Data Description Registry Interoperability*
Working Group: Case Statement

* The alternative title for this group is ‘Cross-Platform Discovery for Research Data’, aiming clarity of the objectives and deliverables.

Version	Date	Edited by	Comment
0.1	30-Oct-13	Amir Aryani	First Draft
0.2	31-Oct-13	Susannah Sabine	Reviewed
0.3	1-Nov-13	Adrian Burton	Reviewed
1	14-Nov-13	Amir Aryani	Addressed Adrian comments and changes to adopting institutions, deliverables and engagement with existing work
1.1	15-Nov-13	Amir Aryani	Minor formatting
1.2	15-Nov-13	Jon Corson-Rikert	Reviewed
1.3	29-Nov-13	Amir Aryani	Revised deliverables and addressed comments by Ross Wilkinson, Jon Corson-Rikert, Simeon Warner, Suenje Dallmeier-Tiessen, Sebastian Peters
1.4	11 Dec 2013	Amir Aryani	Submitted for the RD-A community review
1.5	16 Jan 2014	Amir Aryani	Changed the group title and revised section 1
1.6	18 Jan 2014	Adrian Burton	Revised section 1
1.7	18 Jan 2014	Amir Aryani	Edited section 1
1.7-sc	22 Jan 2014	Simon Cox	Reviewed sections 1, 3, 5 and 6
1.8	23 Jan 2014	Amir Aryani	Improved sections 1-5 and 6
1.8.1	28 Jan 2014	Amir Aryani	Minor formatting and new web links in section 2
1.8.2	5 Feb 2014	Amir Aryani	Revised the title
1.8.3	19 Feb 2014	Amir Aryani	da|ra and Data-Pass joined the group, and Adrian Burton the new co-chair
1.8.4	20 Feb 2014	Amir Aryani	Added “Data in Context IG” as a collaborative community
1.8.5	11 June 2014	Dimitris Gavrilis	Add DCU as a new adopting institution

1. Problem Statement and Scope

In recent years there has been a significant growth of research data repositories and registries. These repositories are fragmented across institutions, countries and research domains. As such, finding research datasets is not a trivial task for many researchers.

This group aims to address the problem of cross-platform discovery through a series of bi-lateral information exchange projects and work toward open, extensible, and flexible cross-platform research data discovery software solutions. Developing such solutions requires answers to problems including author disambiguation, persistent identifiers (requires for identity resolution and disambiguation), authentication (e.g. commercial

publishers), access right management, search optimisation (search ranking), metadata exchange (crosswalks), and creating a connected graph of research datasets, authors, publications and grants.

This group does not aim to develop new standards or create new protocols. Instead this group will leverage existing e-infrastructure and collaborate with other RD-A groups and research communities to build a set of working cross-platform software solutions.

Where research data registries and repositories provide machine-to-machine readable interfaces, the issue of wider discovery is often addressed either by metadata aggregation (collecting records from these registries and making them accessible through an aggregator portal) or federated search (exchanging queries and search results within a federation). However, the main problem is providing scientists search results for datasets that are actually relevant to their research. Such relevance depends on research context, and as a result enabling cross-platform discovery includes providing a connected graph of researchers, research activities (projects and grants), research datasets, publications and other research outcomes and research concepts. Such a linked snapshot of global research depends highly on persistent identifiers as exemplified by the EU-funded ODIN project[1] -- collaboration between ANDS, arXiv, DataCite, CERN, DRYAD and ORCID – where the connections between elements of the research data ecosystem are defined as *Identity Awareness of Research Data*[2].

This working group extends the concept of Identity Awareness of Research Data and will benefit from the outcomes of the following RD-A communities:

- *Data in Context Interest Group*
- *Data Type Registries Working Group*
- *Metadata Standards Directory Working Group*
- *PID Information Types Working Group*
- *Publishing Data Interest Group*
- *Standardisation of Data Categories and Codes WG*

This working group does not aim for a monolithic solution, avoiding a one uber-portal to rule them all. Rather it compiles simple enabling infrastructures based on existing open protocols and standards with a flexible and extensible approach that allows registries to opt-in and enables any third-party to create particular global views of research data.

2. Adopting Institutions

Which institutions will adopt the outcomes of this working group?

The outcome and the deliverables of this working group will be the result of the direct contribution of the following major institutions in Australia, US and Europe:

- Australia:
    - **ANDS** (Australian National Data Service). Research Data Australia[3] -- ANDS' flagship service -- an internet-based multidisciplinary discovery service designed to provide rich connections between data, projects, researchers and institutions, and promote visibility of Australian research data collections in search engines.

