

Ten Quick Tips for Finding Research Data

Kathleen Gregory¹, Siri Jodha Khalsa², William K. Michener³, Fotis E. Psomopoulos⁴, Anita de Waard⁵, Mingfang Wu^{6*}

¹ Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences, The Hague, Netherlands. orcid.org/0000-0001-5475-8632

² National Snow and Ice Data Centre, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA. orcid.org/0000-0001-9217-5550

³ College of University Libraries & Learning Sciences, The University of New Mexico, Albuquerque, New Mexico, USA. orcid.org/0000-0002-8341-1141

⁴ Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece. orcid.org/0000-0002-0222-4273

⁵ Research Data Management Solutions, Elsevier, USA. <https://orcid.org/0000-0002-9034-4119>

⁶ Australia National Data Services, Melbourne, Australia. <http://orcid.org/0000-0003-1206-3431>

*Corresponding author

E-mail: Mingfang.wu@ands.org.au

All authors contributed equally to this work.

Introduction

Over the past decades, science has experienced rapid growth in the volume of data available for research - from a relative paucity of data in many areas to what has been recently described as a data deluge [1]. Data volumes have increased exponentially across all fields of science and human endeavour, including data from sky, earth and ocean observatories; social media such as Facebook and Twitter; wearable health monitoring devices; gene sequences and protein structures; and climate simulations [2]. This brings opportunities to enable more research, especially cross-disciplinary research that could not be done before. However, it also introduces challenges in managing, describing and making data findable, accessible, interoperable and reusable by researchers [3].

When this vast amount and variety of data are made available, finding relevant data to meet a research need is increasingly a challenge. In the past when data were relatively sparse, researchers discovered existing data by searching literature, attending conferences and asking colleagues. In today's data-rich environment, with accompanying advances in computational and networking technologies, researchers increasingly conduct web searches to find research data. The success of such searches varies greatly and depends to a large degree on the expertise of the person looking for data, the tools used and, partially, on luck. This article offers ten quick tips that researchers can follow to more effectively and precisely discover data that meets their specific needs. A list of useful resources and examples from the article are listed in the final Resources section.

Tip 1: Think about the data you need and why you need them

Before embarking on a search for data, consider how you will use the desired data in the context of your overall research question. Are you seeking data for comparison or validation, as the basis for a new study or for another reason? List the characteristics that the data must have in order to fulfil your identified purpose(s), including requirements such as data format, spatial or temporal coverage, availability, and author or research group. In many cases, your initial data requirements and the identified constraints will change as you progress with the search. Pausing to first analyse what you need and why you need it can lead to a more analytic search, save searching time and facilitate the actions described in Tips 2-6.

Tip 2: Select the most appropriate resource

Directories of research data repositories such as re3data.org and FAIRsharing, web search engines, and colleagues can be consulted to discover domain-specific portals in your

discipline. Subject domain is but one criterion to consider when selecting an appropriate data repository. Various certification processes have also been implemented to help develop trustworthiness in repositories and to make their data governing policies more transparent. For example, repositories earning the CoreTrustSeal Trustworthy Data Repository certification must meet sixteen requirements measuring the accessibility, usability, reliability, and long-term stability of their data. Knowing what standards and criteria a repository applies to data and metadata provides more confidence in understanding and reusing the data from that repository.

Domain-specific portals provide ways to quickly narrow your search, offering interfaces and filters tailored to match the data and needs of specific disciplinary domains. Map interfaces for data collected from specific locations (see the National Water Information System) and specific search fields and tools (see the National Centre for Biotechnology Information's, NCBI, complement of databases) facilitate discovering disciplinary data. Other domain-focused repositories, such as the National Snow and Ice Data Centre (NSIDC), collect and apply knowledge about user requirements and incorporate domain semantics into their search engines to help data seekers quickly find appropriate data. Data aggregators, including DataONE for environmental and earth observation data, VertNet and GBIF for museum specimen and biodiversity data, or DataMed for biomedical datasets, enable searching multiple data repositories or collections through a single search interface. The Resources section provides links and more information about these repositories.

