The Journal of Chulabhorn Royal Academy
eISSN 2697-5203 (online)

Invited article

Dataset Search in Research Data Services

Namtip Wipawin

Department of Information Science, School of Liberal Arts, Sukhothai Thammathirat Open University
namtipstou@gmail.com

Received: 9 December 2019; Revised: 23 December 2019; Accepted: 25 December 2019

Abstract

Research data services are services organized in data repositories that are provided by libraries and research centers, for scientists and researchers to find, reuse, and publish data. This paper outlines the types of dataset search used in research data services, and the core criteria for research data services defined by the CoreTrustSeal international standard. In this study, the types of dataset search include search engines, national data services, institutional repositories, and subject repositories. Examples of the four types of research data services are DataSearch, UK Data Service, DataFirst, and OpenICPSR. For research data services to satisfy the core criteria of international standards, they should provide a research data policy, data management plan, data catalog with metadata, data repository, and persistent identifiers, such as digital object identifiers. Datasets are easier to find when they provide structured data in a data catalog. For high-quality research data services, every organization needs to organize a data catalog to help data analysts find and understand data.

Keywords: Dataset Search, Research Data Services

Introduction

The trend in international research communities of scientists and researchers is to publish the final reports of research work as scientific articles, and to share research products—which are the outputs of scientific research—in the form of data, code, workflows, and slide presentations, which we call **datasets**. Datasets are easier to find when research data services (RDSs) have organized metadata in a data catalog[1]. Datasets can then be searched via open data portals for RDSs provided by libraries and research centers. The datasets can be analyzed by machine learning techniques and displayed visually[2]. Scientists and researchers can reuse data from an open data portal free of charge, by acknowledging the source; examples include the EU open data portal and World Bank open data. The OPEN Government Data Act and the Data Sharing Act require data.gov to publish their information online as open data, using standardized, machine-readable data formats, with their metadata included in the data.gov catalog. According to OECD Index data, open government data refers to a set of policies that promotes transparency, accountability, and value creation by making government data available to all. By making datasets available, public institutions become more transparent and accountable to citizens.

© Chulabhorn Royal Academy. Terms and conditions of Creative Commons Attribution 4.0 International (CC BY NC ND 4.0) apply

There is a need to encourage government data centers to manage national data services (NDSs) and RDSs according to international standards, such as CoreTrust-Seal and the Trustworthy Repository ISO 16363. An open data portal is able to provide datasets if there is a consortium with NDSs and RDSs from various research disciplines and academic institutions[3]. The best practices of government open data portals can be found from the Global Open Data Index (https://index.okfn.org/). Ranks 1–5 are occupied by Taiwan, Australia, Great Britain, France, Finland, Canada, and Norway. Thailand is ranked 51st according to the Global Open Data Index in 2019.

According to the data catalog vocabulary[4], a dataset is defined as a collection of data, published or curated by a single agent, and available for access or download in many formats. A dataset is a conceptual entity, and can be represented by one or more distributions that serialize the dataset for transfer. Distributions of a dataset can be provided via data services. A data service typically provides selection, extraction, combination, processing, or transformation operations over datasets that might be hosted locally to the service or remotely. The basic idea is that, to be discovered, each dataset or object should have its own persistent identifier, such as a digital object identifier (DOI) or open researcher and contributor identifier (ORCID). Therefore, RDSs are services that are organized in institutional data repositories, and provided by libraries and research centers for scientists and researchers across all disciplines to find, reuse, and publish data. A dataset can be considered to contain research data if the data are derived from the research process: it contains any information that has been collected, observed, generated, or created to validate original research findings. Research data exists as digital files in many forms, such as documents, spreadsheets, laboratory notebooks, field notebooks, diaries, questionnaires, transcripts, codebooks, audiotapes, videotapes, photographs, films, slides, artefacts, specimens, samples, collections of digital outputs, data files, database contents (video, audio, text, images), models, algorithms, and scripts. Researchers and scientists can use research data from many sources in different disciplines' repositories to support their research analysis and findings.