[1] odin-project.eu

[2] Amir Aryani, Adrian Burton, Identity Awareness: Toward an Invisible e-Infrastructure for Identifying Data and Authors; eResearch Australasia 2012. DOI: 10.6084/m9.figshare.155650

[3] researchdata.ands.org.au

ANDS will contribute to the outcome of this working group by bilateral interoperability projects between Research Data Australia and other research data registries in this working group.

- US:
  - **Data-PASS**[4]. The Data Preservation Alliance for the Social Sciences (Data-PASS) is a voluntary partnership of organizations created to archive, catalog and preserve data used for social science research. Examples of social science data include: opinion polls; voting records; surveys on family growth and income; social network data; government statistics and indices; and GIS data measuring human activity. A National Digital Stewardship Alliance Founding Member, the Data-PASS partnership works to:
    - Archive social science data collections at-risk of being lost.
    - Catalog and promote access to archived collections in the Data-PASS shared catalog.
    - Replicated preservation of archived collections.
    - Advocate best practices in digital preservation.

    Data-PASS will contribute to this working group by collaborating with da|ra in an interoperability project based on DDI model.

  - **Dryad.** The Dryad Digital Repository[5] is a curated resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of datatypes.

    Dryad will contribute to this working group by exploring the existing infrastructures to enable cross-platform discovery of data between Research Data Australia and Dryad.

  - **Thomson Reuters DCI** (Data Citation Index)[6]**.** DCI from Thomson Reuters is designed to be the source of data discovery for the sciences, social sciences and arts and humanities. DCI indexes a significant number of the leading data repositories including over two million data studies and datasets.

    DCI will contribute to this group by creating a data description exchange platform and other research data registries. The function of this platform will be demonstrated through an interoperability project with Research Data Australia.

  - **VIVO Cornell.** VIVO[7] is a research-focused multidisciplinary discovery tool that enables collaboration among researchers across all disciplines. The VIVO project was initiated and resides in the Cornell University Library, with support from the Office of the Provost.

    VIVO Cornell will contribute to this working group by working on an interoperability approach between VIVO sites and other platforms, building on the success of the VIVO platform in US, Central America, Europe and Australia.

- Europe:
  - **CERN** with collaboration of DESY, Fermilab and SLAC have built the next-generation High Energy Physics (HEP) information system, INSPIRE[8]. It combines the successful SPIRES database content, curated at DESY, Fermilab and SLAC, with the Invenio digital library technology developed at CERN. INSPIRE is run by a collaboration of the four labs, and interacts

[4] data-pass.org

[5] datadryad.org/pages/organization

[6] thomsonreuters.com/data-citation-index

[7] vivo.cornell.edu/about

[8] inspirehep.net

closely with HEP publishers, arXiv.org, NASA-ADS, PDG, HEPDATA and other information resources.

CERN will contribute to this working group by exploring the common infrastructure and creating a software solution for interoperability between Research Data Australia and INSPIRE registries.

- **DANS**[9] promotes sustained access to digital research data, and encourages researchers to archive and reuse data in a sustained manner, e.g. through the online archiving system EASY. DANS also provides access, via NARCIS.nl, to thousands of multidisciplinary datasets, e-publications and other research information in the Netherlands.

  DANS will contribute to this working group by implementing a data description exchange service between NARCIS and Research Data Australia.

- **da|ra**[10] is the DOI registration service in Germany (and beyond) for social and economic data jointly run by the GESIS LeibnizInstitute for Social Sciences and the ZBW - Leibniz Information Centre for Economics. This infrastructure lays the foundation for long-term, persistent identification, storage, localization and reliable citation of research data. More than 20 clients have registered about 290.000 DOI names (last update February 2014). Every DOI name is linked to a comprehensive set of metadata, a collection of bibliographical and content information, which refer to the registered dataset (title, author, publication date, copyright etc.). These metadata are compliant with the DDI metadata standard and stored the da|ra database. The da|ra service is provided in cooperation with DataCite, the international initiative to establish easier access to digital research data.

  da|ra will contribute to this working group by Exploring the interoperability of da|ra and Data-PASS registries, and working on cross-platform discovery based on DDI model.