The accessibility of data resources is another important consideration. University librarians can provide advice about particular subscription-based resources available at your institution. Research papers in your field can also point to available data repositories. In domains such as astronomy and genomics, for example, citation of datasets within journal articles are commonplace. These references usually include dataset access information that can be used to locate datasets of interest or to point toward data repositories favoured within a discipline.

Tip 3: Construct your query strategically

Describing your desired data effectively is key to communicating with the search system. Your description will determine if relevant data are retrieved and may inform the order of the hits in the results list. Help pages provide tips on how to construct basic and advanced searches within particular repositories (see for example Research Data Australia - click on Advanced Search -> Help). Note that not all repositories operate in the same manner. Some portals, such as DataONE, use semantic technologies to automatically expand the keywords entered in the search box to include synonyms. If a portal does not use automatic expansion, you may

need to manually add various synonyms to your search query (e.g., in addition to “demography” as a search term, one might also add “population density”, “population growth”, “census”, or “anthropology”).

If you are looking for data that span different disciplines, or if you cannot find a suitable domain repository and opt instead for a general search engine, learn how to make the most of search operators. The “site:” operator, for example, restricts web domains for returned results. For example, using this search string

sea level (site:.edu)

will return results only from sites whose URLs end in .edu when using Google or Bing. You can find out what operators are supported by your selected search engine by searching for them directly. For example, the query google search operators results in pages such as [4] describing all operators that can be used in your query. Constructing a query with search operators is like writing a program script; they allow you to precisely communicate what you are searching for, with the reward being the retrieval of more on-topic results.

Tip 4: Make the repository work for you

Repository developers invest significant time and energy organizing data in ways to make them more discoverable; use their work to your advantage. Familiarize yourself with the controlled vocabularies, subject categories, and search fields used in particular repositories. Searching for and successfully locating data is dependent on the information about the data, termed metadata, that are contained in these fields; this is particularly true for numeric or non-textual data. Browsing subject categories can also help to gauge the appropriateness of a resource, hone in on an area of interest or to find related data that has been classified in the same category.

Researchers can also register or create profiles with many data repositories. By registering, you may be able to indicate your general research data interests, which can be utilized in subsequent searches, or to receive alerts about datasets that you have previously downloaded (see also Tip 7).

Tip 5: Refine your search

In many cases, your initial search may not retrieve relevant data or all of the data that you need. Based on the retrieved results, you may need to broaden or narrow your approach. Apart from rephrasing your search query and using search operators as discussed in Tip 3, facets or filters specific to individual repositories can be used to narrow the scope of your

The screenshot displays the ICPSR search interface. On the left, a 'Filter Results' sidebar includes sections for Subject, Geography, Data Format (listing SAS, SPSS, Stata, Delimited, R, and Online analysis with counts), Type of Analysis, Time Period (with year pickers), Restriction Type (Public and Restricted Use), Recency, Data Availability, and Export Results. The main area is titled 'Search Results' and shows a search box with 'higher education' and a 'REVISE SEARCH' button. Below the search box, there are tabs for 'Studies (1,310)', 'Variables', and 'Publications'. A message states: 'Your query returned 399 studies. Series results are grouped under the most relevant study. [Ungroup series results](#) to view all results.' A 'Sort by' dropdown is set to 'Study Relevance'. The results list includes:

- Higher Education Employees, 1977 (ICPSR 7648)**
United States Department of Education. Institute of Education Sciences. National Center for Education Statistics
[81 more results in Higher Education General Information Survey \(HEGIS\) Series](#)
- Texas Higher Education Opportunity Project (ICPSR 29841)**
Tienda, Marta; Sullivan, Teresa A.
- Carnegie Commission National Survey of Higher Education: Faculty Study, 1969 (ICPSR 7501)**
Trow, Martin, et al.
- Carnegie Commission National Survey of Higher Education: Faculty Study Subsample, 1969 (ICPSR 7078)**
Ladd, Everett; Lipset, S.M.; Trow, Martin
- Carnegie Commission National Survey of Higher Education: Graduate Study Subsample, 1969 (ICPSR 7363)**
Ladd, Everett; Lipset, S.M.; Trow, Martin

Fig 1. Filters available when searching for data in ICPSR

results. Refinements such as data format, types of analysis and data availability, as demonstrated in the Inter-university Consortium for Political and Social Research (ICPSR) database shown in Fig 1, allow users to quickly find usable data.