RDSs also include open data portals at the national level, institutional level, and subject (or discipline) level. At the national level, an NDS is a service organization that builds on data archiving and data sharing concepts for scientists and researchers across all disciplines, to find, reuse, and publish data. Building the collection of datasets, NDS services should have a strong connection to publications and the publishing process, to ensure robust links between published literature and the data. Many countries build an NDS for scientists and researchers across all disciplines to find, reuse, and publish data. The communities of the National Data Service Consortium are collaborating with the Research Data Alliance to ensure transparent access to data around the world. At the institutional level, an RDS is a service with an open data portal that consists of four stages: planning, acquisition, publishing and preservation, and analysis (https://library.princeton.edu/research-data). The planning process starts with writing a data management plan (DMP) and considers which types of data are to be shared in the large dataset. The acquisition process prepares a data use agreement, data security, and data documentation, and finds datasets to work with. Publishing and preservation consider the publisher's requirement to publish the datasets and the DOI to be used. The analysis process considers how to deal with large

datasets, statistics, and the option to share active data.

This article is intended to help researchers to find datasets that relate to their research work for data analysis and for expanding the scope of innovative research questions. It is also intended to help librarians and information scientists identify the core criteria of RDSs according to international standards. The benefit of building research infrastructure by providing research data is that it will lead to the readiness of data repositories to apply FAIR data principles, for Internet of FAIR Data and Services, in the near future[5,6].

Literature Review

Tenopir et al.[7] found that library directors strongly agree on the importance of RDSs. In addition, according to studies of academic libraries in North America, and European libraries are offering, or are planning to offer, consultative or reference RDSs than technical or hands-on RDSs. The majority of libraries provide support for training in skills related to RDS for their staff members. Almost all libraries collaborate with other organizations inside their institutions, or with outside institutions, to offer or develop policy related to RDS.

In addition, Cox et al.[8] presented the results of a survey covering higher education libraries in Australia, Canada, Germany, Ireland, the Netherlands, New Zealand, and the UK. It was found that libraries have provided leadership in RDM, particularly in advocacy and policy development. Service development is still limited, and is focused particularly on advisory and consultancy services (such as data management planning support and data-related training), rather than technical services (such as provision of a data catalog, and curation of active data).

Therefore, research data policy and open data policy are strongly promoted in European countries, as shown by the analysis of the European Research Area (ERA) Progress Report 2016 Country Snapshots[9] (SPARCeurope 2017). Science Europe which is an association of European Research Funding Organizations and the Netherlands Organization for Scientific Research (NWO) launched an initiative for research data management (RDM) policy in 2018. The Trustworthy Data Repository Certification Standards were developed by the international data communities; these standards include the World Data System (WDS) Certification in 2011, the Audit and Certification of Trustworthy Digital Repositories (ISO 16363) in 2012, the Data Management Maturity (DMM) Framework (AGU and CMMI) in 2015, and the Core Trustworthy Data Repositories Requirements in 2016. The Data Seal of Approval (DSA), the World Data System (WDS), DIN 31644, Core Trust Seal (CTS) and ISO 16363 contain criteria in three groups of requirements in organization infrastructure, digital object management, and security risk management infrastructure. In 2015, DSA and WDS shared the Unified Requirements for Core Certification of Trustworthy Data Repositories through a DSA-WDS partnership working group within the Research Data Alliance (RDA).

Digital preservation and trustworthiness are growing interests in the audit and certification of digital repositories. Digital repositories in universities are a set of services that universities offer to members of their communities, as digital archives of the intellectual products, or the storage of knowledge assets, of the institutions. Previous studies in Thailand showed that collections of theses, research reports, and articles were organized by the libraries and research centers of the universities[10,11].

Three organizations at the national level provide RDSs:

1) The Thai Library Integrated System (ThaiLIS), which is run by the Department of University Affairs and Thai Library Integrated System Project; the research database is called Thai Digital Collection (TDC).

2) The Thai National Research Repository (TNRR), which was developed by the National

Research Council of Thailand (NRCT) and Thailand Research Organizations Network (TRON). The latter includes members of various organizations, such as the National Science and Technology Development Agency (NSTDA), Thailand Research Fund (TRF), National Science Technology and Innovation Policy Office (STI), Health Systems Research Institute (HSRI), and Agriculture Research Development Agency (ARDA).

3) Thai Journals Online (ThaiJO), hosted by the Thai-Journal Citation Index Centre (TCI) at King Mongkut University of Technology, Thonburi: an e-Journal system consisting of more than 700 journals in Thailand.

Wipawin[12,13] also found that most Thai Institutional Repositories (IRs) provide research outputs, such as reports and articles, that are not ongoing research data. Because of the lack of institutional policy on open research data, this can limit the accessibility of research and thesis data in institutions. The need for metadata and persistent identifiers also affects long-term data preservation in IRs. Most Thai university libraries provide full-text electronic theses and dissertations for the ThaiLIS online public access catalog (OPAC). IRs and most libraries submit thesis files to ThaiLIS databases under the Thai university library consortium, the Department of University Affairs, and the ThaiLIS Project. The TNRR also provides links to research abstracts and reports so that users have access to each database.

