Top 10 FAIR Data & Software Things

15 September 2019

Sprinters:

Albert Meroño Peñuela, Andrea Scharnhorst, Andrew Mehnert, Aswin Narayanan, Chris Erdmann, Danail Hristozov, Daniel Bangert, Eliane Fankhauser, Ellen Leenarts, Elli Papadopoulou, Emma Lazzeri, Enrico Daga, Evert Rol, Francoise Genova, Frans Huigen, Gerard Coen, Gerry Ryder, Iryna Kuchma, Joanne Yeomans, Jose Manzano Patron, Juande Santander-Vela, Katerina Lenaki, Kathleen Gregory, Kristina Hettne, Leonidas Mouchliadis, Maria Cruz, Marjan Grootveld, Matthew Kenworthy, Matthias Liffers, Natalie Meyers, Paula Andrea Martinez, Ronald Siebes, Spyros Zoupanos, and Stella Stoycheva.

Organisations:

Athena Research Center, Australian Research Data Commons (ARDC), California Digital Library, Centre de Données astronomiques de Strasbourg, Centre for Advanced Imaging, Centre for Microscopy Characterisation and Analysis, Data Archiving & Networked Services (DANS), EIFL, ELIXIR-Europe, EPFL, FORTH-IESL, FOSTER Open Science, GRACIOUS, Greek Ministry of Education, Greendecision, Göttingen State and University Library, Leiden Observatory, Leiden University, Library Carpentry, National Imaging Facility (NIF) and Characterisation Virtual Laboratory (CVL), National Imaging Facility and Centre for Advanced Imaging,National Research Council of Italy, OpenAire, SKA Organisation: Jodrell Bank Observatory, The Open University, The University of Western Australia, Universitaire Bibliotheken Leiden, University of Notre Dame, University of Nottingham, Vrije Universiteit Amsterdam, and Yordas Group.

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. In May 2019, additional sprinters from the 2019 Library Carpentry-Mozilla Global Sprint contributed six new Top 10 FAIR Data & Software Things: Nanotechnology, Astronomy, Linked Open Data, Imaging, Music, and The European Open Science Cloud (EOSC). While sprints are one way to contribute, members of the wider community can submit issues and/or pull requests to the Things to help improve them or add new ones. Published versions of the Things are available via Zenodo and the DOI http://doi.org/10.5281/zenodo.2555498.

Table of Contents

Music

Sprinters

Daniel Bangert (Göttingen State and University Library)

Albert Meroño Peñuela (Vrije Universiteit Amsterdam)

Enrico Daga (The Open University)

Contents

Description

This is a brief guide to ten topics relevant to understanding how the FAIR data principles apply to music research. It includes brief activities designed for self-paced learning or as training ideas. The aim of this document is to help those who wish to find, publish or reuse musical data in adherence with the FAIR data principles.

Audience

- Music students, researchers and scholars
- Librarians
- Data stewards

- Research support staff

Thing 1: What is musical data?

As noted in this paper on Music Information Retrieval, music is multirepresentational, multicultural, multiexperiential, and multidisciplinary. Musical data therefore encompasses a wide range of types and formats, including symbolic representations such as scores, audio recordings, images of manuscripts, and information about works, performances and composers. In many cases, researchers may not think of such resources as 'data', instead referring to primary or secondary sources, reference works, databases, notes or annotations.

Activity 1: Discussion

- What kinds of data do you work with? Do you primarily work with physical or digital objects? Do you work with multiple representations or formats?
- Are some types of data easier or more difficult to make and keep FAIR: findable, accessible, interoperable and reusable?

Activity 2: Go to Digital Resources in Musicology (DRM) and review the top level categories. Which of these are relevant to your work?

Thing 2: Catalogues and repositories of musical data

Musical data is often organised into meaningful collections: groups of musical resources (recordings, scores, transcriptions, biographies, etc.) that make sense together and revolve around a central topic (a period, musician, genre, instrument, culture, etc.). These collections (or catalogues/repositories) are usually hard to find. This is why libraries have index cards, and databases have metadata: so users can browse and search them in order to reach the data they need.

Activity 1: Discussion

- What are the musical data catalogues, collections and repositories most frequently used in your field?
- Where are these typically deposited? Do these depositing services suit your requirements well?
- Have you used institutional, domain-specific or generic repositories (e.g. Zenodo) to share your data?

Activity 2: In this paper on Characterising the Landscape of Musical Data on the Web, the authors tried to find, and describe, as many Web music catalogues as possible. These are published in the musoW (Musical Data on the Web) registry. Are catalogues and collections of your domain covered in this table? If not, please add them at the end of the table.

Thing 3: Metadata

Metadata can be defined as data about data. Metadata commonly describe characteristics such as format, contents, creator and publication date. This information is often captured using a metadata schema, which are designed to capture a common set of information in a structured manner. Whether you are searching for data or depositing a dataset, remember that the quality of the metadata captured influences how easily data can be found and potentially reused. In short, richer metadata increases findability.

Activity 1: Choose one catalogue from Thing 2. This could be, for example, your favourite PDF score collection. Then, browse schema.org and look for properties you could use to accurately describe that collection's metadata (author, time, genre, location, etc.). Most of them will be metadata describing datasets. Are there any music-specific properties in schema.org (or elsewhere) you would use?

Activity 2: Go to Google Dataset Search and try to find your chosen dataset out of Thing 2. Is it there? Why do you think it is (not)?

Thing 4: Persistent identifiers

Persistent identifiers are long-lasting references to a resource, like a document, webpage, file, or music score. They are designed to uniquely identify such resources, and to be actionable upon them: a protocol is typically able to retrieve the content they represent from them (see Thing 5).

There are two important issues about persistent identifiers and musical data: object level identification, and persistent identifier providers. Object level identification refers to the granularity and level of detail for the object for which the identifier is being created. Does the identifier represent a whole musical collection, an item inside that collection, a score within that item, a page of that score, a note within that page, an annotation? Persistent identifier providers refer to the institutional service that generates the identifiers and ensures that they will function permanently. Regular URLs (web addresses) can perform this role with adequate maintenance; but institutionally maintained identifiers (such as DOIs and PURLs) typically do this maintenance externally.

Activity 1: Discussion

- What are the musical objects (collections, databases, database entries, pages, scores, recordings, metadata, etc.) that get permanent identifiers in your domain?
- What are these identifiers typically used for?
- Would identifiers representing more fine-grained musical objects (i.e. notes) be useful?

Activity 2:

- Create an account at purl.org
- Create a new domain
- Create PURLs pointing to musical objects on the Web that you consider important

- Add those PURLs to this sheet

Thing 5: Standards and protocols

We have seen so far that stable, eternal identifiers are useful to name and find musical resources. But how can we use these identifiers to *access* the data they represent? Accessing the data behind identifiers is what we do, for example, when we physically go to a designated library location, or when we write a URL in our Web browser and hit enter. Interestingly, these things can also be done by automated agents (robots, programs). Both humans and machines need a standard, open, free, universally understood and authenticated protocol (so: a systematic sequence of steps) to perform this access. On the Web, URLs are preferred for identifying (musical) things, and the protocol to access the content they represent is the Hypertext Transfer Protocol (HTTP). Despite its initial purpose to transport HTML pages from servers to Web browsers, HTTP can be used to access Web data of any kind.

Activity 1:

- Find the DBpedia URI identifying a famous band or song, by appending its name to the prefix http://dbpedia.org/resource. For example, for The Beatles you would have http://dbpedia.org/resource/The_Beatles
- Paste that URI in the address bar of your browser, and observe the (HTML) results
- Open a terminal in your system (Windows, MacOS, Linux), type the command `'curl -L -H'Accept: text/turtle' http://dbpedia.org/resource/The_Beatles'` (without the quotes and with your chosen band or song), and observe the results. What are the differences with respect to what was shown in the browser? What similitudes?

Thing 6: Encoding standards

Apart from identifying resources uniquely, a key aspect of sharing them is to make them readable and actionable by other users and applications. This is valid for any relevant resource that is published on the Web. However, a large number of music activities depend on some musical content. When reusing musical objects from the Web, a key problem is the compatibility of the format with the target tool. Therefore, a number of standards have been developed by the community of researchers and practitioners to represent music scores and making it usable across applications. These include (but are not limited to):

- MEI - https://music-encoding.org/
- MusicXML - https://www.musicxml.com/
- ABC - http://abcnotation.com/
- LilyPond - http://lilypond.org/
- MIDI and XMF https://www.midi.org/specifications

Activity 1: Collect scores of the same song in different standards and compare them: do they include the same information?

- ABC tune examples - http://abcnotation.com/tunes
- MIDI song examples - https://github.com/albertmeronyo/awesome-midi-sources
- MEI examples - https://github.com/music-encoding/sample-encodings

- MusicXML examples - https://www.musicxml.com/music-in-musicxml/
- LilyPond examples - https://www.mutopiaproject.org/

Activity 2: Choose a tool/application you are familiar with and check which formats are supported and which ones are not. Request the missing feature to the organisation or community that supports the development.

Activity 3: Once a score is encoded according to a particular standard, how is it rendered? One tool for rendering MEI files is Verovio. Go to Verovio's MEI viewer at https://www.verovio.org/mei-viewer.xhtml and use the navigation menu to turn pages, zoom in/out and switch between examples. Verevio is used in a range of projects, including digital editions of Beethoven and Mozart.

Thing 7: Ontologies

Ontologies are representations of concepts and their relations according to the meaning they have in a specific community. Standard formats like the one discussed above have the purpose of encoding information in a symbolic form. However, they generally lack details about the *meaning* of the symbols used, that is specified outside, usually on a documentation manual. Ontologies aim at expressing the meaning of the symbols used with a high degree of formalisation.

Ontologies are defined using Semantic Web standards: URIs, RDF, and OWL. Web ontologies can be useful to publish Linked Data on the Web (see below). Domain ontologies are developed with the purpose of representing concepts which belong to a specific part of the world, such as biology, social media, … or music!

Music ontologies vary from metadata standards to sophisticated schemas to represent music-related objects. Some examples are:

- DOREMUS - https://www.doremus.org/ (http://data.doremus.org/ontology/)
- Music Ontology - http://purl.org/ontology/mo/
- Chord Ontology - http://purl.org/ontology/chord/
- Music Theory Ontology - http://purl.org/ontology/mto/
- Temperament Ontology - http://purl.org/ontology/temperament/
- MusicNote - http://cedric.cnam.fr/isid/ontologies/MusicNote.owl#
- MusicOWL - http://linkeddata.uni-muenster.de/ontology/musicscore/

Activity 1: Find a music ontology on the Web. What is the aspect of Music whose *meaning* it describes? A starting point for your search could be Linked Open Vocabularies.

Activity 2: Find projects using music ontologies. What is the ontology useful/used for? For example, have a look at JazzCats and its data structures

http://jazzcats.cdhr.anu.edu.au/documentation/ Which classes and properties are from existing ontologies?

Thing 8: Linked Data

Linked Data is a way of representing structured data using the Resource Description Framework (RDF), so multiple datasets can be easily connected and queried together via the SPARQL Protocol and RDF Query Language (SPARQL). The Web community has linked so far more than 1,200 datasets and 200 billion statements.

Activity 1:

- Go to https://lod-cloud.net/ and find datasets related to your musical interests. Observe their links to other (perhaps not so musical, but still related) datasets.
- What other datasets do you know from your domain that would be interesting to link to? To what purpose?

Activity 2: Awesome Semantic Web enumerates a large number of Linked Data tools. Which of them do you think would be useful to support linking musical data? Which of them would support FAIR in musical data?

Activity 3: Get to know a few methods and platforms for navigating Linked Data resources:

- Search the DOREMUS catalogue http://overture.doremus.org/
- Explore the Linked Jazz network visualisation https://linkedjazz.org/network/
- Run a sample SPARQL query using the Wikidata SPARQL endpoint https://query.wikidata.org/ For example, run the query to return 'paintings depicting musical instruments with some connection to Hamburg'. See the query help page for more information about querying Wikidata.
- Search the MIDI Linked Data cloud https://midi-ld.github.io/, and send some queries to its SPARQL endpoint through its API http://grlc.io/api/midi-ld/queries/

Thing 9: Licensing and provenance

Licensing is a key topic in music and musicology, since music has historically been a cultural asset with strong ties with industrial exploitation and copyright. At the same time, researchers that investigate music need musical data to be openly available, which sets a whole spectrum of compromise. At the same time, the high availability of musical assets opens questions about the provenance of the data: Who made them and why? When? What instruments and musicians were involved? These questions might be key for trusting musical catalogs and establishing standards of data quality.

Activity 1: Enumerate data licenses that are typically used in your field. What are their limitations? Are there types of musical data for which specific licenses suit better? Are there needs not covered by any such license? Examples of data licenses are:

- Creative Commons (CC0, CC BY-SA, etc.)
- Open Data Commons

- Free Art License
- Open Game License

Activity 2: Do you need guidance on how to license your research data? Read OpenAIRE's guide on how to apply licenses to research data.

Activity 3: Discuss practices and standards in recording provenance of musical data in your field. Is provenance recorded automatically, manually, or not at all? In what situations would provenance of musical data be useful or necessary?

Thing 10: FAIR policies

The FAIR data principles have gained significant traction since their conception. Statements citing the importance of FAIR data can be found in the policies set by funders, higher education institutions, repositories, journals and publishers. For example, the Enabling FAIR Data initiative of the American Geophysical Union has been endorsed by a large number of publishers and repositories. Signatories to the initiative, such as Nature, aim to promote best practices for data sharing and have implemented policies that assist adherence to the FAIR principles. As research transitions towards a FAIRer future, what other policy developments do you expect to see in the next few years?

Activity 1: Discussion * Have you come across the FAIR principles, or aspects of the principles, in any policy documents? * What are the research data policies of your local institution or national funding bodies?

Activity 2:

- Read the brief data policy of the Transactions of the International Society for Information Music Retrieval (TISMIR) at https://transactions.ismir.net/about/editorialpolicies/ (scroll to Reproducibility section).
- Look up the data policy of another journal in your discipline or subdiscipline. What are the similarities or differences between policies? Is a data availability/accessibility statement required? Are there recommendations for depositing data in a trusted digital repository?

Activity 3: Read Tuomas Eerola's blog on Open Data in Music and Science, which includes comments on education and advocacy. What steps could you take to promote good data management and data sharing practices amongst your students and colleagues?

Imaging

Authors

- Paula Andrea Martinez https://orcid.org/0000-0002-8990-1985 - National Imaging Facility (NIF) and Characterisation Virtual Laboratory (CVL)
- Gerry Ryder https://orcid.org/0000-0001-7444-4489 - Australian Research Data Commons (ARDC)
- Aswin Narayanan https://orcid.org/0000-0002-4473-7886 - National Imaging Facility and Centre for Advanced Imaging
- Iryna Kuchma https://orcid.org/0000-0002-2064-3439 - FOSTER and OpenAIRE
- Jose Manzano Patron https://orcid.org/0000-0002-0040-1926 - University of Nottingham
- Andrew Mehnert https://orcid.org/0000-0001-8076-7532 - Centre for Microscopy Characterisation and Analysis and The University of Western Australia
- Matthias Liffers https://orcid.org/0000-0002-3639-2080 - Australian Research Data Commons (ARDC)

Description:

This guide aims to promote the FAIR data principles and to encourage their adoption by the bioimaging and characterisation1 community. The FAIR principles are described in the context of bioimaging and characterisation and the activities are optional. This guide seeks to empower researchers, scientists and health professionals to enable them to adopt best data practices throughout the research lifecycle, to improve the quality, reproducibility and reusability of research outputs.

1"Characterisation is the general process of probing and measuring the structures and properties of materials at the micro, nano and atomic scales. It is essential across natural, agricultural, physical, life and biomedical sciences and engineering."

Audience:

Researchers, neuroscientists, clinicians, microscopists, platform engineers, graduate students and computational and data scientists working on image analysis and processing.

Goals:

To inform data producers and users about the FAIR principles applied to bioimaging/characterisation and suggest activities to apply to their research.

Table of Contents:

1. What is FAIR?

The acronym FAIR, as detailed in 15 principles (GO FAIR 2016) stands for *Findable*, *Accessible*, *Interoperable* and *Reusable*. The FAIR principles (Wilkinson et al. 2016) are guidelines to motivate and enhance *reusability* of data, by facilitating its *discovery*, *integration* and *evaluation*. In this context, "data" refers to all research-oriented digital objects (including data, metadata, software, workflows and packages) (Wilkinson et al. 2017). Wilkinson et al. 2016, pioneered the definition of the guiding principles "emphasising the capacity of computational systems to Find, Access, Interoperate and Reuse data with none or minimal human intervention", this is referred to machine-actionable FAIR principles. FAIR is not separated, but the intersection of research data management and open science, as (Higman, Bangert, and Jones 2019) describe.

"FAIRness is a prerequisite for proper data management and data stewardship"

Communities are motivated to apply the FAIR principles to research activities and to enable people and machines to find, read, use and reuse research data and research outputs. In 2018 a coalition of stakeholders (COPDESS 2018), representing the international Earth and Space science community set out to develop standards to connect researchers, publishers, and data repositories in this community to enable FAIR data on a large scale. This project will accelerate scientific discovery and enhance the integrity, transparency, and reproducibility of this data. In imaging, on 1 March 2019 (Bioimaging 2019) and other research infrastructures including ELIXIR-Europe joined forces as part of The European Open Science Cloud project to publish research data via FAIR databases. Community participation from academia, industry, small and medium-sized enterprises (SMEs) and regional bio-clusters is paramount for the success of this four-year project (starting in 2020). The imminent global uptake of the FAIR principles in different scientific domains, serves to motivate the bioimaging/characterisation community to do likewise and to move forward, promote and apply them.

Activity 1: In 2018, CODATA - The Committee on Data for Science and Technology - released news of the "Enabling FAIR Data Project and Commitment Statement". Take a look at the partners in (COPDESS 2018), do you recognise partners in your discipline?

Activity 2: Can you think of the benefits of making your data FAIR? How can you align your current data practices to the FAIR principles? Consider the following resources when addressing the activity above:

- "How to make your data FAIR" by the Australian Research Data Commons (ARDC).
- "The Research Data Lifecycle - A Model for Data Management" by Alan Turing Institute, 2019, or the Curation Lifecycle Model by the Digital Curation Centre (DCC).
- "What is FAIR data? by LIBER (Europe's Research Library Network) Cavalli, 2018.
- Interpretation of the FAIR data principles the Swiss National Science Foundation (SNSF), 2018.

Back to top

2. What are publishers and funders saying about data access?

Many governments, funders, and publishers around the world have adopted data access policies that either encourage or require researchers to start their journey to FAIR research.

All research papers accepted for publication in *Nature* and an initial 12 other Nature Research titles are required to include information on whether and how others can access the underlying data Nature Announcement 2016: where are the data? (Nature, 2016).

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction at the time of publication (PLOS one, 2017). PLOS suggests using FAIRsharing to index resources, for example their own PLOS list.

The European Commission has drafted Guidelines on FAIR Data Management for the H2020 programme (European Commission, 2013) "those projects funded in this scheme must submit a version of this FAIR Data Management Plan (DMP)".

The Australian Research Council states "Author(s) should consider selecting publishers and research outlets, which have policies supporting the F.A.I.R. principles, as well as immediate or early availability of Publications via Open Access, in order to maximise the availability and impact of their ARC Funded Research."