Examining results that look interesting (for example by clicking on links for “more information”) can be a signal of the type of information that you find relevant. These records can then be linked to related ones, and in subsequent searches, other results algorithmically determined to be related will be brought to the top of the results list.

Some search systems will also suggest query formulations to retrieve better results, and many repositories (e.g. the Commonwealth Scientific and Industrial Research Organisation, CSIRO, Data Access Portal depicted in Fig 2) tailor their advanced search features to enhance query precision. (See the Resources section for more information about these repositories).

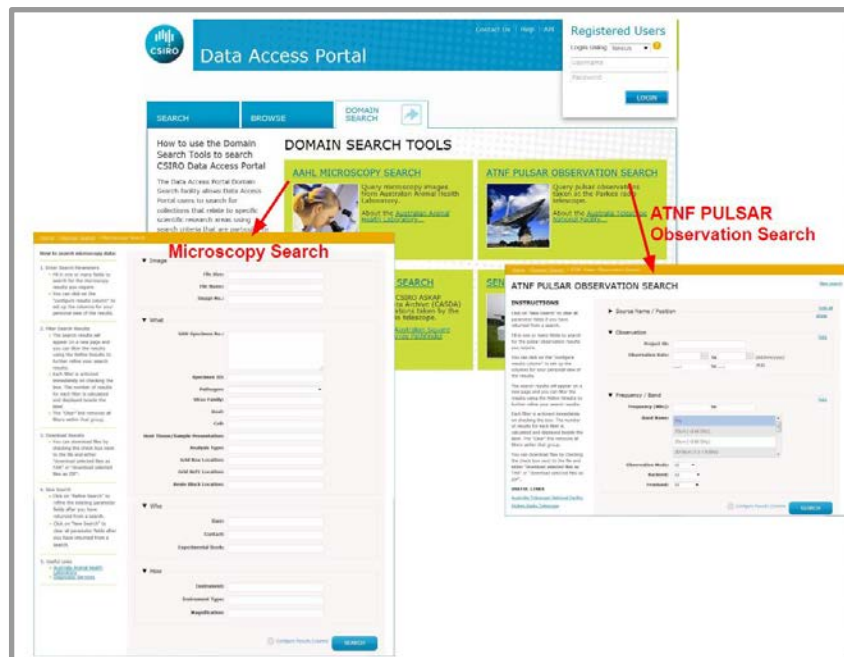


Fig 2. Query interface from CSIRO Data Access Portal, with advanced search customised for each domain

Tip 6: Assess data relevance and fitness-for-use

Conduct a preliminary assessment of the retrieved data prior to investing time in subsequent data download, integration, analytic and visualization efforts. A quick perusal of the metadata (text and/or images) can often enable you to verify that the data satisfy the initial requirements and constraints set forth in Tip 1 (e.g., spatial, temporal and thematic coverage and data sharing restrictions). Ideally, the metadata will also contain documentation sufficient to comprehensively assess the relevance and fitness-for-use of the data, including information about how the data were collected and quality assured, how the data have been previously used, etc. Some data repositories such as the National Science Foundation's Arctic Data Centre enable the data seeker to generate and download a "Metadata Quality Report" that assesses how well the metadata adhere to community best practices for discovery and reusability. For the sample report shown in Fig 3, checks of data discovery indicate 100% completeness, whereas checks of identification and interpretation demonstrate 50% completeness. Clearly, if none of your criteria for data are met, you may not wish to download and use the associated data.

