Chapter 7
Data Discovery

William K. Michener

Abstract Data may be discovered by searching commercially available internet search engines, institutional and public repositories, online data directories, and the content exposed by data aggregators. Chapter 7 describes these various search approaches and presents seven best practices that can promote data discovery and reuse. It further emphasizes the need for data products to be uniquely identifiable and attributable to the data originators who must also be uniquely identifiable.

7.1 Introduction

Data discovery is the act of searching for and finding data that are or may be of particular interest. Prior to the advent of the Internet and World Wide Web, data discovery was often a difficult and laborious process. Researchers discovered that data existed via word-of-mouth and conference presentations as well as through the published literature. Accessing or acquiring such “found” data was often even more difficult as data sharing has only begun to become the norm over the past two to three decades (Michener 2015).

It is presently much easier to discover data. An array of Internet search engines, data repositories, data directories, and data aggregators have been created to facilitate data and information discovery and provide other services. This chapter describes the various tools and approaches that are most commonly used today to discover specific data products (Sect. 7.2) and best practices for promoting data discovery and use (Sect. 7.3).

W.K. Michener (✉)
University of New Mexico, Albuquerque, NM, USA
e-mail: william.michener@gmail.com

F. Recknagel, W.K. Michener (eds.), *Ecological Informatics*,
DOI 10.1007/978-3-319-59928-1_7

7.2 Discovering Data Created by Others

The various tools and approaches that are used to discover data differ widely in their efficacy at precisely finding the data that one is hoping to examine or acquire. Commercial Internet search engines are very effective at discovering web sites and web pages that mention research projects and publications resulting from those studies. For instance, Google Scholar is particularly adept at enabling users to discover publications related to a particular topic and that may include descriptions of data collection and analytical methods. Internet search engines are often less useful for precisely discovering data that are held in institutional and public repositories as such data may be insufficiently described, hidden behind institutional firewalls, or disambiguated from their associated metadata.

Searches of institutional and public repositories may quickly lead one to particular data products as such repositories often provide search tools that are tailored to facilitate searches of their data holdings. It may, however, prove challenging to identify the specific repository where one invests the time and effort in conducting individual searches of the repository holdings. Data directories help address this challenge by enabling one to search for particular data by keywords and then be pointed to specific databases or repositories that may be linked to the online data directory; the user may then be directed to a particular repository where the data and metadata can be further examined. Data aggregators are increasingly being developed to provide a mechanism to discover data that originate from many different sources (e.g., individuals, repositories, and institutions such as museums and research networks). Data aggregators often provide additional value-added services and products such as quality assurance, metadata checks, and access to analytical and visualization tools. The different approaches to data discovery and relevant examples are described below.

7.2.1 Internet Search Engines

Internet search engines are commonly used to search for information, publications, data and other content that is available on web sites that are part of the World Wide Web. Some of the more commonly employed general search engines include Google, Bing, Yahoo! and Baidu. Internet search engines work by: (1) retrieving information about web pages by routinely visiting web sites (i.e., web crawling); (2) indexing the information that is retrieved such as titles and page content based on HTML markup of the content; and (3) allowing users to query the indexed content based on one or more keywords that the user enters. Different search engines use different and, typically, proprietary approaches and algorithms for indexing and caching (i.e., storing content for rapid access and processing) web content. Google's search engine, for example, is based on an algorithm that ranks web pages based on the number and rank of the web sites and pages that link to them. Commercial internet search engines may derive income by assigning higher rankings to pages from web sites that pay to have their content prioritized in searches, by allowing paid advertisements to appear alongside search results, or by both approaches. Search and indexing algorithms may filter and preferentially rank results based on user characteristics such as location and prior user search history.

The usefulness of a web search engine is related to how relevant results are to the user. For instance, a user will often enter one or more search terms or keywords such as "primary production" and retrieve millions of results. Some internet search engines offer various ways to filter large lists of results and more precisely find desired content. Such approaches include specifying a particular range of dates (e.g., retrieving results that are from the current calendar year) and employing Boolean operators (i.e., AND, OR, and NOT) to refine the query. Using the previous example, after retrieving a daunting list of results after querying "primary production," a user may refine the search by entering "primary production AND grassland AND data" to more precisely discover content of interest. Nevertheless, such a search may still retrieve millions of results—some that point to specific data products, others that point to publications dealing with topics such as how to calculate net primary productivity from biomass data and a multitude of other related issues.

