses to the reflection above to: the data custodian, or the comments section at data.gov.au.

Intermediate activity;

If you are a data custodian/researcher, consider your five most important datasets, that you have contributed to or that you manage. Pick the most important dataset to describe.

1. Start with: Title, Author, Year, Institution, Location/URL. This is the minimum description required to get a DOI (a permanent identifier). The URL for a DOI is the home page for the dataset description. If you don't have one, make a person's contact the URL.

- (Hint: if you get stuck with the description, copy the abstract of a paper or conference paper or annual report, which uses or references your dataset. Edit the abstract to talk only about the data.)

Q: What type of data identifier does a government data custodian have?

2. Add more rich description to your data description eg subjects, grant IDs (where applicable - RDA; the Australian National Data Catalogue, has permanent URLs for Australian ARC and NHMRC grants). Include a significant statement about why the dataset is important.
3. Ask a colleague in a related field if they can understand your description. This helps the description be broadly readable by someone who is not deeply knowledgeable in your field. This will ensure that your description is more broadly understood.

Advanced activity:

Publish your data description on your resume, especially if online e.g. LinkedIn. Send your data description to your data librarian, for addition to your Institutional Repository or Data Portal. Alternatively, post your description to a public cloud service, such as Zenodo, Figshare or Data Dryad. No data need be included. A description record is valuable in itself as it reveals the existence of data, previously unknown and inaccessible.

Thing 5: Identifiers

To make data findable, It has to be uniquely and persistently stored with an identifier. A digital object identifier (DOI) is a unique, case-insensitive, alphanumeric character sequence and can be very helpful for this purpose. See also [ANDS Guide: Digital Object Identifiers (DOI) System for Research Data]](https://www.ands.org.au/__data/assets/pdf_file/0006/715155/Digital-Object-Identifiers.pdf).

See who mints ANDS DOIs, including NSW Office of Heritage and Environment, Bureau of Meteorology, CSIRO, Geoscience Australia, Dept of Environment.

Types of persistent identifiers:

- DOI
- Handle
- IGSN

Videos

Watch the video Persistent identifiers and data citation explained by Research Data Netherlands - https://youtu.be/PgqtiY7oZ6k

Read about persistent identifiers on a very general level (awareness). DOI requires five fields; author, title, year, publisher, URL of DOI landing page.

Beginner activity:

Visit http://www.doi.org/ and try resolving these DOI numbers:

10.26179/5bf63428ea2a1 10.26186/5b76556b396c0

Thing 6: Licensing

See the licensing guide: what is the appropriate licence for data produced by a government agency?

Refer to Australian Government Data Statement: “At a minimum, Australian Government entities will publish appropriately anonymised government data by default: ...under a ***Creative Commons By Attribution licence*** (ie CC_BY licence) unless a clear case is made to the Department of the Prime Minister and Cabinet for another open licence.”

Specific CC licences, which require DPC approval, include NC - non-commercial, SA - share alike, and the very restrictive (and not-recommended ANDS) ND - no derivatives allowed.

Examples of licensing statements:

http://www.bom.gov.au/waterdata/index.shtml?selected=Copyright

Thing 7: Dirty data

Why is ”clean” data important? Public policy, changes to medical protocols and economic decisions all depend on accurate and complete data. See further at ECU resource which looks at the why and what of “dirty data.”

https://ecu.au.libguides.com/10-marine-science-rdm-things/Thing10

Beginner activity:

Read this case study. The Data Retriever automates the tasks of finding, downloading, and cleaning up publicly available data, and then stores them in a variety of databases and file formats. This lets data analysts spend less time cleaning up and managing data, and more time analysing it. https://frictionlessdata.io/articles/the-data-retriever/

- Bad data guide - https://github.com/Quartz/bad-data-guide
- Releasing data or statistics in spreadsheets - http://www.clean-sheet.org/
- How to share data with a statistician - https://github.com/jtleek/datasharing
- A gentle introduction to data cleaning - https://schoolofdata.org/courses/#IntroDataCleaning
- Tidy data for librarians - https://librarycarpentry.org/lc-spreadsheets/

Advanced activity:

- Open refine - https://librarycarpentry.org/lc-open-refine/
- Clean Your Data: Getting Started with OpenRefine [video] - https://www.youtube.com/watch?v=wGVtycv3SS0

Thing 8: Working with sensitive data

What is sensitive data?