Objectives and methodologies

This article aims to classify the types of dataset search used in RDSs, and to identify the core criteria of RDSs according to the CoreTrustSeal international standard. The samples are four RDSs—DataSearch, UK Data Service, DataFirst, and OpenICPSR—which are analyzed using documentation research and the content analysis method. The findings are described below.

1. The types of dataset search in research data services

In the public domain, four types of dataset search can be found in RDSs: search engine, NDSs, institutional repositories, and subject repositories, as shown in Table 1.

Table 1: Types of dataset search in RDSs.

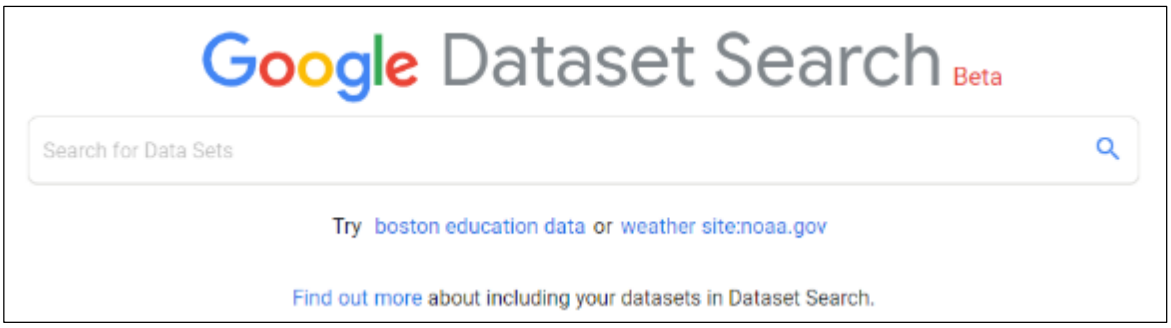
Search engines	NDSs	Institutional Repositories	Subject Repositories
Google Dataset Search	UK Data Service	DataFirst	OpenICPSR
DataSearch by Elsevier	World Bank Data Service	DSpace@MIT	DataDRYAD.org
	Swedish Data Services	4TU.nl	World Data Center for Climate
	Research Data Australia	Harvard Dataverse	Databrary
	Research Data Canada	CUIR	Archaeology Data Services

Table 1: Types of RDSs include search engine, NDSs, institutional repositories, and subject repositories.

1.1 Search engines

Search engines for datasets can be found at Google Dataset Search, which was launched in 2018 by Natasha Noy, a research scientist at Google AI; its aim was to unify the tens of thousands of different repositories for datasets online. Institutions that publish their data online, such as universities and governments, need to include, in their web pages, metadata tags that describe their data, including who created it, when it was published, and how it was collected. This information is then indexed by Dataset Search and combined with input from Google's Knowledge Graph.

Another search engine for data is DataSearch from the Elsevier publishing company, which was launched in 2016 (https://datasearch.elsevier.com). This is a data search engine that allows scientists and researchers to search for many different data types and formats across a variety of domain-specific and cross-domain institutional data repositories and other data sources, such as Mendeley Data, Dryad, Zenodo, Harvard Dataverse, NeuroElectro, The Inter-University Consortium for Political and Social Research (ICPSR), ThermoML at NIST Thermodynamic Research Center (TRC), PANGAEA, EarthChem Portal, ScienceDirect, and ArXiv. DataSearch indexes both metadata and data, to facilitate the matching of queries to objects described in the research.



Source: https://toolbox.google.com/datasetsearch



Source: https://datasearch.elsevier.com/#/

1.2 National Data Services

NDSs are organized as data repositories by national and international data centers, such as World Bank Data Services, UK Data Service, Swedish Data Services, Research Data Australia, and Research Data Canada. Some examples are shown below.

World Bank Data Service
(https://data.worldbank.org/)
Comprises of Data Bank, Microdata Library and Open Data Catalog which provides datasets according to the datatypes in the Data Catalog such as Geospatial or GIS data, Microdata or Survey data and Time Series data.

The UK Data Service
(https://www.ukdataservice.ac.uk/)
Provides datasets from survey data, longitudinal data, aggregate data from national and international databases such as IMF, World Bank, UNESCO, OECD and other Statistical agencies.

