Australian National University

Australian Research Data Commons

Australian Government

NCRIS National Research Infrastructure for Australia

ARDC is enabled by NCRIS

HASS and Indigenous Research Data Commons (HASS&I RDC) Research Data Management (RDM): A Conceptual Overview

RSE-CEP is a co-investment partnership with the Australian Research Data Commons (ARDC) through the HASS and Indigenous Research Data Commons (DOI: 10.3565/x5f6-mw53). The ARDC is enabled by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS).

This RDM Conceptual Overview emerged from identified needs across the HASS and Indigenous RDC focus areas. The draft Overview is being released for feedback from the HASS, research data management and research software engineering communities, and may be formally adopted by the HASS&I RDC in the future. It represents a genuinely collaborative articulation of shared research data management needs across the full breadth of HASS research infrastructure.

Version	Draft for public review
Last modified	2026-08-06
Lead Author	Peter (PT) Sefton
Technical Editor	James Smithies
Reviewers	Nick Jenkins, Nichola Burton, Mat Bettinson, Wolfgang Barth, Junran Lei, Jenny Fewster, Jamie Archer, Robin Burgess

Table of Contents

Purpose

This document provides a conceptual overview of research data management activities and infrastructure for the Humanities, Arts, Social Sciences and Indigenous Research Data Commons (HASS&I RDC). It draws on widely used "Research Data Lifecycle" frameworks for planning and promoting best practice Research Data Management (RDM), adding a simple informal software architecture layer to show how RDM practice in a research data commons interacts with and requires broad classes of infrastructure. The document is designed to complement the concept of a research data commons defined by the Australian Research Data Commons (ARDC) as a site that "brings together people, skills, data, and related resources such as storage, compute, software, and models into one system". It also aligns with the "Global Open Research Commons: Creating an International Model for Improved Interoperability and Collaboration" (Treloar & Woodford 2024), expanding on its underpinning elements ICT infrastructure and services & tools, and showing how these interact to enable the creation use and reuse of GORC's Research Objects via interoperability and standards.

By providing integrated resources and easy-to-use interfaces, a research data commons enables researchers to speed up their existing research, and to undertake research that wasn't possible before. We intentionally avoid close technical specification of research data commons, preferring instead to provide a high-level overview highlighting core components, common workflows, and an emphasis on the importance of incorporating infrastructure into our mental model. Engineering standards and models are provided as references, for readers interested in considering how the idea of a research data commons might be implemented.

This RDM Conceptual Overview emerged from identified needs across the HASS and Indigenous RDC focus areas. The draft Overview is being released for feedback from the HASS, research data management and research software engineering communities, and may be formally adopted by the HASS&I RDC in the future. It represents a genuinely collaborative articulation of shared research data management needs across the full breadth of HASS research infrastructure.

Provenance

"Research Lifecycle" and "Research Data Lifecycle" frameworks are commonly used by service and infrastructure providers such as university libraries, research offices and IT departments to talk about research processes and to design software architectures (Jiang, Maurici-Pollock & Tang 2025) The document builds on the ARDC Research data management framework for institutions and The Shared Analytic Framework for the Environment (SAFE 2.0) Framework (The Western Australian Biodiversity Science Institute and Western Australian Marine Science Institution 2023). In a departure from those frameworks, this document explicitly ties research activities to infrastructure, to show the flow of data; from collection, to analysis, to publication or deposit in an archive, and subsequent re-use. This contrasts with many common frameworks which focus on activities without reference to infrastructure. The terms used in Jiang, Maurici-Pollock & Tang inform this document.

Audience

This document is designed to support infrastructure providers, Research Software Engineers (RSEs) both experienced and emerging, researchers-who-code, and their client researchers and research support staff, and IT and Library staff, who are building and supporting a research data commons and want to understand how it can benefit the humanities and social science (HASS) community. Software architects interested in implementing or extending a research data commons might use the document to orient their analysis and understand the high-level requirements of HASS researchers and infrastructure providers, and communicate using a common vocabulary. The document is intended as a conceptual guide rather than a formal engineering specification or reference model. Readers responsible for designing or evaluating systems should complement their requirements definition process with relevant formal standards including but not limited to Application Programming Interfaces (APIs);

Modelling Frameworks; Computer Language Specifications; Interoperability Specifications and Cybersecurity models.