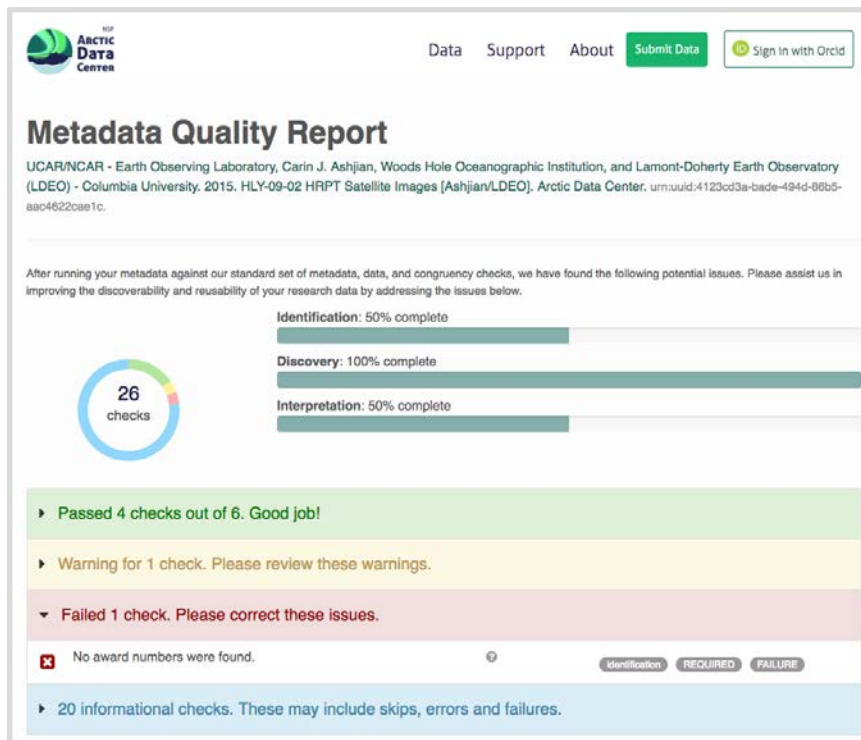


Fig 3. Example of Metadata Quality Report from the NSF Arctic Data Centre

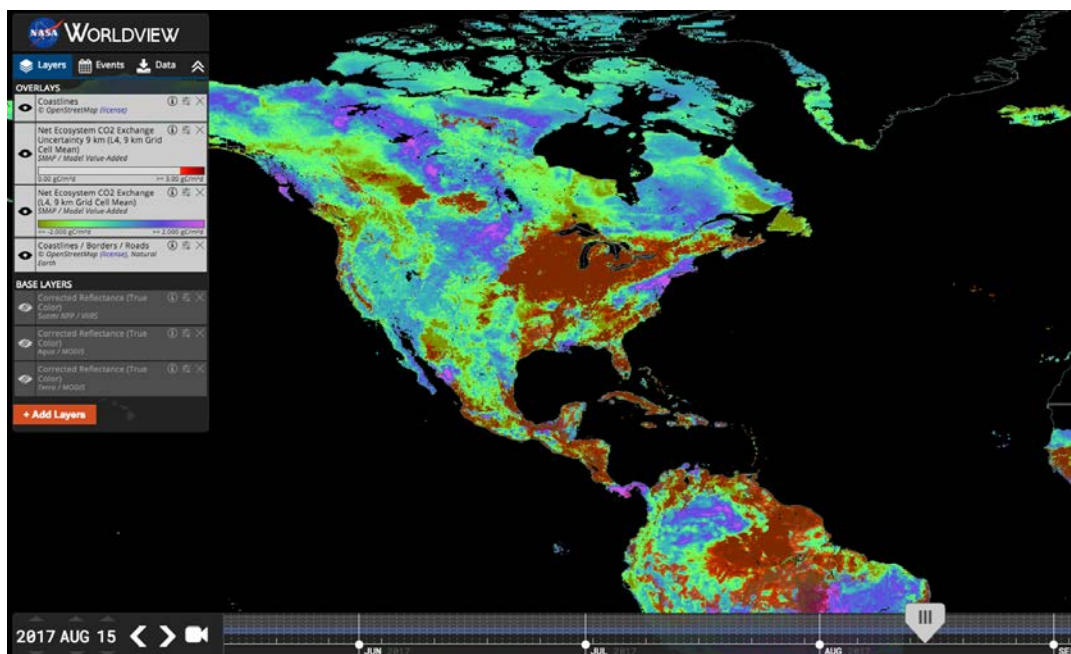


Fig 4. NASA Worldview showing areas of high uncertainty (red) over Net Ecosystem Exchange (green-blue) values for 15 August 2017

