TE PŌKAI TARA
UNIVERSITIES
NEW ZEALAND

RESEARCH DATA MANAGEMENT FEASIBILITY STUDY REPORT

CONZUL Working Group

This document details a current state, opportunity and recommendations for CONZUL members to consider when crafting a CONZUL-wide position on research data management (RDM)

Authors: Max Wilkinson, Howard Amos, Brian Flaherty, Shari Hearne, Helen Lynch, Heather Lamond, Natalie Dewson, Mike Kmiec, Janette Nicolle, Erin Talia-Skinner and Gillian Elliot.

1 July 2016

New Zealand Vice-Chancellors' Committee | Level 9, 142 Lambton Quay | PO Box 11915 | Wellington 6142 | New Zealand

T 64 4 381 8500 | F 64 4 381 8501 | **W** www.universitiesnz.ac.nz

CONZUL Research Data Management Working Group

Feasibility Study: National Research Data Registry

Project Report: July 2016

Members

Howard AMOS: CONZUL Sponsor
Max WILKINSON: Project Manager
Brian FLAHERTY: The University of AUCKLAND
Shari HEARNE: AUCKLAND University of Technology
Helen LYNCH: The University of WAIKATO
Heather LAMOND: MASSEY University
Natalie DEWSON: MASSEY University (to replace Heather LAMOND)
Mike KMIEC: Victoria University of WELLINGTON
Janette NICOLLE: University of CANTERBURY
Erin TALIA-SKINNER: LINCOLN University
Gillian ELLIOT: University of OTAGO

F2F Meeting Schedule

THURSDAY 25th February 2016 Universities New Zealand Wellington
FRIDAY 8th April 2016, Universities New Zealand Wellington
FRIDAY 3rd June 2016 Universities New Zealand Wellington

Table of Contents

Executive Summary

The CONZUL Research Data Management (RDM) Working Group (WG) was re-convened to undertake a feasibility study arising from recommendation 4 from the CONZUL RDM Framework Report of 2015. The group considered a National Research Data Registry (NRDR) and the benefit it would deliver to New Zealand University members, their researchers and other stakeholders. The group investigated 5 project elements; (1) metadata collection, (2) technology platforms, (3) metadata standards, (4) use cases and (5) governance. The working group demonstrated a proof of concept instance by presenting harvestable metadata in one university system and then harvesting these metadata into an independent system at another university.

The Working Group found that collecting and presenting appropriate metadata would be a significant effort for member institutions. Many institutions did not have processes or systems in place to locate, negotiate access and manage metadata aggregation. While it was not considered necessary to have exactly the same databases or repositories in each member's institution, it was considered important to have some solution to manage metadata that a harvesting application programme interface (API) or distributed search protocol could access.

The Working Group found that technology was not a significant burden. It was further agreed that the most sensible approach would be working with existing local systems and infrastructures as much as possible rather than replicating one system across all eight universities. This would minimise any non-technical issues such as expectations and co-location of similar research content. An existing harvesting service such as NZ Research would provide a window of opportunity that could demonstrate a harvesting service while members prepared for a longer term solution that had fewer governance risks and greater benefit to more stakeholders, including non-university institutions.

The Working Group found that metadata standards existed that could fulfill many of the requirements of a NRDR. A set of metadata principles and minimal metadata were suggested and a brief exercise in mapping desired elements to established schemas was undertaken. The value in re-using existing metadata schemas is that schema maintenance can be delegated to established bodies, in particular Open Research and Contributor Identifier (ORCiD) and/or DataCite even commercial products such as Symplectic Elements. Such an approach would result in a simple schema rather than a detailed, discipline specific schema, which is not considered critical.

The Working Group established that university users (both academic and administrative) continued to value the idea of an NRDR but also there was some confusion concerning its purpose and features; most of those interviewed continued to assume that the existence of a catalogue meant that the research data could be accessed via the catalogue, while those users interested in meta analyses had no real need to access the data underlying the metadata. Many users were concerned with the sustainability of any such NRDR.

The Working Group found that governance was a complex challenge. Existing structures that were examined ranged from large international consortia to national representative committees. Considering the lack of readiness of many CONZUL members to manage and present metadata the working group suggested a small, focused and manageable structure over large and comprehensive governance. If the primary service infrastructure could also be delegated to third parties then the existing CONZUL committee, or some other similarly senior committee may be willing to take on the strategic and oversight roles required.