The (Australian) National Health and Medical Research Council (NHMRC) promotes the highest quality in the research that it funds, based on international best practice. The NHMRC lists the FAIR principles under useful resources for publication and reporting of research outcomes.

Other governments, funders, and publishers adopting FAIR principles include: * The Australian government's 2016 National Research Infrastructure Roadmap. * European Commission - Turning FAIR into reality. * The Genomics Health Futures Mission (GHFM)

(funder) - Projects Grant Opportunity guidelines 2019. * eLife Journals (publisher) - data policy. * Wiley (publisher) - data sharing and citation.

Back to top

3. Data sharing and discovery

Why sharing?

"Both researchers and the broader community stand to benefit from the knowledge produced through publicly funded research" (ARC Open Access Policy Version 2017.1). Hence, why data sharing is well connected with the concepts of reproducibility and reusability.

"Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them. Researchers must be certain that data held in archives remain useful and meaningful into the future. Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in Data Management Plans. Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository" (Core Trust Seal, 2016).

Activity 1: (Infographic) Research data may be discovered (*findable*) and shared (*accessible*) in many ways. Start by looking at some data sharing trends across countries and research disciplines (Wiley, 2014). Consider your own current data sharing practices, and those of your project team(s), as yourself the question: How FAIR are those practices?

Activity 2: How can data be shared and discovered? Think about open, mediated, restricted access data repositories. What examples of these types of repositories are you aware of? Discuss with others about their answers.

Back to top

4. Reusable data repositories for the image community

How to walk towards FAIR?

Imagine if you were able to obtain extra datasets for your existing research project, start new collaboration based on common research questions, or start a new project reusing publicly available datasets. You can do this by exploring the following reusable data repositories for the imaging community divided by topic.

Reusable data repositories 1. **Neurosciences** 1. **Microscopy** 1. **Biomedical sciences** 1. **Non-domain specific data registries and catalogues**

This and the previous section intend to show that it is becoming more common for funding agents and publishers to require research data to be made *accessible* via appropriate repositories. This list is a starting point for you to *find* out what data already exists in your research area. If you want to share your data, or find data relevant to your research take a

detailed look at the examples provided in the next sections. Also, most if not all the listed repositories will have guides on how to share data.

Activity 1: Find repositories for imaging in FAIRsharing.org and search for repositories relevant to your research. Try for example, searching on "neuroimaging". Explore at least one repository you find. How well does it support the FAIR data principles? Tip: look for things such as persistent identifiers, clear descriptions, licence information, download options, file formats.

Back to top

5. Managing and sharing sensitive data

Clarification, FAIR data is not necessarily "open" data. There are some good reasons why some data should not be open. For example, to protect intellectual property, commercialisation, national security, personal privacy or endangered species. However, it may still be possible to provide mediated *access* to such data, or to publish a description of the data so that others can *discover* its existence. To align with FAIR principles your "*research data should be as open as possible, as closed as necessary*".

The FAIR principles encourage us to disseminate data as widely as possible, in the most effective manner and at the earliest opportunity. This statement takes into account any restrictions relating to privacy, confidentiality, intellectual property, embargo period, or cultural sensitivities, that need to be addressed, discussed and clarified before sharing any data. In the planning phase of a research project, researchers are encouraged to consider at least making project metadata publicly accessible (read more about metadata in section 7).

If you need examples and more information, check OpenAIRE sensitive data guide (OpenAIRE, 2017), ANDS, 2018 guide to publishing and sharing sensitive data, Earth Science Information Partners (ESPI) Handling sensitive data tutorial (Downs, 2012). In addition, The Australian Bureau of Statistics (ABS) informs of the five safes framework and Table 2 provides examples at different levels of accessibility.

Activity 1: Promoting FAIR principles in the healthcare field (DCC, 2019). Highlights: The sensitive nature of patient data and additional concerns for these data include security and anonymisation of data subjects as major components considered. For more information on FAIR related to healthcare visit FAIR4health.eu.

Activity 2: Think about when and how people can share data along the research cycle. Keeping in mind that it is strongly recommended to release metadata (description) of the project to comply with FAIR principles, even if you cannot share the data itself. Institutional repositories or domain specific repository should be able to store metadata of your project and then link that information via registries (have a look at the reusable repositories section).

De-identification / Anonymisation

In the case of sensitive data, the aim is to *minimise the risk of exposing confidential information*. Sometimes restrictions on sharing can be resolved by de-identification or

anonymisation of data. Anonymisation is sometimes used interchangeably with de-identification, ANDS, 2018 makes a clarification of these terms in the De-identification guide.

- De-identification is the removal of identifying information from a dataset, and this data could potentially be re-identified e.g. if the identifying information is kept (as a key) and recombined with the de-identified dataset.
- Anonymisation is the permanent removal of identifying information, with no retention of the identifying information separately.

Activity 3: Look at The Future of Privacy Forum's visual guide to practical data de-identification, what examples can you take from it and apply to your field?.

Optional extra information. 1)Open de-identification tools from Open Brain Consent (Halchenko, 2018). 2) A blog by Latanya Sweeney about The HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule (US), which establishes national standards to protect individuals' medical records and personal health information. 3) Guidance about methods for de-identification by the US Department of Helath and Human Services (HHS). 4) Anonymization of DICOM Electronic Medical Records (Newhauser, et al. 2014).

Back to top

6. Persistent identifiers

Identifiers are essential to the human-machine interoperation (*F1* FAIR principle). Assigning globally unique persistent identifiers "is arguably the most important FAIR principle, because it will be hard to achieve other aspects of FAIR without them" (GO FAIR, 2017). Persistent identifiers or PIDs help find and collect data accurately, enable proper citation by collecting citation *metrics* about the use of a dataset, article or data generator (e.g. instrument, software, workflow). For the researcher, persistent identifiers enable disambiguation of people, and enable linking existing works as well as promoting wider dissemination of research.

For individuals: * ORCID Open Researcher and Contributor ID is a persistent digital identifier for an individual researcher. Activity: Distinguish yourself with ORCID by WILEY. * Web of Science ResearcherID is a unique identifier that connects researchers with works across the Web of Science ecosystem (*Web of Science*, *Publons*, and *InCites*).

For digital objects (files, datasets, publications, software, etc.): * DOI stands for Digital Object Identifier, which is a unique persistent identifier for a published digital object, issued by the DOI Foundation and its registered agencies using the handle system. * RAiD Research Activity Identifier RAiD for research activities and projects Persistent Uniform Resource Locator (PURL) currently in the process of becoming a universal PID.

Disclaimer, there are a wide range of PIDs available, we only cited two examples for each type.

Activity 1: OpenAIRE/FREYA/ORCID guide for researchers "How can identifiers improve the dissemination of your research outputs?".

Activity 2: Six Ways to Make Your ORCID ID Work for You! (Meadows, 2017). If you already have an ORCID, check this video to link publications to your ORCID profile.

Activity 3: Discussion the points highlighted by The Joint Declaration of Data Citation Principles from FORCE11 (Martone (ed), 2014).

To learn more about persistent identifiers visit GO-FAIR (F1 Principle) or the ARDC identifiers examples.

Back to top

7. Describing data: metadata

"Metadata (information about data) provides means for discovering data objects as well as providing other useful information about the data objects such as experimental parameters, creation conditions, etc." (Rajasekar, 2001). Unofficially, metadata can be grouped in two types by the way it has been created: automatically or manually. Read more about metadata in Working with Data by ARDC.

Why is building and using metadata relevant? It is a long-term mediator that supports the discovery, understanding and organisation of the process of research data. Across different communities, usually metadata follows **standards** see some examples gathered by DCC in collaboration with the Research Data Alliance. To optimise the reuse of data, metadata and data should be well-described so that they can be replicated and/or combined in different settings. Moreover, the FAIR principles give clear descriptors on what metadata should contain:

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource
- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
- A2. Metadata are accessible, even when the data are no longer available
- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data
- R1. Meta(data) are richly described with a plurality of accurate and relevant attributes a) a clear and accessible data usage license, b) associated with detailed provenance, c) meeting domain-relevant community standards.

These are aspects of metadata to keep in mind whether you produce, read or reuse metadata. In order to properly be interpreted by either humans or software, metadata needs to follow a standard vocabulary to precisely define what it should include. Metadata for imaging should include a standard terminology for describing the topic of study: physiological, clinical, demographic and genetic changes, and tools and instruments used for data capture and generation. The main recommendation is to share metadata per

project whenever possible, even if the data is not yet available (due to case by case restrictions).

a. Why ontologies?

By expressing image/characterisation annotation in machine computable form as a formal ontology, human knowledge can be brought to bear on effective search and interpretation of image data, especially across multiple disciplines, scales, and modalities" (Eliceiri, et al. 2012). Keep in mind that due to privacy restrictions any (meta)data can be listed under embargo or by limited access (go to section 5, if this is the case). Implementation, adoption and harvesting of metadata, requires defined ontologies. Due to increased demand for quantitative analysis and robust curation and sharing of the image/ characterisation data, the need for full ontologies and annotations is growing.

Ontologies for Neuroscience describe three domain specific ontologies and how they build on top of each other (Larson and Martone, 2009). They also note that existing domain specific vocabularies were built with the help of the Open Biological Ontologies (OBO) community (Smith, et al 2007). For example a subset of OBO is the EDAM Ontology which includes bio-imaging (Kalaš, et al. 2019). In addition. the Neuroscience Information Framework has developed the NIF Standard ontology (NIFSTD) for annotating and searching neuroscience resources. Plant, et al. 2011 provide an overview of what is needed to implement metadata that follows domain specific ontologies, they use as example microscopy cell image data. The National Center for Biomedical Ontology (NCBO) NCBO's BioPortal provides access to more than 270 biomedical ontologies and controlled terminologies (Musen, et al. 2012), and include some of those cited before. Also, the Ontology for Biomedical Investigations OBI Ontology.org enables communication between existing ontologies (Bandrowski, et al. 2016).

b. Controlled vocabularies

A defined list of agreed terms constitutes a controlled vocabulary, which is usually led by a user-community. Controlled vocabularies help data integration when, for example, ambiguities may exist on the terms used in the different datasets and across different repositories. If the data are to be re-used outside this community additional information may be required. Controlled vocabularies are part of a model called an ontology. An ontology has controlled vocabularies and the glue to link the terms providing an effective means whereby human and electronic agents can communicate unambiguously about concepts. This is relevant to the *Interoperability* principle of FAIR *I1* (GO FAIR). The goal of making data *interoperable* is to enable members of disparate communities to reuse and understand digital information over time.

Domain specific controlled vocabularies might be a wider landscape than ontologies to cover here, hence some more generic vocabulary examples are given. Schema.org widely used to build controlled vocabularies, a more specific example is bioschemas.org a collection of specifications that provide guidelines to facilitate a more consistent adoption of schema.org within the life sciences. Research vocabularies Australia is a public database of controlled vocabularies, at the time of writing this guide, no specific bioimaging vocabularies were found, maybe that is something you can help with?

c. Storing and publishing metadata

Where to store and publish metadata? The short answer is, depends which institution you are from (we recommend enquiring the university library, research officer or data steward), some options are:

1. Institutional repositories
2. Domain specific repositories
3. Generic repositories

Keeping in mind the FAIR principle *A2 metadata should be accessible, even when the data are no longer available*, reinforces the need of having at least shared metadata. To answer the question of where to publish (meta)data?, start with section 4 "Reusable data repositories for the image community". For a broader view, look at FAIRsharing.org databases for imaging. The ARDC - Research Data Archive (RDA) harvests institutional repositories, hence it can be the link between multiple repositories. The CSIRO - data access portal (is another option for projects related to CSIRO). DataCite metadata store allows users to register DataCite DOIs and associated metadata in a more generic context. As well as Zenodo which provides a DOI and versioning capabilities.

Activity 1: For discussion. Have a look at the metadata stored at Research Data Australia for the 7T Magnetom instrument (CAI, 2017), it contains simple but important public metadata and a PID.

Activity 2: For more information read where to store metadata? from ARDC.

Back to top

8. Reusable data best practices

Here is a suggested list of data best practices to adopt in your research outputs. These will improve data and software reusability by others, which includes yourself in the future. Remember, making data/software available for others to re-use publicly is the goal, but not all data must be shared to all. Adding terms and conditions of accessibility is an option to consider. To share data, you can make use of public infrastructures already mentioned (section 4 "Reusable data repositories") or use your institutionally provided data repository. To get started, there are a few things you should keep in mind.

a. Provenance - Usually provenance is a manually produced metadata file (it can also be automatically produced). It is important for the reuse of data in the future, it should contain descriptors such as data producer, date history (log of changes), data dictionary. **Primary data ought to be read only.**

b. File formats - Most file formats are defined by the data producer (e.g. instrument or software), whenever possible you should try to convert data to formats that are publicly accessible (open formats).

For example, DICOM (Digital Imaging and Communications in Medicine) format mostly used in neurosciences, can be converted to NIfTI (Neuroimaging Informatics Technology

Initiative) or BIDS format. Another example is the Hierarchical Data Format version 5 (HDF5) (Dougherty 2009), an open source file format that supports large, complex, heterogeneous data [HDF5] used by MINC and Huygens Software. In addition, you can read more about why metadata matters and a discussion about propietary formats by Linkert, et al. 2010, introducing multidimensional microscopy image data and formats like TIFF and OME TIFF.

c. Data structures Keep consistent file and folder naming conventions across linked projects. * Brain imaging data structure (BIDS) BIDS website (Gorgolewski, et al. 2016), BIDS fairsharing link. * IDR metadata example from the open microscopy community - Image Data Resource (IDR). * Datacrate example a more generic specification for packaging research data.

d. Data curation Should be included in your data quality workflow as part of the process, ideally this will be automated.

e. Data versioning To keep the provenance of your data you might use data versioning tools: Git or GitHub (for code). Git annex and Datalad are other options for data, you should investigate whether your repository of choice, has the capability to do this.

f. Containerisation For data processing pipelines, e.g. Singularity, Docker, or use Virtual environments, such as the Characterisation Virtual Laboratory.

g. Protocols Search for imaging protocols publicly shared by Protocol exchange an open resource where the community of scientists pool their experimental know-how to help accelerate research, e.g. protocol exchange for imaging.

h. Create documentation Write and describe **everything** you would need to understand a project or dataset in a few months. *A README file* helps ensure that your data can be correctly interpreted and reanalysed by others. For example, the DataDryad Readme is an example of minimum documentation. Write the docs writethedocs.org is also a great innitiative of people who care about documentation, we recommend you to use it!

i. Benchmarks or checksums Checksums are used to make sure that transferred or stored copies of data match the data that was originally created. Read more about data integrity checksums.

Activity 1: Recommended reading. A brain imaging case study that provides direct evidence of the impact of open sharing on data use and resulting publications over a seven-year period (2010-2017) stated: “We dispel the myth that scientific findings using shared data cannot be published in high-impact journals and demonstrate rapid growth in the publication of such journal articles”. You can pick to read the (*pre-print* Milham, et al. 2017) or the paper (Milham, et al. 2018), what are your thoughts on that, and conclusions after reading the paper?

Activity 2 (Discussion + Action): What Can You Do?

Contribute your data – Previously published datasets.

Release some or all of the project metadata – your call, as a simple rule, the more the better!

Curate existing datasets to make available in the future - you set the upload schedule.

Contribute your scripts/code.

Have discussions with your team members about licensing and sharing.

Create a data management plan.

Activity 3: Go through the questions from the Horizon2020 guide to create a FAIR Data Management Plan and see if you can already answer any of the questions.

Recommended extra reading: Best Practices in Data Analysis and Sharing in Neuroimaging using MRI, Ten Simple Rules for Creating a Good Data Management Plan, Ten Simple Rules for Reproducible Computational Research and Ten principles for machine-actionable data management plans, these papers will help you connect all the concepts that you have learned so far.

Back to top

9. Licensing your work

Licensing your work / research outputs to be open access (research output here means data, metadata, code, workflows) allows you as author or contributor to enable reuse and appropriate attribution of that work. If there is no licence attached to your work, you are actually stopping anyone to legally reuse it. Did you know that *No licence = No permissions?*. Also, if you find research outputs that you want to reuse, you should only reuse it according to their licence.

Be aware that you have the right to choose a licence that best suits your purpose. There are multiple different licences and versions of these, to be applied to data and software. Some licences are applicable only in certain countries, so think of applying an international licence. Be aware that the data repository that you use might ask you to accept their "terms and conditions" which affects how you might use or share data, by expanding, modifying or limiting the intended purpose or your own licence. Also, you can have multiple licences, for different purposes or different audiences. Finally, not every part of your work/ research outputs needs to be publicly available or be licensed, but the more you share with clear permissions the better.

Activity 1: What if you don't choose a licence?, explains and gives you a few reasons to think about licensing your work. If you are interested in reading about GitHub terms and conditions take 5 extra minutes.

Activity 2: (flowcharts as a survey) The ARDC has a guides about licensing for three specific scenarios: a) Data creator flowchart b) Data supplier flowchart and c) Data users flowchart. If you want to know more about licensing and copyright for data reuse visit the ANDS page about this.

A few types of licences: *Creative Commons (CC)* is, so far, very easy to apply and it is broadly being reused. It is strongly promoted in the United States, however it is an internationally recognised licence creator. CC is good for: a) very simple, factual data sets b) data to be used automatically. You should watch out for the version in use, recommended to use version 4 or later. CC has attribution stacking Non Commercial (NC), Shared Alike (SA) and Non derivatives (ND). The NC condition: only to be used with dual licensing. The SA condition reduces interoperability. The ND condition severely restricts reuse. To help you decide, use this https://creativecommons.org/choose. Another licence is *Copyleft* a general method for making a program (or other work) free (in the sense of freedom, not "zero price"), and requiring all modified and extended versions of the program to be free as well. Following on, *Open Data commons*, also provides licences specifically for open data, good for most databases and datasets, e.g. Open Data Commons Open Database Licence (ODC-ODbL) or Open Data Commons attribution licence (ODC-By). Licences specific for software: Furthermore, *Mozilla Public Licence (MPL)*, *MIT Licence*, *the GNU General Public Licence (GPL)* and a list of open source licences by category are other options you might want to investigate. To help you choose a licence for software, look at the descriptions: https://choosealicense.com/. *Acknowledgement, most of the cited licences on this section, were first mentioned by "License Research Data from the Digital Curation Centre" (Ball, 2014).*

Back to top

10. Data citation for access and attribution

Citation analysis and citation metrics are important to the academic community, which gives recognition to the researchers and their work. Data citation continues the tradition of acknowledging other people's work and ideas. It also helps make research data more *findable* and *accessible*. It is now common practice for authors to formally cite the research datasets and associated software that underpin their research findings.

Activity 1: (Video, 12 mins) Responsible Data Use: Citation and Credit (Mayernik, 2013).

Activity 2: *How to cite data and software?* This example from Dryad clearly shows how to cite the dataset that underpins a journal article as well as the article itself. Note that both citations include a Digital Object Identifier (DOI).

Activity 3: *What to cite and why?* For data and software from ARDC.

Back to top

Acknowledgements

We acknowledge Chris Erdmann for reviewing the first version of this document, and all collaborators now listed as authors for useful comments and the editing sections of this document. Paula Andrea Martinez also acknowledged the National Imaging Facility and the Australian Research Data Commons for funding this research.