7.2.2 Data Repositories

Numerous data repositories exist worldwide that hold ecological and environmental data. The Registry of Research Data Repositories (also known as re3data.org; re3data.org Project Consortium 2016) is a global registry of research data repositories where one can search for and discover relevant research data repositories from various academic disciplines. re3data.org lists hundreds of repositories and includes many data directories and data aggregators. The listed repositories vary widely in size and scope from archives such as the Macaulay Library[1] which is the largest scientific archive of biodiversity audio and video recordings collected worldwide (Cornell University 2016), to the HJ Andrews Experimental Forest[2] which hosts ecological, environmental and related research data that are primarily associated with a large forest research site in Oregon's Cascades Mountains in the U.S. Pacific Northwest (Andrews Experimental Forest LTER 2016), to the Dryad

[1] re3data.org: Macaulay Library; editing status 2014-06-25; re3data.org—Registry of Research Data Repositories. http://doi.org/10.17616/R3CS4N last accessed: 2016-01-14.

[2] re3data.org: HJ Andrews Experimental Forest; editing status 2015-05-28; re3data.org—Registry of Research Data Repositories. http://doi.org/10.17616/R3591T last accessed: 2016-01-14.

Digital Repository[3] which is a large general purpose, international, curated archive that holds data that underlie scientific and medical publications from hundreds of journals, professional societies and publishers (Dryad 2016).

In addition to variable size and scope, data repositories offer different approaches for discovering and acquiring data. HJ Andrews Experimental Forest data, for example, may be searched by: (1) using a simple “string search” where a word or phrase is entered; (2) “advanced search” where one can specify data associated with a particular researcher, a subset of theme keywords selected from a list and specific study sites that are also selected from a list; or by (3) browsing the list of all data products. Once one has identified data of interest, the metadata and other descriptive information can typically be downloaded immediately, but acquisition of the data requires that one register as a user and state the purpose for which the data are being requested.

Biodiversity audio and video recordings can be easily discovered in the Macaulay Library by searching catalog numbers or common names or species names of the organism(s) of interest. A search for “bluebird” generates a web page that includes a listing of available audio and video recordings and other information about the recordings including links to most of the recordings so they may be listened to or viewed. Acquiring the recordings requires that one license the media and place an order for the recordings that includes catalog number and/or species, a description of the recording, requested data format, and delivery details; use for research and education purposes is free, but commercial and other users may be required to pay a license fee and studio fee for preparing the media.

Data may be discovered in the Dryad repository via several mechanisms including simple text string search or a more advanced search that allows the user to narrow the results set by title, author, subject, publication date and publication name. For example, one may search for “wood density” and then narrow the search further by specifying “Ecology Letters” as the publication name which leads to a seminal paper by Chave and colleagues (2009) and the associated Dryad data package (Zanne et al. 2009); note that the lead author of the journal article and the Dryad data package are different individuals. The Dryad web page describes the contents of the data package (i.e., downloadable file names and file sizes, title, and other details) as well as links to the full metadata, the number of times the data package contents have been downloaded and instructions for citing both the journal article and the data package. The inclusion of a digital object identifier (DOI) in the data package citation makes it possible to easily link to the data from data package citations that are included in the Literature Cited sections of papers by other authors (e.g., Mascaro et al. 2012) that have cited the journal publication that is based on the data (i.e., Chave et al. 2009 in this case) and that have used and cited the data (i.e., Zanne et al. 2009).

[3] re3data.org: DRYAD; editing status 2015-11-18; re3data.org—Registry of Research Data Repositories. http://doi.org/10.17616/R34S33 last accessed: 2016-01-14.