The Working Group considered all project elements and recommends that a phased approach to a NRDR should be adopted where existing services are leveraged for the purpose of establishing an NRDR while a longer term plan to establish a more robust and sustainable NRDR is agreed. In recommending this approach the Working Group believes that risks will be offset and minimised, expectations can be managed and service experiences can be positive for the majority of stakeholders.

Background

A Working Group was established in 2015 to report on the development of a CONZUL-wide position on research data management. The Working Group examined RDM activity across CONZUL member institutions and considered a series of university-focused benefits of RDM before submitting a series of recommendations. The group sought to identify and promote those areas of RDM that would benefit from a national perspective on RDM and in doing so recognised that some RDM issues are better managed locally. In addition, the group recommended facilitating a sharing of ideas amongst CONZUL members such that all universities can benefit from the experiences of others, including any solutions which might be implemented in response to RDM challenges.

CONZUL accepted the Framework Report and recommendations of the WG and agreed to the WG reconvening to address two of these:

> *RECOMMENDATION 4: "CONZUL should undertake a feasibility study to investigate and appraise potential national data registry platforms in two phases. First, a six-month project to investigate and test approaches for a registry and discovery service and second, a pilot of the preferred option. The study should investigate the extensibility of local solutions as both a metadata store for an institutional data registry and as harvestable metadata sources."*
>
> *RECOMMENDATION 5: CONZUL should establish a position statement on research data licencing that encourages data sharing and reuse to the widest possible audience. This may be via an existing initiative, committee or national programme like eResearch2020 or Universities New Zealand Copyright Working Group. The impact of licencing is such that a limited stakeholder group should be consulted to focus licencing concerns on specific needs of NZ research organisations promoting research data sharing and reuse.*

Recommendation 5 is provided in a separate discussion paper 'Ownership and Licensing of Research Data'.

The reconvened WG addressed recommendation 4 and undertook a feasibility study to assess the potential benefits of a NRDR. In the first instance the WG considered Universities only, excluding other stakeholders in the NZ data landscape like Crown Research Institutes (CRI's), polytechnics, government agencies or infrastructure providers, most notably REAANZ, NeSI and NZGL. This was to contain the scope of the feasibility study but with the acknowledgment that expansion of any such NRDR would need to engage these and other organisations.

The goal of this feasibility study is to investigate effort and to demonstrate ability to harvest defined metadata elements relating to research data from distributed locations and present them as a National Research Data Registry. The feasibility study was divided into separate work packages and each work package was assigned an individual lead. Working Group members were encouraged to actively contribute to all five work packages, where they felt able to do so.

Study Brief

A NRDR suggests a catalogue of the metadata and data created by New Zealand research organisations. Such a registry should store metadata about datasets and possibly a pointer to the underlying datasets, but not necessarily the data themselves. Any registry solution would need to:

- provide a discovery service to promote visibility and increase exposure;
- support interoperability, reuse, repurposing and increase collaboration;
- facilitate the verification and assessment of research data value and impact by research funding agencies.

Any NRDR must be underpinned by an agreed and appropriate metadata standard and governance structure that is sustainable and affordable. The creation of a new, or reuse of an existing standard requires evidence-backed guidance for the description of research data together with evidence on which supports those metadata elements, which it is argued, make the greatest contribution to discovery and reuse.

There are numerous sources of research data in the New Zealand research and government landscape and some, for example data.govt.nz have established registries of metadata and data[1]. Others provide services internal to the organisation, e.g. many of the CRIs, while still others such as NZGL, do not provide any metadata registry services for the data they generate at all. This study will consider Universities NZ stakeholders only, but the Working Group recognised that to be truly national, an NRDR would need to include these and other institutions as potential sources of research metadata and data.

Deliverables

1. Detailed report on possible approaches and recommendation for a National Research Data Register to support benefits outlined the RDM Framework report (2015).
2. Metadata specification for each participating institution and Proof of Concept (PoC) and/or feasibility analysis of a federated metadata registry

[1] https://data.govt.nz/

The benefit of using any such NRDR to calculate metrics or other quantitative or qualitative measures was not considered the role of the working group and placed out of scope for this feasibility study. Excluding metrics does not preclude any NRDR from being used for such assessments, simply that the working group felt it was not their role to undertake such activity.