Implementation Context

Research takes place in a complex socio-technical context, informed by a wide array of principles and processes. Four important areas include:

- Legal & policy constraints
- Communication and communities of practice
- Metadata Standards and Data Protocols
- Technical Standards & Specifications

Legal and Policy Constraints

Project leaders should consider legal and policy constraints when facilitating conversations with software architects and engineers about the norms of research practice. Requirements definition should be broad and attend to as many factors as possible that affect system design and implementation. These include but are not limited to:

- Indigenous Cultural and Intellectual Property (ICIP) as described in the Framework for the Governance of Indigenous Data: HASS and Indigenous Research Data Commons (Langton et al. 2026) .
- The need for research integrity, research quality, excellence, and impact as defined by the Australian Chief Scientist (Foley 2023).
- Codes of practice, laws, institutional policies including the Australian Code for the Responsible Conduct of Research "The Code" (National Health and Medical Research Council, Universities Australia, & Australian Research Council 2018).
- The CARE principles that address data governance concerns related to Indigenous data sovereignty and cultural stewardship (Carroll et al. 2020).
- The FAIR Guiding Principles for scientific data management and stewardship(Wilkinson et al. 2016).

Consideration should extend to funder agreements, institutional policies and procedures (research policies, data management policies, intellectual property policies); project-specific considerations such as research ethics agreements and data licensing.

Preliminary business analysis and requirements definition should ensure all relevant global, local, and institutional policies and governance assumptions are identified and enabled and their importance communicated to the teams implementing software solutions.

Communication and Communities of Practice

The ARDC definition of research data commons foregrounds bringing people together as the primary purpose of the commons. This implies a range of activities that are chiefly concerned with how people are supported in undertaking these activities, enabled by skills development and resources including:

- Training programmes to help researchers and their support teams work in a research data commons.
- Data Guides such as the Trove Data Guide (https://tdg.glam-workbench.net) and the GLAM Workbench (https://glam-workbench.net/).
- Digital methodologies (computational thinking, collections as data (Padilla et al. 2019).
- Support for research communities of practice.
- Best practice guides and principles.

Metadata Standards & Data Protocols

For a research data commons to function, data needs to be re-usable, to enable new kinds of research to take place. This depends on adherence to metadata standards and other protocols including:

- Metadata Standards and specifications such as Markup languages, Linked Data (Berners-Lee) and RO-Crate (Carragáin et al. 2019).
- Vocabularies, Ontologies, Schemas, such as Dublin Core.
- Discipline metadata profiles such as the Language data commons schema https://w3id.org/ldac/terms and profile https://w3id.org/ldac/profile.
- Persistent Identifier Schemes such as Digital Object Identifiers (DOIs) which build on the Handle (Sun 2001) protocol.
- Provenance records describing how data has been created and derived from other data, using tools such as scripts or applications.

Component Views

The articulation of the research data commons concept provided below includes views of three core components:

Component	Description
Implementation context	The context in which these activities take place, and in which the infrastructure is created, managed and governed.
Infrastructure functions	A description of three broad classes of research infrastructure: • Administrative Systems that support and evaluate research including planning, metrics, authentication • Research "Workspace" systems where the day-to-day practice of research takes place • The data persistence layer consisting of Archival Repositories.
Activity definitions	A set of five core research and data management activities that are common to all best practice research and familiar from lifecycle frameworks (Plan, Collect, Process/Describe, Analyse, Deposit/Publish).

Infrastructure Functions

Figure 1 classifies types of infrastructure systems that perform broad functions. Some actual systems (software platforms, tools or services) may contain components from two, or all three. As is usual with high level architectural descriptions, additional diagram views would provide different perspectives. A prose description of the diagram is offered below.