- **DataCite**[11] is an international consortium which aims to improve data citation in order to enable easier access to research data on the Internet, increase acceptance of research data as legitimate, citable contributions to the scholarly record, and support data archiving that will permit results to be verified and re-purposed for future study.

  DataCite will contribute to this working group by investigating the potential opportunities for exposing and supporting standard query services which enable data description exchange between DataCite members, associated data centres and research data registries by using DataCite APIs.

- **Digital Curation Unit (DCU)**[12] is part of the Institute for the Management of Information Systems of the "Athena" Research and Innovation Centre in Information, Communication and Knowledge Technologies. DCU's mission is to conduct research, develop technologies and applications, provide services and training and act as a national focus point in the field of digital curation.
  - DCU will contribute to this working group by creating a de-duplication service for humanity related research. The service will focus on author disambiguation and dataset de-duplication. This work enables an interoperability project between DCU and DANS.

[9] dans.knaw.nl

[10] da-ra.de/en/home

[11] datacite.org

[12] dcu.gr

In addition, the researchers, practitioners and eResearch experts from the following institutions will engage in the conversations of this working group and provide feedback on the relevance of the outcomes to the broader community: arXiv, Griffith University, RMIT University, University of Adelaide and University of Melbourne.

Please note that the projects and deliverables that are described in this document represent only the initial plan for the working group. We envisage that the scope of these projects and deliverables will be extensible and new adopting institutions will join the group in the near future.

3. Deliverables

Within a 12 month time frame what will its "deliverables" or outcomes be?

This working group investigates existing infrastructures, standards and protocols in order to enable cross-platform discovery of research data. In order to evaluate the function of these infrastructures and protocols, this group conducts a set of bilateral interoperability projects across major research data repositories. These projects would enable this group to identify some of the opportunities or gaps for creating global research infrastructure, and outcome of these project will be presented to the Research Data Alliance community as a gap analysis report. Specifically, in the next 12 months this working group provide the following deliverables:

D1. Kick-off meeting (M1)

A virtual meeting (video conference), hosted by ANDS.

> The outcome of this meeting will be the confirmed timelines of the projects. The minutes will be published as a brief open report to the the Research Data Alliance community.

D2. Plenary Workshop 1 (M1-3)

Host a workshop at the third Research Data Alliance plenary conference to discuss the interoperability platforms, existing infrastructures, standards and protocols.

> A brief report of this workshop will be published that shows the preliminary work on the interoperability approach.

D3. Investigating the Existing Standards and Protocols (M1-3)

Explore existing infrastructures, standards and protocols in order to identify a lightweight interoperability approach and cross-platform discover solutions that can leverage these existing platforms.

> The outcome of this deliverable is a collaborative report by adopting institutions to describe the common platforms and opportunities for leveraging existing infrastructures for creating discovery software solutions. The engaging partners will provide feedback on this report, describing the relevance of the proposed approach to their research data infrastructure.

D4. INSPIRE-HEP & Research Data Australia Interoperability Project (M3-9)

Explore the possible solutions for interoperability between ANDS Research Data Australia and CERN INSPIRE-HEP research data registries. There are a noticeable number of dataset metadata records in INSPIRE-HEP from the Australian high energy physics community. The aim of this project is to connect these records to Australian research grants, and other related research activities in Research Data Australia; hence, improve the connectivity, re-usability of data, and enable cross-platform discovery.

The INSPIRE-HEP and Research Data Australia are data aggregators that are connected to a number of research institutions, as such this project can provide a great value-added service to the international high energy physics community.

The outcome of this project will be a workflow for interoperability and cross-platform discovery between Research Data Australia and INSPIRE-HEP; in addition, this project will deliver the preliminary software solution, and a roadmap for future development.

D5. Dryad & Research Data Australia Interoperability Project (M3-9)

Explore the existing infrastructures for enabling cross-platform discovery of research data between Research Data Australia and Dryad. This project provides the preliminary web services that demonstrate the function of existing interoperability platforms and provide a report for future area of development.

The outcomes of this project are the preliminary software solutions and a report that explores the potential extension of these services by leveraging the existing common infrastructures such as DataCite DOIs and ORCID.

D6. NARCIS & Research Data Australia Interoperability Project (M3-9)

Create a data description exchange solution between the ANDS Research Data Australia and DANS NARCIS registries. As part of this deliverable ANDS and DANS will explore existing standards and protocols, and create software web services that enables cross-platform discovery of research data across Australian and Dutch research sectors.

The outcome of this project are the preliminary software solutions and a roadmap for future development of the interoperability platform between Research Data Australia and NARCIS.