Attention should also be paid to quality parameters or flags within the data files. NASA's Worldview enables visualization of embedded quality layers for some products, such as Net

Ecosystem CO2 Exchange. Fig 4 shows how a semi-opaque layer representing uncertainty can be viewed on top of the associated data variable.

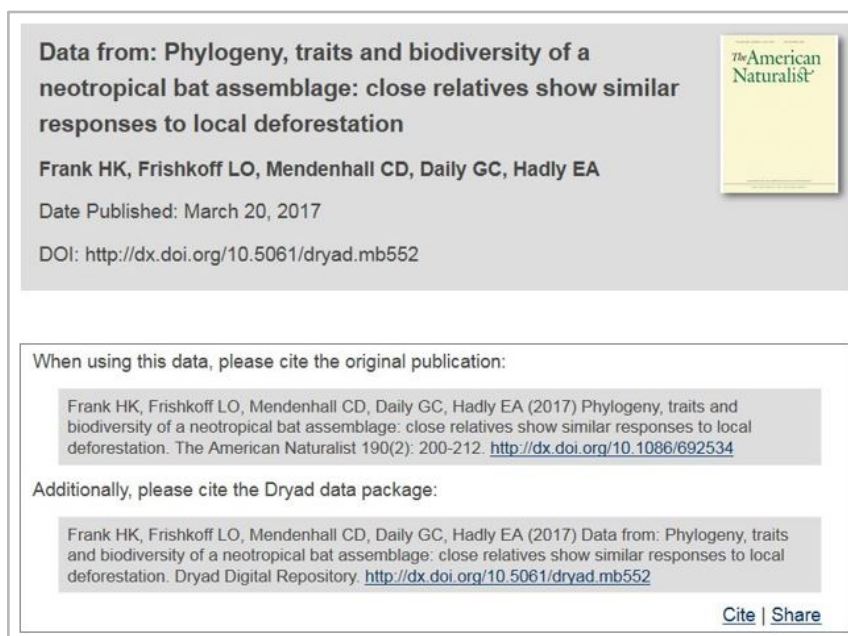
Tip 7: Save your search and data source details

Record the data source and data version if you access or download a data product. This may be accomplished by noting the persistent identifier, such as a digital object identifier (DOI) or another Global Unique Identifier (GUID) assigned to the data. Recording the URL from which you obtained the data can be a quick way of returning to it, but should not be trusted in the long term for providing access to the data, as URLs can change. It is also a good practice to save a copy of any original data products that you downloaded [5]. You may, for example, need to go back to original data sources and check if there have been any changes or corrections to data. Registering with the data portal (as described in Tip 3), or registering as a user of a specific data product allows the repository to contact you when necessary. Such information may be needed when you publish a paper that builds on the data you accessed. If there are any errors found in the original data, registering with the data service allows them to contact you, to see if there is an impact on any research conclusions that you have drawn from this data.

If you have registered with a portal, it may also be possible to save your searches, allowing you to resume your data search at a later time with all previously defined search criteria. Some portals use RESTful search interfaces, which means you can bookmark a results set or dataset and return to it later, simply by going to the bookmark.