WP1 Collection

It was assumed that each institution collects and host the metadata associated with its own research data. While the minimum metadata suggested in Work Package 3 is quite general, the real benefit for discovery and reuse comes with richer and standardised descriptions. The Framework analysis indicated that such metadata existed numerous systems across partner institutions, from dedicated data repositories, data management planning tools, HR and grant management systems. Because many internal systems are not available externally, collecting metadata required an 'internal-to-institution' approach to locating and negotiating metadata collection.

It is good information management practice that metadata is created once and reused. A cursory investigation of the research process reveals that metadata is created throughout the research life-cycle; from grant applications, to data management plans, data creation and review and finally publishing. Institutional data registries need to identify useful workflows and negotiate harvesting from these sources rather than defining new metadata to be captured.

There was a make use of a common research information management system Symplectic Elements[2], which is used by seven of the eight universities. The current use of Elements is focused on published outputs, such as papers, book chapters, conference papers, exhibitions etc., but it could also be used to describe published research datasets by incorporating common data elements from personal identifiers, ORCiD, or data citation frameworks such as DataCite. Independent data repositories are increasingly available as data sources to Elements through a rich set of APIs[3]. Data sources such as Figshare is already a metadata source in the latest release of Elements. However, it was noted that while many NZ universities implement Elements, they do so under separate operational structures and for slightly separate purposes. A simple extract from any Elements report may well be different for each university.

[2] http://symplectic.co.uk/products/elements/
[3] http://symplectic.co.uk/services/vivo-network/

Whatever the source of metadata, there will be non-trivial effort in modifying existing university or Institutional processes to enable the collection of appropriate metadata concepts and then additional effort from library resources to cross-walk the metadata into a schema required by the NRDR. Deposit workflows need to be created by institutions and managed in a local database to insure accurate and stable representations of collected metadata. An alternative strategy was identified where the effort in cross walking (but not collection) could be transferred to an external party such as DigitalNZ[4] who collect metadata about diverse NZ-focused content, including research outputs. The Digital NZ service would then collect metadata in any format specified and present it through its search portal, which also presents other content types, including publications. One clear benefit for this approach is that the research data content can be co-located with more traditional research output like publications, or other research content like images and historical artefacts.

Once metadata is collected from across the university, it would be managed in a dedicated database as this would make any harvesting efficient and under the control of a single authority, generally the institutional library.

The user experiences and expectations from Work Package 4 (use cases) indicated quite strongly that even with a comprehensive metadata catalogue, there was a strong expectation from researchers that the underlying research data would also be accessible. That said, other stakeholder groups concerned with meta-analysis (for example grant administrators, funding bodies and institutional review bodies) suggest that the metadata-only registries were sufficient for their expectations. Thus any such metadata registry should be capable of hosting metadata for datasets but also indicate access to those datasets, whether stored locally like the Landcare's Datastore[5] or in external discipline-specific repositories such as Dryad[6] and EarthStat[7]. The key issue here is the institutions ability to provide metadata for administrative benefits and the underlying research for researcher benefit. To fully realise the benefits of RDM to 'all' stakeholders, both metadata and the underlying data should, where appropriate, be made available, although it was noted that providing both was unrealistic for all CONZUL members.

In summary, while the specification of metadata could harmonise representation of research data across all partner institutions, three challenges in collection were identified; first locating and negotiating access to various metadata across the institutional systems which may not be under the

[4] http://www.digitalnz.org/
[5] http://datastore.landcareresearch.co.nz/
[6] http://datadryad.org/
[7] http://www.earthstat.org/