Back to top

Pre-print

This document is also available via the Open Science Framework as a pre-print and it is citable with the following DOI 10.17605/OSF.IO/ZKJ4R where versions of it in .docx, .odt and .md have been saved. This document links back to the website of Top 10 FAIR for Imaging.

Back to top

Supplementary Information

Characterisation

"Characterisation is the general process of probing and measuring the structures and properties of materials at the micro, nano and atomic scales. It is essential across natural, agricultural, physical, life and biomedical sciences and engineering." Back to top

Reusable data repositories Section 4

Section 4 lists various public repositories which we have collected in the following list.

Neurosciences

Data repositories recommended by the Scientific Data Journal which accept human-derived data, in addition NeuroMorpho.org and G-Node also accept data from other organisms. Please note that human-subject data submitted to OpenNeuro must be de-identified, while Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (FCP/INDI) can handle sensitive patient data.

- Neuromorpho.org is a centrally curated inventory of reconstruction data for the neuroscience community associated with peer-reviewed publications.
- G-Node is a modern Research Data Management for Neuroscience with git-annex version control. Inspired by Github a rebuilt for distributed file synchronization system dealing with files larger than git can currently easily handle.
- Datasets from Datalad.org have built-in support for metadata extraction and search. It allows to search through a large collection of readily available datasets. It is the master collection for openneuro/openfmri and neurovault.
- Openneuro.org (formerly openfMRI) a free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data. Get access via orcid or google account. Also accessible via GitHub OpenNeuro Datasets
- Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (FCP/INDI) can handle sensitive patient data! Upon successful registration users have the right to unrestricted usage of the datasets for non-commercial purposes. Current goal: To make the aggregation and sharing of well-phenotyped datasets a cultural norm for the imaging community.

- Central XNAT is a database for sharing neuroimaging and related data with select collaborators or the general community. Available repository options: Public, Protected and Private.
- Neurovault.org is a public online repository for statistical maps, parcellations and atlases of the brain. It is registered in identifiers.org.
- Brain Map Portal This portal provides access to high quality data and web-based applications created for the benefit of the global research community studying brain sciences.
- Human Connectome.org The human connectome houses and distributes public research data for a series of study aspects of how age, growth, disease, and other factors can affect the ever-changing connections in the human brain.
- LORIS (Longitudinal Online Research and Imaging System) for heterogeneous data acquisition e.g. imaging, clinical, behavior, and genetics, storage, processing, and ultimately dissemination. Read more.
- Child Mind Institute Healthy Brain Network, shares a biobank of data from 10,000 young participants. The HBN Biobank houses data about psychiatric, behavioral, cognitive, and lifestyle phenotypes, as well as multimodal brain imaging (fMRI, diffusion MRI, morphometric MRI), electroencephalography, eye-tracking, voice and video recordings, genetics and actigraphy read more.
- The Alzheimer's Disease Neuroimaging Initiative (ADNI) collects, validate and utilises data, including MRI and PET images, genetics, cognitive tests, CSF and blood biomarkers as predictors of the disease.
- Brain Genomics Superstruct
- Brain base is a collaborative research and data management platform for the Human Neuroscience community.
- NIH ABCD Adolescent Brain Cognitive Development Study
- OASIS The Open Access Series of Imaging Studies aims making neuroimaging data sets of the brain freely available to the scientific community. with > 1000 participants, over the course of 30 years, totalling 1500 raw imaging scans.
- Neurosynth is a platform for large-scale automated synthesis of fMRI data that allows meta-analysis.
- DaRIS (Distributed and Reflective Informatics System) is a framework for managing data and meta-data primarily for biomedical imaging.
- MyTardis is an Australian solution that provides data transfer, manages data storage and provides mechanisms to access and share the data and it is domain agnostic.

Microscopy

- Image Data Resource from Open Microscopy (IDR) is a public data integration and publication platform.
- OMERO from Open Microscopy is a central repository, OMERO supports over 140 image file formats, including all major microscope formats. It uses bio-formats.
- The Cell: An Image Library is a freely accessible, public repository of reviewed and annotated images, videos, and animations of cells from a variety of organisms,

showcasing cell architecture, intracellular functionalities. It is registered in identifiers.org.
- Electron microscopy Protein data bank.
- EMPIAR (Electron Microscope Public Image Archive) for 2D images.

Biomedical sciences

- UK Biobank provides health information of over 500,000 volunteer participants de-identified, to approved researchers in the UK and overseas, from academia and industry.
- The Cancer Imaging archive (TCIA) is a service which de-identifies and hosts a large archive of medical images of cancer accessible for public download. The data are organized as "Collections", typically patients related by a common disease, image modality (MRI, CT, etc) or research focus. DICOM is the primary file format used by TCIA for image storage.
- The Cardiac Atlas Project comprises cardiac imaging data from relevant studies such as MESA or DETERMINE.
- Siscas Medical imaging repository for CT and microCT images. It offers controlled access.
- Coherent X-ray Imaging Data Bank (CXIDB) uses CXI file format.
- DeepLesion, a project from the NIH that released a dataset of 32,000 CT chest images with different type of lesions.
- Content manager SciCrunch, a data sharing and display platform and training materials for searching and sharing in biomedical sciences across hundreds of databases.
- BioStudies The database can accept a wide range of types of studies, and its used to describe it. It also enables manuscript authors to submit supplementary information and link to it from the publication.

Non-domain specific

- Australian Data Archive ADA it has a Core Trust Seal Certification. It is mainly for digital data relating to social, political and economic affairs which might include social and health studies.
- Data dryad
- Dataverse.org
- Zenodo.org

Data registries and catalogues

re3data.org - a registry of some 2000 data repositories. Research Data Australia- RDA or read more about their services. Also, FAIRSharing.org offers a catalogue of databases, described according to the BioDBcore guidelines. OpenAIRE content provider, European Open Science Cloud, Google Public Data, Google Dataset Share, for open access publications Open knowledge maps.

Back to top

References

Alan Turing Institute. 2019. “Research Data Management.” The Turing Way. https://the-turing-way.netlify.com/rdm/rdm.html.

ANDS. 2017. “The FAIR Data Principles.” ANDS. https://www.ands.org.au/working-with-data/fairdata.

ANDS. 2018. “De-Identification.” ANDS. http://www.ands.org.au/working-with-data/sensitive-data/de-identifying-data.

ANDS. 2018 "Publishing and sharing sensitive data". https://www.ands.org.au/guides/sensitivedata.

ANDS. “Licensing and Copyright for Data Reuse.” ANDS. Accessed July 17, 2019. https://www.ands.org.au/working-with-data/publishing-and-reusing-data/licensing-for-reuse.

ANDS. “Storing Metadata.” Working with Data. Accessed July 17, 2019. https://www.ands.org.au/working-with-data/metadata/storing-metadata.

ARDC. 2017a. “Citation and Identifiers.” ARDC. https://ardc.edu.au/resources/working-with-data/citation-identifiers/.

ARDC. 2017b. “Data Citation.” https://ardc.edu.au/resources/working-with-data/citation-identifiers/data-citation/.

ARDC. “Metadata.” Working with Data. Accessed July 17, 2019. https://ardc.edu.au/resources/working-with-data/metadata/.

Ball, Alex. 2014. “How to License Research Data.” DCC http://www.dcc.ac.uk/resources/how-guides/license-research-data.

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, et al. 2016. “The Ontology for Biomedical Investigations.” PLOS ONE 11 (4): e0154556. https://doi.org/10.1371/journal.pone.0154556.

Bioimaging. 2019. “EOSC-Life: Developing an Open Collaborative Space for Digital Biology in Europe Euro-BioImaging.” http://www.eurobioimaging.eu/content-news/eosc-life-developing-open-collaborative-space-digital-biology-europe.

CAI. 2017. “7T Magnetom Metadata.” Research Data Australia. https://researchdata.ands.org.au/7t-magnetom/1305790.

Cavalli, Valentino. 2018. “Open Consultation on FAIR Data Action Plan.” LIBER. https://libereurope.eu/blog/2018/07/13/fairdataconsultation/.

CODATA. 2018. “Enabling FAIR Data Project and Commitment Statement - CODATA.” http://www.codata.org/news/299/62/Enabling-FAIR-Data-Project-and-Commitment-Statement.

Commonwealth of Australia - Australian Bureau of Statistics. 2017. “Managing the Risk of Disclosure: The Five Safes Framework.” https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1160.0Main%20Features4Aug%202017.

COPDESS. 2018. “Enabling FAIR Data Project – COPDESS.” http://www.copdess.org/enabling-fair-data-project/.

Core Trust Seal. 2016. “An Introduction to the Core Trustworthy Data Repositories Requirements.” https://www.coretrustseal.org/wp-content/uploads/2017/01/Intro_To_Core_Trustworthy_Data_Repositories_Requirements_2016-11.pdf.

Council Australian Research. 2018. “ARC Open Access Policy Version 2017.1.” https://www.arc.gov.au/policies-strategies/policy/arc-open-access-policy-version-20171.

2019. “Promoting FAIR Principles in the Healthcare Field Digital Curation Centre.” http://www.dcc.ac.uk/blog/promoting-fair-principles-healthcare-field.

DCC.“DCC Curation Lifecycle Model Digital Curation Centre.” Accessed July 26, 2019. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Dougherty, Matthew T., Michael J. Folk, Erez Zadok, Herbert J. Bernstein, Frances C. Bernstein, Kevin W. Eliceiri, Werner Benger, And Christoph Best. 2009. “Unifying Biological Image Formats with HDF5.” Commun ACM 52 (10): 42–47. https://doi.org/10.1145/1562764.1562781.

Downs, Robert. 2012. “Providing Access to Your Data: Handling Sensitive Data.” https://doi.org/10.7269/p3mk69t8.

Eliceiri, Kevin W., Michael R. Berthold, Ilya G. Goldberg, Luis Ibáñez, B. S. Manjunath, Maryann E. Martone, Robert F. Murphy, et al. 2012. “Biological Imaging Software Tools.” Nat Methods 9 (7): 697–710. https://doi.org/10.1038/nmeth.2084.

European Commission. 2013. “Guidelines on FAIR Data Management in Horizon 2020.” http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

FAIRsharing. 2015a. “DICOM Digital Imaging and COmmunications in Medicine.” https://doi.org/10.25504/fairsharing.b7z8by.

FAIRsharing. 2015b. “NIfTI-1 Data Format.” https://doi.org/10.25504/fairsharing.jgzts3.

Future of Privacy Forum. 2017. “A Visual Guide to Practical Data de-Identification.” https://fpf.org/wp-content/uploads/2016/04/FPF_Visual-Guide-to-Practical-Data-DeID.pdf.

GO FAIR. 2016. “FAIR Principles.” GO FAIR. https://www.go-fair.org/fair-principles/.

GO FAIR. 2017a. “A2: Metadata Should Be Accessible Even When the Data Is No Longer Available.” GO FAIR. https://www.go-fair.org/fair-principles/a2-metadata-accessible-even-data-no-longer-available/.

GO FAIR. 2017b. “F1: (Meta) Data Are Assigned Globally Unique and Persistent Identifiers.” GO FAIR. https://www.go-fair.org/fair-principles/f1-meta-data-assigned-globally-unique-persistent-identifiers/.

GO FAIR. 2017c. “I1: (Meta)data Use a Formal, Accessible, Shared, and Broadly Applicable Language for Knowledge Representation.” GO FAIR. Accessed July 17, 2019. https://www.go-fair.org/fair-principles/i1-metadata-use-formal-accessible-shared-broadly-applicable-language-knowledge-representation/.

Gorgolewski, Krzysztof J., Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, et al. 2016. “The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments.” Scientific Data 3 (June): 160044. https://doi.org/10.1038/sdata.2016.44.

Halchenko, Y. 2018. “Anonymization Tools — Open Brain Consent 0.1.dev1 Documentation.” https://open-brain-consent.readthedocs.io/en/latest/anon_tools.html.

Health Information Service. 2012. “Methods for de-Identification of Protected Health Information.”. HHS.gov. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

Higman, Rosie, Daniel Bangert, and Sarah Jones. 2019. “Three Camps, One Destination: The Intersections of Research Data Management, FAIR and Open.” Insights 32 (1): 18. https://doi.org/10.1629/uksg.468.

Kalaš, Matúš, Nataša Sladoje, Laure Plantard, Martin Jones, Leandro Aluisio Scholz, Joakim Lindblad, and contributors. 2019. “Edamontology/Edam-Bioimaging: Alpha05.” Zenodo. https://doi.org/10.5281/zenodo.2557012.

Larson, Stephen D., and Maryann E. Martone. 2009. “Ontologies for Neuroscience: What Are They and What Are They Good for?” Front. Neurosci. 3. https://doi.org/10.3389/neuro.01.007.2009.

Linkert, Melissa, Curtis T. Rueden, Chris Allan, Jean-Marie Burel, Will Moore, Andrew Patterson, Brian Loranger, et al. 2010. “Metadata Matters: Access to Image Data in the Real World.” J Cell Biol 189 (5): 777–82. https://doi.org/10.1083/jcb.201004104.

Martone, M. 2014. “Joint Declaration of Data Citation Principles - FINAL.” Force11. https://www.force11.org/datacitationprinciples.

Meadows, Alice. 2017. “Six Ways to Make Your ORCID iD Work for You!”. https://orcid.org/blog/2018/07/27/six-ways-make-your-orcid-id-work-you.

Milham, Michael P., R. Cameron Craddock, Michael Fleischmann, Jake Son, Jon Clucas, Helen Xu, Bonhwang Koo, et al. 2017. “Assessment of the Impact of Shared Data on the Scientific Literature.” bioRxiv, September, 183814. https://doi.org/10.1101/183814.

Milham, Michael P., R. Cameron Craddock, Jake J. Son, Michael Fleischmann, Jon Clucas, Helen Xu, Bonhwang Koo, et al. 2018. "Assessment of the Impact of Shared Brain Imaging Data on the Scientific Literature." Nature Communications 9 (1): 2818. https://doi.org/10.1038/s41467-018-04976-1.

Musen, Mark A., Natalya F. Noy, Nigam H. Shah, Patricia L. Whetzel, Christopher G. Chute, Margaret-Anne Story, and Barry Smith. 2012. "The National Center for Biomedical Ontology." J Am Med Inform Assoc 19 (2): 190–95. https://doi.org/10.1136/amiajnl-2011-000523.

Nature. 2016. "Where Are the Data?" Nature News 537 (7619): 138. https://doi.org/10.1038/537138a.

Newhauser, Wayne, Timothy Jones, Stuart Swerdloff, Warren Newhauser, Mark Cilia, Robert Carver, Andy Halloran, and Rui Zhang. 2014. "Anonymization of DICOM Electronic Medical Records for Radiation Therapy." Comput Biol Med 0 (October): 134–40. ttps://doi.org/10.1016/j.compbiomed.2014.07.010.

NHMRC. "Research Quality NHMRC." Accessed May 30, 2019. https://www.nhmrc.gov.au/research-policy/research-quality.

OpenAIRE. 2017. "How to Deal with Sensitive Data." https://www.openaire.eu/sensitive-data-guide.

OpenAIRE. "How Can Identifiers Improve the Dissemination of Your Research Outputs?". Accessed July 17, 2019. https://www.openaire.eu/how-can-identifiers-improve-the-dissemination-of-your-research-outputs.

ORCID. 2012. "ORCID Overview for Researchers." https://orcid.org/content/orcid-overview-researchers.

Plant, Anne L., John T. Elliott, and Talapady N. Bhat. 2011. "New Concepts for Building Vocabulary for Cell Image Ontologies." BMC Bioinformatics 12 (1): 487. https://doi.org/10.1186/1471-2105-12-487.

PLOS ONE. 2017. "PLOS ONE: Accelerating the Publication of Peer-Reviewed Science.". https://journals.plos.org/plosone/s/data-availability#loc-acceptable-data-sharing-methods.

PLOSData. 2017. "FAIRsharing Recommendation: PLOS." https://fairsharing.org/recommendation/PLOS.

Protocol exchange. "Protocol Exchange Research Square Subject Imaging." Accessed July 17, 2019. https://protocolexchange.researchsquare.com/?journal=protocol-exchange&limit=10&offset=0&status=all&subjectArea=Imaging.

Rajasekar, Arcot K., and Reagan W. Moore. 2001. "Data and Metadata Collections for Scientific Applications." In High-Performance Computing and Networking, edited by Bob Hertzberger, Alfons Hoekstra, and Roy Williams, 72–80. Lecture Notes in Computer

Science. Springer Berlin Heidelberg https://link.springer.com/chapter/10.1007/3-540-48228-8_8.

Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, et al. 2007. "The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration." Nature Biotechnology 25 (11): 1251–5. https://doi.org/10.1038/nbt1346.

Sweeney, Latanya. "Identifiability of de-Identified Data." Accessed July 17, 2019. http://latanyasweeney.org/work/identifiability.html.

Swiss National Science Foundation. 2018. "Explanation of the FAIR Data Principles." http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf.

Web of Science. "Web of Science ResearcherID." Accessed July 17, 2019. https://www.researcherid.com/#rid-for-researchers.

Wiley. 2014. "Researcher Data Sharing Insights.". http://www.acscinf.org/PDF/Giffi-%20Researcher%20Data%20Insights%20--%20Infographic%20FINAL%20REVISED.pdf.

Wiley. "Distinguish Yourself with ORCID Wiley." Accessed July 17, 2019. https://authorservices.wiley.com/author-resources/Journal-Authors/submission-peer-review/orcid.html.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (March): 160018. https://doi.org/10.1038/sdata.2016.18.

Wilkinson, Mark D., Ruben Verborgh, Luiz Olavo Bonino da Silva Santos, Tim Clark, Morris A. Swertz, Fleur D. L. Kelpin, Alasdair J. G. Gray, et al. 2017. "Interoperability and FAIRness Through a Novel Combination of Web Technologies." PeerJ Comput. Sci. 3 (April): e110. https://doi.org/10.7717/peerj-cs.110.

Back to top

Linked Open Data

Sprinters

- Ronald Siebes / ORCID: 0000-0001-8772-7904 / Data Archiving & Networked Services (DANS)
- Gerard Coen / ORCID: 0000-0001-9915-9721 / Data Archiving & Networked Services (DANS)
- Kathleen Gregory / ORCID: 0000-0001-5475-8632 / Data Archiving & Networked Services (DANS)
- Andrea Scharnhorst / ORCID: 0000-0001-8879-8798 / Data Archiving & Networked Services (DANS)

Audience

We aim to provide a document which is understandable to non-experts, but that also provides specific technical references and does not downplay some of the complexities of LOD.

- Researchers (especially from the social sciences & humanities)
- Anyone interested in publishing Linked Open Data (LOD)
- Anyone interested in supporting use of LOD in research

Description

Linked Open Data (LOD) are inherently interoperable and have the potential to play a key role in implementing the "I" in FAIR. They are machine-readable, based on a standard data representation (RDF - Resource Description Format) and are seen as epitomizing the ideals of open data (see https://5stardata.info/en/). They offer great promise in helping to achieve a specific type of machine-executable interoperability known as semantic interoperability, which relies on linking data via common vocabularies or Knowledge Organisation Systems (KOS). This document attempts to demystify LOD and presents_ Ten Things_ to help anyone wanting to publish LOD.

Although this list of_ "Things"_ are presented in a roughly linear order, preparing and publishing LOD are iterative processes. Expect to go back and forth a bit between the_ Things_, and take the time to double check that your progress matches your desired end result. Some_ Things_ can be executed in parallel; you will also notice recurring themes (e.g. sustainability and licensing concerns) that need to be considered throughout the workflow.