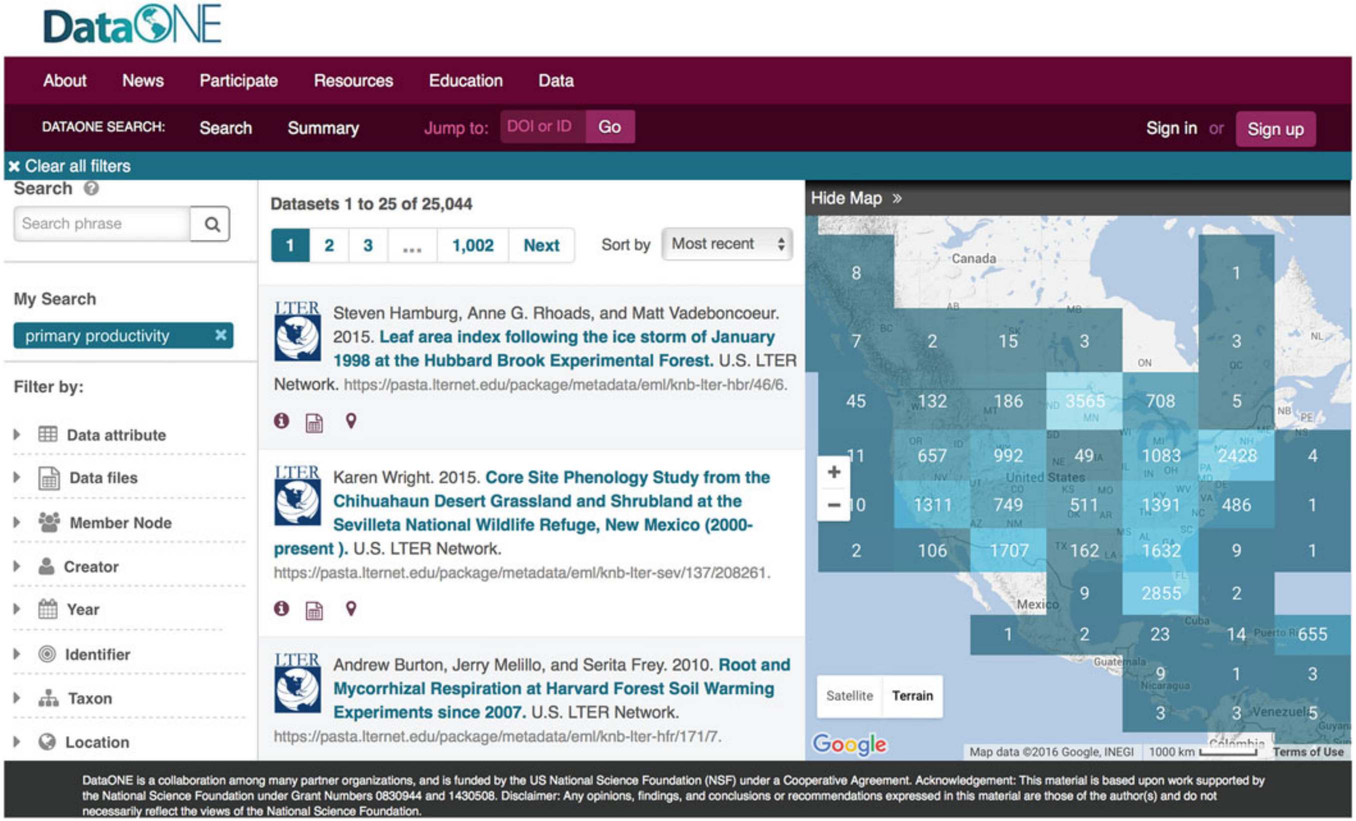
7.2.3 Data Directories

The U.S. National Aeronautics and Space Administration's (NASA) Global Change Master Directory (GCMD) makes it easy for scientists and the public to discover and access data relevant to climate change (NASA 2016). The GCMD contains descriptions of tens of thousands of data sets from the Earth and environmental sciences. One can perform searches of science keywords (e.g., atmosphere, biosphere, oceans, paleoclimate), instruments (e.g., Earth remote sensing instruments, in situ/laboratory instruments), platforms (e.g., aircraft, Earth observation satellites, in situ ocean-based platforms), locations (e.g., continent, geographic region, vertical location), providers (e.g., academic, government agencies, non-government organizations), project name or acronym, and free text. Searches lead to records that include project titles and brief abstracts and the records individually link to the more complete metadata file and, frequently, to the data.

The GCMD also provides access to authoring tools and other services that data and service providers can use to describe and facilitate discovery of their data products. Keyword vocabularies are central to the GCMD search capability and provide "controlled" lists of keywords that are accepted by the broader scientific community. The vocabularies enable data providers to describe their data products using standardized terms and are continually being expanded and revised.

The GCMD enables research organizations and other partners to create portals that support discovery of the portion of the GCMD content that is associated with a particular organization or partner (e.g., Antarctic Master Directory, World Water Forum). The GCMD also serves as one of NASA's contributions to the international Committee on Earth Observation Satellites (CEOS), through which it is named the CEOS International Directory Network (IDN) Master Directory (CEOS 2016). The IDN Master Directory provides links to numerous GCMD-associated portals.

DataONE is another related type of service that supports discovery of Earth and environmental science data (DataONE 2016). DataONE harvests and indexes metadata from a large international network of data repositories (Michener et al. 2011, 2012). It provides direct links to 100 s of thousands of data products that are stored in various repositories worldwide. To discover data, a researcher typically enters a keyword or phrase (e.g., "primary productivity") in the DataONE search bar which links to a visual display that enumerates the number of data products that exist in different geographic regions worldwide and includes more advanced search capabilities. The user can then easily narrow down the result set by searching for particular data attributes (e.g., density, length), repositories (e.g., data only from the Dryad Digital Repository), data creators, years, identifiers (e.g., DOIs), taxa (e.g., class, family), and locations (Fig. 7.1).



Steven Hamburg, Anne G. Rhoads, and Matt Vadeboncoeur. 2015. **Leaf area index following the ice storm of January 1998 at the Hubbard Brook Experimental Forest.** U.S. LTER Network. https://pasta.lternet.edu/package/metadata/eml/knb-lter-hbr/46/6.

Karen Wright. 2015. **Core Site Phenology Study from the Chihuahaun Desert Grassland and Shrubland at the Sevilleta National Wildlife Refuge, New Mexico (2000-present).** U.S. LTER Network. https://pasta.lternet.edu/package/metadata/eml/knb-lter-sev/137/208261.