Tip 8: Look for data services, not just data

The data you seek may be available only via an application programming interface (API) or as linked data [6]. That is, instead of a file residing on a server, the data that best suits your purposes is provided as a service through an API. Examples of such services include the climate change projection data available through the NSW Climate Data Portal where data is dynamically generated from a simulation model, Google Earth Engine or Amazon Web Services (AWS) public datasets (see Resources). Data made available from these services may not be searchable from general web search engines, but data services may be registered to data catalogues or federations such as Research Data Australia, DataONE, and other resources listed in re3data.org and FAIRsharing. Many repositories that host extremely large volumes of data such as sequencing, environmental observatory, and remotely sensed data provide access to tools, workflows and computing resources that allow one to access, visualize, process and download manageable subsets of the data. Often, the processing



Data from: Phylogeny, traits and biodiversity of a neotropical bat assemblage: close relatives show similar responses to local deforestation

Frank HK, Frishkoff LO, Mendenhall CD, Daily GC, Hadly EA

Date Published: March 20, 2017

DOI: <http://dx.doi.org/10.5061/dryad.mb552>

When using this data, please cite the original publication:

Frank HK, Frishkoff LO, Mendenhall CD, Daily GC, Hadly EA (2017) Phylogeny, traits and biodiversity of a neotropical bat assemblage: close relatives show similar responses to local deforestation. *The American Naturalist* 190(2): 200-212. <http://dx.doi.org/10.1086/692534>

Additionally, please cite the Dryad data package:

Frank HK, Frishkoff LO, Mendenhall CD, Daily GC, Hadly EA (2017) Data from: Phylogeny, traits and biodiversity of a neotropical bat assemblage: close relatives show similar responses to local deforestation. Dryad Digital Repository. <http://dx.doi.org/10.5061/dryad.mb552>

[Cite](#) | [Share](#)

Fig 5. Data and code supporting the publication are archived and made available from Dryad

workflows that one might use to process and download a dataset can also be downloaded, saved and used again in subsequent searches.

Tip 9: Monitor the latest data

One of the most effective ways to identify new data submissions is to monitor the latest literature. Many journals such as *Nature*, *PLOS*, *Science* and others require that the data underlying a publication also be published in a public (e.g. Dataverse, Dryad, or Zenodo) or discipline-based repository (e.g. EASY from DANS, GenBank, or PubChem) (The Resources section provides further details about these repositories). Fig 5, for example, shows a recent journal article in *American Naturalist* that includes a citation to the associated data package (the data, metadata and analysis scripts), which is uniquely identified via a DOI and stored in the Dryad Digital Repository.

In addition, many domain-based repositories such as environmental observatories and sequencing databases are constantly accepting similar types of data submissions. Fig 6 shows an example where the same type of data is collected at different time intervals and regularly added to a collection. Publishers and some digital repositories also offer alerting services when new publications or data products are submitted. Depending on the resource, it may be possible to set up a recurring search API or an RSS feed to automatically monitor specific resources. For example, the NSIDC offers a subscription service where new data



Fig 6. Example of evolving time series data, two expeditions in 2016 [7] and 2017 [8]

meeting a list of user-generated specifications is automatically pushed to a location specified by the user.

Tip 10: Give back (cite and share data)

There are three ways to give back to the community once you have sought, discovered and used an existing data product. First, it is essential that you give proper attribution to the data creators (in some cases, the data owners) if you use others' data for research, education, decision making, or other purposes [9]. Proper attribution benefits both data creators/providers and data seekers/users. Data creators/providers receive credit for their work and their practice of sharing data is thus further encouraged. Data seekers/users make their own work more transparent and, potentially, reproducible, by uniquely identifying and citing data used in their research.

Many digital repositories provide specific guidelines for properly citing data products, Figure 5 shows such an example from the Dryad Digital Repository. Furthermore, the Dryad Digital Repository recommends that the journal article based on the data product also be cited by subsequent data users, thereby giving full attribution to the intellectual contributions by the data creators and the authors of the journal article, who may or may not be the same individuals.

Many data creators and institutions adopt standard licenses from organizations such as Creative Commons that govern how their data products may be shared and used. Creative Commons recommends that a proper attribution should include Title, Author, Source and License [10].

Second, provide feedback to the data creators or the data repository about any issues associated with data accessibility, data quality, or metadata completeness and interpretability. Data creators and repositories benefit from knowing that their data products are understandable and usable by others, as well as knowing how the data were used. Future users of the data will also benefit from your feedback as well.