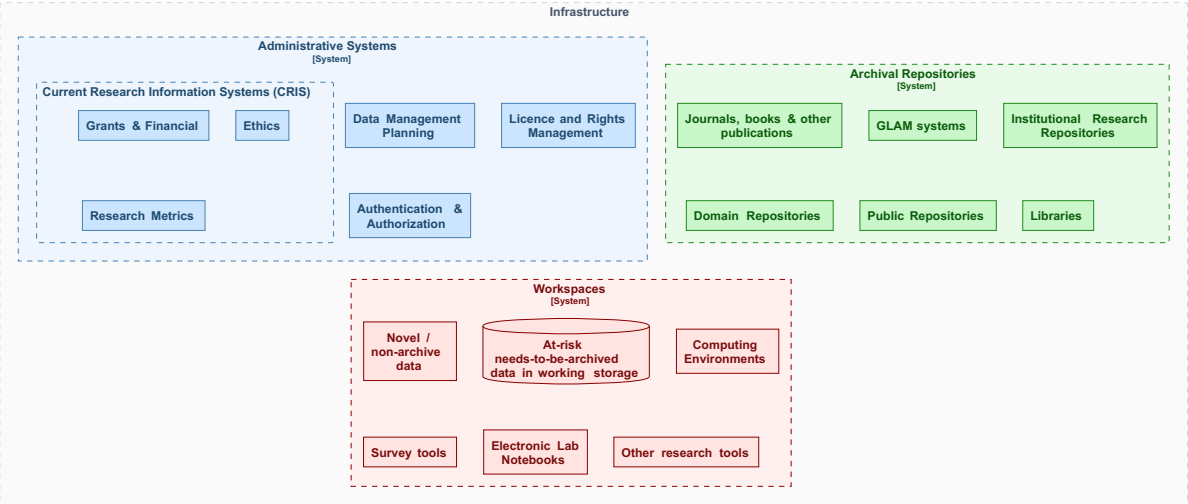

Figure 1: Infrastructure functions

Administrative Systems

Research Administrative systems are crucial to the conduct of research as they control the flow of resources to research activities through funding and staffing and measure research performance. They include Research Information Systems (grants, publication counts etc), often called Current Research Information Systems CRIS systems in Europe, less commonly in Australia. Research Data Management Planning (RDMP) and Research Ethics Management systems; Administrative Records Management, for records related to the research process such as forms; Authentication systems for people and other agents; License and rights management. For the purposes of a research data commons these key systems interact with the infrastructure for doing research data management.

Research "Workspace" Systems

This is a large and diverse category including computing services, working data storage, tools for data collection, processing, analysis as well as a vast array of applications and services used by researchers. This is where data is collected using tools, created from other data or code, where code is run and written. In the diagrams used in this guide, Workspaces are shown as a system boundary rather than a fixed container, reflecting the fact that they span institutional, commercial,

and personal environments and have no hard edges. Workspaces include but are not limited to:

Workspace system	Description
Computing services	Computing services at national, institutional and commercial levels, including public and private cloud computing and High-Performance Computing services.
Code notebooks	Interactive environments (e.g. Jupyter, R Markdown) for writing, running, and sharing code alongside narrative documentation and visualisations.
Workflow systems	Tools for orchestrating and automating multi-step research data processing pipelines and computational tasks.
Discipline data sharing networks	Domain-specific platforms for sharing and exchanging research data among communities of practice. Sometimes these also fulfil Archival Repository functions.

Workspace system	Description
Electronic lab notebooks (ELN)	Digital systems for recording, organising, and sharing experimental procedures, observations, and results.
Survey tools	Platforms for designing, distributing, and collecting responses from structured questionnaires and surveys (e.g. REDCap, Qualtrics).
Qualitative analysis tools	Software for coding, annotating, and analysing qualitative data such as interview transcripts, field notes, and multimedia (e.g. NVivo).
Working storage — institutional research stores	Managed network drives ("R Drives") provided by the institution for active research data.
Working storage — cloud services	Commercial cloud storage (OneDrive, Google Drive, etc.) used for day-to-day file access and collaboration.
Working storage — personal devices	Personal computers and hard drives used for local storage. High risk due to lack of backup, access control, and institutional oversight.
Working storage — HPC/cloud compute storage	Storage provisioned alongside cloud computing and High-Performance Computing facilities for active computational workloads.

Archival Repositories

Archival Repositories are a critical infrastructure component, required by institutional policies for research integrity purposes. They must underpin any implementation of the FAIR principles and form the backbone of a research data commons. For data to be findable, and accessible there must be well managed services and infrastructures for keeping data and gatekeeping so the appropriate people can access it. The term Archival Repositories was chosen based on its use in the PILARS (Sefton et al. 2024) as a general-purpose way of covering a range of specific practices in the archive, library and general research software domains.

Archival Repository services for research data need to provide the following services:

- Long term data management - which necessarily includes metadata services.
- Access control and clear data licensing so that data can be securely managed in the long term.
- Governance for sustainability and persistence.
- Persistent Identifiers for data, which implies not just persistence of an identifier scheme, but also persistence of data assets.

There are four main types of repository infrastructure:

- Institutional.
- Discipline / national repositories.
- Collections at galleries, archives, libraries, and museums.
- Public archives such as Zenodo and the Internet Archive.

As with any software, it is preferable to re-use existing repository services where they exist, redeploy existing software stacks where they do not, and build new software only where it is required.