FAIR data doesn’t need to be published as open data. See Thing 2.

Reuse: https://www.ands.org.au/working-with-data/skills/23-research-data-things/10-medical-and-health-things/m-and-h-thing-4

Useful resource: CSIRO Data 61 The De-Identification Decision-Making Framework - https://publications.csiro.au/rpr/download?pid=csiro:EP173122&dsid=DS3

Indigenous Knowledge: Issues for protection and management - https://www.ipaustralia.gov.au/sites/g/files/net856/f/ipaust_ikdiscussionpaper_28march2018.pdf

Additional resources (from Library-Research-Support-Top-10-FAIR-Things_DRAFT)

- Despite being written for Human Research Ethics Committees, the ANDS Human Research Ethics Committees guide is a handy overview for people interested in making personal data FAIR: https://www.ands.org.au/__data/assets/pdf_file/0009/748737/HREC_Guide.pdf Key points: "....."
- NHMRC National Statement on Ethical Conduct of Human Research (2018) - CH3.1 Element 4 https://nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018#element_4__data_collection_and_management Key points: "...."
- Guiding Principles for Ethical Research (U.S National Institutes of Health) - https://www.nih.gov/health-information/nih-clinical-research-trials-you/guiding-principles-ethical-research

Thing 9: Vocabularies - Assisting with interoperability

Beginner activity:

Controlled vocabularies for data description

In addition to selecting a metadata standard or schema, whenever possible you should also use a controlled vocabulary. A controlled vocabulary provides a consistent way to describe data - location, time, place name, and subject.

Controlled vocabularies significantly improve data discovery. It makes data more shareable with researchers in the same discipline because everyone is 'talking the same language' when searching for specific data e.g. plants, animals, medical conditions, places etc

1. Start by browsing Controlling your Language: a Directory of Metadata Vocabularies from JISC in the UK. Make sure you scroll down to 5. Conclusion - it's worth a read.

Advanced activity:

Have a browse around the stunning level of data description and data contained in the Atlas of Living Australia.

Other examples: * Geosciences Australia - http://ldweb.ga.gov.au/def/voc/ga/ * National Environmental Information Infrastructure - http://www.neii.gov.au/vocabulary/vocabulary-providers * Australian Governments' Interactive Functions Thesaurus (AGIFT) - http://www.naa.gov.au/information-management/managing-information-and-records/describing/AGIFT/index.aspx (of interest to Australian Government Linked Open Data working group)

Data Dictionaries Standardised, accepted terms and protocols used for data collection

- Australian Institute of Health and Welfare - http://meteor.aihw.gov.au/content/index.phtml/itemId/274816
- Australian Business Register - https://abr.gov.au/For-Government-agencies/Accessing-ABR-data/ABR-Data-Dictionary/
- Health.VIC - https://www2.health.vic.gov.au/about/reporting-planning-data/data-dictionaries
- South Australian electronic forms data dictionary - https://www.sa.gov.au/editors/electronic-forms-platform/data-dictionary
- Growing up in Australia Data Dictionary - https://growingupinaustralia.gov.au/data-and-documentation/data-dictionary
- Department of Social Services Settlement Database Data Dictionary - https://www.dss.gov.au/our-responsibilities/settlement-services/programs-policy/settlement-services/settlement-reporting-facility/help-for-settlement-reports/data-dictionary

Thing 10

Data impact:

Data reuse - It is hard to check/track when you don’t have persistent identifiers and there’s not much of a data citation culture.