D7. VIVO Data Description Exchange Solutions (M3-9)

Explore an interoperability approach between VIVO sites and other platforms, building on the success of the VIVO platform in US, Central America, Europe and Australia. As part of this project ANDS and VIVO Cornell explore how the VIVO technology, RDF and Linked Data can be leveraged for interoperability between VIVO sites and other platforms such as Research Data Australia.

The outcome of this deliverable will include preliminary working services for cross-platform discovery and research data description exchange between VIVO Cornell and ANDS Research Data Australia. In addition, Dryad, arXiv and other members of the group will provide feedback on future extension of these services to other VIVO and non-VIVO platforms.

D8. DCI & Research Data Australia Interoperability Project (M3-9)

Thomson Reuters is collaborating with ANDS to build an efficient data description exchange service between ANDS Research Data Australia and Data Citation Index (DCI). This project examines the opportunities for leveraging the commercial platforms such as DCI to increase the discoverability of research data.

The outcome will include open data description exchange services that enables research data registries push their data descriptions to the DCI platform and retrieve information about data citation, citation

counts and bibliographic records. In addition, this project provides a report that describes the existing DCI interoperability platform and future areas of development.

D9. DataCite Data Description Exchange Solutions (M3-9)

Investigate the potential opportunities for exposing and supporting standard query services which enable data description exchange between DataCite members, associated data centres and research data registries by using DataCite APIs.

The outcome will be a report on the technical and operational experience of creating cross-platform discovery software solutions based on DataCite services, and a roadmap for future development of DataCite discovery services.

D10. da|ra and Data-PASS Interoperability Project (M3-9)

Exploring the interoperability of da|ra and Data-PASS registries. In the present situation there are a noticeable number of metadata records in both registries. Our goal is to improve the connectivity of the two inventories and enable cross-platform discovery. In the course of this project da|ra and Data PASS will examine the opportunities of a metadata exchange solution based on DDI-lite.

The outcomes of the project are a preliminary software solution (web service) and a roadmap for future development of the interoperability between da|ra and Data-PASS.

D11. DCU and DANS Interoperability Project (M3-9)

This project involves a de-duplication service for humanities related content. This de-duplication service will focus on Authors and Datasets. It will take into account available information such as subject terms, descriptions, possibly geo-spatial information in order to identify first authors and secondly datasets. The datasets that are going to be used is those of DANS (NARCIS and possibly EASY) and the focus will be on authors. The service will work on XML encoded data and will be able to “understand” metadata schemas that are found in the humanities domain.

The outcome of this project will be software services that demonstrate the author disambiguation and datasets de-duplication using DANS NARCIS and EASY platforms.

D12. Enabling Infrastructure Prototype (M6 - 12)

The goal of this deliverable is to explore the enabling infrastructure for interoperability at the global level. This deliverable will build on the experience of the projects in this working group and create a prototype of a platform that enables cross-platform discovery using different technologies (e.g legacy data and state of the art registries).

The outcome will be a prototype of the enabling infrastructure and a report that describes how such an infrastructure can assist extending the outcome of this working group at the global level.

D13. Plenary Workshop 2 (M9)

Host a workshop at the fourth Research Data Alliance plenary conference to discuss the commonalities and lessons learnt from the bilateral interoperability projects.

The report of the discussions in this workshop will be published to reflect the group members' understandings and lessons learnt from the interoperability projects and outline future areas of investigation.

D14. Gap Analysis and Roadmap (M11-12)

This deliverable will build on the outcome of the plenary workshop and provide a report on the outcomes of this working group, gap analysis and a roadmap for future work in this area.

4. Value Proposition

Who will benefit from the adoption or implementation of the WG outcomes?

For the research community, there is a significant value in discovery and easy access to research data. Although enabling interoperability at the global level is beyond the scope of this working group, the deliverables from this group are the initial steps toward global data infrastructure. Specifically the outcomes of this working group will demonstrate how registries can improve the discovery and visibility of research data by enabling researchers to find research datasets with a single search across multiple infrastructures and platforms.

Value for participating research data registries comes from the network-benefit effect, where any given registry is more valuable when it is part of a global view of research data. Research is an inherently global enterprise and researchers quite naturally tend toward more global views on research data. The research data cross-platform discovery initiative proposed here is designed to allow data registered in one domain to be more easily visible in another, thus increasing the value of registering data in any of the participating registries.