Activity Definitions

The research activities described in this guide closely align with the ARDC Data Management Framework for institutions (ARDC 2023)and other common frameworks (Jiang, Maurici-Pollock & Tang 2025).

Activity	Notes
Plan	Planning includes several activity functions supported by administrative systems, the key one here being the production of Research Data Management Plans (RDMP) which are informed by ethical and policy frameworks 'upstream'. RDMPs are mandated for most Australian researchers and are often expected to be considered 'live' documents which are updated as circumstances change. Also included in this would be other activities needed to get a project started including research ethics approvals. It is important for research ethics to be considered in building systems as approvals often have an impact on how data may be collected and analysed and commitments may be made to deposit/publish data in a certain system.
Collect	Collect is a generic term for acquiring data to work on. In classic desk-based humanities research, this may involve getting books out of the library but would also include video and audio recording, surveying, photographing, fetching data from repositories; any activity to create or procure data. Collecting data includes both novel data and re-use of existing data. It must be noted that some past practices for data collection are no longer considered acceptable. Where the results of such practices have been archived, there are questions about ethical reuse which must be addressed.
Process / Describe	Processing data for use is a key activity for a lot of research. Data is rarely ready for analysis in the form in which it is collected, so data goes through preparation processes before it can be meaningfully used. In best practice research, data also needs to be described, to enable future FAIR functions: Findability via consistent metadata, Interoperability via documentation of how data is structured.
Analyse	Taking a very broad view here, analysis is any research use of data, including close reading, building (curating) collections, visualization, annotation, qualitative analysis, computation methods, training of large language or other AI models etc. Different disciplines have quite different approaches to analysis, which would be described in more detail in patterns and architectures in use in various domains.
Deposit / Publish	This activity is the one that closes the loop in the life cycle and allows for the re-use of data in further research or in other activities such as policymaking, education and product development. Depending on the Archival Repository system that is being used this may be seen as an act of publishing (submitting to a journal repository) or depositing (to an institutional research data repository).

Researchers engage in activities in all sorts of orders and combinations but for the purposes of this guide, and to align with common ways of representing research activities, they are presented in Figure 2 within a temporal sequence. In the tradition of research lifecycle frameworks, there is a feedback loop representing data re-use: data is described, deposited or

published (into infrastructure that implements the Archival Repository function), then (re)collected into new research activities (after being planned, etc.). The infrastructure function is indicated by the colour coding as used in the previous diagram.

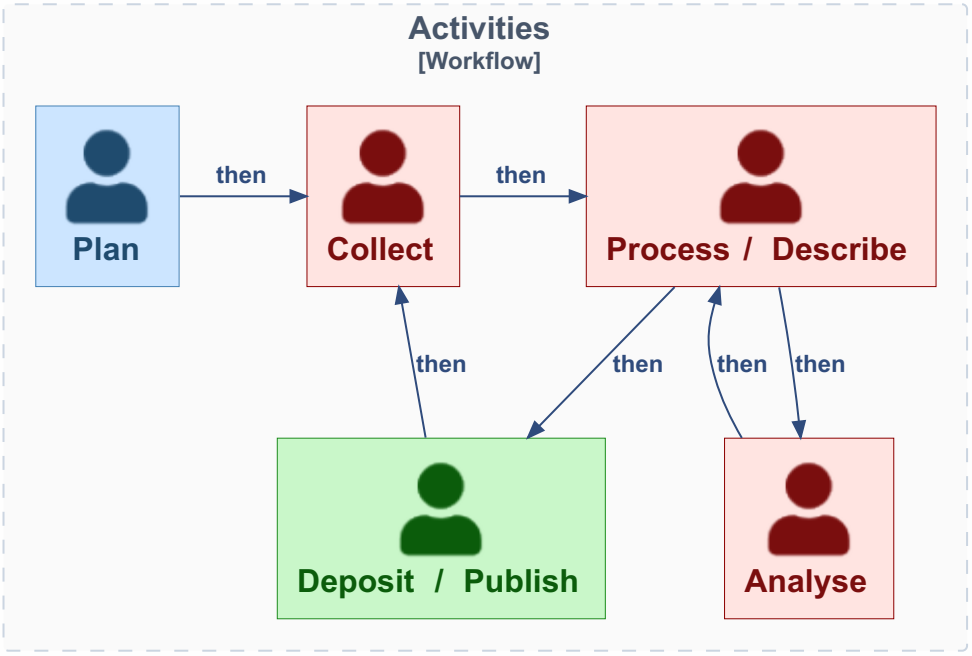

Figure 2: Activity definitions in sequence, showing the "lifecycle" loop of data reuse

Combined View

Figure 3 shows a different view of the activities interacting with the infrastructure, with the four contextual areas shown as well, indicating how they should inform systems design. In small research groups the actors may all be the same few people undertaking different activities in different roles, while in large research projects roles may be much more demarcated.