Web stats Selected data.gov.au web analytics - https://search.data.gov.au/dataset/ds-dga-9fa9bfda-96b3-4214-8a09-497af105524b/details?q=data.gov.au

Some old uses of open data: https://data.gov.au/showcase

Use in GovHack(AU) - https://twitter.com/govhackau?lang=en

Tracking identifiers - data citation

Beginner activity:

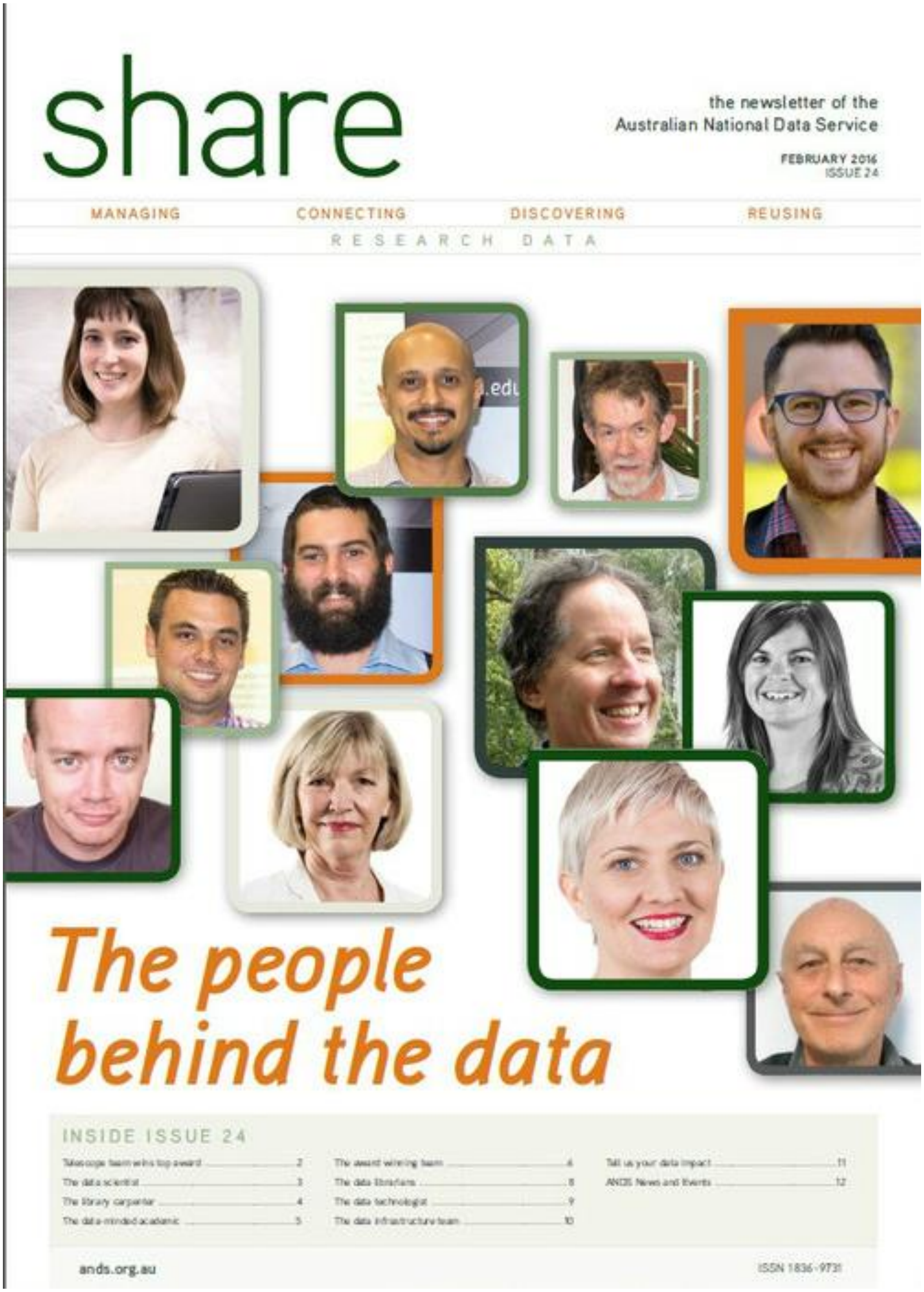
Looking at the broader impact of how the data has been used and the benefits it has brought to society, industry, economy, etc. is a richer source of impact evidence than just looking at citations.

https://www.ands.org.au/working-with-data/articulating-the-value-of-open-data/data-engagement-and-impact

Postscript: Other topics to consider:

- **Data People** - data technologists, data librarians, data trainers, data leaders, data scientists
- **Data Governance** - policy, procedure, planning, improving systems, request funding, build business cases for change
- **Data Training** - when: induction, checkups, when problems occur; what? Store, describe, how and why do data. Advanced topics eg sensitive data, spatial data, vocabularies, provenance.