Third, virtually all data seekers and data users also generate data. The ultimate “give-back” is to also share your data with the broader community.

Conclusion

This paper highlights ten quick tips that, if followed, should make it easier for a data seeker to discover data that meet a particular need. Regardless of whether you are acting as a data seeker or a data creator, remember that “data discovery and reuse are most easily accomplished when: (1) data are logically and clearly organized; (2) data quality is assured; (3) data are preserved and discoverable via an open data repository; (4) data are accompanied by comprehensive metadata; (5) algorithms and code used to create data products are readily available; (6) data products can be uniquely identified and associated with specific data originator(s); and (7) the data originator(s) or data repository have provided recommendations for citation of the data product(s).” [11].

Resources

This section provides links and further information about the resources and examples discussed in the article.

A: Public and Multidisciplinary Data Repositories

- CSIRO Data Access Portal (<https://data.csiro.au/dap/home>): multidisciplinary data repository from Australia’s national science agency with multiple searching and browsing possibilities, including an API.
- Dataverse (<https://dataverse.org>): open source web application to share, preserve, cite, explore, and analyse research data; the Harvard Dataverse (<https://dataverse.harvard.edu>) is openly available to researchers and data collectors worldwide from all disciplines, to deposit data and search.
- Dryad Digital Repository (<http://datadryad.org>): aims to make data of different types and from different disciplines freely discoverable, usable, and citable.

- Zenodo (<https://zenodo.org>): an open data repository service maintained by CERN, Geneva, providing repositing services for research data outputs in all scientific disciplines. Datasets can be located via the Zenodo search engine.

B: Disciplinary Data Repositories and Aggregators

- DataONE: Data Observation Network for Earth (<https://www.dataone.org>): aggregates data from member repositories in the earth and environmental sciences
- DataMed (<https://datamed.org>): a prototype biomedical data search engine to discover data sets across data repositories and aggregators.
- Data Archiving and Networked Services (DANS): EASY Repository (<https://easy.dans.knaw.nl>): data archive with focus on social sciences and humanities data, particularly archaeology.
- GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>): genetic sequence database containing annotated collections of publicly available DNA sequences.
- Global Biodiversity Information Facility (GBIF) (<https://www.gbif.org>): a research data infrastructure providing open access to data about all types of life on earth.
- Inter-university Consortium for Political and Social Research (ICPSR) (<https://www.icpsr.umich.edu/icpsrweb/>): international consortium of research organizations and institutions providing access to social and behavioural sciences data.
- NASA Worldview Net Ecosystem CO2 Exchange (<https://go.nasa.gov/2eY0vJp>): example of resource allowing the addition of layers indicating uncertainty to assist in assessing data quality.
- National Centre for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/guide/all/>): complete list of NCBI's biomedical databases.
- National Science Foundation Arctic Data Centre (<https://arcticdata.io>): provides access to all research outputs resulting from NSF-funded research in the Arctic; an example of a system allowing the creation of metadata quality reports to assist in data assessment
- National Snow and Ice Data Centre (NSIDC) (<http://nsidc.org/data/search/>): provides access to data about the cryosphere and its interactions from numerous funders
- National Water Information System: Mapper (<https://maps.waterdata.usgs.gov/mapper/index.html>): an example of a map-based search interface to locate hydrological data in the United States

- PubChem (<https://pubchem.ncbi.nlm.nih.gov>): a database of chemical molecules and their activities against biological assays. The system is maintained by the National Centre for Biotechnology information (NCBI, <https://www.ncbi.nlm.nih.gov/>).
- VertNet (<http://vertnet.org>): aggregates data from other vertebrate networks to make biodiversity data freely available via a single data portal.