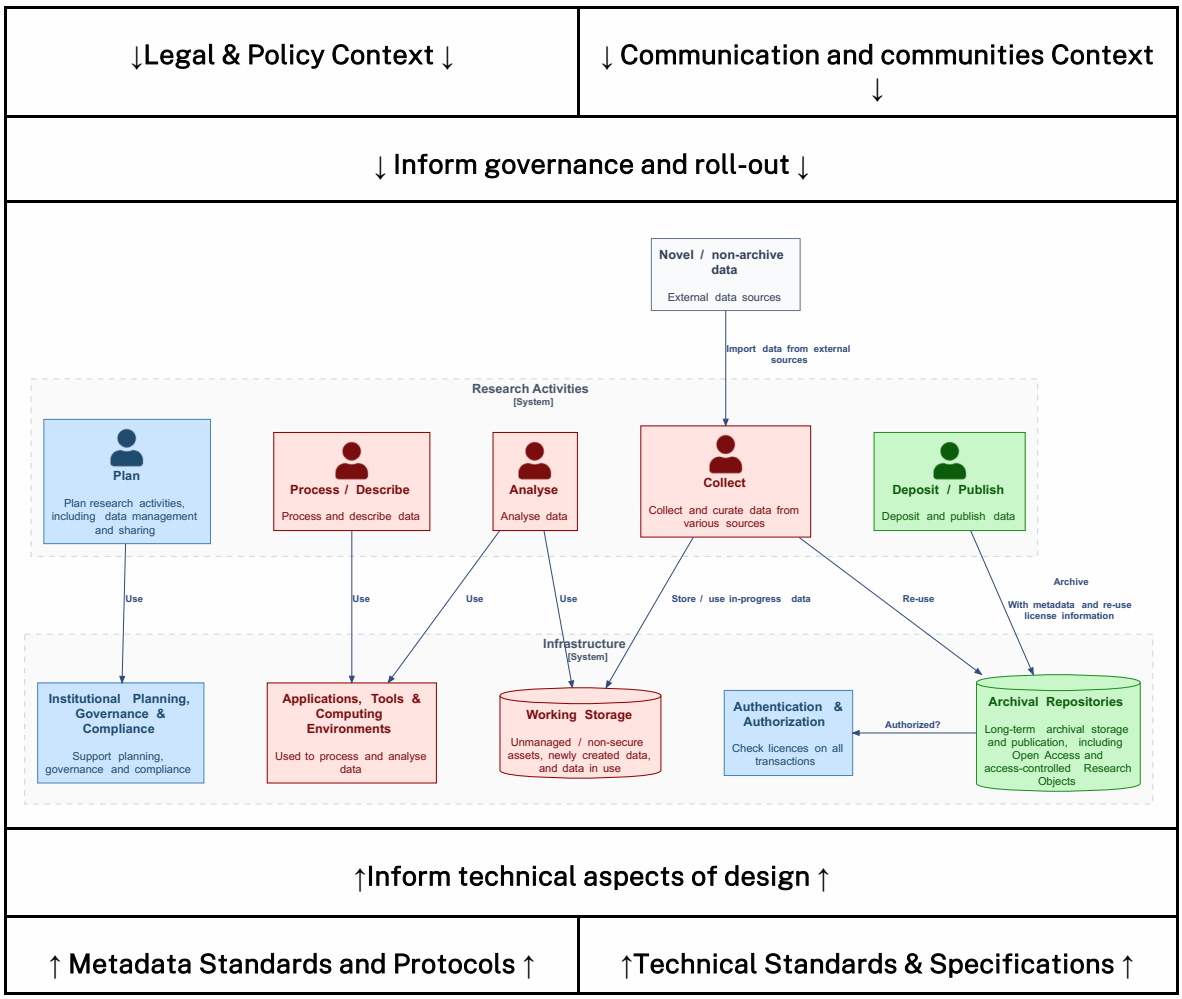

Figure 3: Research data management activities in relation to infrastructure

Figure 3 shows how activities are tied to infrastructure. The following table provides context for the diagram:

Step	Description
Plan	Research takes place in an institutional and global context of policy and funding that allows research to start and governs its progress. The most salient infrastructure systems here are Research Data Management Planning (RDMP) and provisioning systems.
Collect	The data management cycle begins with data collection. This happens in three ways: **a)** Novel data is obtained to work on, by observation, survey, synthesis etc. **b)** As yet unpublished or archived working data is used, subject to licensing conditions, sometimes under an Open Access license, or subject to other conditions set out in a license or data sharing agreement. **c)** Data is reused (the R in FAIR) from Archival Repository infrastructure such as a library or archive.