This infrastructure is designed also to create value for third-party data discovery portals, internet indexing engines, etc by making it cheaper, better, and easier to connect and aggregate information about research data.

In addition, the outcome of this working group contributes to the practice of data management by enabling system designers and engineers to create robust interoperable services that can exchange data descriptions across institutions. Such interoperability can reduce the need for entering duplicated records in multiple platforms, and subsequently reduces the admin work by data managers.

Value for researchers and research organisations who register research data in participating registries comes from increased global exposure with subsequent increased re-use, acknowledgement, impact, and opportunities for collaboration.

Value for funders of research and those who design the research system as a whole comes from increased visibility of the outputs of funded research projects and decreased replication of data generation with subsequent increased return on research investment and greater ability to address global challenges.

5. Engagement with Existing Work

Related work and plan for engagement with other activities in the area

This working group will benefit from the collaboration and outcomes of the following RD-A communities:

- *Data Type Registries Working Group*
- *Metadata Standards Directory Working Group*
- *PID Information Types Working Group*

- *Publishing Data* Interest Group
- *Standardisation of Data Categories and Codes WG*

In addition, this working group build on the success of ODIN project (ORCID and DataCite Interoperability Network), a European Commission funded project to investigate identifier interoperability between two global identifier initiatives, DataCite and ORCID.

6. Invited Members

The following list shows the initial participants in this working group. If you are interested to get involved in this group, please send an email to Amir Aryani (amir.aryani@ands.org.au).

Name	**Institution**	**Contact Information**	**Status**
Amir Aryani (co-chair)	ANDS	amir.aryani@ands.org.au	Confirmed
Adrian Burton (co-hair)	ANDS	adrian.burton@ands.org.au	Confirmed
Abe Lederman	Deep Web Technologies	abe@deepwebtech.com	Confirmed
Ben Greenwood	ANDS	ben.greenwood@anu.edu.au	Confirmed
Brigitte Hausstein	da|ra	brigitte.hausstein@gesis.org	Confirmed
Cathy Miller	University of Adelaide	cathy.miller@adelaide.edu.au	Confirmed
Gudmundur Thorisson	ORCID	gthorisson@gmail.com	Confirmed
Jan Brase	DataCite	jan.brase@tib.uni-hannover.de	Confirmed
Jared Lyle	Data PASS	lyle@umich.edu	Confirmed
John Kaye	British Library	john.kaye@bl.uk	Confirmed
Jon Corson-Rikert	Cornell	jc55@cornell.edu	Confirmed
Jonathan Hodge	CSIRO	jonathan.hodge@csiro.au	Confirmed
Laurents Sesink	DANS	laurents.sesink@dans.knaw.nl	Confirmed
Laure Haak	ORCID	L.Haak@orcid.org	Confirmed
Laura Paglione	ORCID	l.paglione@orcid.org	Confirmed
Leo Monus	ANDS	leo.monus@ands.org.au	Confirmed
Liz Woods	ANDS	liz.woods@ands.org.au	Confirmed
Mark Fallu	Griffith	m.fallu@griffith.edu.au	Confirmed
Mary Vardigan	Data PASS	vardigan@umich.edu	Confirmed
Minh Nguyen	ANDS	minh.nguyen@ands.org.au	Confirmed
Mohsen Laali	RMIT	mn.laali@gmail.com	Confirmed
Natasha Simons	Griffith University	n.simons@griffith.edu.au	Confirmed
Nigel Robinson	Thomson Reuters	nigel.robinson@thomsonreuters.com	Confirmed

Prof. Heinz Schmidt	RMIT	heinz.schmidt@rmit.edu.au	Confirmed
Salvatore Mele	CERN	salvatore.mele@cern.ch	Confirmed
Sergio Ruiz	DataCite	Sergio.Ruiz@datacite.org	Confirmed
Simon Porter	University of Melbourne	simon.porter@unimelb.edu.au	Confirmed
Simeon Warner	arXiv	simeon.warner@cornell.edu	Confirmed
Sunje Dallmeier-Tiessen	CERN	sunje.dallmeier-tiessen@cern.ch	Confirmed
Todd Vision	Dryad	tjv@bio.unc.edu	Confirmed

Conclusion

This document presented an overview of the new RD-A working group for Data Description Registry Interoperability. This working group focuses on the challenge of cross-platform discovery and interoperability between research data registries. The deliverables of this group will include working software services and pragmatic methods that enable finding datasets across multiple registry systems.