Andrew Burton, Jerry Melillo, and Serita Frey. 2010. **Root and Mycorrhizal Respiration at Harvard Forest Soil Warming Experiments since 2007.** U.S. LTER Network. https://pasta.lternet.edu/package/metadata/eml/knb-lter-hfr/171/7.

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Fig. 7.1 The DataONE search interface showing search criteria (*left*), datasets matching "primary productivity" (*center*) and distribution of data sets in portions of North America (*right*)

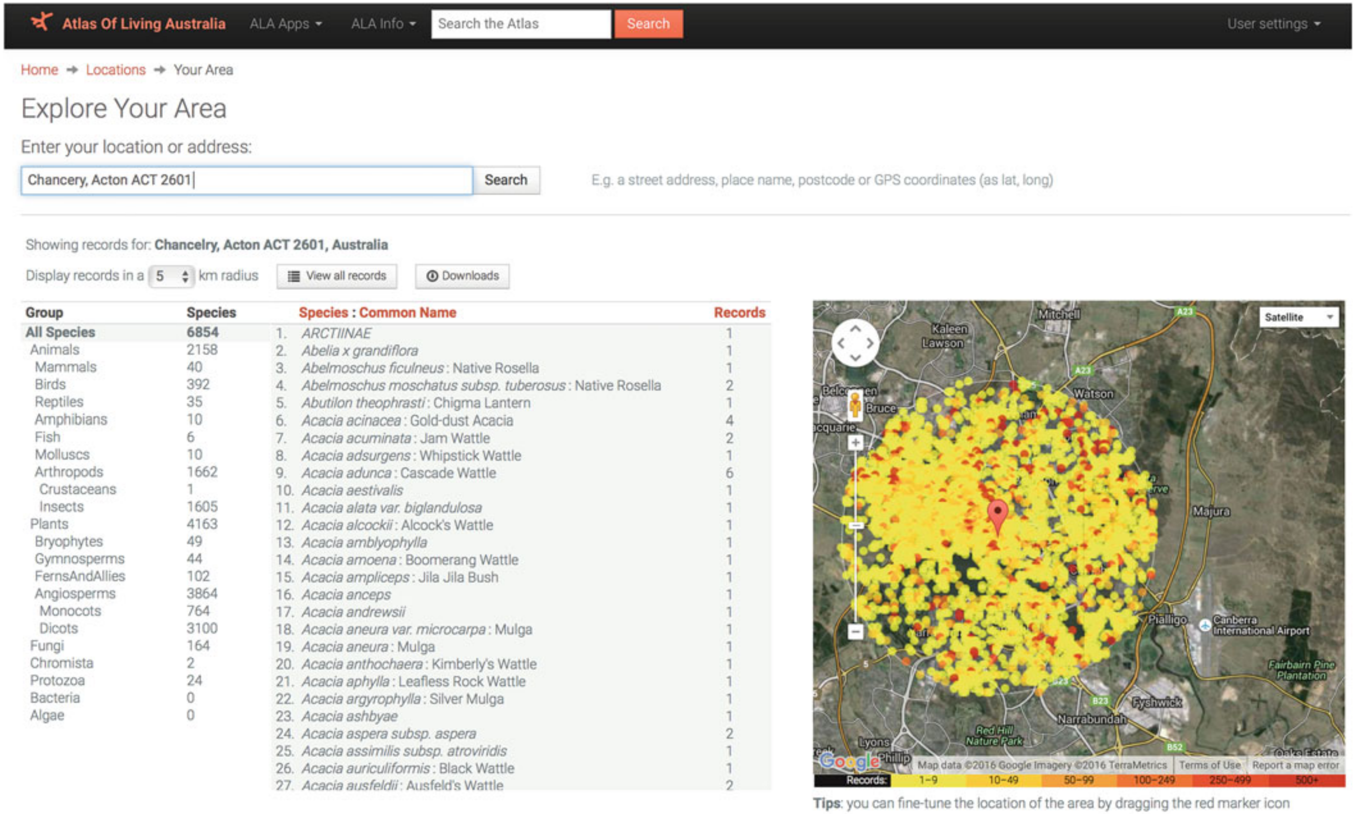
7.2.4 Data Aggregators

Data aggregation is the process whereby data are gathered from multiple sources and then, typically, presented in a standardized format to users. Some data aggregators perform minimal or no additional processing of the data whereas others provide numerous value-added services to benefit users. Value-added services can include data reformatting, performing quality assurance checks (e.g., duplicate detection, invalid entries), adding taxonomic or geographical location information, and providing statistical and graphical summaries. Those aggregators that provide significant value-added services often work with specific types of data such as meteorological data or data pertaining to particular groups of organisms.