Research Data Australia
(https://researchdata.ands.org.au/)
Providers more than 100,000 datasets from data collections in over one hundred Australian research organizations in many fields of research

Swedish National Data (SND)
(https://snd.gu.se/en)
Provides datasets in accordance with the FAIR data principles from more than 30 Swedish higher education institutions and research institutes which called SND Network.

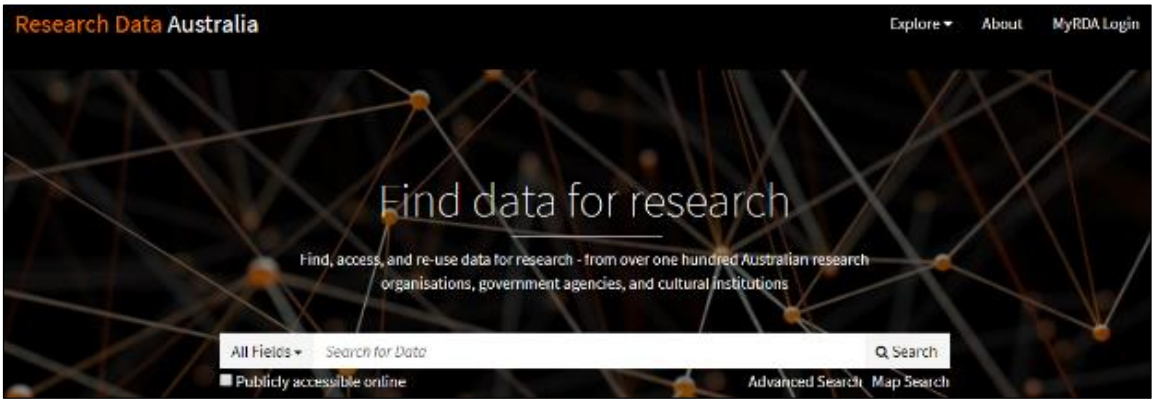
Research Data Canada
(https://www.rdc-drc.ca/)
Provides datasets from Statistics Canada microdata in the Research Data Centers Network such as 29 universities in Canada.



Source: https://data.worldbank.org/



Source: https://www.ukdataservice.ac.uk/



Source: https://researchdata.ands.org.au/



Source: https://snd.gu.se/en



Source: https://www.rdc-drc.ca/

1.3 Institutional Repositories

Institutional Repository (IRs) can be defined as a digital repository or digital archive of the university publications such as thesis, dissertations, research articles, project reports and slide presentations. It is the intellectual assets of the academic or research institutions. IRs can be searched from OpenDOAR (Directory of Open Access Journals) and re3data (Registry of Research Data Repositories)

4TU.ResearchData
(http://data.4tu.nl)
Provides datasets from 45 Open Access Collections from the university consortium in the Netherlands in the fields of science, engineering and technology.

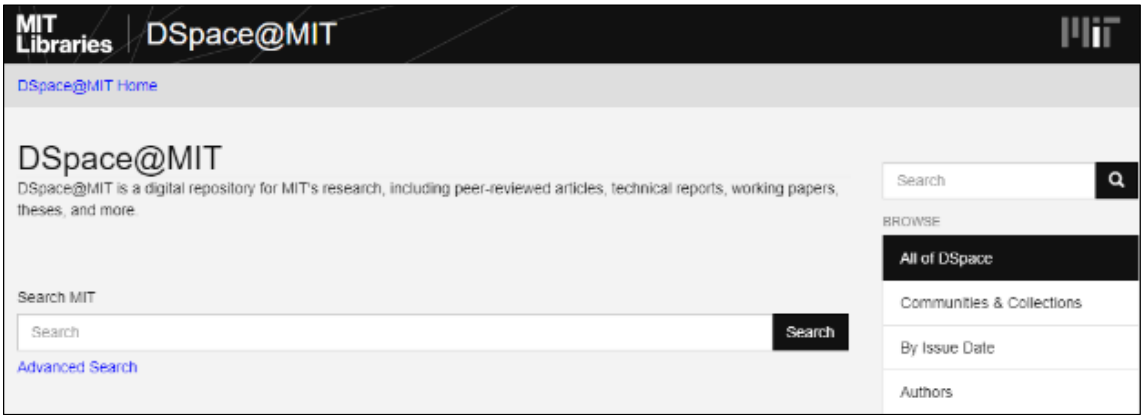
DSpace@MIT
(https://dspace.mit.edu/)
DSpace is the Open Access Collection of scholarly articles by MIT authors and related publishers. From 2009-2019, there are more than 100 million article downloads from MIT open access articles, MIT Thesis, MIT Courseware and other research laboratories.