These *Things* are based on our own practical experiences in publishing LOD in various interdisciplinary settings, e.g. the Digging into the Knowledge Graph project (http://di4kg.org). Our goal is to complement existing scholarly reports on LOD implementations (e.g. Hyvönen, 2012; Hyvönen, 2019; Meroño-Peñuela et al, 2019), other workflow models (see W3C Step #1), and the authoritative "Best Practices for Publishing Linked Data" of the W3C, which we cross-reference (as W3C Step #X) wherever appropriate. We include visualisations, suggest readings, and highlight other projects to

make this guide understandable and usable for people across disciplines and levels of expertise. However, it is important to note that semantic web technology is a complex scientific field, and that you may need to consult a semantic web expert along the way.

Overview

Thing 1: Learning - Understand and practice the Semantic Web and LOD basics.

Thing 2: Exploring - Inventory of your data.

Thing 3: Defining - Define the URI (Uniform Resource Identifier) naming strategy.

Thing 4: Resolving - Consider resolvability when a person or machine visits the URI.

Thing 5: Transforming - Generate the URIs for the selected concepts and relations according to the URI naming strategy.

Thing 6: Mapping - Map your Linked Data from your newly defined namespace to similar concepts and relations within the LOD.

Thing 7: Enriching - Enrich your data with information from the LOD.

Thing 8: Exposing - Define how people can get access to your LD: a data-dump, a SPARQL endpoint or a Web API.

Thing 9: Promoting - Publish and disseminate the value of your data via visualisations and workflows.

Thing 10: Sustaining - Ensure sustainability of your data.

Things

Thing 1: Learning - Understand and practice the Semantic Web and LOD basics

Semantic web technology (which underlies LOD) is complex. It requires not only a new data model (Resource Description Framework, RDF) but also infrastructures for storing and linking data as well as algorithms for retrieving, enriching and reasoning across those data.

Understanding LOD begins with understanding the Resource Description Framework. RDF is a standard format defined by the World Wide Web Consortium (W3C) that can be easily interpreted by machines.

RDF statements are called “triples” because they contain three pieces - **subject : predicate : object**. RDF data is modelled as a “labeled graph” which links description of resources together. **Subjects** and **objects** are nodes, **predicates** are links.

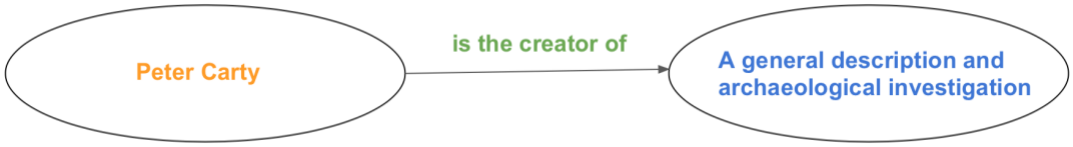


Figure 1

Figure 1: An example of a triple

RDF is notated as a list of these statements (the triples) that describe each piece of knowledge in your dataset. This list of statements can be thought of as a large indexing file to your data.

There are different formats for creating RDF. Popular formats include RDFa, RDF/XML, Turtle and N-Triples. Although these formats are slightly different, the meaning of the RDF statements written with them remains the same. In our examples, we use the Turtle format and wrote the code using Atom, a collaborative text editor. (See W3C Step #3)

This is a section of RDF representing an example dataset, which we use throughout the text.

```
1
2  # Ourxiv Records
3
4  # Prefixes
5  @prefix dc: <http://purl.org/dc/elements/1.1/> .
6  @prefix rds: <http://rdfs.org/ns/void#> .
7  @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
8  @base <http://www.ourxiv.com/ui/datasets/id/ourxiv-dataset> .
9
10 ourxiv:98261 dc:identifier "ourxiv:98261" .
11 ourxiv:98261 dc:creator "Peter Carty" .
12 ourxiv:98261 dc:title "A general description and archaeological investigation" .
13 ourxiv:98261 dc:date "2002-08-18" .
14 ourxiv:98261 dc:coverage "Dublin" .
15 ourxiv:98261 dc:type "Text" .
```

Figure 2

Figure 2: A screenshot of an RDF representation

The graph structure of RDF offers benefits over typical database structures. Creating new subjects and predicates is far less tedious than creating new fields and linking tables, as is common in database design. Storage in RDF is also more compact. Perhaps most importantly, RDF enables specific ways of questioning your data that are not possible with other structures. In a triple, the predicates (the links) also have meaning and thus are semantically encoded; this facilitates executing more complex operations (known as "semantic reasoning") on the graph. In our example (see Figure 1; Figure 2), the role of "Peter Carty" as "creator" of the dataset is spelled out, and so can be differentiated from

other possible roles - such as being a contributor or a collaborator. In the end, RDF is simply another way of expressing your data.

Activity: Who better to introduce you to the concepts ot LOD than Tim Berners-Lee, founder of the World Wide Web? View these videos for an overview of the topic:

- Tim Berners-Lee on "The Next Web": http://www.ted.com/talks/tim_berners_lee_on_the_next_web
- Tim Berners-Lee at the GOV 2.0 expo ("bag of crisps"): https://youtu.be/ga1aSJXCFe0
- Then, check out how Google uses this same technology in the Knowledge Graph: https://www.youtube.com/watch?v=mmQl6VGvX-c

Putting data into RDF is one ingredient in working toward LOD. RDF statements must also be expressed as Uniform Resource Identifiers (URIs - see Steps 3-6) in order to link them to other data. It is possible to have Linked Data (LD) living on internal servers that are not a part of the larger Linked Open Data Cloud (LOD Cloud). In order to publish data to the LOD Cloud, URIs must be readable, or resolvable (see Step 4), not only internally, but also to outside sources.

Activity: Visit the LOD Vocabularies https://lov.linkeddata.es/dataset/lov/ and see the different types of things that can be linked to. Eventually you will be linking your data to some of these vocabularies (see Thing 7). Can you identify datasets that contain concepts similar to those in your own dataset? (For example, if you have data about cities, you may want to look for datasets with information about cities, e.g. DBPedia).

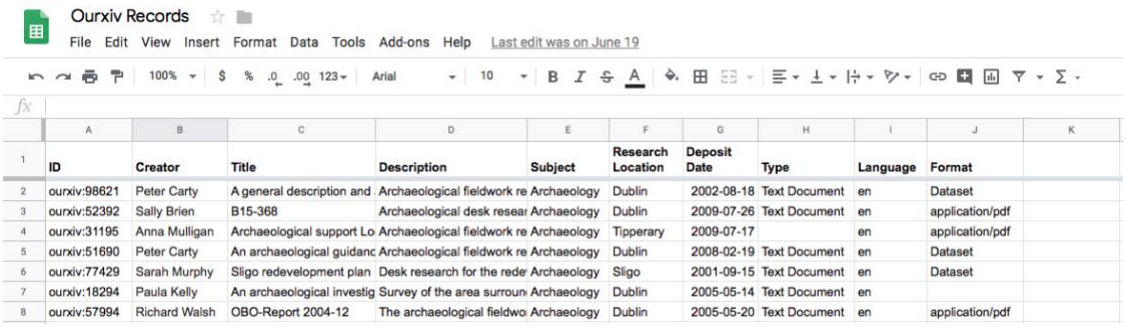
Thing 2: Exploring - Inventory of your data

2a: Identify relevant concepts from your datasets that you want to expose as Linked Open Data (LOD)

The first step in expressing your data in RDF is to identify the concepts in your dataset that you would eventually like to expose and link to the LOD Cloud. It will most likely NOT make sense to expose all of the concepts that exist in your data; you will need to be selective. (See W3C Step #2)

Activity: Take an inventory of your dataset. What type of data do you have? What structure (e.g. XML, a database) are they in? Based on your exploration of the LOD Vocabularies in Thing 1, think about which concepts it makes sense to eventually link.

We take as our example an archeological dataset based on data found in the EASY data repository of DANS. The dataset contains archaeological records with ten attributes listed for each record; the dataset has been anonymised for the purpose of this example. As we will discuss later, not all of the concepts in the dataset are interesting to share.



	A	B	C	D	E	F	G	H	I	J	K
1	ID	Creator	Title	Description	Subject	Research Location	Deposit Date	Type	Language	Format	
2	ourxiv:98621	Peter Carty	A general description and	Archaeological fieldwork re	Archaeology	Dublin	2002-08-18	Text Document	en	Dataset	
3	ourxiv:52392	Sally Brien	B15-368	Archaeological desk resear	Archaeology	Dublin	2009-07-26	Text Document	en	application/pdf	
4	ourxiv:31195	Anna Mulligan	Archaeological support Lo	Archaeological fieldwork re	Archaeology	Tipperary	2009-07-17		en	application/pdf	
5	ourxiv:51690	Peter Carty	An archaeological guidanc	Archaeological fieldwork re	Archaeology	Dublin	2008-02-19	Text Document	en	Dataset	
6	ourxiv:77429	Sarah Murphy	Sligo redevelopment plan	Desk research for the rede	Archaeology	Sligo	2001-09-15	Text Document	en	Dataset	
7	ourxiv:18294	Paula Kelly	An archaeological investig	Survey of the area surroun	Archaeology	Dublin	2005-05-14	Text Document	en		
8	ourxiv:57994	Richard Walsh	OBO-Report 2004-12	The archaeological fieldwo	Archaeology	Dublin	2005-05-20	Text Document	en	application/pdf	

Figure 3

Figure 3: The original data in tabular form

It is also important to consider who owns the data in this step. In an archive, each dataset can have information about the license status. Are the data listed as being open, or is there a requirement for you to request permission or acknowledgement in order to use the data? Ownership and licensing are also important to consider for your own data.

Activity: Think about your own data. Are you the data rights holder for all parts of your data? If not, identify any licenses that may restrict if you can expose and link the data to the LOD Cloud. (See W3C Step #1)

Identifying relevant concepts and relationships is a step that is vital for everyone, both novices and experienced computer scientists. Computer scientists use visual tools to "model" their data, such as Visio and Draw.io, but you could also create a simple list of the concepts that you would like to expose and link.

Activity: View the introductory tutorial for Draw.io and experiment using the tool by visiting this link: https://www.draw.io.

One side remark: The term 'concept' has a specific meaning in various disciplines. In the context of computer science, 'concept' means a class, i.e. a knowledge representation (of objects, individuals, actions, etc.). In general, 'concepts' are abstractions made to order things. They are represented by terms (Dextre Clarke, 2019) An ensemble of concepts is often represented in the form of controlled vocabularies, schemas, or ontologies, more generally known as Knowledge Organisation Systems (KOS).

This means that, if your data are structured in a database format, the headers or fields represent your concepts.The actual cell values are the concrete instantiation (or specific examples) such as phenomena or observations for these concepts. This is a general rule of thumb; you could also have cases, however, in which your cells are concepts themselves.

Often not all of the elements in the dataset are things you wish to share. (We discuss this further in Thing 5 when we address how to filter your dataset). We have selected five concepts from our example dataset to demonstrate the difference between concepts and instances.

Concepts from out dataset:

ID	Creator	Title	Research Location	Date	Type
Instances of each of these concepts:					
ourxiv:98621	Peter Carty	A general description and archaeological investigation	Dublin	2002-08-18	Text Document

Figure 4: Examples of concepts and instances in our example dataset

2b: Identify relevant relations from your datasets that you want to expose as Linked Data (LD)

Having identified the concepts in your dataset, you now need to identify the relations (i.e. what later becomes predicates or links) between those concepts.

The relations are often determined by the structure of the database itself; however, sometimes the column or row headers can also express relations.

Here is a data model which shows some of the relationships between the concepts in our example from 2a. We can see the **subject : predicate : object** structure.

Peter Carty created A general description and archaeological investigation

Visualising your data model in this way can also help with understanding where there are relationships which may have been hidden. Once you have a model of the concepts and relationships you would like to link and expose, you are ready to begin defining them as URIs, which will be explained in Thing 3.

Activity: Examine the below model of our dataset. What might a possible relationship between 'Peter Carty' and 'Dublin' be?

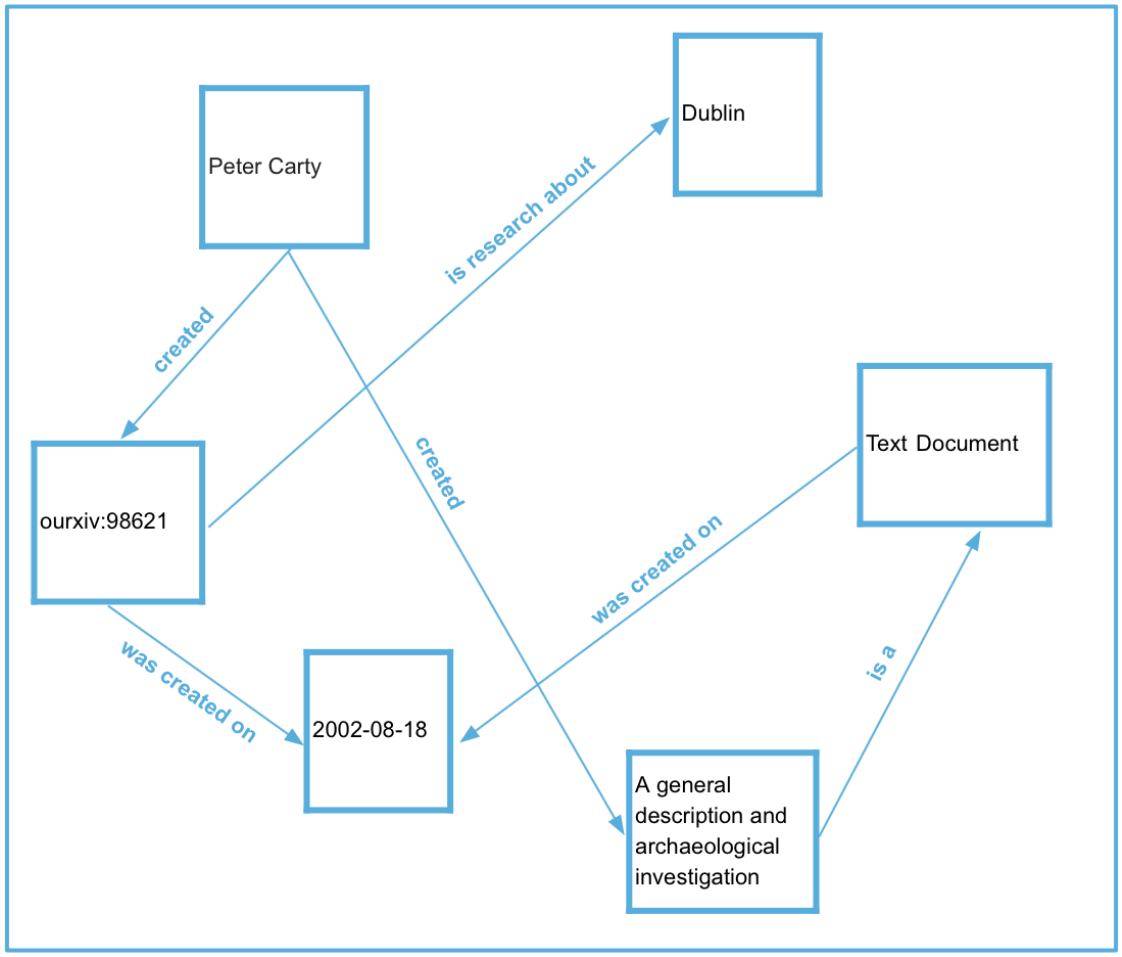


Figure 5

Figure 5: An example of modelling data

Thing 3: Defining - Define the URI (Uniform Resource Identifier) naming strategy

3a: Define a suitable and durable namespace

The URI is an address that a machine can use to find exactly the right piece of information on the World Wide Web. (You are familiar with this idea already - think of the URL of any website). A URI consists of three parts: **a scheme, a domain name and a path.** The domain name plus the path are known together as the namespace. Defining a namespace is extremely important in LOD, as it allows machines (and humans) to tell the difference between identically named elements from multiple datasets. (See W3C Step#5)

Figure 6

Figure 6: Example of the three parts of a URI

You will first need to choose a domain name to use for the URIs you will create. It is important to think about the sustainability of the domain name that you use. URIs should be persistent and not change over time. If you plan on using a domain name that is part of a project, think about how (or if) that website will be maintained after the end of the project. It is always better to choose something that you are sure will stand the test of time (see W3C Step #5), such as institutional domains. Institutional domain names have the added benefit of conferring a sense of authority. That is, a domain name of "*harvard.edu*" suggests more authority than a domain name of *"jane_smith.com*". (If you are unsure about your institutional options, check with your local IT professional for guidance).

The remainder of the URI is the path. You can think of the slashes in the path like folders and sub-folders that are used to organize information in a manner that is understandable to both people and search engines. We have further recommendations for constructing the path in steps 3b and 3c.

Activity: To further prepare for creating your own URIs, read "The Role of Good URIs for Linked Data" from the W3C guidelines.

3b: Consider a versioning strategy that reflects past and future modifications of your LD in the URI Path

Datasets are not static; they are often updated and modified with new versions. We recommend that you include versioning as part of your namespace (in the path) to make it perfectly clear which version of your data you are referring to.

The W3C also recommends using vocabularies that provide versioning control. Vocabularies are the definition of concepts, relations and their mutual order. Vocabularies also change, as they are developed and edited. Using a vocabulary with versioning control ensures that if the vocabulary changes, you point to the _right _version of it. (See W3C Step #6). Concrete observations or instantiations of a vocabulary can also be reclassified. If this happens, you also need to indicate the change at this level.

There is no firm policy on this problem yet. Although in most cases only a small percentage of concepts and relations change between different versions, our proposal is to include the version information in the URI, such as in this example:

http://ourxiv.com/resource/**v1.1**/archaeology/fossil

We believe that this strategy will help your "audience" - those who map to your versioned vocabularies - from doing unnecessary updates. We imagine the following scenario: when anybody uses the URI without the version tag (shown in red in our example above) a "smart" lookup service would return versioning information about the concept 'fossil', it would also return the preferred version, and the versions where there were changes to this concept.

Changes in vocabulary or instantiations, but also any other changes that you make (i.e. mapping and enriching, which we will discuss in later *things*) can be documented or "logged" and described with a commonly shared vocabulary (Moreau & Groth, 2013). This is also called provenance information. Compared to versioning, it is like a meta operation on changes concerning the whole or parts of the graph. The versioning we discuss above concerns what has been changed in the knowledge representation structure itself, but does not focus that much on who did it, when and by which process.

3c: Decide how the concepts and relations are represented by its unique identifiers which are part of the URI.

We also recommend that you construct your URI in a way that it reflects the meaning of the concepts and relations that you identified in Step 2. This will make it much easier for people to interpret the URI and understand the link.

Rather than using a long string of numbers in our example URI, we used the URI to indicate the relationship between the thing that our data describes (a fossil) and the description of that thing.

http://ourxiv.com/resource/v1.1/archaeology**/fossil**

This involves thinking about how to distinguish between objects in the real world and the webpages describing those objects. Use specific patterns to represent properties, individuals, and classes. For example:

http://ourxiv.com**/resource/**v1.1/archaeology/fossil ← Thing

http://ourxiv.com**/data/**v1.1/archaeology/fossil ← RDF data

http://ourxiv.com**/page/**archaeology/fossil ← HTML page

Activity: Using what we have discussed about best practices for creating a URI, draft a few URIs to describe your own data.

