



Deliverable 1:

The creation of a harmonised research data lifecycle (RDL) model and crosswalk to existing models

Working Group name: [RDA-OfR Mapping the Landscape of Digital Research Tools](#)

Working Group co-chairs: Emmanuel Adamolekun ([0000-0003-2992-5448](#)), Francis P. Crawley ([0000-0002-6893-5916](#)), Rory Macneil ([0000-0002-8429-096X](#)), Adam Vials Moore ([0000-0002-2085-1908](#)) and Hea Lim Rhee ([0000-0002-4171-5710](#))

Assisting co-chairs: Marcelo Garcia ([0000-0002-2927-2371](#)) & Richard Pitts ([0000-0002-2037-3360](#))

Working Group Facilitator and Editor: Connie Clare ([0000-0002-4369-196X](#))

Version: July 2024 | [CC-BY 4.0 International](#) | DOI:

Table of Contents

Executive Summary.....	1
Aims and Objectives.....	1
Methodology and Results.....	2
Mapping the Landscape of Digital Research Tools Harmonised (MaLDReTH) Research Data Lifecycle (RDL) Stages and Definitions.....	8
Contributors.....	10

The creation of a research data lifecycle (RDL) model and crosswalk to existing models

Executive Summary

As there is not one accepted universal Research Data Lifecycle (RDL) model, the first deliverable by the [RDA-OfR Mapping the Landscape of Digital Research Tools Working Group \(WG\)](#) involved the creation of a harmonised RDL that could be used as the foundational framework for underpinning later work on the categorisation and characterisation of digital research tools.

To achieve this, a landscape review was carried out to identify a list of existing RDLs. A qualitative analysis process was conducted to determine the top five best characterised and most comprehensive RDLs. Their stages and definitions were aligned using a semantic distance methodology over an etymological/taxonomic net to produce the Mapping the Landscape of Digital Research Tools Harmonised (MaLDReTH) RDL model.

Aims and Objectives

Research data management (RDM) is a critical aspect of modern scientific endeavours, ensuring the integrity, reproducibility, and long-term preservation of valuable research outputs. However, the lack of a universally accepted RDL model poses challenges to establishing standardised practices and facilitating effective RDM across various research domains.

The Mapping the Landscape of Digital Research Tools WG examined and identified different stages of the research data lifecycle (RDL), and developed a harmonised RDL to serve as the foundational framework for categorising and characterising various types of digital research tools. Since numerous different RDL models exist that have been conceptualised for specific research paradigms and audiences, the WG conducted a landscape review to research and consult existing models and identify common stages of the RDL for use as the framework to guide the research tool categorisation. A harmonised RDL was created to provide a common language and reference point for researchers, data stewards, and tool



ORACLE
for Research

developers alike, facilitating collaboration and interoperability within the research data ecosystem. Each RDL stage is supported by definitions. The WG created a crosswalk to demonstrate connections between the chosen model and existing models.

Methodology and Results

A comprehensive landscape review was conducted to identify existing RDL models from various sources, including scholarly publications, institutional guidelines, and domain-specific best practices. Initially, [20 candidate RDLs](#) were identified, representing diverse perspectives and approaches to RDM (Fig. 1).




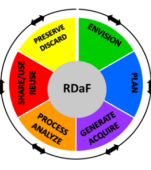
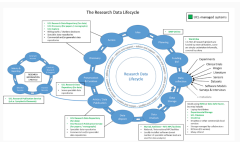
A qualitative analysis process was undertaken to assess the comprehensiveness, clarity, and applicability of these candidate RDLs across different research domains. This evaluation considered factors such as the completeness of the lifecycle stages, the preciseness of stage definitions, and overall suitability of the model for addressing the nuances of RDM in various disciplines.

The [top five best characterised and most comprehensive RDL models](#) were selected for harmonisation (Table 1). To reconcile any potential overlaps, gaps, or inconsistencies in the terminology and concepts, stages of each selected RDL model were aligned using a semantic distance methodology over an etymological/taxonomic net (Fig. 2). This approach, inspired by the work of [Thompson et al., 2020](#), leveraged etymological and taxonomic networks to align 12 common stages and definitions of each RDL model, ensuring conceptual coherence and consistency.