See for example slide 54 in this Data Readiness slideshow as well as the 24th edition of Share (cover shown below).



share

the newsletter of the
Australian National Data Service

FEBRUARY 2016
ISSUE 24

MANAGING CONNECTING DISCOVERING REUSING

RESEARCH DATA

The people behind the data

INSIDE ISSUE 24

ands.org.au

ISSN 1836-9731

People in Data

References:

- Government data links
- Public Records Office Victoria

Appendix:

List of Australian state/territory government open data policies:

Australian Federal Government: Refer policy at Dept of Prime Minister and Cabinet. See also National Data Commissioner, ”responsible for implementing a simpler data sharing and release framework”.

Victoria Data Access Policy

“The Victorian Government recognises the benefits from and encourages the availability of Victorian government data for the public good. The DataVic Access Policy has been developed to support this recognition.”

New South Wales Policy (NSW)

“The objectives of this policy are to assist NSW Government agencies to: release data for use by the community, research, business and industry accelerate the use of data to derive new insights for better public services embed open data into business-as-usual…”

Queensland Policy

Tasmania Policy

South Australia Policy

Western Australia Policy

Australian Capital Territory Policy

Northern Territory Policy (Darwin)

TOP 10 FAIR DATA & SOFTWARE THINGS:

Archaeology

Sprinters:

Deidre Whitmore, Tim Dennis (UCLA)

Description:

This guide brings concepts surrounding FAIR data principles and the 23 (research data) Things program to the archaeological research domain with the aim of fostering better data practices and stewardship throughout the discipline.

Audience:

Researchers, scholars, employees, students, volunteers -- anyone working with or around data collected for archaeological research and management.

How to use this guide?

You don’t have to do all of the Things, and in fact, you may not be able to do every Thing. However, familiarize yourself with each Thing and implement those which suit your work and interests. Try to schedule time to learn more about a Thing regularly and work through how you could integrate it into your own research practices.

Why this guide?

Archaeological data is costly to collect, difficult or impossible to re-collect, and frequently lacks the context or documentation to reuse. Because of this, the domain has not yet coalesced around standards, though guidelines and data services are gaining traction. This guide helps introduce these services and calls out resources that can facilitate the adoption of leading practices.

Data in archaeology:

Archaeologists collect and work with a wide range of data types: textual, visual (raster, vector), tabular (spreadsheets, databases), spatial, audio, 3D, etc. This makes the creation and adoption of standards surrounding data management challenging but also even more necessary as these varied types frequently need to be analyzed together and shared among collaborators.

After working through the 10 Things below you'll know how to:

- plan and prepare for data collections so that the data that are collected are FAIR
- document collection processing analyses to support FAIR data
- draft and refine a data model
- find training or data specialists that can assist you in your work
- identify the multiple roles in the interdisciplinary project
- plan for a field season that integrates best practices for data management
- cite data, publish your data so that it can be cited, and why it is important to do so
- write a good data management plan
- identify the major data repositories in Archaeology
- reference the Guides to Good Practice and when to do so (at the start of a project and prior to collecting data!)
- evaluate tools that exist and can be used for humanities data

Things

Thing 1: Understanding the lifecycle of research data

Getting started
* Read Planning for the Creation of Digital Data in the Digital Antiquity Guides to Good Practice. * Consider the types of data collected and used within your own work. How many file formats do you work with regularly? How many files have become inaccessible to you over the years? To your colleagues or collaborators?

Learn more
* Watch the short film on the lifecycle of research data at https://www.ukdataservice.ac.uk/manage-data/lifecycle. * Map out the lifecycle of data on your most recent project. What processes and workflows have gotten you to the stage you are at currently? What can you do to facilitate the ongoing use and reuse of your data?