Thing 4: Resolving - Consider resolvability when a person or machine 'visits' the URI.

If someone were to put your URI into a browser, what would she get back? A URI is "resolvable" if anyone, regardless of their own domain, can put it into a browser and see a result. Please note, that the example URI's we have constructed so far are not resolvable!

Activity: Take a look at example.org. Is this domain resolvable? Why or why not?

Not every domain is resolvable. The domain in the above activity is not resolvable, but is rather just a placeholder. Remember, you gain authority and trust from other users when your URIs are resolvable and lead to information.

In terms of LOD, it is important that the information that is returned describes the concept in the URI entered in the browser. The information returned could be a snippet of RDF with, for example, information about properties, classes or provenance.

A basic implementation of an RDF URI resolver is the Urisolve server. The Urisolve server takes a URI as input and returns a simple list of triples that all have the URI somewhere in each statement. This implementation assumes that there is an HDT (Header, Dictionary, Triples) or SPARQL endpoint that hosts your RDF data. Virtuoso is a well known open-source RDF datastore that includes a SPARQL endpoint. HDT is a binary format for RDF which has major performance benefits.

Activity: Visit http://www.rdfhdt.org/ to learn more about HDT and supporting tools.

Thing 5: Transforming - Generate the URIs for the selected concepts and relations according to the URI naming strategy.

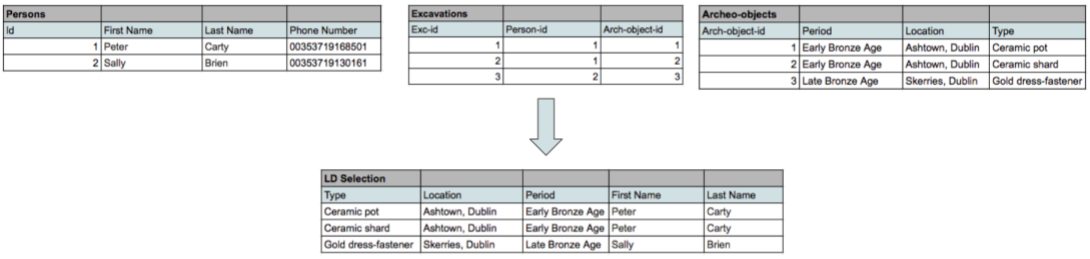
Steps 1-4 are primarily planning steps; in principle, you could actually do them on paper. Step 5 requires software, tools and/or scripts to transform your data into Linked Data.

Your exact approach depends a lot on your particular situation; the format of your data, the size and the available (programming) expertise are the main factors.

The below workflow suits many situations:

1: Filter your data

In Thing 2, we mentioned that it will most likely not make sense to share all of your data. Filtering your data involves creating a new temporary dataset that contains only those concepts and relations that you want to expose as LD. If your data is in a database like PostgreSQL or MySQL, it is often easiest to write a SQL command that generates one new temporal table containing the union of selected columns from the various tables. If your data is in a spreadsheet like Excel, you can create a new sheet via macros and filters. Note that you should try to keep this filtering and generation process as automated as possible and save the macros or SQL for future version conversions.



Persons			
Id	First Name	Last Name	Phone Number
1	Peter	Carty	00353719168501
2	Sally	Brien	00353719130161

Excavations		
Exc-id	Person-id	Arch-object-id
1	1	1
2	1	2
3	2	3

Archeo-objects			
Arch-object-id	Period	Location	Type
1	Early Bronze Age	Ashtown, Dublin	Ceramic pot
2	Early Bronze Age	Ashtown, Dublin	Ceramic shard
3	Late Bronze Age	Skerries, Dublin	Gold dress-fastener

LD Selection				
Type	Location	Period	First Name	Last Name
Ceramic pot	Ashtown, Dublin	Early Bronze Age	Peter	Carty
Ceramic shard	Ashtown, Dublin	Early Bronze Age	Peter	Carty
Gold dress-fastener	Skerries, Dublin	Late Bronze Age	Sally	Brien

Figure 7

Figure 7: An example of filtering data

Activity: Based on your work in Thing 2, create a temporary dataset containing only those concepts and relations that you want to expose as LD.

2: Bridge your prepared data to your tool

There are basically two ways for an RDF generation tool to work with the table from the previous step: 1) set up a connection between the data store and the tool, or 2) serialize the data to a format that the tool can use.

2a: set up a DB connection

Tools like Ontop can connect directly to your database and uses transformation rules to create Linked Data, or even a SPARQL endpoint to your live data.

2b: Serialize your data

Serialization is turning your data from the format you usually interact with to a series of bits. Based on your data format, most tools and databases have the functionality to store tables in CSV format. Please be aware that encoding can be tricky especially with special character sets.

3: Use tools to transform your serialized/connected data into LD

Depending on the previous step, the selected tool directs the way how your data will be transformed into Linked Data.

Thing 6: Mapping - Map your Linked Data from your newly defined namespace to similar concepts and relations within the LOD

Most likely there will be concepts and relations in your fresh LD dataset that are similar to concepts and relations in the LOD. The challenge in this step is to 1) find them, 2) make a selection based on a quality metric and 3) select the schema to express these mappings.

1: Finding related concepts and relations

The 'Linked' aspect of Linked Data is the focus of this point. In this exercise you browse online resources to find vocabularies and schemas that have concepts and relations similar to those you have created.

Activity:

- Explore the following public sources to find Linked Data related to your own: The LOD Laundromat, the Linked Open Vocabularies portal and BARTOC.
- Next, look at the following domain-specific resources: GeoName for locations and Getty AAT for excavational objects like Etruscian Pottery.
- Are there any other domain-specific sources for vocabularies that you know of that could be relevant for your data?

Note that, although preferred otherwise, the external concepts you wish to link to themselves do not need to be designed as Linked Data. For example, a researcher mentioned in your database can have a persistent identifier in ORCID and a publication can have a DOI.

2: Sort and make a selection from the sources found in the previous step

The decision depends on many factors, such as your audience (if, e.g., it needs to be multilingual), the coverage with your own concepts (i.e., exact match is preferred to broad superclasses), the authority of the external source (who developed it and maintains it), etc.

3: Selection of the mapping schema

There are different 'flavours' regarding mapping concepts and relations. The choice is made primarily based on the inferencing and other logical reasoning requirements, as we detail below.

3a: RDF and RDFS

The de-facto linked data format is RDF. In RDF one can specify that an instance is of a certain class, like a *cat* is a type of *animal*. RDFS is an extra logical expressive schema that allows one to bind a property to a domain and range, for example the *employer* relation is between the domain: *person* and range: *organisation*. RDFS provides the means to specify sub-classes, e.g. *student rdfs:subClassOf person*, and sub-properties, e.g. *hasSibling rdfs:subPropertyOf hasRelative*. Unfortunately, neither RDF nor RDFS offer an option to state equality between concepts or relations. For that we have OWL and SKOS which we cover next.

3b: The OWL variants

The W3C-OWL stack (e.g. OWL-Lite, OWL-full OWL-DL) extends RDFS with additional reasoning options grounded in formal logic, which has as an advantage that more automated checking and derivations can be done but is also for many people difficult to learn and adds more computational demands to the reasoning backend. The most popular

owl statement is the property *owl:sameAs* which as expected expresses equality between two instances (e.g. “Bill Clinton” and “William_Jefferson_Clinton”) or classes (e.g. “Area” and “Region”).

3c: SKOS

The popularity of SKOS perhaps lies in the fact that it has **no** formal grounding and people use it to express all kinds containment relations. For example the *skos:broader* property is used to express a subclass relation (*mammal skos:broader animal*), a subregion (*Texas skos:broader USA*), subperiod (*baby-boom-period skos:broader 20thCentury*) etc. Despite the lack of formal grounding, most humans do understand the inherent reasoning and can develop in retrospect applications that properly deal with these mappings.

Having a good idea about which concepts in your own data you want to expose and which concepts are already published as Linked Data helps in the decision making process for selecting a mapping schema. Likewise you can revisit your data model and see if there are better ways to define the relationships between your concepts.

Activity: Study the below figure. On the left is the data model from 2b and on the right we can see some concepts which we have identified as relevant to map to our dataset. Can you identify which concepts on the right would be mapped to which parts of our data model on the left?

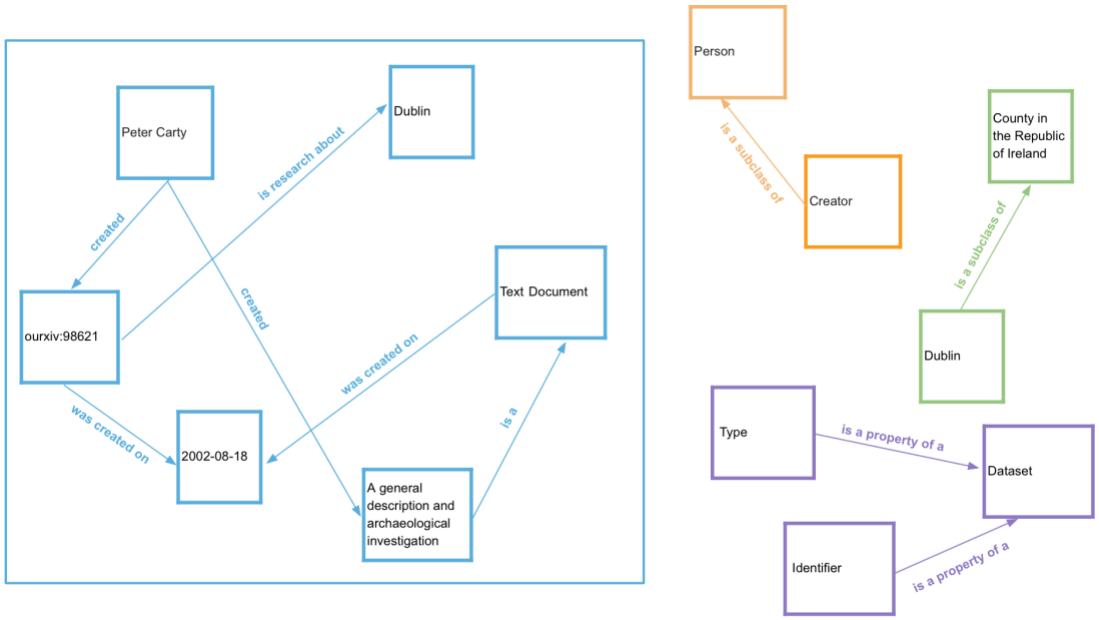


Figure 8

Figure 8: An example of a modelled dataset (right), with some potential external concepts for the data to be linked to (left)

Thing 7: Enriching - Enrich your data with information from the LOD

The enrichment process is very similar to the mapping, with the subtle difference that the goal of mapping is to *connect* your data to existing Linked Data, and enrichment is to *describe* your data with Linked Data. Although not set in stone, the **mapping** process uses a well known set of properties that results in a linkset of similarities in RDFS (e.g subClass), SKOS (e.g. exactMatch) and OWL (e.g. sameAs). The **enrichment** process has a wider scope on both the selection of properties and objects. Key is that the enrichments are relevant for the goal of sharing your data.

Activity: Imagine that you are a producer of chemical compounds. The molecular weight, structure, boiling point, etc. for different compounds may be relevant properties for your data. Take a look at ChEMBL and explore how it could be useful to you.

Similarly, if you work for a library, you can enrich your collection with concepts from library classification systems like LCC, UDC and DDC.

Even using our tiny example (shown again in Figure 9) the power of Linked Data becomes apparent. By linking your dataset it is possible to enrich it with new meaning.

Activity: Take a close look at the below figure where we have now labelled the relationships between the two sides of the diagram. Does this match what you were thinking of in the earlier activity with this diagram? Through enriching our data, we now know that the Dublin in our dataset is Dublin, Ireland and not Dublin, Ohio (where the Dublin Core metadata schema originated).

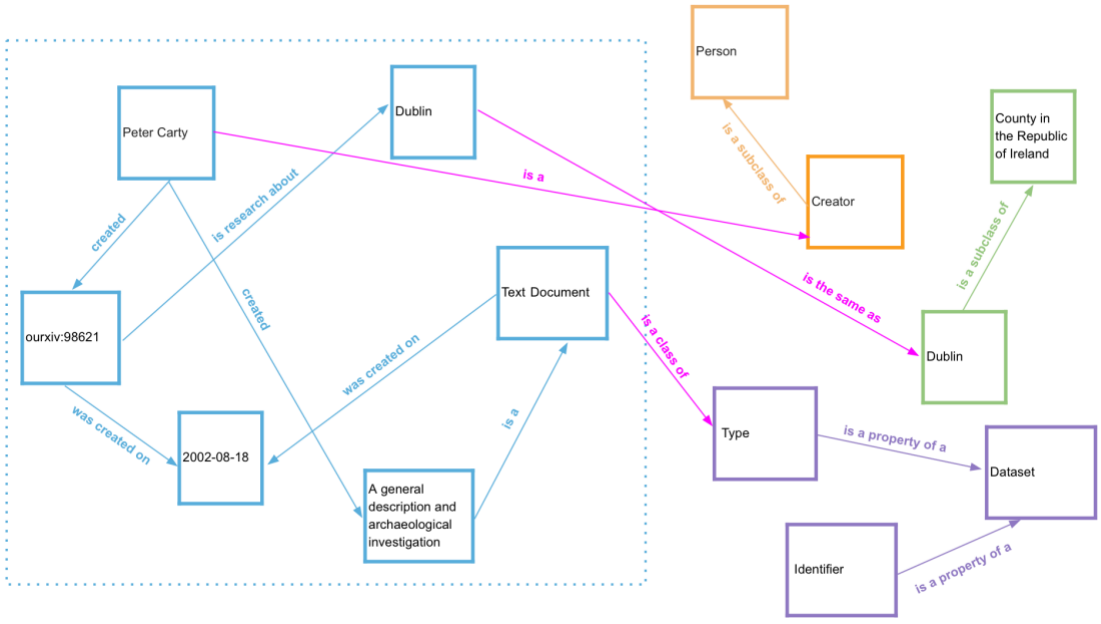


Figure 9

Figure 9: An example of a modelled dataset (right), linked to some external concepts (left)

Thing 8: Exposing - Define how people can get access to your LD: a data-dump, a SPARQL endpoint or a Web API.

After you expose your data, the next step is to think about your intended audience and how people will use and access the data. You will need to consider if and how you are going to expose your Linked Data "graph" as a whole. You have a few options for exposing your data and making it part of the LOD cloud. Access to the entire LD dataset must be possible via either RDF exploration, an RDF dump or a SPARQL endpoint. These options are further described below.

- **Via RDF exploration**: This refers to the ability to manually navigate the graph. It allows you to save the "breadcrumb trail" links from document to document and gather the results for searching.
- **As an RDF (data) dump**: RDF/Turtle is a human friendly serialization format, and one can describe the graph with provenance metadata (e.g. W3C-PROV) and accessibility information (W3C-VOID).
- **As a SPARQL endpoint**: Be careful because SPARQL is not very easy. It requires a background in query languages and one can easily get lost in the graph; the wrong queries can also put a very heavy load on the server. Initiatives like Puelia-PHP, RISIS-SMS and GRLC shield the SPARQL complexity by offering an abstraction layer (e.g. as a RESTful service) or visual components for pre-defined query templates.

_Activity: If you are curious to learn more about how queries are formed, visit the Wikidata Query Service. This service provides user friendly query examples which allow you to see how queries are formed and how the results are presented.

An alternative option is to use Linked Data fragments: Linked Data Fragments is a conceptual framework that provides a uniform view on all possible interfaces to RDF. A Linked Data Fragment (LDF) is characterized by a specific selector (subject URI, SPARQL query, …), metadata (variable names, counts, …), and controls (links or URIs to other fragments).

Thing 9: Promoting - Publish and disseminate the value of your data via visualisations and workflows

Once your data are out in the open, you can continue to link each of your statements (objects, subjects, predicates) to other statements in the LOD cloud (see Thing 7). But, you can also create other services on top of your data to tell the world how your data are equal, similar or different to other existing data

Activity: Visit https://www.cedar-project.nl/about/ to learn more about how linked open data are being used in the CEDAR project. Then, take a look at the below map, which Ashkan Ashkpour and Albert Meroño Peñuela created as a part of this project. To make the map, they combined linked data from the Dutch census with openly available geographic data to bring their research to life.

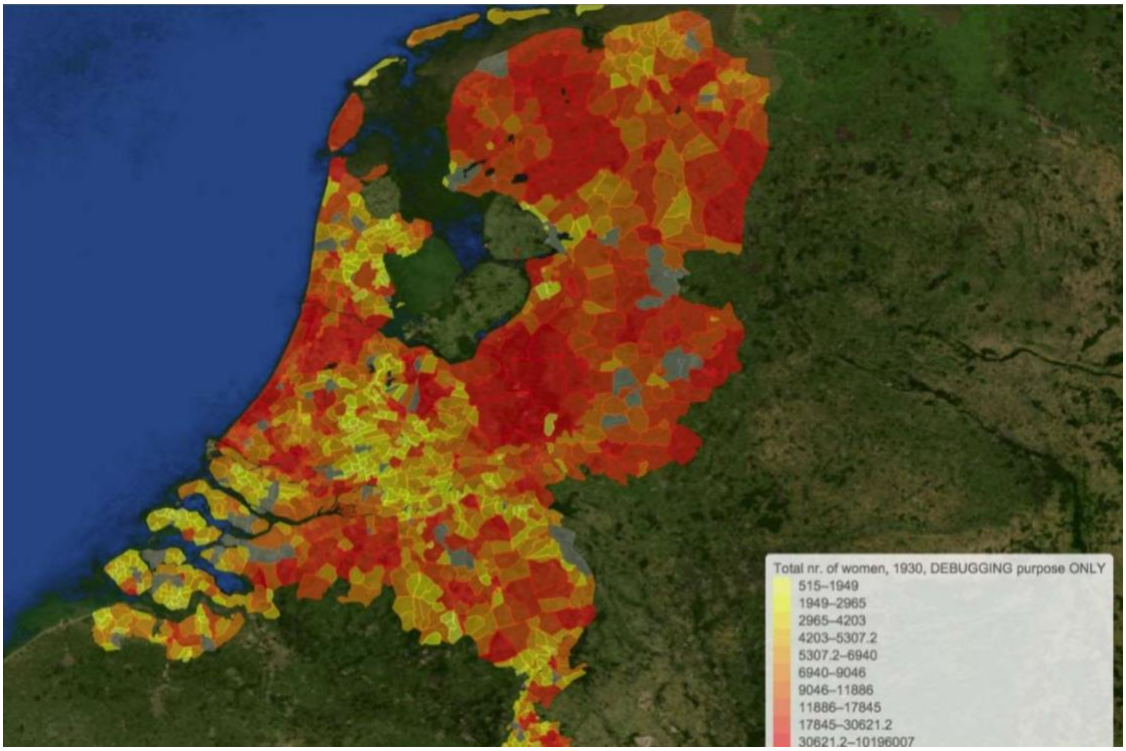


Figure 10

Figure 10: A map showing a combination of Linked Data from the Dutch census with geographic data - the heatmap shows the total number of female inhabitants

Thing 10: Sustaining - Ensure sustainability of your data

Publishing Linked Data into the LOD cloud is one specific instance of dealing with data on the web. The W3C recommendation "Data on the Web Best Practices" provides further pointers and considerations for many of the issues that we have raised here, such as the persistence of URI's, version policy, or the reuse of vocabularies.