Step	Description
Process / describe	Once collected, data usually needs to be processed to make it interoperable with Research Workspace tools and analytical methods. To enable archiving and subsequent reuse, it is imperative to describe data, making it FAIR-Findable and Interoperable. Processed data is also stored in working storage during this phase.
Analyse	Processed data can be analysed, producing outputs. There is a short iteration cycle here where a large proportion of research time is actually spent — trying analyses which create new data assets (findings and derived products). The wide range of Research Workspace services used for analysis belongs in solution architectures for specific analytical services and tools, but the connection between the Analyse activity and Workspace infrastructure emphasises just how much goes on in this space.
Deposit / Publish	To enable the reuse of data — a key aim of a research data commons — it must be made available so it can be Found and Accessed, and form part of the scholarly record. The Deposit/Publish activity represents this step and enables the information flow: data must be put in a managed repository with the requisite services for identifiers, discovery services, and access control for licensed re-use. Following the GORC model, deposits of data are considered to be Research Objects, representing a (re)usable well-described bundle or package of data.

Conclusion

This conceptual overview document expands on the aforementioned research lifecycle models by anchoring HASS research activities directly to the infrastructure layer. It demonstrates how administrative systems, active workspaces, and access controlled archival repositories connect and interact through the research lifecycle. It aims to establish a common vocabulary for researchers, research software engineers (RSEs), researcher support groups and infrastructure providers.

Whilst this document is tailored to the infrastructural and operational context of the Australian Research Data Commons (ARDC), it rightly contains commitments to standards-based design, governance, and data principles (e.g. FAIR/CARE), that are relevant to implementation more broadly. We encourage readers to use this conceptual overview as a basis for their requirements definition for RDM for RDC whilst tailoring their analysis and design to their local institutional environments.

References

ARDC. 2023. ***Research data management framework for institutions***. Zenodo. https://doi.org/10.5281/zenodo.8433246.

Berners-Lee, Tim. 2009. Linked Data - Design Issues. http://www.w3.org/DesignIssues/LinkedData.html. (25 June, 2009).

Carragáin, Eoghan Ó, Carole Goble, Peter Sefton & Stian Soiland-Reyes. 2019. RO-Crate: A lightweight approach to Research Object data packaging. ***F1000Research*** 8. https://doi.org/10.7490/f1000research.1117129.1.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. The CARE Principles for Indigenous Data Governance. https://doi.org/10.5334/dsj-2020-043.

Foley, Cathy. 2023. Trust in science: clarifying the distinctions between research integrity, research quality, excellence, and impact. Office of the Chief Scientist (Australia). https://apo.org.au/node/324072. (19 February, 2026).

Jiang, Jie, Danielle Maurici-Pollock & Rong Tang. 2025. Research Data Lifecycle (RDLC): An Investigation into the Disciplinary Focus, Use Cases, Creator Backgrounds, Stages and Shapes of RDLC Models. ***International Journal of Digital Curation*** 19(1). 30–30. https://doi.org/10.2218/ijdc.v19i1.860.

Langton, Marcia, Robert McLellan, Jennifer Fewster & KRISTEN SMITH. 2026. Framework for the Governance of Indigenous Data: HASS and Indigenous Research Data Commons. Zenodo. https://doi.org/10.5281/zenodo.11135962.

National Health and Medical Research Council, Universities Australia, & Australian Research Council. 2018. ***Australian Code for Responsible Conduct of Research***. Canberra: Commonwealth of Australia.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke & Stewart Varner. 2019. Always Already Computational: Collections as Data: Final Report. ***Copyright, Fair Use, Scholarly Communication, etc.*** https://digitalcommons.unl.edu/scholcom/181.

Sun, S. 2001. Establishing persistent identity using the handle system. In ***Proceedings of the Tenth International World Wide Web Conference***.

The Western Australian Biodiversity Science Institute and Western Australian Marine Science Institution. 2023. Shared Analytic Framework for the Environment (SAFE 2.0). The Western Australian Biodiversity Science Institute and Western Australian Marine Science Institution.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. ***Scientific Data*** 3. 160018.