Data aggregators are typically grouped in with data repositories (e.g., listed in the Registry of Research Data Repositories; re3data.org Project Consortium 2016) although they differ from most repositories with respect to the services provided. For instance, a typical institutional data repository may archive a wide range of data from a large number of contributors; services may be limited to activities such as the provision of a metadata entry tool, addition of a DOI, citation recommendations, and periodic backup. A data aggregator, on the other hand, may accept limited types of data that are in one or a small number of specific formats; the aggregator may then further process the data by summarizing the data, adding additional metadata descriptors, and so on. The examples below highlight a subset of non-commercial data aggregators indicating the types of data they aggregate and some of the services that are provided.

The Atlas of Living Australia (ALA 2016) is an online repository that contains data and information about Australia's plants, animals and microbes (e.g., species occurrence records, photos, sound recordings, maps, molecular data and links to pertinent literature). ALA aggregates records and datasets submitted from thousands of sources including citizens, governmental agencies and other groups. It provides access to keys as well as tools that enable data and metadata to be entered in standardized formats. In addition, a variety of value added services and features are provided including: (1) a spatial portal that allows one to view and create maps that show species occurrences relative to climate and numerous other features; (2) "fishmap" which allows one to find Australia's marine fishes; (3) "Explore Your Area" which allows one to see all species within a user-specified radius of your home location (Fig. 7.2); and (4) a "dashboard" that provides updates on numbers of occurrence records and datasets, records submitted by institution/data provider, conservation status, and numbers of records by state and territory, date and taxonomic grouping.

The Advanced Ecological Knowledge and Observation System (ÆKOS) Data Portal (ÆKOS 2016) is a data portal that allows one to discover data about Australian plants, animals and their environment (Fig. 7.3). The portal provides detailed information about the research methods employed to facilitate understanding and reuse of the data; such information is associated with various icons that



Explore Your Area

Enter your location or address:

Chancery, Acton ACT 2601| Search

E.g. a street address, place name, postcode or GPS coordinates (as lat, long)