The outcome of this harmonisation process was the MaLDReTH RDL (Fig. 3 & 4), a synthesised model that captures the comprehensive aspects of the 5 selected RDLs. The MaLDReTH RDL model represents a standardised and harmonised framework for RDM, facilitating effective collaboration and tool development across various research domains.

By providing a common language and reference point, the MaLDReTH RDL model addresses a long-standing gap in the RDM landscape, promoting data stewardship, reproducibility, and the long-term preservation of valuable research outputs. For the purpose of this WG, the MaLDReTH RDL model serves as the foundational framework for the production of [Deliverable 2](#) (The identification, categorisation, and mapping of different types of research tools: A categorisation schema) and [Deliverable 3](#) (The creation of a preliminary structural framework for an online open access 'map of the digital research tool landscape'), enabling the systematic categorisation, characterisation, and mapping of digital research tools.

Table 1. Top five Research Data Lifecycle (RDL) models selected for harmonisation.

Research Data Lifecycle (RDL)	Source	Rationale
	DCC Data Curation Lifecycle Digital Data Curation Centre	Focused on curation, workflow, data management centric.
	RDMkit Data Life Cycle ELIXIR	Popular, well known, simple, uses commonly referenced stages.
	Best Practice Data Life Cycle Approaches for the Life Sciences EMBL-ABR Data Life Cycle Workshop Series	Best practice RDL for researchers in the Life Sciences/Bioinformatics.
	Research Data Framework (RDaF) National Institute of Standards and Technology (NIST)	Broad, discipline agnostic.
	The Research Data Lifecycle University College London (UCL)	Comprehensive, uses commonly referenced stages of the lifecycle, useful references to activities relating to stages.