Many of the best practices listed in the W3C recommendation touch upon the importance of ensuring the sustainability of data publications in the immediate, mid- and long-term. These are also important for you to consider when publishing your data to the LOD cloud.

For example, it is important to associate a clear and preferably well-known standard license with your data and to present it clearly to the audience. You should also indicate if you maintain the right to change the license in the future. Standard content licences such as Creative Commons can be used for this purpose; licence information should be included in the served content. (See W3C Step #4)

Archiving a version of your RDF dataset as a static data dump in a certified, long-term stable data repository might also be a good option to help ensure the long-term sustainability of your data. This provides a way for you to preserve and potentially reuse all of the work that you have already invested. (see for an example Beek et al. 2016)

Activity: Examples of how to archive your RDF dataset can be found in the DANS EASY_1 data archive. Explore the below examples, paying particular attention to the associated readme-file instructions.

- The deposit of an RDF dataset from the CEDAR project
- The deposit of the Laundromat dataset

Publishing data as Linked Open Data is new for many researchers. Hopefully the steps, recommendations and references that we have presented here will help you to begin your own journey into the realm of Linked Open Data.

References

Beek, MSc W.G.J. (VU University Amsterdam); Rietveld, MSc L. (VU University Amsterdam); Schlobach, Dr. S. (VU University Amsterdam) (2016): LOD Laundromat (archival package 2016/06). DANS. https://doi.org/10.17026/dans-znh-bcg3

Dextre Clarke, S. Entry 'Thesaurus (for information retrieval). *Encyclopedia of Knowledge Organisation*. Available at: https://www.isko.org/cyclo/thesaurus.htm

Hyvönen, E. (2012). Publishing and using cultural heritage linked data on the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *2*(1), 1-159. Available at: https://doi.org/10.2200/S00452ED1V01Y201210WBE003

Hyvönen, E. (2019). Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Knowledge Discovery. submitted to *Semantic Web Journal* (under review). See: http://semantic-web-journal.net/system/files/swj2214.pdf;

Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure,L., Scharnhorst, A., Schlobach, S., van Harmelen, F. (2019). *Semantic Technologies for Historical Research: A Survey*. submitted to *Semantic Web Journal* (under review). See: http://www.semantic-web-journal.net/content/semantic-technologies-historical-research-survey

Moreau, L., & Groth, P. (2013). Provenance: an introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*, *3*(4), 1-129. https://doi.org/10.2200/S00528ED1V01Y201308WBE007

1 Crystallographic Information Framework (CIF) "A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals. Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration." More information, tools and use cases are available here: https://rd-alliance.github.io/metadata-directory/standards/cif-crystallographic-information-framework.html

Acknowledgements

We are grateful for pointers from and discussion with the DANS Research group, in particular from Herbert van de Sompel; and also for the valuable contributions of Esther Plomp, Marjan Grootveld and Enrico Daga. This document has been informed by the 'GO FAIR Implementation Network Manifesto: Cross-Domain Interoperability of Heterogeneous Research Data (Go Inter)' (Ed. by Peter Mutschke) https://www.go-fair.org/implementation-networks/overview/go-inter/ . The grants "Digging into the Knowledge Graph" and "Re-search: Contextual search for scientific research data" (NWO) have enabled part of this work.

Astronomy

Sprinters:

- Maria Cruz https://orcid.org/0000-0001-9111-182X
- Chris Erdmann https://orcid.org/0000-0003-2554-180X
- Kristina Hettne https://orcid.org/0000-0002-4182-7560
- Francoise Genova https://orcid.org/0000-0002-6318-5028
- Matthew Kenworthy https://orcid.org/0000-0002-7064-8270
- Natalie Meyers https://orcid.org/0000-0001-6441-6716
- Evert Rol https://orcid.org/0000-0001-8357-4453
- Juande Santander-Vela https://orcid.org/0000-0002-2660-510X
- Joanne Yeomans https://orcid.org/0000-0002-0738-7661

Brief description

Much has already been done in the area of Astronomy with regards to data management and software management. For example, the International Virtual Observatory Alliance (IVOA) is the collegiate organisation that supports the Virtual Observatory as a federation of astronomical archives and datasets through the standardisation of data formats and data access protocols, and also underlying data models. The goal for the Virtual Observatory is that astronomical datasets and other resources should work together in a seamless whole. The IVOA develops and agrees on the standards underlying the Virtual Observatory.

This document intends to connect a few of these efforts to the **FAIR** principles - **F**indable, **A**ccessible, **I**nteroperable and **R**eusable - and introduce general activities that can be undertaken towards "FAIRness". They can be used for researchers to become more FAIR aware, or by librarians or other research data support staff as an inspiration for training. The document is structured into informative small pieces of text (so-called "Things") to jump-start activities a researcher can do to make their data and software more FAIR. They do not have to be followed in a particular order, you can just pick and choose. We have sorted the "Things" under the respective FAIR category they belong to.

A related previous effort by the Australian National Data Service is the 10 Astronomy Things.

Audience

Early Career researchers, Research Data support staff

Goals

The goals of this document are to raise awareness and give practical advice and exercises as a starting point towards "FAIRness". The aim is not to be comprehensive, but to provide a teaser for those wanting to know more.

Findable

Thing 1: Finding and sharing data and software

Relates to F4: (Meta)data are registered or indexed in a searchable resource

Astronomy has a long tradition of sharing and reusing data. Astronomical data can be found and accessed in many different forms and ways. Raw and processed data can be accessed from major observatory archives (e.g. the ESO Archive). There are also collections of astronomical catalogues (e.g. VizieR) providing tables and associated data published in academic journals, and databases (e.g. SIMBAD Astronomical Database), providing information on astronomical objects of interest. The Astrophysics Source Code Library (ASCL) is a free online registry for source codes of interest to astronomers and astrophysicists and lists codes that have been used in research and that have appeared in, or have been submitted to, peer-reviewed publications.

Activity 1: Go to there3data.org registry of research data repositories and search for astronomical data or software repositories that may be relevant to your area of research. Try it out by typing "astronomy" into the search box. Compare the query form with the web interface of the IVOA registry of resources. The IVOA registry of resources is mostly to enable discovery of and access to data and services available through the Virtual Observatory by programmes.

Activity 2: Learn how to search for data through filtering your search in the NASA Astrophysics Data System (ADS). Find out how you can navigate to data products for records/papers via the data/database icon in the ADS.

Activity 3: Consider attending a dotastronomy.com conference or an AstroHackWeek and learn about tools for collaboration and sharing. If not feasible, try perusing the conference videos ofdotastronomy.com, or those from the AstroHackWeek 2015 or 2016.

Thing 2: Metadata

Relates to F2: Data are described with rich metadata

Metadata is "data about your data". It provides context about the actual data, makes the data more findable and categorizable, provides some idea about the quality of data, etc. (see Wikipedia summary). For example, the license for the data is also metadata. Metadata often exists in headers of the actual data files. Additionally, metadata exists in the database of, for example, observatories or journals. If metadata is recorded in the file itself, it is still preserved even if the other metadata resources are not available.

The types of metadata will be dependent on the field, data format, used software tools, and possible related publications (or publishers). Separate them as necessary, and include relevant standards when possible (for example, in a data header, the metadata standard may be mentioned first, then a list of metadata), so that the meaning of keywords can be looked up. Avoid using filenames to record metadata: these can accidentally change, meaning that metadata will be lost.

The FITS format is widely used in astronomy, in particular for observational data. A FITS file contains data and the relevant metadata that describes, at a minimum, the instrument used and the observation parameters, such as its sky coordinates. Another example of how metadata are organised in astronomy is the standard description of a "catalogue" (a table or a set of tables) in VizieR, stored in the ReadMe file, which contains information about the dataset and for each table a description of the quantity contained in each column.

Specific astronomy-related types of metadata: * IVOA standards * FITS keyword dictionaries * World Coordinate System * Observatory, telescope & instrument information & configuration * Astrophysical simulations * Astronomy Visualisation Metadata

Activity 1: Verify that your current data contains at least the items listed in the Wikipedia section.

Activity 2: Examine the ReadMe file of a VizieR "catalogue", for instance J/A+A/620/A89. The dataset is a table from a paper published in *Astronomy and Astrophysics* (you can have a look at the paper from the catalogue page). Could you reuse a similar template to describe your own data, even if you are not an astronomer? (See Thing 5, Activity3 if you want a description of the template)

Activity 3: Sign up for the IVOA newsletter to stay informed about recent developments

Thing 3: Persistent identifiers

Relates to F1: (Meta)data are assigned globally unique and persistent identifiers

Persistent identifiers prevent "link rot", the process by which web links stop referring to the original sources over time, rendering these sources unavailable. If you store your data or software on your own or your institute's website you are likely to fall foul of this problem. Assigning a globally unique, persistent identifier (such as a DOI) to your data, software, and their metadata is one of the first and most important steps towards FAIR, as it will make your research outputs findable and accessible to both humans and machines in the long term. It will also make it easier to cite your research outputs as independent citable objects and to link them to related publications.

Depositing your data and/or code in a repository that assigns a persistent identifier (e.g. Zenodo) is the easiest way to get one. Disciplinary repositories more and more provide persistent identifiers. Persistent identifiers usually come packaged with metadata — often following metadata standards, e.g. the DataCite's Metadata Schema. And these metadata are usually registered or indexed in searchable resources, such as the DataCite.

Activity 1: Some major observatories, such as ESO and MAST, are starting to move to using DataCite DOIs. Read more about some of these initiatives to get familiar with the concept of assigning DOIs to datasets in the context of astronomy and astrophysics: The ESO Digital Object Identifier Service and About DOIs at MAST.

Activity 2: Identifiers exist for researchers as well. ORCID is a persistent digital identifier for researchers that helps connect research to researchers. It helps to distinguish you from every other contributor, particularly useful if you have a common name, and it supports

automatic links between your various research outputs and activities. Go to the ORCID website and create an ORCID profile if you do not have one already. Then learn how to claim your papers using the NASA ADS.

Accessible

Thing 4: Access regulation

Relates to A1.2: The protocol allows for an authentication and authorisation when necessary

It is crucial to know that the "Accessible" requirement in FAIR does not have to mean "open" or "free". Or in other words, it can be "as open as possible, as closed as necessary." According to the GO FAIR website: *"Rather, it implies that one should provide the exact conditions under which the data are accessible. Hence, even heavily protected and private data can be FAIR."* The actual protocol for accessibility is often taken care of by the repositories and infrastructures that offer data storage and is less of a concern for individual researchers. However, as a researcher, you should think about *who* can access the actual data and how, so that when you deposit your data you can pick the right level of access.

Embargoes in astronomy are one way that data are not necessarily open from the start. Embargoes provide researchers who got observation time on a telescope in a competitive process with a proprietary period of time to work with the data they have obtained before it is opened to the general research community and public. See the ESO Archive Frequently Asked Questions: Getting Data: How is the proprietary period of data regulated?

Activity: Read the European Southern Observatory data access policy. When do data become publicly available?

Interoperable

Thing 5: Data structuring and organization

Relates to I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

Ensure the data is in an accessible format; standard file formats such as CSV, JSON, HDF5, FITS (astronomy only) work. Many tools exist that can read such files, even for data online. This makes the data Findable, potentially even by search engines, as well as Interoperable with (other) software. On the GO FAIR home page the following data formats that satisfy the requirements for Linked Data are mentioned: Resource Description Framework (RDF) and JavaScript Object Notation for Linked Data (JSON LD).

Ensure the data are logically structured: this may be within the data itself (multi-extension FITS files; HDF5 is inherently structured), or through the use of directories. This makes it easier (for humans, mainly) to find and use (parts of) the data. Be careful not to just rely on filenames, since these are (too) easy to change (see also Thing 2 about metadata).

Units should not be forgotten. Often, units will be indicated through the metadata, and when done according to relevant standards, tools and libraries should automatically use them. For FITS, section 4.3 of the FITS standard specifies the use of units. In HDF5, units are often defined using a separate (one-dimensional) dataset. For representation of units, see also the IVOA units recommendation.

Note: data published in a table in an article is *also* data. Publish the table data separately in a standard file format (e.g., CSV); don't rely on, say, just a LaTeX table, even if journals, ADS, or CDS are able to scrape and read this. Never forget that you should provide enough metadata with the file so that it is reusable and its origin is known (See Thing 2).

Activity 1: Have a critical look at your own data files: are they well structured and in an interoperable format?

Activity 2: Make a list of the different data formats that you are using, and try to see if they correspond to well (preferably online) documented formats, and collect those links.

Activity 3: Look into your data/data tables. Do you have everything ready for them to be published? You can use a checklist such as this one from VizieR table submission.

Thing 6: Terminology

Relates to I1: (Meta)data uses vocabularies that follow the FAIR principles

The Unified Astronomy Thesaurus (UAT), a reference work that lists words grouped together according to similarity of meaning, builds on prior controlled vocabularies used in astronomy. The aim of the UAT is to create a high quality open, interoperable and community-supported thesaurus that is freely-available and formalizes astronomical concepts and their inter-relationships. The development and maintenance of the UAT is stewarded by a broad group of parties having a direct stake in it, including professional associations (IVOA, IAU), learned societies (AAS, RAS), publishers (IOP, AIP), librarians and other curators working for major astronomy institutes and data archives. The goal of the UAT is to ultimately improve the discovery of astronomy research including papers, software, data products and services.

Activity 1: Explore the UAT and/or select concepts that you can use to describe your papers, software, and/or data.

Activity 2: Learn more about the UAT by reading Katie Frey and Alberto Accomazzi (2018): The Unified Astronomy Thesaurus: Semantic Metadata for Astronomy and Astrophysics.

Thing 7: FAIR data modelling

Relates to I1: (Meta)data use formal, accessible, shared, and broadly applicable language for knowledge representation, R1: (Meta)data are described with a plurality of accurate and relevant attributes and R1.3: (Meta)data meet domain-relevant community standards

Data modelling is the process by which terminology and a framework for the terminology comes together. From Juande Santander-Vela (2009): *"Data models are the detailed description of the set of entities needed for information storage in a particular field, and specify both the data being stored, and the relationships between them. Data models are part of the hidden VO infrastructure astronomers would normally never be involved with, but knowledge of data models can enhance the opportunities for the exploitation of the VO."*

The Virtual Observatory has two standards that have to do with data modelling: Universal Content Descriptors (UCDs) and UTypes. Universal Content Descriptors describe in general terms, by means of a restricted but expandable vocabulary, which kind of astronomical concept a particular value (and unit) corresponds to. For instance, a column in a table can have the UCD `pos.eq.ra`, indicating that a value corresponds to a right ascension in equatorial coordinates (one of the sky coordinate systems). If each row contains the position of multiple objects (for instance, of the Sun, the moon, and the target object), additional UCDs can be composed, so that the more precise `pos.eq.ra; meta.main` can be used for the observed source's Right Ascension, and the rest of the right ascensions are codified with just `pos.eq.ra`. The latest list of the UCD vocabulary is provided by Andrea Preite Martinez et al. (2018). The CDS has a page with lots of useful tools for UCDs, such as those to decode UCDs, or to generate UCDs from natural language descriptions, among other things.

UTypes are a feature of the VOTable —see section 4.6 of Ochsenbein et al. (2013)— which can indicate exactly to which piece of a given Data Model (expressed as an XML Schema) a particular metadata item refers to. This is much more technical than the UCD, but it can also be very useful in order to generate machine-readable astronomical datasets.

For **FAIR** data modelling it is important that the terms have resolvable, globally unique and persistent identifiers and that the framework is well-defined. See also Things 2, 5 and 6.

Activity 1: Check out this presentation on data modelling within the IVOA by M. Louys, which describes the work of the IVOA Data Model Working group. Why is interoperability important for VO users?

Activity 2: Go to the page of CDS' useful tools for UCDs and try to use it to build UCDs for different concepts, such as photometric magnitude in the *i* band, or the difference in *g-i* colours. Can you find concepts that are not easily expressed as UCDs? Can you imagine a similar tool in fields outside of astronomy?

Activity 3 On the GO FAIR home page the following formats are mentioned: Resource Description Framework (RDF), W3C Web Ontology Language (OWL), DARPA Agent Markup Language (DAML+OIL), JavaScript Object Notation for Linked Data (JSON LD). Note that RDF and JSON LD can be used to model the data as well as the metadata. Have a look at this document on Vocabularies in the Virtual Observatory. Which format is recommended by the IVOA to use for modelling vocabularies?

Reusable

Thing 8: Licensing

Relates to R1.1: (Meta)data are released with a clear and accessible data usage license

If you don't apply a licence to your data telling people up front how they can use it, then you may have to deal with emails from people asking this question in the years to come. And if a future researcher doesn't know how they can use your data, then they may not want to use it at all. This is especially important when it comes to software as others may wish to adapt and make derivatives of the original code which they legally cannot do without permission. For this reason it's useful to apply a licence to improve the re-usability of your data and software.

Licences can be hand-written by anyone (although this is not recommended!) and you will commonly find such 'licences' on many existing websites for astronomy software and data. They are often not called a licence, and are often not complete from a legal standpoint. It may be, however, that by contributing to certain software you are agreeing to make it available without restriction, which is by implication, an open source or "all rights waived" licence. Astronomy has been sharing data openly, eventually after an embargo period, for decades, and even before the foundation of Creative Commons in 2001. In addition to the 'hand-written licences' provided in some cases, "data provided by others must be cited appropriately, even if obtained from a public database", as stated in the Ethics Statement of the American Astronomical Society. The absence of formal licence practices has not been an obstacle to widespread and mostly fair usage of shared data in the discipline, but the astronomy case cannot be generalised.

The lack of clarity is why it's advisable to use a pre-defined licence like those offered by Open Source Initiative or Creative Commons (CC). These are often written by legal experts and allow you to choose the terms that best match the conditions you wish to apply.

Activity 1: Take a look at this comprehensive introduction to licensing research data from the Digital Curation Centre. How can you add a licence to your data?

Activity 2: If you are developing software, read the blogpost on "The Whys and Hows of Licensing Scientific Code".

Activity 3: Journals are increasingly encouraging authors to archive the data behind their publications but the uptake amongst authors is slow. Licences are therefore something you may not realise you'll need to think about. Take a look at some examples to see how others are doing it:

- License file associated with the code at https://github.com/omsharansalafia/radiogw17
- Data and analysis script in Zenodo with licence indicated in the Zenodo metadata: https://doi.org/10.5281/zenodo.1320054
- Open Source Initiative range of possible licences for software: https://opensource.org/licenses

Activity 4: Bearing your own software and data in mind, which licence might you choose: https://choosealicense.com (mainly software) or https://creativecommons.org/choose/ (data)?

Thing 9: Data and software citation

Relates to R1.2: (Meta)data are associated with detailed provenance

A number of authors have found a citation advantage related to sharing astrophysical data. See Sharing data increases citations and Linking to Data: Effect on Citation Rates in Astronomy. This is also the case outside of astronomy, for instance, see Data reuse and the open data citation advantage. While data citation and linking has been more prevalent, a recent project called Asclepias aims to improve software citation across astronomy. The Astrophysics Source Code Library is a good place to start to register your source code and to preserve and publish the code (via a DOI) using GitHub to Zenodo. See also the Top 10 FAIR Data & Software Things for Research Software and Matthew Kenworthy's guide on how to submit your astronomy software and data reduction scripts for a more comprehensive look.