Showing records for: **Chancelry, Acton ACT 2601, Australia**

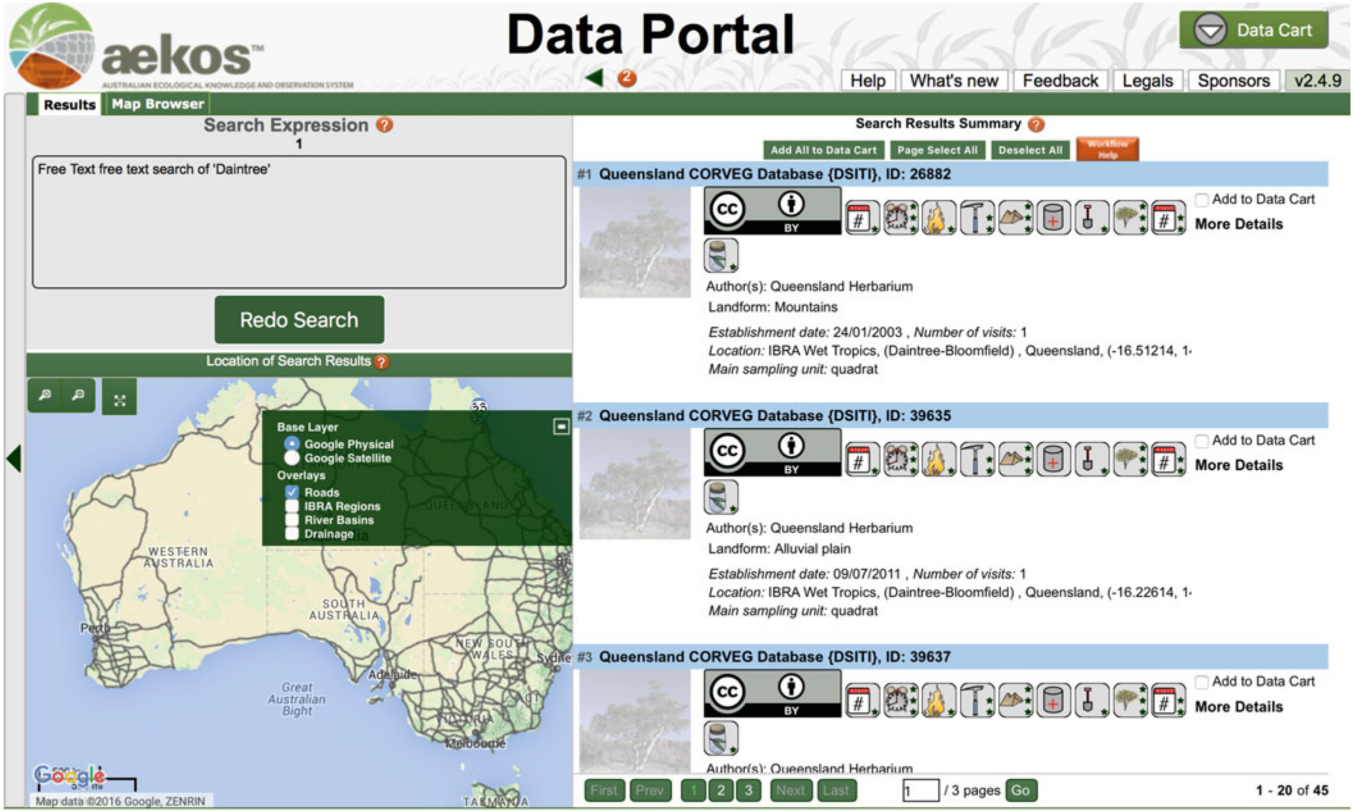
Display records in a 5 km radius

Group	Species
All Species	**6854**
Animals	2158
Mammals	40
Birds	392
Reptiles	35
Amphibians	10
Fish	6
Molluscs	10
Arthropods	1662
Crustaceans	1
Insects	1605
Plants	4163
Bryophytes	49
Gymnosperms	44
FernsAndAllies	102
Angiosperms	3864
Monocots	764
Dicots	3100
Fungi	164
Chromista	2
Protozoa	24
Bacteria	0
Algae	0

Species : Common Name	Records
1. *ARCTIINAE*	1
2. *Abelia x grandiflora*	1
3. *Abelmoschus ficulneus* : Native Rosella	1
4. *Abelmoschus moschatus subsp. tuberosus* : Native Rosella	2
5. *Abutilon theophrasti* : Chigma Lantern	1
6. *Acacia acinacea* : Gold-dust Acacia	4
7. *Acacia acuminata* : Jam Wattle	2
8. *Acacia adsurgens* : Whipstick Wattle	1
9. *Acacia adunca* : Cascade Wattle	6
10. *Acacia aestivalis*	1
11. *Acacia alata var. biglandulosa*	1
12. *Acacia alcockii* : Alcock's Wattle	1
13. *Acacia amblyophylla*	1
14. *Acacia amoena* : Boomerang Wattle	1
15. *Acacia ampliceps* : Jila Jila Bush	1
16. *Acacia anceps*	1
17. *Acacia andrewsii*	1
18. *Acacia aneura var. microcarpa* : Mulga	1
19. *Acacia aneura* : Mulga	1
20. *Acacia anthochaera* : Kimberly's Wattle	1
21. *Acacia aphylla* : Leafless Rock Wattle	1
22. *Acacia argyrophylla* : Silver Mulga	1
23. *Acacia ashbyae*	1
24. *Acacia aspera subsp. aspera*	2
25. *Acacia assimilis subsp. atroviridis*	1
26. *Acacia auriculiformis* : Black Wattle	1
27. *Acacia ausfeldii* : Ausfeld's Wattle	2

Tips: you can fine-tune the location of the area by dragging the red marker icon

Fig. 7.2 The Atlas of Living Australia repository website illustrating the tool "Explore Your Area" which lists and pinpoints the locations of species within a user-specified radius of a particular location.



Search Expression 1

Free Text free text search of 'Daintree'

#1 Queensland CORVEG Database {DSITI}, ID: 26882

Author(s): Queensland Herbarium
Landform: Mountains
Establishment date: 24/01/2003 , *Number of visits:* 1
Location: IBRA Wet Tropics, (Daintree-Bloomfield) , Queensland, (-16.51214, 1-
Main sampling unit: quadrat

#2 Queensland CORVEG Database {DSITI}, ID: 39635

Author(s): Queensland Herbarium
Landform: Alluvial plain
Establishment date: 09/07/2011 , *Number of visits:* 1
Location: IBRA Wet Tropics, (Daintree-Bloomfield) , Queensland, (-16.22614, 1-
Main sampling unit: quadrat

#3 Queensland CORVEG Database {DSITI}, ID: 39637

Author(s): Queensland Herbarium

Fig. 7.3 The result set obtained by performing a simple search for "Daintree" of the Advanced Ecological Knowledge and Observation System (ÆKOS) Data Portal. Each resulting dataset (*right panel*) is accompanied by icons that represent conditions of use, the types of variables included in the data, duration of the study, and research methods employed

accompany each data set discovered during a user's search. The portal includes numerous features and services that support researchers, educators and resource managers. One can search by location and ecological data themes, and create complex Boolean searches. Figure 7.3 illustrates the result set obtained by performing a simple search for "Daintree." Each resulting dataset is accompanied by icons that represent conditions of use, the types of variables included in the data, duration of the study, and research methods employed. By selecting "More Details", one is taken to a webpage where: (1) an "Observation Diagram" provides a visual representation of the types of observations that are recorded; (2) a "Methods Diagram" that similarly illustrates the sampling methods that are employed with links to methodological details; and (3) "Metadata" where detailed information about all aspects of the data is available. If the data appear suitable, then a user can easily download the data in .csv format.