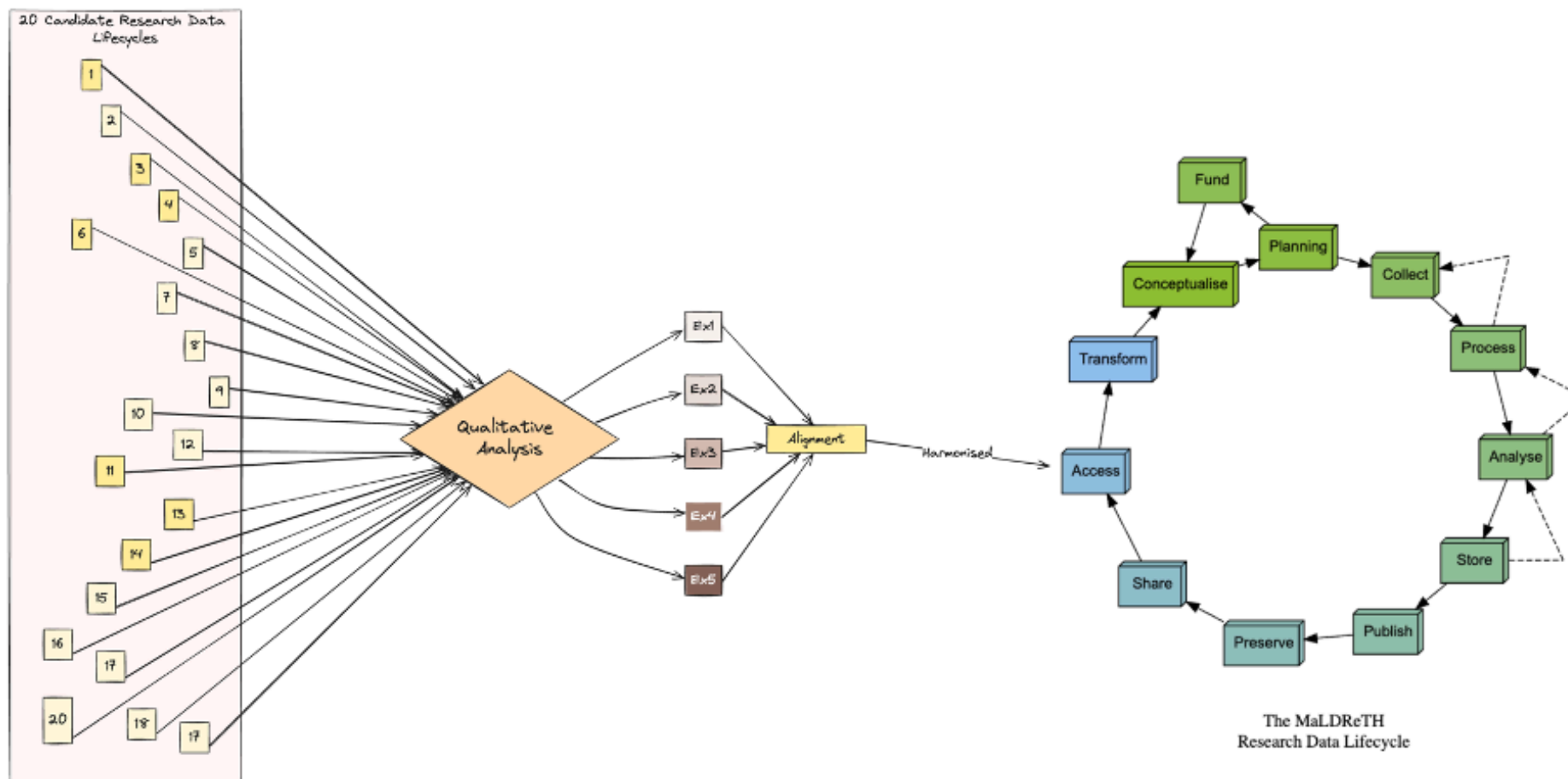


Figure 1. Qualitative analysis, selection and alignment of 20 Research Data Lifecycle (RDL) Models to produce the MaLDReTH model.