Harvard Dataverse
(https://dataverse.harvard.edu/)
provides more than 93,000 datasets from Open Access repositories in dataverses from many countries in many fields. There is the Global Dataverse Community Consortium which supports dataverse repositories around the world.

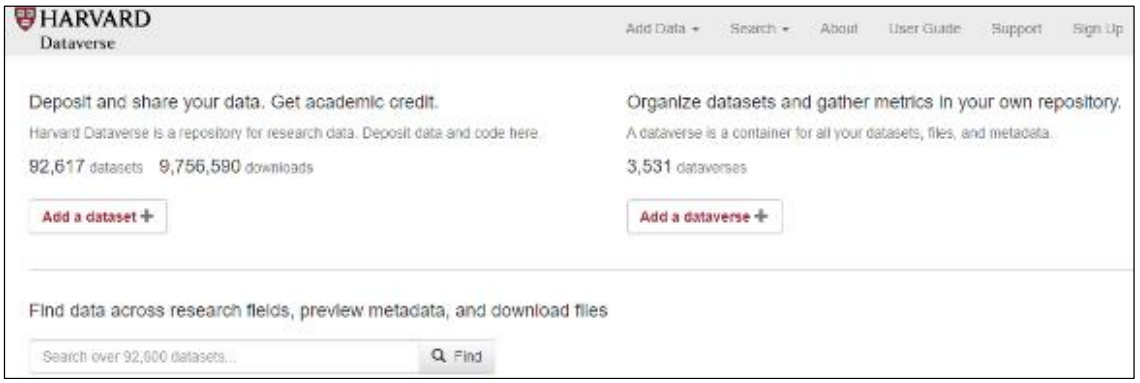
CUIR
(https://cuir.car.chula.ac.th/)
Chulalongkorn University Intellectual Repository (CUIR) provides full text of thesis, dissertations, academic papers, and other learning materials from faculties at Chulalongkorn University, Thailand.



Source: http://data.4tu.nl



Source: https://dspace.mit.edu/



Source: https://dataverse.harvard.edu/



Source: https://cuir.car.chula.ac.th/

1.4 Subject Repositories

Subject repositories can be defined as discipline repositories or Open Access subject repositories of scholarly articles, thesis, area dissertations and other working papers in a particular subject area from any institution such as AgEcon search (Research in Agricultural & Applied Economics) and arXiv.org (e-prints in the Fields of physics, mathematics, computer science and science)

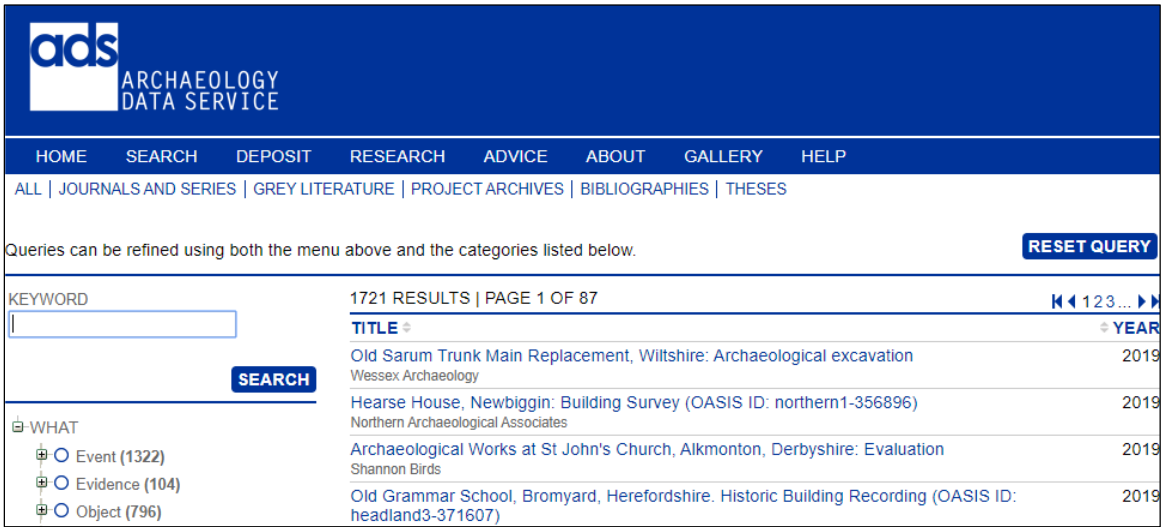
Archaeology Data Services (ADS)
(https://archaeologydataservice.ac.uk/)
ArchSearch provides datasets in digital research of historic archives and databases in United Kingdom. There are 282 ADS collections this data service with the partnership of Natural Environment Research Council and Marine Environmental Data and Information Network.