Challenge me
* Read Project Documentation and Project Metadata in the Digital Antiquity Guides to Good Practice. * Draft documentation for your most recent project or a forthcoming project. Include information about the background, methodology employed or to be employed, a narrative on the site and its context (historically, archaeologically, culturally, etc.). This documentation will not only facilitate the eventual dissemination of your data but also any proposals or publications about the work itself. * Review the metadata for this project, document in a single location what metadata you currently record or plan to record and compare it to the metadata tables at

http://guides.archaeologydataservice.ac.uk/g2gp/CreateData_1-2. Are you missing any Project Metadata? File-Level Metadata (general and technical)? How can you fill in any gaps?

Thing 2: Preservation

Getting started
* Browse the websites for archaeological data repositories and preservation services (Archaeology Data Service, tDAR, Open Context). * Identify which service(s) contain data of interest to your work. Get familiar searching the services. * Read Why Deposit Data and consider what is significant about your data, what requirements you need to meet, and which reasons resonate with your work and beliefs.

Learn more
* Dig into the deposit instructions and criteria for each repository and service and identify which is the best fit for your own data. * Contact the service and discuss your project and data with them. Document their recommendations and determine how you can update your current workflow to support deposit.

Challenge me
* Select a dataset you can deposit and go through the process of depositing in a repository.

Thing 3: Training and community

Getting started
* Review online resources and training materials for archaeological data management such as DataTrain's 'Open Access Post-Graduate Teaching Materials in Managing Research Data in Archaeology' at http://archaeologydataservice.ac.uk/learning/DataTrain.xhtml.

Learn more
* Attend a workshop at an upcoming Archaeology conference that focuses on data management or a session on the topic.

Challenge me
* Attend a conference or program on data and scholarly communication such as FORCE11's Scholarly Communication Institute and/or ASIST's Annual Meeting. * If you are in a position to do so, incorporate archaeological data management and preservation lessons into courses you teach. Consider inviting a data librarian or information specialist that is familiar with archaeological data to be a guest speaker.

Thing 4: Data Management Plan (DMP) tools

Getting started
* Review the guidelines for DMPs from funding agencies you are considering or have applied

to in the past: NEH, NSF, AIA, etc. See SPARC's Browse Data Sharing Requirements by Federal Agency.

Learn more
* Check if your institution is participating in the DMP Tool (meaning they have customized the tool to point to institutional resources and services) at https://dmptool.org/public_orgs. * Read through publicly available DMPs at https://dmptool.org/public_plans and consider what makes them strong/weak. Take notes on what aspects are important to include when writing your own.

Challenge me
* Use a DMP Tool to create a DMP for a project you are currently working on or planning to start. * Ask a data librarian or specialist at your institution to review your DMP.

Thing 5: Describing data

Getting started
* Learn more about metadata schema, controlled vocabularies and why describing data is a good practice. Read *What are Metadata Standards* from the Digital Curation Center at http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards and *Preparing Datasets - Metadata* from ADS at http://archaeologydataservice.ac.uk/advice/PreparingDatasets.xhtml#Metadata0. * Consider your current metadata practices - do they follow any schema or incorporate any vocabularies? Are your metadata fields described and documented explicitly?

Learn more
* Review some of the vocabularies and thesauri related to archaeological data including Getty Vocabularies at http://www.getty.edu/research/tools/vocabularies/index.html and PeriodO at http://perio.do/en/. * Consider whether these vocabularies could be incorporated into your data practices and workflow.

Challenge me
* Create a data dictionary (metadata field, type, definition, controlled vocabulary status) for a current or future project based on the metadata recommendations in the Guides to Good Practice. * Do this for each type of data you plan to or have collected that has an associated Guide (i.e. Raster images, Geophysics, GIS).

Thing 6: Cleaning, processing, and documentation

Getting started
* Learn about processing and documentation in ‘Data Selection: Preservation Intervention Points’ at http://guides.archaeologydataservice.ac.uk/g2gp/ArchivalStrat_1-3. * Consider your own workflow and the different stages at which your data is transformed. Write down the

equipment and instruments you use to collect data and the process for obtaining the data from those instruments (i.e. calibrating, exporting)

Learn more
* Investigate tools that facilitate data cleaning and documentation such as Open Refine at http://guides.archaeologydataservice.ac.uk/g2gp/ArchivalStrat_1-3. * Attend a workshop or go through a tutorial to learn how to use the tool and its features including exporting out the record of the cleaning, etc.