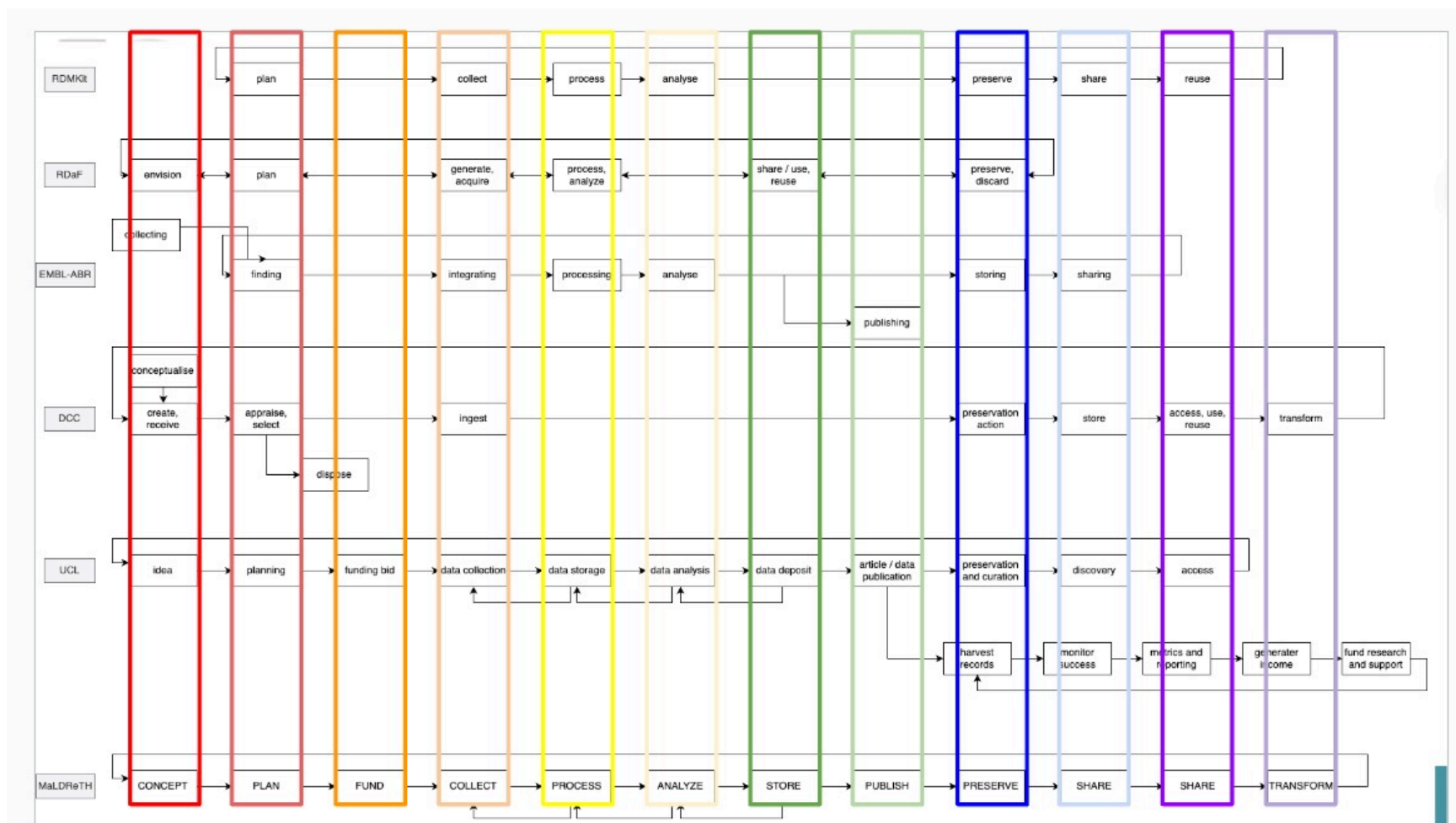


Figure 2. Research Data Lifecycle (RDL) stage alignment using a semantic distance methodology over an etymological/taxonomic net.