VertNet (2016) aggregates a wide variety of vertebrate biodiversity data from natural history collections worldwide and provides tools that facilitate data discovery, acquisition and publication (Constable et al. 2010; Guralnick and Constable 2010). VertNet has integrated, standardized and "cleaned" data derived from previously existing vertebrate data aggregators [i.e., Mammal Networked Information System (MaNIS 2016); Ornithological Information System (ORNIS 2016); HerpNET (2016); and FishNet2 (2016)]. VertNet supports publication, indexing, and georeferencing of data and provides training as well as a clear and concise set of norms for data use and publication.

7.3 Best Practices for Promoting Data Discovery and Reuse

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). Data organization, data quality, metadata and data preservation were discussed in detail in Porter (2017), Michener (2017a, b) and Cook et al. (2017), respectively.

Good data archiving and sharing policies promote long-term discoverability and accessibility of data and do so in a way that benefits both the data producers and consumers (Duke and Porter 2013; Whitlock et al. 2016). The following discussion focuses on simple steps that can be taken to ensure that data products and scientific code can be easily discovered, reused and cited.

7.3.1 Data Products

Data Products Should Be Uniquely Identifiable and Attributable to Their Originators "Consumers" or users of data benefit from knowing that a data product exists, that it can be used and cited, that the data originators receive proper attribution, and that others can subsequently discover and use the same data product (e.g., for research transparency and data verification purposes). "Producers" or originators of data benefit from having their data products cited and used by others much like the peer-recognition that is associated with having publications cited by others in the literature.

Persistent Identifiers (PIDs) have emerged as the principal mechanisms to provide a long-lasting reference to datasets and other digital resources. PIDs make it easy to uniquely cite and access research data and other digital resources. Some of the more common PIDs include Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), Life Science Identifiers (LSIs), and Universal Resource Names (URNs). DOIs are increasingly becoming the norm for citing all types of digital resources and various organizations have emerged to facilitate the creation and management of DOIs. Crossref (2016), for example, commonly provides DOIs for journal articles, books, reports and datasets.

Likewise, DataCite (2016) creates and supports standards for PIDs for data and other digital resources. DataCite member institutions are globally distributed data centers, national libraries, universities and other organizations that serve users by assigning DOIs to data and other objects. DataCite provides specific recommendations for how to cite data (Box 7.1). DataCite also provides various tools for users such as (1) DOI Citation Formatter which creates different citation formats for DataCite and Crossref DOIs; (2) Metadata Search Tool that allows one to search the metadata of datasets registered with DataCite; and (3) Metadata Stats service that provides statistics on datasets that have been uploaded and accessed.

Box 7.1 DataCite Recommendations for Data Citation

The DataCite "recommended format for data citation is as follows:

Creator (PublicationYear): Title. Publisher. Identifier.

It may also be desirable to include information about two optional properties, Version and ResourceType (as appropriate). If so, the recommended form is as follows:

Creator (PublicationYear): Title. Version. Publisher. ResourceType. Identifier".

(http://www.datacite.org/); accessed 20 Jan 2016).

Many data repositories work with DataCite member institutions to assign DOIs and also provide specific guidelines for citing datasets that are housed in their repository. For example, the dataset citation recommendations for the Dryad Digital Repository are listed in Box 7.2.

Box 7.2 Dryad Digital Repository Data Citation Recommendations

"When referencing data in the text, we recommend the following as a template (substitute your DOI suffix for the xxxxx):

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.xxxxx

In the Bibliography, we recommend a citation similar to:

Heneghan C, Thompson M, Billingsley M, Cohen, D (2011) Data from: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.585t4".

(https://datadryad.org/pages/faq; accessed 20 Jan 2016).

Data Originators and Users Should Be Uniquely Identifiable It is highly unlikely that any given personal name can be resolved to a single individual; further, some common names like "J. Smith" may be associated with thousands of individuals and many researchers undergo name changes over the course of their careers (e.g., through marriage). This situation presents a real challenge when the goal is to make sure that individuals receive proper attribution for the output of their scholarly and research endeavors. ORCID (Open Research and Contributor ID; ORCID 2016) provides a valuable service that enables researchers to be uniquely identified. ORCID identifiers are unique alphanumeric codes that resolve to a specific individual and can be easily linked to publications, grant proposals, and other outputs and activities. The ORCID organization maintains a registry of unique researcher identifiers and supports mechanisms that enable researchers to link their identifiers to research products. Increasingly, research sponsors and publishers are encouraging or requiring that individuals associate their works with an ORCID identifier.

7.3.2 Scientific Code

Scientific code such as custom software and scripts (e.g., R, Matlab) is used in statistical and graphical analysis, modeling, detecting and correcting errors in data, and creating figures and visualizations. Code precisely records what has been done with the data and the availability of code makes it possible for other scientists to more easily understand and, potentially, reproduce data processing and analytical steps (Maslan et al. 2016; Peng 2011; Barnes 2010; Ince et al. 2012). It is good practice to deposit scientific code in long-term repositories such as Dryad, Figshare, PANGAEA, or Zenodo that provide licenses (e.g. CC0, CC-By) and that assign DOIs so that code is preserved and may be used and properly cited by others (Maslan et al. 2016).