Challenge me
* Choose a recent dataset you've collected and go through the processing and cleaning workflow. Be sure to document every step and follow conventions for file names, file formats, and backup creation.

Thing 7: Sharing

Getting started
* Learn more about why sharing data matters in archaeology. Explore publications on archaeological data, reuse, and publishing including *Openness and archaeology's information ecosystem* at https://escholarship.org/uc/item/9tq378jg and *Other People's Data: A Demonstration of the Imperative of Publishing Primary Data* at https://escholarship.org/uc/item/1nt1v9n2 * Consider times you haven't been able to access data associated with your research. How did you address this issue? * Consider times you have tried to use collaborators' or colleagues' data in your own research. What steps did you have to take to make sense of the data, to incorporate it into your own dataset, or to analyze it? What might have made this process easier?

Learn more
* Learn more about sensitive data and what you can do to protect while still making it accessible from resources such as ANDS (https://www.ands.org.au/working-with-data/sensitive-data/sharing-sensitive-data) and ADS (http://archaeologydataservice.ac.uk/advice/sensitiveDataPolicy.xhtml) * Consider whether there are any ethical or legal restrictions around data in your own work. Discuss these considerations with the appropriate representatives and determine what the best plan for sharing data is for all relevant parties.

Challenge me
* Learn more about the differences between publishing and sharing data then either: * Prepare a dataset of your own for sharing with a colleague or collaborator and ask them to report back on any issues they faced understanding the data, accessing files or information, and what you could have done to simplify their use of the dataset. * Or publish a dataset of your own. This can be done either in association with an article or book, as a data paper with a journal that specializes in data publication, or through a data publishing service. Consider

the challenges you faced as your prepared the dataset and what you can do to simply the process next time, then incorporate these practices into your workflow.

Thing 8: Citation

Getting started
* Data citation continues the tradition of acknowledging other people's work and ideas. Along with books, journals and other scholarly works, it is now possible to formally cite research datasets and even the software that was used to create or analyze the data. Consider if there are times your data was reused by someone else and whether you received scholarly credit. * Read the Force11 Joint Declaration of Data Citation Principles at https://www.force11.org/datacitationprinciples. * Watch this video on persistent identifiers and data citation at https://www.youtube.com/watch?v=PgqtiY7oZ6k. * Search data repositories and services such as ADS, tDAR, and Open Context and see how their recommended citations are formatted.

Learn more
* Consider how many times you've read research papers and felt the data was either insufficient or inaccessible and how this impacted your interpretation. * Have a discussion with your colleagues about their perspectives on publishing data so that it is findable, in formats that are accessible, and with enough descriptive metadata and documentation to be reusable. Have any of them ever cited a dataset? Why or why not? What would be needed for this to become a common practice in archaeology?

Challenge me
* Include citations to datasets, not just scholarly articles and books, relevant to your work in your next publication. * Consider whether persistent identifiers (PIDs) should be routinely applied to all research outputs. Remember that PIDs carry an expectation of persistence (maintenance costs, etc.) but can be used to collect metrics as well as link articles and data (evidence of impact).

Thing 9: Licensing

Getting started
* Research licensing research data in your country and what set of licenses is used most commonly. * Discuss with colleagues if they have licensed their data and what their experience has been.

Learn more
* Read through the licensing agreements and policies for data services and repositories, starting with ADS, tDAR, and Open Context. Consider whether these policies align with your datasets and obligations.

Challenge me
* Determine which license is appropriate for your data and if possible, release one of your own datasets by depositing into an archive or repository. Consider consulting with a data service representative or data librarian about your selection.

Thing 10: FAIR in archaeology

Getting started
* Read through the FAIR data principles at https://www.go-fair.org/fair-principles/. * Consider what these principles mean in practice and how each of the Things you are implementing support FAIR archaeological data. What would it mean if every archaeologist followed these principles?

Learn more
* Watch the webinar *Enabling FAIR Data* at https://www.dataone.org/webinars/enabling-fair-data or *Are we FAIR yet?* at https://rd-alliance.org/webinar-are-we-fair-yet.

Challenge me
* Assess the *FAIRness* of one of your recent datasets using the FAIR self-assessment tool from ARDC. What did you learn about your data? How can you do better?