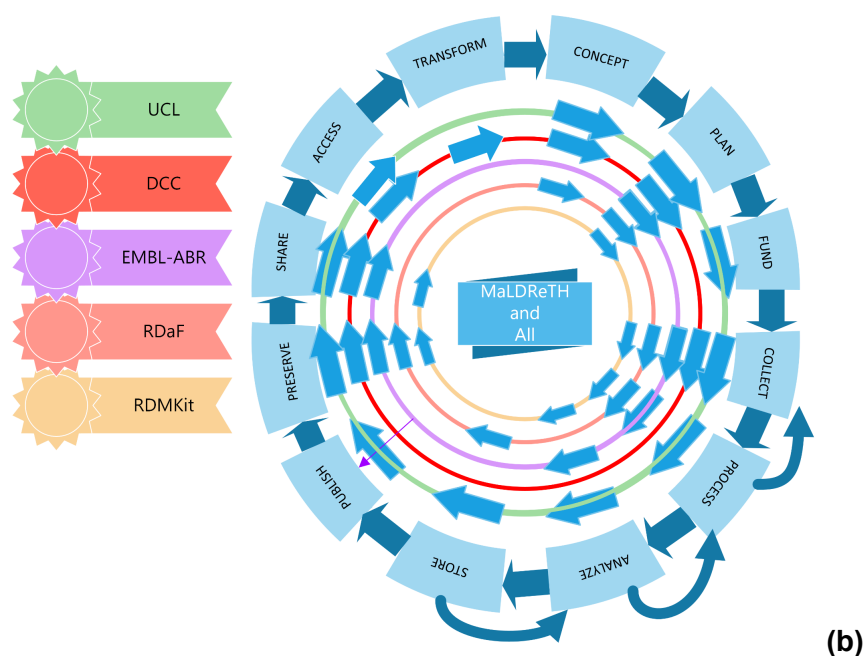
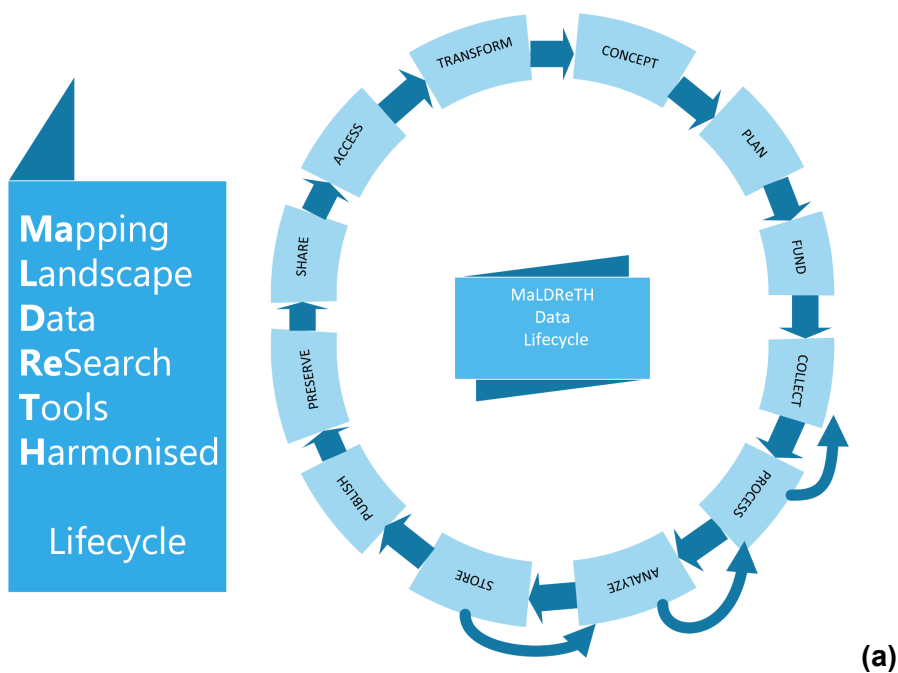
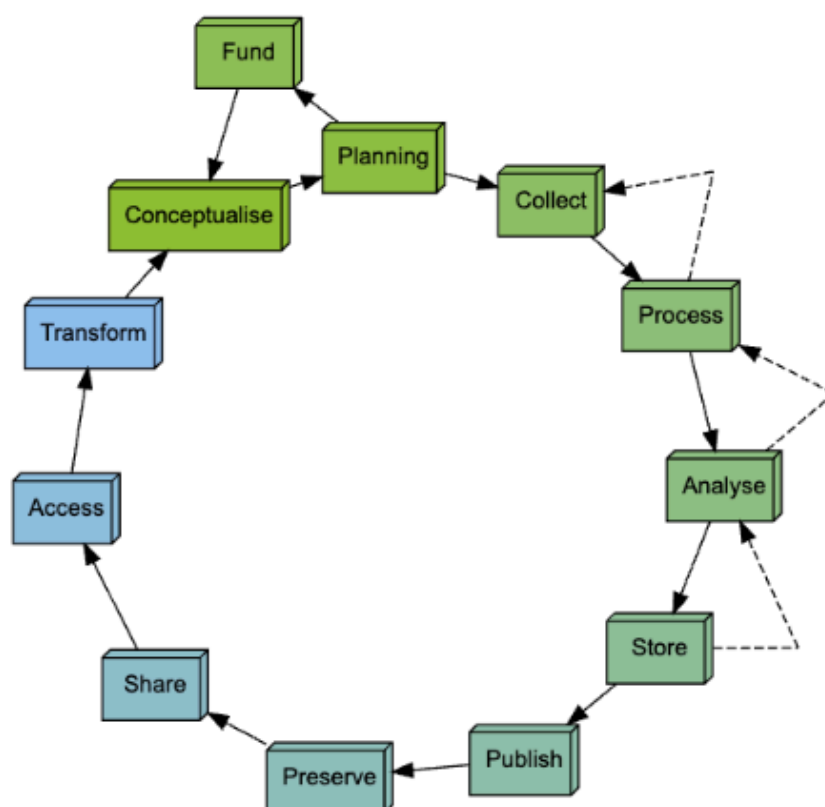


Figure 3. Initial Mapping the Landscape of Digital Research Tools Harmonised (MaLDReTH) Research Data Lifecycle Model (RDL) (a) and crosswalk to existing models created using Figma (b).



The MaLDReTH
Research Data Lifecycle

Figure 4. Mapping the Landscape of Digital Research Tools Harmonised (MaLDReTH) Research Data Lifecycle (RDL) Model created using GraphViz in the R prototype app.

Mapping the Landscape of Digital Research Tools Harmonised (MaLDReTH) Research Data Lifecycle (RDL) Stages and Definitions

CONCEPTUALISE: To formulate the initial research idea or hypothesis, and define the scope of the research project and the data component/requirements of that project.

PLAN: To establish a structured strategic framework for management of the research project, outlining aims, objectives, methodologies, and resources required for data collection, management and analysis. Data management plans (DMP) should be established for this phase of the lifecycle.

FUND: To identify and acquire financial resources to support the research project, including data collection, management, analysis, sharing, publishing and preservation.