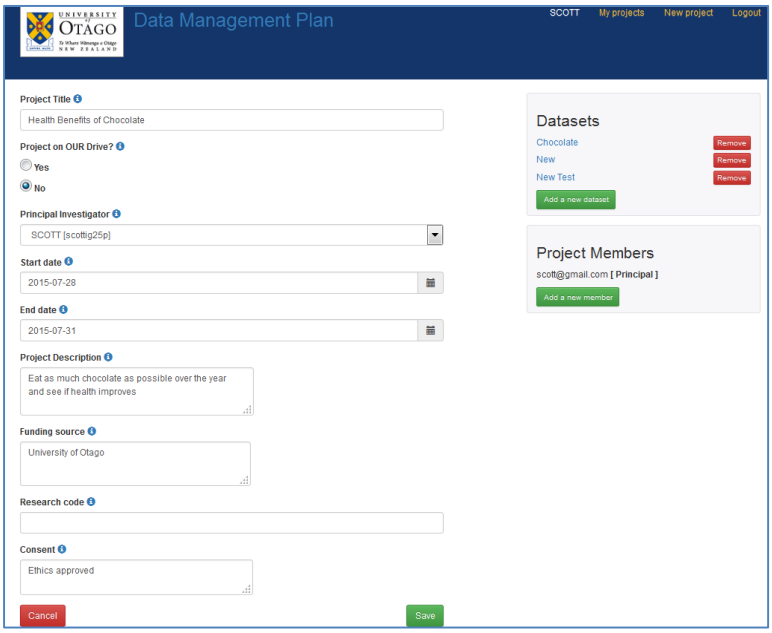
operational control of the library; second, the mapping or cross-walking those metadata to the standard agreed in any NRDR metadata model and; finally, making underlying research data available whether in local or external repositories.

Proof of Concept: Federated harvesting using Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH)

As an example of a federated harvest the University of Otago Library undertook a simple implementation of an OAI PMH API that could be used to harvest metadata into an external registry, in this case the Dspace repository located at the University of Auckland. There were four steps to establish the service:

1. University of Otago Data Management Planning (DMP) service was used to gather and present the metadata about research datasets (See Figure 1 below): https://dmp-test.otago.ac.nz/

Figure 1 List of records



2. The University of Otago DMP metadata was mapped to minimum Dublin Figure 2 University of Otago DMP Core for OAI-PMH harvesting (see Figure 2 below)

Figure 3 University of Otago DMP to Dublin Core crosswalk

DMP element	DC element	DC meaning and [DMP notes]
Dataset Title	dc: title	The name given to the resource by the CREATOR or PUBLISHER.
Creator	dc: creator	The person(s) or organization(s) primarily responsible for the intellectual content of the resource; the author.
Keywords	dc: subject	The topic of the resource; also keywords, phrases or classification descriptors that describe the subject or content of the resource.

DMP element	DC element	DC meaning and [DMP notes]
Description	dc: description	A textual description of the content of the resource, including abstracts in the case of document-like objects; also may be a content description in the case of visual resources.
Owner	dc: publisher	The entity responsible for making the resource available in its present form, such as a publisher, university department or corporate entity.
No DMP element	dc: contributor	Person(s) or organisation(s) in addition to those specified in the CREATOR element, who have made significant intellectual contributions to the resource but on a secondary basis.
Release date	dc: date	The date the resource was made available in its present form.
No DMP element	dc: type	The resource type, such as home page, novel, poem, working paper, technical report, essay or dictionary. It is expected that TYPE will be chosen from an enumerated list of types.
Format	dc: format	The data representation of the resource, such as text/html, ASCII, Postscript file, executable application or JPG image. FORMAT will be assigned from enumerated lists such as registered Internet Media Types (MIME types). MIME types are defined according to the RFC2046 standard.
Dataset id	dc: identifier	A string or number used to uniquely identify the resource. Examples from networked resources include URLs and URNs (when implemented).
No DMP element	dc: source	The work, either print or electronic, from which the resource is delivered (if applicable).
No DMP element	dc: language	The language(s) of the intellectual content of the resource
Citations	dc: relation	The relationship to other resources. Formal specification of RELATION is currently under development. [DMP notes: Publications which cite the dataset]
Coverage start date	dc: coverage	The spatial locations and temporal duration characteristics of the resource. Formal specification of COVERAGE is also now being developed. [DMP notes: Start of the data coverage; beginning of the date range (data may relate to current or historic date range]
Coverage end date	dc: coverage	The spatial locations and temporal duration characteristics of the resource. Formal specification of COVERAGE is also now being developed. [DMP notes: End of the data coverage; end of the date range (data may relate to current or historic date range]
Access permission	dc: rights	A link (URL or other suitable URI as appropriate) to a copyright notice, a rights-management statement or perhaps a server that would provide such information in a dynamic way. [DMP notes: Amount of information that is to be made OPENLY available… 4 options listed]

3. A developer coded and implemented an OAI-PMH service using the metadata from the University of Otago DMP following the OAI PMH developer documentation
4. Auckland harvested the metadata using a standard OAI PMH command (ListRecord) from their Dspace instance (See Figures 3 and 4 below)