DataDRYAD
(https://datadryad.org/stash)
provides datasets from digital repositories in scientific and medical publications. All datasets in DataDRYAD are Indexed by Web of Science, Data Citation Index, SCOPUS and Google Dataset Search.

World Data Center for Climate
(https://www.dkrz.de/up/systems/wdcc)
provides datasets from ICSU World Data System, WDC Cluster Earth System Research, CERA-2 Database, WDCC Archiving workflow and other research projects in Germany.

DataBrary
(https://nyu.databrary.org/)
DataBrary are digital repository for sharing video data in the developmental and educational research communities. Databrary is complemented by Datavyu (and open source video-coding software) and Labnanny (a data management system for data-sharing)

OpenICPSR
(https://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp)
OpenICPSR is a data archive of an international consortium of more than 700 academic institutions. It contains 16 collections of data in an international consortium of more than 700 academic institutes and research organizations.



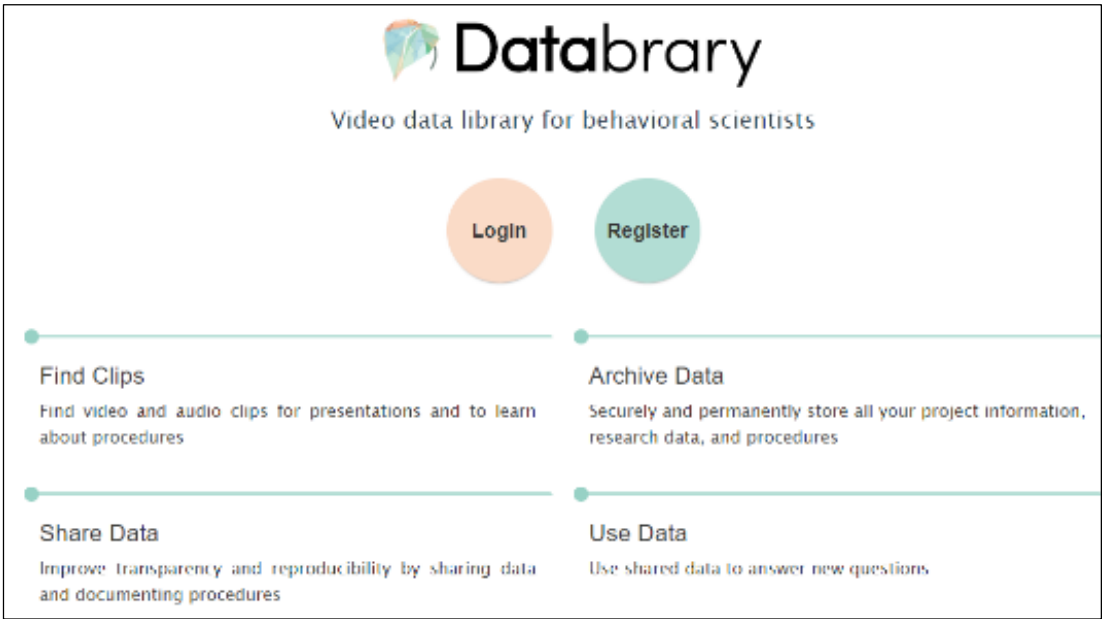
Source: https://archaeologydataservice.ac.uk/