COLLECT: To use predefined procedures, methodologies and instruments to acquire and store data that is reliable, fit for purpose and of sufficient quality to test the research hypothesis.

PROCESS: To make new and existing data analysis-ready. This may involve standardised pre-processing, cleaning, reformatting, structuring, filtering, and performing quality control checks on data. It may also involve the creation and definition of metadata for use during analysis, such as acquiring provenance from instruments and tools used during data collection.

ANALYSE: To derive insights, knowledge, and understanding from processed data. Data analysis involves iterative exploration and interpretation of experimental or computational results, often utilising mathematical models and formulae to investigate relationships between experimental variables. Distinct data analysis techniques and methodologies are applied according to the data type (quantitative vs qualitative).

STORE: To record data using technological media appropriate for processing and analysis whilst maintaining data integrity and security.

Note **COLLECT > PROCESS > ANALYSE > STORE** may be a repeating cycle.

PUBLISH: To release research data in published form for use by others with appropriate metadata for citation (including a unique persistent identifier) based on FAIR principles.

PRESERVE: To ensure the safety, integrity, and accessibility of data for as long as necessary so that data is as FAIR as possible. Data preservation is more than data storage and backup, since data can be stored and backed up without being preserved. Preservation should include curation activities such as data cleaning, validation, assigning preservation metadata, assigning representation information, and ensuring acceptable data structures and file formats. At a minimum, data and associated metadata should be published in a trustworthy digital repository and clearly cited in the accompanying journal article unless this is not possible (e.g. due to the privacy or safety concerns).

SHARE: To make data available and accessible to humans and/or machines. Data may be



ORACLE
for Research

shared with project collaborators or published to share it with the wider research community and society at large. Data sharing is not limited to open data or public data, and can be done during various stages of the research data lifecycle. At a minimum, data and associated metadata should be published in a trustworthy digital repository and clearly cited in the accompanying journal article.

ACCESS: To control and manage data access by designated users and reusers. This may be in the form of publicly available published information. Necessary access control and authentication methods are applied.

TRANSFORM: To create new data from the original, for example: (i) by migration into a different format; (ii) by creating a subset, by selection or query, to create newly derived results, perhaps for publication; or, iii) combining or appending with other data

Contributors

	Name	Affiliation	ORCID
1	Wolmar Nyberg Åkerström	Uppsala University / ELIXIR Sweden / National Bioinformatics Infrastructure Sweden (NBIS)	0000-0002-3890-6620
2	Mohammad Akhlaghi	Centro de Estudios de Física del Cosmos de Aragón (CEFCA)	0000-0003-1710-6613
3	Louise Bezuidenhout	Leiden University	0000-0003-4328-3963
4	Francis P. Crawley	Good Clinical Practice Alliance - Europe (GCPA) & Strategic Initiative for Developing Capacity in Ethical Review (SIDCER)	0000-0002-6893-5916
5	Gavin Farrell	ELIXIR (EMBL-EBI)	0000-0001-5166-8551
6	Marcelo Garcia	King Abdullah University of Science and Technology (KAUST)	0000-0002-2927-2371
7	Margareta Hellström	Lund University	0000-0002-4154-2610
8	Malgorzata Lagisz	UNSW Sydney	0000-0002-3993-6127
9	Hea Lim Rhee	Korea Institute of Science and Technology Information (KISTI)	0000-0002-4171-5710
10	Rory Macneil	Research Space	0000-0002-8429-096X
11	Lauren Maxwell	University of Heidelberg, World Health Organization (WHO)	
12	Andrea Medina-Smith	National Institute of Standards and Technology (NIST)	0000-0002-1217-701X
13	Richard Pitts	Oracle for Research	0000-0002-2037-3360
14	Marina Razmadze	CHU Sainte-Justine	



ORACLE
for Research

15	Martina Stockhause	German Climate Computing Center (DKRZ)	0000-0001-6636-4972
16	Ville Tenhunen	EGI Foundation	0000-0003-0217-0831
17	Adam Vials Moore	Jisc	0000-0002-2085-1908
18	James Wilson	University College London (UCL)	0000-0002-8546-1142
19	Nina Leonie Weisweiler	Helmholtz Association	0000-0001-6967-9443