References

ÆKOS (2016) ÆKOS: Advanced Ecological Knowledge and Observation System Data Portal. http://www.aekos.org.au/home. Accessed 22 Apr 2016

ALA (2016) Atlas of Living Australia. http://www.ala.org.au. Accessed 22 Apr 2016

Andrews Experimental Forest LTER (2016) HJ Andrews experimental forest long term ecological research. http://andrewsforest.oregonstate.edu. Accessed 22 Apr 2016

Barnes N (2010) Publish your computer code: it is good enough. Nature 467:753

CEOS (2016) CEOS Committee on Earth Observation Satellites. http://ceos.org/ourwork/workinggroups/wgiss/current-activities/idn/. Accessed 22 Apr 2016

Chave J, Coomes DA, Jansen S et al (2009) Towards a worldwide wood economics spectrum. Ecol Lett 12:351–366. doi:10.1111/j.1461-0248.2009.01285.x

Constable H, Guralnick R, Wieczorek J et al (2010) VertNet: a new model for biodiversity data sharing. PLoS Biol 8:e1000309. doi:10.1371/journal.pbio.1000309

Cook RB, Wie Y, Hook LA et al (2017) Preserve: protecting data for long-term use, Chapter 6. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Cornell University (2016) The Cornell Lab of Ornithology Macaulay Library. http://macaulaylibrary.org/. Accessed 22 Apr 2016

Crossref (2016) Crossref.org. http://www.crossref.org. Accessed 22 Apr 2016

DataCite (2016) DataCite. https://www.datacite.org. Accessed 22 Apr 2016

DataONE (2016) DataONE: Data Observation Network for Earth. http://dataone.org. Accessed 22 Apr 2016

Dryad (2016) Dryad. http://datadryad.org. Accessed 22 Apr 2016

Duke CS, Porter JH (2013) The ethics of data sharing and reuse in biology. BioSci 63:483–489

FishNet2 (2016) FishNet2. http://www.fishnet2.net/aboutFishNet.html. Accessed 22 Apr 2016

Guralnick R, Constable H (2010) VertNet: creating a data-sharing community. BioSci 60:258–259. doi:10.1525/bio.2010.60.4.2

HerpNet (2016) HerpNet. http://herpnet.org. Accessed 22 Apr 2016

Ince DC, Hatton L, Graham-Cumming J (2012) The case for open computer programs. Nature 482:485–488

MaNIS (2016) MaNIS: Mammal Networked Information System. http://manisnet.org. Accessed 22 Apr 2016

Mascaro J, Hughes RF, Schnitzer SA (2012) Novel forests maintain ecosystem processes after the decline of native tree species. Ecol Monogr 82:221–228

Maslan KAS, Heer JM, White EP (2016) Elevating the status of code in ecology. Trends Ecol Evol 31:4–7. doi:10.1016/j.tree.2015.11.006

Michener WK (2015) Ecological data sharing. Ecol Inf 29:33–44. doi:10.1016/j.ecoinf.2015.06.010

Michener WK (2017a) Quality assurance and quality control (QA/QC), Chapter 4. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Michener WK (2017b) Creating and managing metadata, Chapter 5. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

Michener W, Vieglais D, Vision T et al (2011) DataONE: Data Observation Network for Earth – preserving data and enabling innovation in the biological and environmental sciences. D-Lib Mag17 (Jan/Feb 2011). doi:10.1045/january2011-michener

Michener WK, Allard S, Budden A et al (2012) Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. Ecol Inf 11:5–15

NASA (2016) NASA Global Change Master Directory. http://gcmd.nasa.gov/. Accessed 22 Apr 2016

ORCID (2016) ORCID. http://orcid.org. Accessed 22 Apr 2016

ORNIS (2016) ORNIS. http://www.ornisnet.org. Accessed 22 Apr 2016

Peng RD (2011) Reproducible research in computational science. Science 334:1226–1227

Porter JH (2017) Scientific databases for environmental research, Chapter 3. In: Recknagel F, Michener W (eds) Ecological informatics. Data management and knowledge discovery. Springer, Heidelberg

re3data.org Project Consortium (2016) re3data.org Registry of Research Data Repositories. http://www.re3data.org. Accessed 22 Apr 2016

VertNet (2016) VertNet. http://www.vertnet.org/. Accessed 22 Apr 2016

Whitlock MC, Bronstein JL, Bruna EM et al (2016) A balanced data archiving policy for long-term studies. Trends Ecol Evol 31(2):84–85. doi:10.1016/j.tree.2015.12.001

Zanne AE, Lopez-Gonzalez G, Coomes DA, et al (2009) Data from: towards a worldwide wood economics spectrum. Dryad Digital Repository. doi:10.5061/dryad.234