C: Directories of Data Repositories

- CoreTrustSeal (<https://www.coretrustseal.org/about>): replaces the Data Seal of Approval for repository certification and provides a browsable and searchable list of core certified data repositories.
- FAIRsharing.org: Standards, Databases, Policies (<https://fairsharing.org>): a directory that brings together links to data and metadata standards, databases and repositories, and journal and funder data policies.
- re3data.org: Registry of Research Data Repositories (<http://www.re3data.org>): a global registry of data repositories across academic disciplines developed for researchers, funders, publishers, and institutions.
- Research Data Australia (RDA) (<https://researchdata.andis.org.au>): a registry of research data from research organizations, governmental agencies, and cultural institutions in Australia. The RDA services are maintained by Australian National Data Services (<http://www.andis.org.au/>).

D: Data Services and Repository Services

- Amazon Web Services Public Datasets (<https://aws.amazon.com/public-datasets/>): freely available public datasets that can be analysed using Amazon Web Services without needing to be locally downloaded and stored; collections include geospatial and environmental datasets, genomic and life science datasets, datasets for machine learning, and regulatory and statistical datasets.
- Google Earth Engine (<https://earthengine.google.com>): a cloud computing platform for processing, analysing and visualising satellite imagery and other observation data at the petabyte-scale.
- NSW Government Climate Data Portal (<http://climatechange.environment.nsw.gov.au/Climate-projections-for-NSW/Download-datasets>): service that provides access to dynamically generated climate change projection data available from the New South Wales government.

- National Snow and Ice Data Centre Data Subscription Request (https://nsidc.org/forms/subscription_form.html): an example of a repository service that can be used to receive data matching a list of user-generated specifications.

Acknowledgement:

This paper is an outcome of the work of the Research Data Alliance Data Discovery Paradigms Interest Group. We would like to thank members of the group for their support, especially Andrea Perego, Mustapha Mokrane, Susanna-Assunta Sansone, Peter McQuilton, and Michel Dumontier who read this paper and provided constructive suggestions.

References:

1. Gray J. Jim Gray on eScience: A transformed scientific method. In: Hey T, Tansley S, Tolle K, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Richmond, WA: Microsoft Research; 2009. p.xvii–xxxi. Available from: <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>.
2. Fox G, Hey T, Trefethen A. Where does all the data come from? In: Kleese van Dam K, editor. *Data-Intensive Science*. Chapman and Hall/CRC; 2013. p. 15–51.
3. Wilkinson, M.D. et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* 3, 15 March 2016, doi:10.1038/sdata.2016.18, retrieved online on 25 October 2017 from <http://www.nature.com/articles/sdata201618>.
4. Warner, R. Google Advanced Search: A Comprehensive List of Google Search Operators [Internet]. 2015 [cited 2017 Oct 26]. Available from: <https://bynd.com/news-ideas/google-advanced-search-comprehensive-list-google-search-operators/>.
5. Sandve G, Nekrutenko A, Taylor J, E H. Ten simple rules for reproducible computational research. *PLoS Comput Biol*. 2013;9(10):e1003285.
6. Heath T, Bizer C. Linked Data: Evolving the Web into a global data space. In: Hendler J, van Harmelen F, editors. *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool; 2011. p. 1–136.
7. Australian Ocean Data Network. RV Investigator Voyage IN2017_V02 EM710 MKII Multibeam Echosounder Data. 2017. Available from: <http://www.marlin.csiro.au/geonetwork/srv/eng/search?uuid=e1e6b802-1df0-4703-8813-9fe38f8ee5c3>.

8. Australian Ocean Data Network. RV Investigator Voyage IN2016_V05 EM710 MKII Multibeam Echosounder Data. 2016. Available from: <http://www.marlin.csiro.au/geonetwork/srv/eng/search?uuid=e08c7c03-0d88-45e7-a4c4-e91191b19335>.
9. Martone (ed.) M. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. FORCE11. San Diego CA: 2014. Available from: <https://www.force11.org/group/joint-declaration-data-citation-principles-final>
10. Creative Commons. Best practices for attribution [Internet]. 2014 [cited 2017 Sep 10.]. Available from: https://wiki.creativecommons.org/wiki/Best_practices_for_attribution.
11. Michener, W. K. Data discovery. In: Recknagel F, Michener WK, editors. Ecological informatics: Data management and knowledge discovery. Springer International Publishing; 2017.