The major journals in astronomy include information on data & software citation, for example, see AAS Journal Reference Instructions and the AAS Policy Statement on Software. For citing and linking to data and software post publication, the NASA ADS has a section on

Data and Curation Frequently Asked Questions where abstract submissions or corrections can be made.

One step in the process of linking papers to data and software is leveraging the "Related Identifiers" feature when depositing data and software. The Zenodo Sandbox is a helpful space to test this feature without depositing data or software permanently.

Additionally, your institution might have a policy and instructions for citing data and software similar to the Data Access Policy for ESO Data held in the ESO Science Archive Facility. For example, this policy allows ESO to curate and add links post publication via the NASA ADS and their Telescope Bibliography repository.

Activity 1: Examine how publications and data are linked in the ADS. For instance, find the link to telescope archives (XMM, HEASARC, CHANDRA) and the SIMBAD database for the paper Advan et al., 2019.

Activity 2: Learn how you can deposit, cite, and link your own data and/or software by exploring the Zenodo Sandbox.

Thing 10: Data management plans

This Thing is not related to a specific principle, but is the first step in the process to make data FAIR.

The purpose of a data management plan (DMP) is to document the management of data during a research project, from the time of data collection to their entry into permanent archives. The goal of a DMP is to consider the different aspects of data management including metadata generation, data preservation, and analysis so that the data are well managed from the start. Some of the main components for a DMP include:

- The types of data
- The standards to be used for data and metadata format and content
- Policies for access and sharing
- Policies and provisions for re-use
- Plans for archiving data

Many of the main funding agencies require DMPs as part of the application process and/or provide guidance. For instance, see Directorate of Mathematical and Physical Sciences Division of Astronomical Sciences (AST) Advice to PIs on Data Management Plans.

There are tools available to assist researchers with developing DMPs such as the DMPTool and DMPonline. Using tools like these are a good first step towards making your DMP more FAIR, but there is ongoing work to improve DMPs so that they are machine actionable. For instance, see Ten principles for machine-actionable data management plans and Ten Simple Rules for Creating a Good Data Management Plan.

Activity 1: Review the Data Management and Access Plan from the University of Florida Research Computing and determine if there are steps you can take to make this DMP more FAIR (based on the Ten Principles/Simple Rules).

Activity 2: Using a tool such as the DMPTool or DMPonline, try to identify the steps that would benefit from having the process documented in a FAIR way.

Nanotechnology

Sprinters:

- Elli Papadopoulou (Athena Research Center / OpenAIRE)
  https://orcid.org/0000-0002-0893-8509
- Emma Lazzeri (National Research Council of Italy / OpenAIRE)
  https://orcid.org/0000-0003-0506-046X
- Iryna Kuchma (EIFL / OpenAIRE)
  https://orcid.org/0000-0002-2064-3439
- Dr Leonidas Mouchliadis (FORTH-IESL, Laser and Applications Division)
  https://orcid.org/0000-0001-7255-9397
- Katerina Lenaki (Greek Ministry of Education, open science enthusiast)
  https://orcid.org/0000-0001-7984-8888
- Spyros Zoupanos (EPFL)
  https://orcid.org/0000-0002-6069-5241
- Danail Hristozov (Greendecision / GRACIOUS)
  https://orcid.org/0000-0002-2386-7366

Contributors:

- Stella Stoycheva (Yordas Group / GRACIOUS)
  https://orcid.org/0000-0001-8025-565X
- Ellen Leenarts (DANS / OpenAIRE)
  https://orcid.org/0000-0002-5589-756X

Description:

This brief guide is based on FAIR principles (findability, accessibility, interopera_bility, reusability) and describes data management in the field of nanotechnology. It consists of ten steps/chapters with information on data discovery and publication, practices and resources as well as activity ideas.

Audience:

It is addressed to nanotechnology students, researchers, librarians and research support staff.

Acknowledgments:

OpenAIRE Research Data Management Task Force

Things

Thing 1 - Nanotechnology: overview and current trends

Nanotechnology is part of the broader scientific discipline of material science and, particularly, it is the area of research in science, engineering and technology which deals

with the manipulation and manufacturing of nano-dimensional materials at a scale of 10-9m. Nanotechnology applies in diverse disciplines, such as but not limited to physics, chemistry, engineering, energy, medicine, space, agriculture, information, communication etc.

"Some areas are more mature than others. For example, the entire semiconductor industry is now based on nanotechnology. Transistors with critical dimensions of 30nm are being built today and put together into circuits composed of over one billion devices, on one single chip, roughly the size of a thumbnail. In medicine, nanotechnology has led to the development of drug delivery vehicles and diagnostic devices for the detection and treatment of cancer. It is used in tissue engineering to repair damaged tissue and organs. There are advanced uses of nanotechnology in areas of storage, conversion, and renewal of energy – from LEDs to fuel cells to solar cells. Most high-tech information and communication devices use nanoscale production processes. Nanotechnology is also found in everyday consumer goods, such as stain-resistant fibers for clothes, tennis balls, running shoes, cosmetics, and numerous other day-to-day products" (Source: https://nanohub.org/about/nano)

Depending on how nanomaterials are manufactured, they bear different properties, however the main attribute of nanoscale materials, structures, devices and systems is electricity production and binary information production, storage and transmission. Nanotechnology materials can be used in mobile applications and the emerging flexible electronics, in detectors/sensors of security systems and for health monitoring, in everyday objects to make them resilient and durable (e.g. baseball bats, tennis rackets) etc.

Current research trends include the exploration of two dimensional (2D) materials' surface capabilities, like graphene for use in the areas of flexible electronics and valleytronics. 2D materials are atomically thins, exhibit remarkable surface properties and capabilities and their output vary when compiled by different angles.

In addition to the processes of materials manipulation and manufacturing, there is research around nanotechnology which concerns complementary aspects such as risk assessment and governance, physicochemical characterization, (eco)toxicity testing, exposure, life-cycle impact assessment and decision support for sustainability of nanotechnologies among others. In fact, the European project GRACIOUS aims to provide the means to more efficiently assess risk and obtain safety information for the diverse in size, morphology and surface characteristics nanomaterials/nanoforms to ultimately develop a grouping framework.

Activity 1 - Discussion

* Which is the scientific field/ discipline of primary focus of your nanotechnology research endeavours? What is the application of the outputs of your nanotechnology research? * Are you manipulating or manufacturing nanoscale materials, structures, devices, systems or both? What are the nanomaterials' properties that you most commonly come across?

Thing 2 - Workflow and Methods

Nanotechnology research is based on hypotheses (theory) and experiments. Hypotheses explore ways of producing structures that are 2D and simulations confirm or deny these hypotheses[2]. Theoretical and experimental researchers work together to produce new materials and their communication is bidirectional:

- Experimentals might notice properties that are different from the expected outcome after running the simulation
- Theoreticians can propose alternative ways of proceeding with the hypothesis.

There are different methods for the sample preparation and characterization. The former mainly focus on the thin-film deposition of single crystals and include, but are not limited to, Molecular-Beam Epitaxy (MBE) and Chemical Vapor Deposition (CVD). The latter include a variety of characterization methods such as Raman and TEM spectroscopy, atomic force microscopy (AFM) and nonlinear microscopy (second harmonic generation (SHG) and two-photon photoluminescence (2p-PL)).

In addition there are free access packages for the calculation of the electronic, optical, mechanical and thermal properties of 2D materials, based on first principle calculations. The most widely used packages are Quantum espresso, VASP, Yambo and Wien2K.IATA (Integrated Testing and Assessment Strategy) usually defines which methods and hypotheses could be used for the given purpose.

Activity 1 - Discussion
Based on your field of application and your experience, what kind of workflows seem to work best? Why?

Thing 3 - Data types, outputs and formats

As already addressed, nanotechnology is a multidisciplinary scientific field by nature. It can be applied to many disciplines and thus, the types of data produced by nanotechnology research can be from simulations to chemical data and more. Some data types are, but are not limited to, the following:

1. chemical data on intrinsic properties you measure the properties: aspect ratio, chemical composition, size, surface area, surface properties, coating
2. chemical data on extrinsic properties : particle interacting within the environment
3. crystallographic data about the real space positions of the atoms in 2D crystals such as graphene, transition metal dichalcogenides, hexagonal boron nitride and black phosphorous.

[2] Check also p.44 “B. Life Cycle of Nanomaterials” in *CODATA-VAMAS Working Group On the Description of Nanomaterials, ., & Rumble, J. (2016, June 30). Uniform Description System for Materials on the Nanoscale, Version 2.0. Zenodo. http://doi.org/10.5281/zenodo.56720*

4. data related to stratified structures comprising the same (homostructures) or different (heterostructures) 2D materials[L1] , such as the number of layers and the relative orientation (twist angle, poire patterns)
5. output data of the density functional theory, e.g. band structure, direct and indirect band gaps, excitonic resonances

Open and standardised file formats are essential for accessing data by providing freely available specification documents necessary to open and read their corpus. Most common file formats used in Nanotechnology are CIF[3], UPF[4].

Activity 1 - Defining your data discussion
* Where does your data come from? What types of data do you produce or consume? Can you find it in the list above? What formats are your data in? How often do you get new data? How much data do you generate? * Where would these data be useful?

Activity 2 - Other uses for your data
Would your data be of use in any other research?

Thing 4 - Describing data: Metadata

Metadata is data about data and is an essential set of information describing scientific outputs, in the form of either physical or digital objects, in a machine-readable format. According to the expected use, metadata can be given different attributes. Most common type which enables discovery and identification are *descriptive metadata.* Descriptive metadata contain information about key aspects needed to search for and successfully find a given scientific output, e.g. by its title, author/creator, abstract, keywords. Moreover, metadata may be used for describing a service or a scientific instrument.

Depending on the area of focus, in nanotechnology there are few metadata standards, vocabularies and ontologies to facilitate standardised interpretation.

[3] Crystallographic Information Framework (CIF) "A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals. Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under consideration." More information, tools and use cases are available here: https://rd-alliance.github.io/metadata-directory/standards/cif-crystallographic-information-framework.html

[4] UPF stands for "Unified Pseudopotential Format" and it is used to describe pseudopotentials (initial values to facilitate the simulations performed on e.g. crystal structures that can be described in a e.g. CIF file). UPF is widely used by the nanotechnology community. Recent developments on the format structure show that there are efforts in converting UPF to XML: http://www.quantum-espresso.org/pseudopotentials/unified-pseudopotential-format

- NeXus

"NeXus is an international standard for the storage and exchange of neutron, x-ray, and muon experiment data. The structure of NeXus files is extremely flexible, allowing the storage of both simple data sets, such as a single data array and its axes, and highly complex data and their associated metadata, such as measurements on a multi-component instrument or numerical simulations. NeXus is built on top of the container format HDF5, and adds domain-specific rules for organizing data within HDF5 files in addition to a dictionary of well-defined domain-specific field names."[5]

- ISA-TAB-Nano

"ISA-TAB-Nano specifies the format for representing and sharing information about nanomaterials, small molecules and biological specimens along with their assay characterization data (including metadata, and summary data) using spreadsheet or TAB-delimited files."[6]

- NanoParticle Ontology

"NanoParticle Ontology "represents the basic knowledge of physical, chemical and functional characteristics of nanotechnology as used in cancer diagnosis and therapy."[7]

Apart from metadata records describing datasets, documentation is equally essential when writing code. As the minimum example, a README file helps ensure that your data can be correctly interpreted and reanalysed by others. A README plain text file should contain the following information:

For each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication; for tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units; any data processing steps, especially if not described in the publication, or provenance file, that may affect interpretation of results; a description of what associated datasets are stored elsewhere, if applicable; whom to contact with questions, read morehttps://datadryad.org/pages/readme.

Activity 1
Go to the Research Data Alliance Metadata Directory and search for a metadata standard relevant to the focus of your nanotechnology work. Is there any relevant to your focus?

Activity 2
Have a look at the metadata record for the 7T Magnetom instrument at Research Data

5 https://rd-alliance.github.io/metadata-directory/standards/

6 https://doi.org/10.25504/FAIRsharing.njqq5b

7 https://doi.org/10.25504/FAIRsharing.vy0p71

Australia. It contains simple but important public metadata and a persistent identifier. How could this record be enhanced to assist you as a researcher?

Activity 3
Go to OpenAIRE and search with a keyword about your research interests. What are the nanotechnology projects you find? How many publications, data or software did you get? What are some of the projects associated with these outputs?

Thing 5 - Identifiers

Persistent and/or Permanent Identifiers (PIDs) uniquely identify objects, people, organisations and activities and can ensure that the scientific output is accessible even when the URL of the website has changed. PIDs can be assigned to research outputs including publications, data and software/code. PIDs can also be assigned to researchers, samples, organisations and projects. A PID may be connected to a metadata record describing an item rather than the item itself.

The repository used for data or software deposit should make use of a PID service and assign PIDs to its outputs, also in compliance with the FAIR principles. PIDs can be resolved by the use of tools, such as the DOI Resolver.

PIDs used in research include:

- (Digital Object Identifier](http://www.doi.org/) (DOI)
- Persistent Uniform Resource Locator(PURL)
- Uniform Resource Name (URN)
- Archival Resource Key (ARK)
- Handle
- Research Activity Identifier (RAID).
- Open Research and Contributor Identifier (ORCID)
- International Geo Sample Number (IGSN)

To learn more about persistent identifiers visit Go-FAIR F1 Principle.

Activity 1
Search for papers and data in nanotechnology and observe the use of ORCIDs. Is it widely known in your community? Why do you think that is?

Activity 2
(Read 5 min) OpenAIRE/FREYA/ORCID guide for researchers "How can identifiers improve the dissemination of your research outputs?"

Activity 3
(Watch 4 min) Six Ways to Make Your ORCID iD Work for You! If you already have an ORCID, check this video https://www.youtube.com/watch?v=h92bUZ5T_vA to link publications to your ORCID profile.

Activity 4
(Discuss in pairs 5 min) The Joint Declaration of Data Citation Principles from FORCE11.

Thing 6 - Interoperability

Interoperability enables data and metadata to flow between different systems with the use of standard vocabularies and references to other data and metadata. That is why standardised formats of protocols are important. In nanotechnology, you may find protocols for:

- different surfaces to interoperate
- preparation of multilayered structures: The preparation of functional multilayered structures comprising different kind of 2D materials, entails complex procedures and fine treatment both mechanically and chemically. These recipes are encoded into protocols that are available to the entire community of sample manufacturers and allow the repetition and improvement of the sample preparation procedure and methodologies.

In materials science, as researchers create independent materials databases, much can be gained from retrieving data from multiple databases. However, the retrieval process is difficult if each database has a different API which is a common case. To address this challenge, there are initiatives such as the Open Databases Integration for Materials Design (OPTiMaDe) consortium which aim to make materials databases interoperational by developing a common REST API.

Activity 1
Protocol Exchange from Nature Protocols is an open repository of community-contributed protocols. Search for nanotechnology protocols publicly shared.

Thing 7 - Licenses and provenance of data for reusability

Licenses grant specific permissions for researchers other than the owner to use scientific output, such as publication, data or software following following each time the specified set of exploitation requirements/recommendations tied with the license.

OpenAIRE Guide for Researchers "How do I license my research data?" provides information about licenses for research data and how to apply them. You could also check OpenAIRE Guide for researchers "Can I reuse someone else's research data?" and "How do I know if my research data is protected?".

You may find useful information about specific licenses for data and software/code below:

- Open Source Initiative variety of open licenses for free and open source software
- Creative Commons a set of widely-used machine-, human-, and lawyer-readable licenses appropriate for use with data

Some tools which assist selection processes when applying licenses to your outputs are:

- OpenMinted Compatibility Matrix tool: understand the use of licenses in software, services and other works, and check the outcome of combination of multiple licenses.
- EUDAT licence selector helps to choose the right licence for your data and/or software by answering simple questions or using the search to find the license you want.

One of the key challenges of the materials science domain is the need to automatically prepare, execute and monitor workflows of calculations as well as to transparently retrieve and store the results in a format which is easy to browse and query. AiiDA's design is based on directed graphs to track the provenance of data, and ensure preservation and searchability. Last, complex sequences of calculations can be encoded into scientific workflows. Sharing capabilities of AiiDA have greatly helped scientific repositories like Materials Cloud.

Activity 1
Select a dataset or code you have been working on lately and choose a license from the list. What did you choose? Why? Have discussions with your team members about licensing and sharing.

Activity 2
Go to AiiDA and load a calculation that you have recently created or that you are currently working on. What do you see? Make some changes. Can you find the previous version? Can you find the first set of calculations that you did in this sequence?

Thing 8 - Services and tools to store, publish and analyse data

Open science for nanotechnology, means open resources, databases and other platforms. Scientific output is often hosted in public and interoperable infrastructures to use, ideally freely available, but sometimes with a small /reasonable reproduction cost resulting from creating, maintaining and publishing data. So, nanotechnology data may be published and stored in a laboratory website or repository, an institutional website or repository, subject specific databases, servers or repositories or it may be collocated with the publication/ journal article it relates to.

Indicative resources:

Registries:

- Re3data.org Registry of generic and thematic Research Data Repositories
- Nanomaterial Registry. The Nanomaterial Registry is an authoritative, fully curated resource that archives research data on nanomaterials and their biological and environmental implications. The data is curated from a variety of data sources, including public databases, manufacturers, regulatory and standards agencies, published literature

Hubs:

- nanoHUB nanoHUB.org is a place for computational nanotechnology research, education, and collaboration. The site hosts a rapidly growing collection of simulation

tools for nanoscale phenomena that run in the cloud and are accessible through a web browser. In addition, nanoHUB provides online presentations, cutting-edge nanoHUB-U short courses, animations, teaching materials, and more.

Databases:

- Nanotechnology knowledge base Database by the European Chemicals Agency (ECHA)
- Crystallography Open Database Open-access collection of crystal structures of organic, inorganic, metal-organics compounds and minerals, excluding biopolymers.
- Bilbao Crystallographic Server Database offering crystallographic and solid state programs and utilities
- Enanomapper prototype database is part of the computational infrastructure for toxicological data management of engineered nanomaterials, funded by the EC and is used to collect and openly share data among researchers.

Repositories:

- Zenodo a generic purpose repository for publications, data, software.
- Materials Cloud , adopts the "repository of repositories" model of Github et al. It aims at creating an ecosystem supporting researchers in various tasks faced with during their work. Sections include educational material, tools to generate and analyze data via the browser and interfaces to facilitate the FAIR sharing of simulation data.
- NOMAD, a data repository that collects a large number of individual computational materials science calculations in one place making them available for a longer period. It enables the confirmatory analysis of materials data, their reuse, and repurposing. NOMAD makes scientific data citable as one can request digital object identifiers.

Several platforms integrate centralized data repositories with software frameworks used to compute data such as:

- AFLOWlib, with Aflow (Automatic Flow). Aflow is a software framework for high-throughput calculation of crystal structure properties of alloys, intermetallics and inorganic compounds.
- OQMD (Open Quantum Materials Database) is a database of DFT-calculated thermodynamic and structural properties based on qmpy, a toolkit for storing crystal structure data, automating calculations, handing computational resources and performing thermodynamic analysis.
- Open Materials Database with the high-throughput toolkit (httk). httk is a toolkit for preparing and running calculations, analyzing the results, and storing the results and outcome in a global and/or in a personalized database.