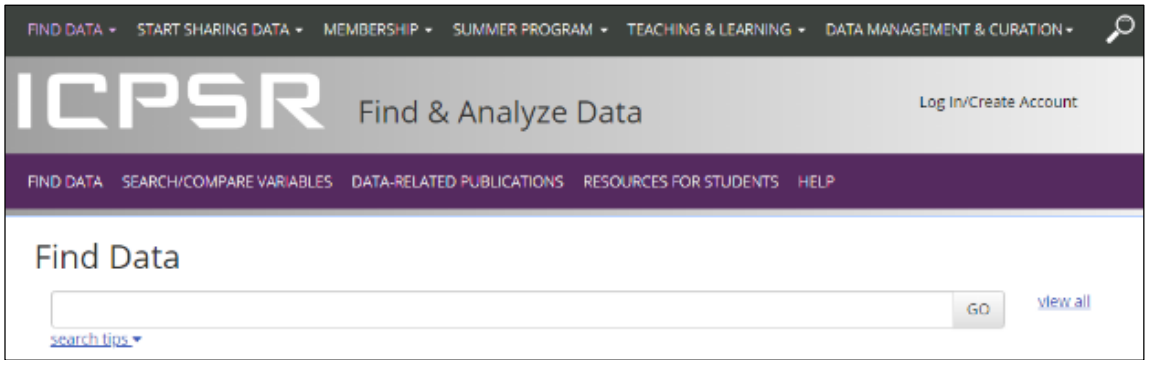
Source: https://datadryad.org/stash



Source: https://www.dkrz.de/up/systems/wdcc



Source: https://nyu.databrary.org/



Source: https://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp

Table 2: RDSs according to CoreTrustSeal international standard.

CoreTrustSeal International Standards	Search Engine	NDS	Institutional Repositories	Subject Repositories
	DataSearch	*UK Data Service*	*DataFirst*	OpenICPSR
1 Mission and scope	✔	✔	✔	✔
2 Licenses	✔	✔	✔	✔
3 Continuity of access	✔	✔	✔	✔
4 Confidentiality and ethics	✔	✔	✔	✔
5 Organizational infrastructure		✔	✔	✔
6 Expert guidance		✔	✔	✔
7 Data integrity and authenticity		✔	✔	✔
8 Appraisal		✔	✔	✔
9 Documented storage procedures	✔	✔	✔	✔
10 Preservation plan	✔	✔	✔	✔
11 Data and metadata quality	✔	✔	✔	✔
12 Workflows	✔	✔	✔	✔
13 Data discovery and identification	✔	✔	✔	✔
14 Data reuse	✔	✔	✔	✔
15 Technical infrastructure		✔	✔	✔
16 Data security		✔	✔	✔

Table 2: RDSs provide tools, resources, and training to help users with data management, organization, discovery, sharing, preservation, and access. Most RDSs meet the core criteria of the CoreTrustSeal international standard are Criteria 1-16 according to the examples from UK Data Service, DataFirst, and Open ICPSR. According to Wipawin[14], the most appropriate standard for research output management in digital repositories is CoreTrustSeal, which is a fundamental standard to develop the national research output management in digital repositories for long-term digital preservation. The recommendation is a policy statement issued by research funders for depositing research outputs in an open-access digital repository system within a year after the research has been completed.

According to Mendeley Data (https://data.mendeley.com/), which can be searched by the DataSearch search engine, RDSs need research data policy in the planning process for criterion 1 (mission and scope). Criteria 2, 3, 4, and 9 (licenses, continuity of access, confidentiality and ethics, and documented storage procedures) are addressed by the DMP for the data repository. Criteria 11, 12, 13, and 14 (data and metadata quality, workflows, data discovery and identification, and data reuse) can be addressed by a data catalog with metadata and persistent identifier, such as a Digital Object Identifier (DOI). Therefore, managing RDSs requires a data catalog: a metadata management tool designed to help organizations find and manage large amounts of data, including tables, files, and databases. For example, DataFinder (a research data catalog for Oxford)[15] comprises of the institutional repository called DataBank, additional metadata from DataStage and metadata captured in DMP online. The core minimum metadata for data deposit are at least 13 elements such as:

1. Data creator name
2. Creator affiliation
3. Data owner
4. Output a product of funded researcher
5. Title
6. Description
7. Subject headings
8. Publication year
9. Publisher
10. Earliest access date (open access/embargo)
11. Digital format
12. Terms and conditions
13. ID + other automatically assigned metadata

Finally, there are the best practices according to the international standard in DataSearch, UK Data Service, DataFirst, and OpenICPSR. For users in all disciplines to access datasets, the World Bank has an open data catalog, which is designed to make the World Bank's development data easy to find, download, use, and share by using keywords (dataset, indicator, or country name) to find relevant data. The data catalog matches these keywords to terms found in the titles, descriptions, countries, regions, and tags; for example:



Source: https://datacatalog.worldbank.org/

Conclusion

This article has highlighted the types of dataset search in RDSs, and the core criteria for RDSs according to the CoreTrustSeal international standard. Most RDSs are provided by libraries and research centers for scientists and researchers to find, reuse, and publish data. In this study, the types of dataset search in research services are search engines, NDSs, institutional repositories, and subject repositories. The core criteria of the CoreTrustSeal international standards are mission and scope, licenses, continuity of access, confidentiality and ethics, organizational infrastructure, expert guidance, data integrity and authenticity, appraisal, documented storage procedures, preserva-tion plan, data and metadata quality, workflows, data discovery and identification, data reuse, technical infrastructure, and data security. RDSs need to prepare a data catalog for datasets so that data can be discovered and preserved for long-term access. Four example RDSs—DataSearch, UK Data Service, DataFirst, and OpenICPSR—satisfy the criteria of the CoreTrustSeal international standard because of their research data policy, DMP, data catalog with metadata, data repository, and a persistent identifier, such as a DOI. Datasets are easy to find when providing structured data in a data catalog; therefore, RDSs require Data Catalog Vocabulary (DCAT) to facilitate interoperability between data catalogs published on the web.