Educational Resources:

- nanoHUB nanoHUB provides online presentations, cutting-edge nanoHUB-U short courses, animations, teaching materials, and more.
- The typical research cycle which starts from learning to simulating and finally to publishing the final and curated results is mirrored in Materials Cloud via its five main

sections: LEARN, WORK, DISCOVER, EXPLORE and ARCHIVE. At the learning section, users can find educational materials and videos. Moreover, simulation services, turnkey solutions and data analytics tools are provided in other sections helping the scientists to perform their simulations.

Activity 1 - re3data
Go to re3data and search for a data repository based on your area and nanotechnology focus. What are the additional information provided by their functions? Are they useful to you? Are they relevant to the FAIR principles?

Activity 2
Looking after your data: Where do you usually store your data? Do you identify any major differences with those repositories listed in re3data for example?

Activity 3
Archiving your data: What data should be kept or destroyed after the end of your project? For how long should data be kept after the end of your project? Where will the data you keep be archived? When will data be moved into the archive? Who is responsible for moving data to the archive and maintaining them?

Activity 4
Sharing your data: Who else has a right to see or use this data during the project? What data should or shouldn't be shared openly and why? Who should have access to the final dataset and under what conditions? How will you share your final dataset?

Thing 9 - Nanotechnology and High Performance Computing (HPC)

In physics, there are two methods used for manipulation of nano-materials:

- Models: that use specific parameters, have the same results but are not that precise in terms of accuracy in measuring quantities
- First principle calculation: where analysis takes place directly on the atoms that are concerned and more precisely than with models

The most popular method is the first principle calculation, however it is computationally intensive and is best undertaken within HPC environments.

In Europe, there are initiatives aiming to tackle big data and intensive analysis issues in a uniform way, such as HELIX (Hellenic Data Service) in Greece, a convergence e-infrastructure with Virtual Machines, cloud computing and HPC capabilities which lowers expected time for analysis to just a few minutes. The European Open Science Cloud (EOSC) which is currently under development, will eventually become a complete and trusted environment of such services and infrastructures serving the whole research lifecycle.

Activity 1
Discussion Have you experienced any challenges when managing and analysing your data relevant to the time of analysis required to take place? How did you overcome them?

Thing 10 - More best practices

These are some additional best practices to follow in order to improve data and software reusability by others, including oneself when accessing data that have been generated long time ago. Adding terms and conditions of accessibility is an option to consider when data can't be shared completely open. To share data, consult Thing 8 - Services and tools to store, publish and analyse data. To get started, some issues to consider are:

Data structures: Keep consistent file and folder naming conventions across linked projects.

Containerisation: For data processing pipelines

- Singularity
- Docker
- Or use Virtual environments provides, such as the Characterisation Virtual Lab

Activity 1
How do you structure and name your folders and files? How do you manage different versions of your files? What additional information is required to understand the data?

Activity 2 - Discussion + Action
What Can You Do?

- Release some or all of the project metadata – your call, as a simple rule, the more the better!
- Curate existing datasets to make available in the future - you set the upload schedule.
- Contribute your scripts/code
- Have discussions with your team members about licensing and sharing.
- Create a data management plan.

Activity 2 - Questions
Go through the questions from the Horizon2020 guide to create a FAIR Data Management Plan and see if you can already answer many of them. Then check the NanoCommons Data Management Plan and compare with your responses. Do you identify any differences?

Recommended extra reading Ten Simple Rules for Creating a Good Data Management Plan, Ten Simple Rules for Reproducible Computational Research and Ten principles for machine-actionable data management plans, these papers will help you connect all the concepts that you have learned so far.

Notes

1 Crystallographic Information Framework (CIF) "A well-established standard file structure for the archiving and distribution of crystallographic information, CIF is in regular use for reporting crystal structure determinations to Acta Crystallographica and other journals. Sponsored by the International Union of Crystallography, the current standard dates from 1997. As of July 2011, a new version of the CIF standard is under

consideration." More information, tools and use cases are available here: https://rd-alliance.github.io/metadata-directory/standards/cif-crystallographic-information-framework.html

The European Open Science Cloud (EOSC)

Sprinters:

Marjan Grootveld / Data Archiving and Networked Services (DANS)

Frans Huigen / Data Archiving and Networked Services (DANS)

Eliane Fankhauser / Data Archiving and Networked Services (DANS)

Ellen Leenarts / Data Archiving and Networked Services (DANS)

Paula Andrea Martinez / National Imaging Facility (former ELIXIR Europe)

Audience:

- Library staff who provide research support
- Researchers and research communities

Description:

In the ideal world, everyone would have access to all research outcomes and available knowledge, to use it, build upon it, and expand it to serve societal goals as well as private and personal interests. A way to do so is through the EOSC, but what exactly is it? Find out through these ten things!

An overview of Things:

1. Introducing EOSC
2. Buzzword busting
3. FAIR principles
4. Infrastructures
5. FAIR in EOSC
6. EOSC for research domains
7. EOSC training
8. EOSC services
9. Scientific integrity and trust
10. Open Science: the rest of the world

Thing 1 — Introducing

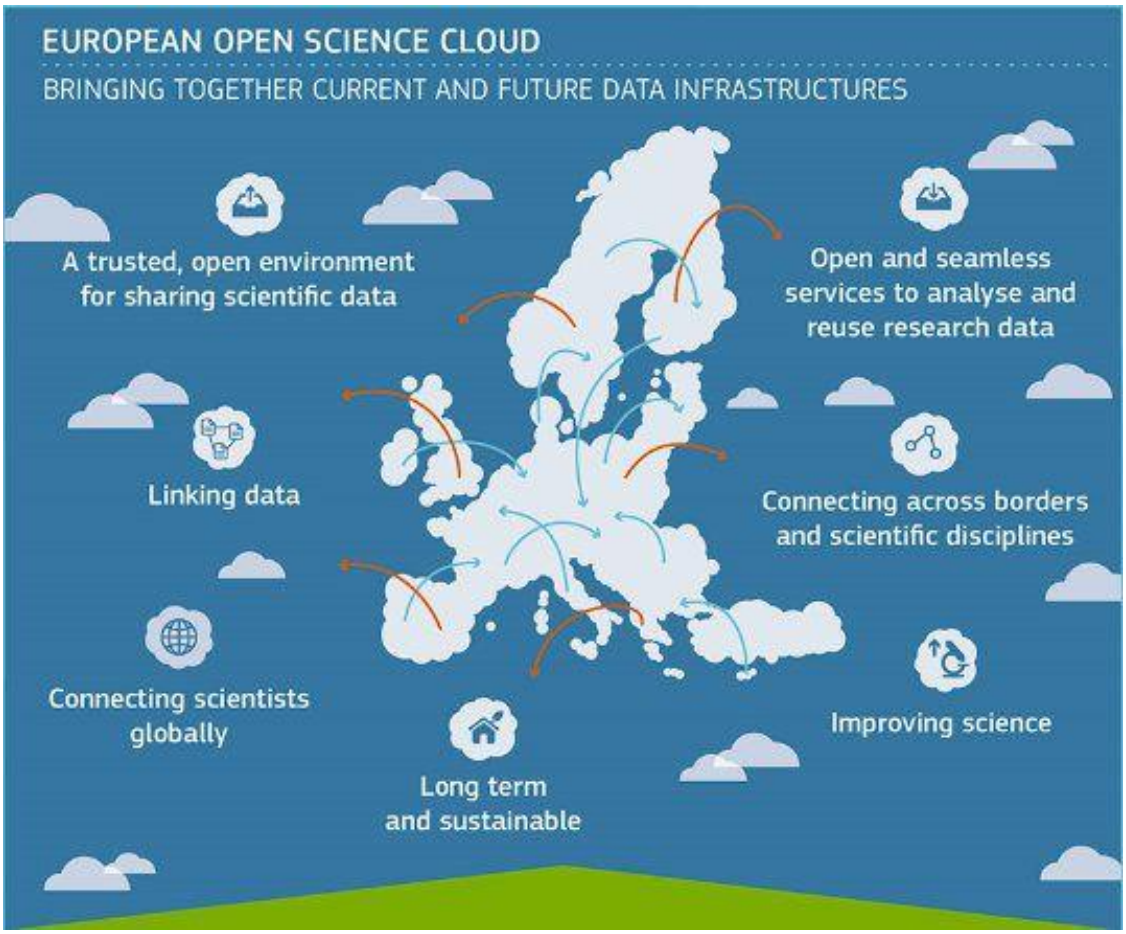
Funded through the Horizon 2020 (H2020) initiative, with 40 countries involved, EOSC offers 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States.

Activity 1:

Watch this 5 min video from the Open Science MOOC on "An introduction to the European Open Science Cloud (EOSC)" by Jean-Claude Burgelman.

Activity 2:

Familiarise yourself with the key EOSC concepts presented in the figure below.



European Open Science Cloud

Figure 1: European Open Science Cloud becomes a reality (European Commission, 23 November 2018)

Thing 2 — Buzzword busting

Buzzwords can't be completely avoided but we will attempt to clarify and scope a handful of them, for the remainder of this Top 10:

EOSC: right, this is the European Open Science Cloud. However,

- **E:** it is not limited to Europe, witness the outward arrows in figure 1;

- **O:** it also supports research findings that cannot be openly shared with everyone, for instance because they relate to individuals, security, or commercial interests;
- **S:** stimulating citizen science and bringing ideas to the market are express goals of the European Commission;
- **C:** the cloud is already a metaphor on its own.

FAIR: this acronym stands for Findable, Accessible, Interoperable and Reusable. 15 FAIR principles (see the next Thing) provide guidance on making your research output more machine actionable. Since the principles were developed in 2016, they've gotten worldwide traction as different scientific disciplines try to operationalise them and measure the FAIRness level of data, software and other related concepts.

Activity 1:

Communication-wise, "FAIR" has been brilliant. However, some argue that other essential research aspects are missing. For instance, check out the plea for Responsible Data Science. Ensuring Fairness, Accuracy, Confidentially, Transparency (FACT). Formulate your opinion on these acronyms, do you think they are helpful?

Infrastructure: while not really a buzzword, this term is frequently used and conveniently vague. Think of it as the basic systems and services that an organization uses in order to work effectively. Keep in mind that an infrastructure can - or even should? - include human experts and support staff. For example, to make this explicit the OpenAIRE project calls itself a socio-technical infrastructure.

Thing 3 — FAIR principles

The FAIR Data Guiding Principles - for Findable, Accessible, Interoperable and Reusable data - came into existence during a workshop in Leiden in 2014 where a broad range of stakeholders in the field of research data management and stewardship came together to discuss the improvement of the reusability of research data. They published the first paper on these principles in 2016.

The principles have become an indispensable part of improving research data and software for the community. According to a 2018 report, The cost of not having FAIR research data, could mean, at a minimum, a €10.2 billion per year loss to the European economy (!). FAIR is equally important to EOSC where there has been an uptake of projects such as FAIRsFAIR, FAIRplus and the ESFRI projects.

Thing 4 — Infrastructures

EOSC facilitates open science through "vertical" or "horizontal" infrastructures:

- Research Infrastructures or RIs cater to researchers in one or a few related disciplines (or see the formal definition):
  - DARIAH for arts and humanities
  - LTER for ecologists.

  - ESFRI is the linking pin of these RIs.

  Such RIs have been around for forty years, and this video sketches the interaction between ESFRI and EOSC. You can call RIs discipline-specific or "vertical", as opposed to…

- e-Infrastructures: projects like…
  - … OpenAIRE, advancing Open Science in the broadest sense;
  - … FAIRsFAIR, promoting everything FAIR;
  - … EOSC-hub, enabling you to for example run large computational analyses;
  - … and FREYA, building the infrastructure for persistent identifiers (PIDs);

  … develop and provide digital services which are cross-disciplinary. The expertise these infrastructures share and the services they offer are generic or "horizontal".

Activity 1:

Compare research infrastructure and e-infrastructure as defined by Science Europe. Which supports you best? Consider how you could benefit more from them.

Activity 2:

Another example of an e-Infrastructure, have a look at the *Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic* and browse through 20.000 letters that were written by and sent to 17th century scholars who lived in the Dutch Republic. Dutch history is at your fingertips thanks to e-infrastructure.

Thing 5 — FAIR in EOSC

Early 2017, the European Commission (EC) embraced the FAIR principles. An innovative element of the Horizon 2020 grant scheme at the time was (and still is, in 2019) an Open Research Data Pilot, asking funded projects to make the data underpinning their publications available or "Open". As EC representatives put it in April 2017: "We are now seeing openness as one component of FAIR data and aim to address all of the FAIR aspects in Horizon 2020".

The FAIRsFAIR project contributes to the implementation of these recommendations. For instance, two work packages and a competence framework with several trainings address the skills-related recommendations. This is done by linking with other parties active in open science and FAIR such as:

- GO-FAIR
- FOSTER
- CODATA
- Research Data Alliance
- European University Association

FAIRsFAIR also addresses certification of FAIR services by strengthening the network of trustworthy digital repositories. Outcomes will feed into the work of the EC's EOSC FAIR

Working group and the Working Group Rules of Participation in the EOSC. These rules will guarantee an open, secure and cost-effective federated EOSC.

Activity 1:

This publication explores how open, FAIR, and research data management (RDM) connect: "The boundaries and intersections between RDM, FAIR and open cover important elements that risk being overlooked if we only focus on one concept." Do you agree with this?

Thing 6 — EOSC for research domains

Research infrastructures (RIs) offer digital services at the domain level, for multiple domains, and/or community-wide. Many of them also provide training, for early-career researchers, on how to use their services.

Several RIs are successful in supporting cross-disciplinary work, for example the Digital Research Infrastructure for the Arts and Humanities (DARIAH). It aims to support transnational researchers in all phases of their work: data acquisition, analysis, publication and data archiving.

DARIAH meets the needs of arts and humanities researchers across Europe including the musicologist analysing digital recordings, the archaeologist digitally recreating ancient buildings, and the historian studying digitised texts to investigate how place names change over time. Cross-disciplinary collaboration also supports the growth of communities, like...

- … CLARIN research infrastructure for language resources and technology;
- … OPERAS for the development of open scholarly communication;
- … CESSDA Consortium of European Social Science Data Archives.

Activity 1:

An archaeological study from The Netherlands, carried out before the EOSC era, presents a nice example of cross-domain research: results are preserved in the 4TU ResearchData long-term repository for technical sciences and the DANS long-term repository, which catered to the social sciences and humanities at the time. Can you find the two datasets and the study?

Thing 7 — EOSC training

EOSC training can be several things:

- **Services & resources**: You can find out about training & support by contacting the service or the infrastructure that provides them. Examples of training support connected to services are Jupyter Hub training, B2Find metadata-based data discovery, and IT service management FitSM Foundation Training.
- **Open science, research data management and data management planning**: Targets the research communities and is provided by research institutes or by projects like FOSTER, and OpenAIRE Advance by their network of National Open Access Desks.

For instance, training for librarians and data support staff, including data stewards, is provided by:
  - CODATA/RDA summer school
  - RDNL Essentials 4 Data Support
  - The FOSTER Open science training book has been translated into Spanish and Portuguese as well.
- **Domain specific training**: Examples of training offered by (discipline) research infrastructures include:
  - ELIXIR's Training and e-Support System for the life sciences
  - #dariah Teach platform for the digital arts and humanities and the Digital humanities Course registry (CLARIN and DARIAH)
  - Data Management Expert Guide for social scientists (CESSDA)
  - SoBigData training registry for big data and social mining
- Training of trainers. Sharing best practices on training in EOSC and beyond is supported by the RDA Interest Group Education and Training on research data and the Community of Practice of training coordinators. These are both global networks of training coordinators and they organize regular online and offline events on training topics such as improving impact of training, improving visibility of training events and resources, skills and competences for open science and data stewards, and training portals.

Thing 8 — EOSC services

The EOSC-hub service catalogue is an obvious place to go to, as long as you're aware that some services target researchers and research communities directly, while other services require administrator expertise. Examples include:

- B2DROP - a solution to store and exchange data with your team members.
- B2NOTE - allows you to easily create searchable annotations on research data hosted in the EUDAT Collaborative Data Infrastructure (managed by EOSC-hub): semantic tags, free-text keywords or free-text comments.
- EGI Check-in - a proxy service that operates as a central hub to connect federated Identity Providers (IdPs) with EGI service providers.

EOSC-building projects have started to jointly present their services. See for instance this use case aboutcomplying with open science ambitions and the GDPR: how best to manage and share person-related data? OpenAIRE'sAmnesia anonymization tool can help with removing identifying information from data.

Activity 1:

Have a look at B2FIND and Zenodo. Both are cross-domain repositories for research output. Consider their respective strengths; how can you benefit from using them?

Activity 2:

Watch this webinar on how to manage your data to make them open and FAIR.

Thing 9 — Scientific integrity and trust

EOSC aims to make science more open. Replicability of research is one important aspect of openness and is greatly improved if research data, software, and methods are explicit and publicly available. You enable fellow future researchers to learn from what you have done.

The blog *Retraction Watch* reports on retractions of scientific papers and erroneous research data practices. But how do we prevent such cases? One method is by using trustworthy repositories, which help to **make** and to **keep** data FAIR:

"**make**": by providing a persistent identifier, supporting metadata standards, supporting findability through their public catalogue, providing clear licences.

"**keep**": by preserving the data, documenting them, and keeping them usable in the long run (through sustainable formats, repositories). In theGuidelines on FAIR Data Management in Horizon 2020, the European Commission states: "*Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.*" (cited from the OpenAIRE initiative). One way to make sure that your data is stored at such a certified repository, is to look at the CoreTrustSeal certification.

Core Trust Seal

Figure 2: CoreTrustSeal Certification Launched (Research Data Alliance, 11 September 2017)

EOSC aims to be a trusted environment for the storage, processing, and reuse of research data. Where FAIR tells us something about the research data, these need to be preserved somewhere trustworthy. So, trust and FAIR go hand in hand.

Activity 1:

Watch this video tutorial about FAIR data in trustworthy repositories. Do you agree with the recommendations? And which of the CoreTrustSeal requirements do you think are most important?

Thing 10 — Open Science: the rest of the world

With all this *"Thinking"* about E-OSC, there are clearly no Schengen-like borders around it, so let's look at similar examples beyond Europe:

- In Latin America, there is the *Consejo Latinoamericano de Ciencias Sociales* (CLACSO). Their organizational structure is similar to that of the EOSC: about 700 institutions from 52 countries collaborate and cooperate in empowering open science on their continent. The main drivers of this initiative are universities and government organisations.
- The UNESCO Global Open Access Portal (GOAP) is funded by the governments of Colombia, Denmark, Norway, and the United States Department of State. The portal presents a current snapshot of the status of open access (OA) to scientific information in 158 countries worldwide.
- Consider the Open and Collaborative Science in Development (OCSD) Network as well, organised by the Global South countries - a network similar to CLACSO. The organisation formulated four specific goals in the framework of open science.
- In addition to these publicly funded initiatives, there is the not-for-profit example, the Open Data Science Cloud, part of the Open Commons Consortium. Their cloud provides infrastructures and tools for researchers to analyse terabytes and petabytes worth of data.

Activity 1:

We've covered initiatives and services on some of the continents but not all of them. One continent where there is a lot happening regarding research data management is Australia. Can you find some of the research data initiatives and services happening down under?