Acknowledgment

We thank Edanz Group (https://en-author-services.edanzgroup.com/) for editing a draft of this manuscript.

References

1. Chen J, Wang X, Cheng G, Kharlamov E, Qu Y. Towards More Usable Dataset Search: From Query Characterization to Snippet Generation. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. New York, NY, USA: ACM; 2019:2445–2448. doi:10.1145/3357384.3358096
2. Dua D, Graff C. *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science; 2019. http://archive.ics.uci.edu/ml/index.php. Accessed December 9, 2019.
3. Aryani A, Poblet M, Unsworth K, et al. A Research Graph dataset for connecting research data repositories using RD-Switchboard. *Sci Data*. 2018;5(1):1-9. doi:10.1038/sdata.2018.99
4. Dave Wells. The Ultimate Guide to Data Catalogs: Key Things to Consider WhenSelecting a Data Catalog. Presented at the: March 2018; Eckerson Group. https://go.waterlinedata.com/hubfs/a/ultimate-guide-to-data-catalogs.pdf.
5. FAIR Principles. GO FAIR. https://www.go-fair.org/fair-principles/. Accessed December 26, 2019.
6. Mons B. FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services. *Data Intell*. 2019;1(1):22-42. doi:10.1162/dint_a_00002

7. Tenopir C, Talja S, Horstmann W, et al. Research Data Services in European Academic Research Libraries. *Liber Q*. 2017;27(1):23-44. doi:10.18352/lq.10180
8. Cox AM, Kennan MA, Lyon L, Pinfield S. Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *J Assoc Inf Sci Technol*. 2017;68(9):2182-2200. doi:10.1002/asi.23781
9. Open Data. SPARC. https://sparcopen.org/open-data/. Accessed October 15, 2018.
10. Wipawin N. The Readiness of Thai Universities towards the Research Data Quality Accreditation Standards in Institutional Repositories. In: ; 2016. https://hal.univ-lille3.fr/hal-01432971. Accessed December 26, 2019.
11. Klungthanaboon W. Stakeholders' perspectives of institutional repositories in National Research Universities in Thailand. 2015. https://eleanor.lib.gla.ac.uk/record=b3110620. Accessed December 30, 2019.
12. Wipawin N. Research Data Quality Accreditation Standards in Institutional Repositories :A Case Study of Thailand. In: Korea: Kongju National University; 2017:25-33. http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07176242.
13. Wipawin N, Wanna A. Institutional Repositories in Thai Universities. In: Tuamsuk K, Jatowt A, Rasmussen E, eds. *The Emergence of Digital Libraries – Research and Practices*. Lecture Notes in Computer Science. Cham: Springer International Publishing; 2014:385-392. doi:10.1007/978-3-319-12823-8_41
14. Wipawin N. *International Standards Selection for Research Output Management in Thai Digital Repositories*. Bangkok: Research Council of Thailand; 2019.
15. Rumsey S, Jefferies N. DataFinder: A Research Data Catalogue for Oxford. *Ariadne*. 2013;(71). http://www.ariadne.ac.uk/issue/71/rumsey-jefferies/. Accessed December 26, 2019.

License, Supplementary Material and Copyright

This is an open-access article distributed under the terms of the Creative Commons Attribution (CC BY NC ND 4.0) License. You may share the material, but must give appropriate credit to the source, provide a link to the license and indicate if changes were made. You may not use the material for commercial purposes. If you remix, transform, or build upon the material, you may not distribute the modified material

Any supplementary material referenced in the article can be found in the online version.

This article is copyright of the Chulabhorn Royal Academy, 2019.

Citation

Wipawin N. Dataset Search in Research Data Services. J Chulabhorn Royal Acad .2019;1(2):1-15. https://he02.tci-thaijo.org/index.php/jcra/article/view/233180

Online Access

https://he02.tci-thaijo.org/index.php/jcra/article/view/233180