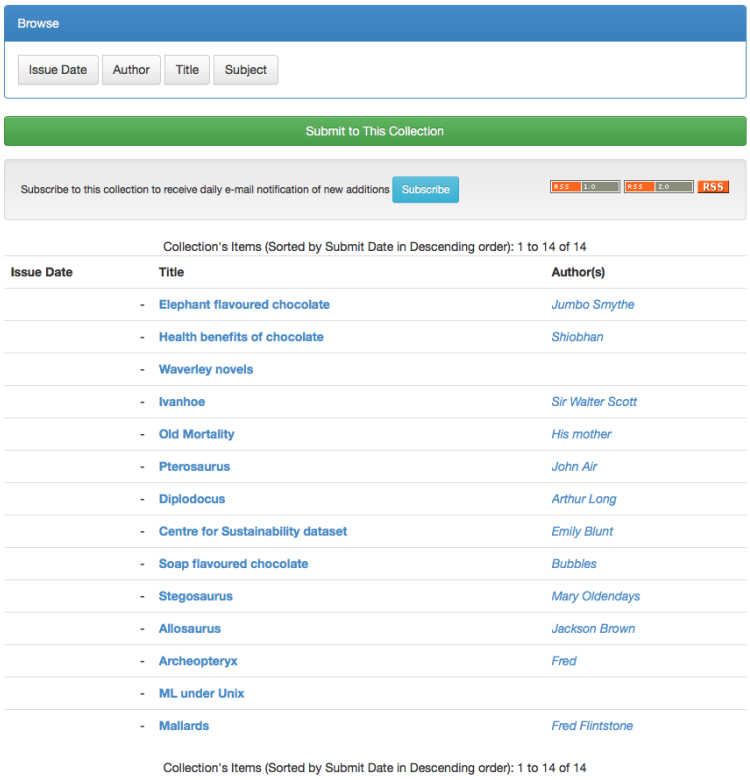


Figure 4 List of records

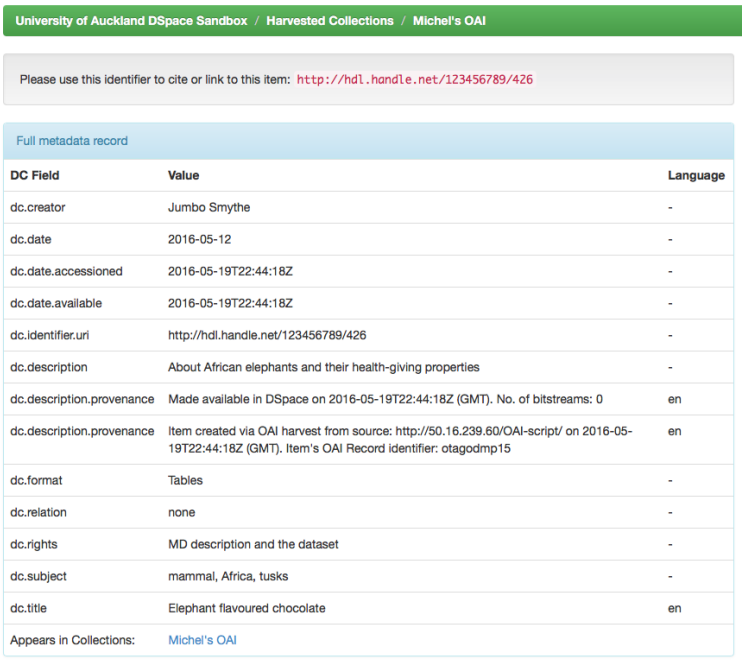


Figure 5 Record Metadata

Observations

The OAI-PMH is a well-established web protocol and can be readily implemented. The Otago developer was quite surprised how little effort was needed to code and implement the API for the local DMP metadata. The total effort required starting from a position of being completely

unfamiliar with OAI PMH through to implementing a basic harvest end point API was estimated as 10-15 hrs per week for 5 weeks. Further implementation and tuning would extend this to several more days.

As this was a proof of concept most, but not the entire protocol was implemented (e.g. had no 'groups of documents' so no need for sets at this stage); the developer estimated that an additional day or two would be required to implement the remaining part of OAI-PMH.

The metadata (about research datasets) was sourced from University of Otago's DMP. Metadata can potentially be harvested from any system e.g. Elements at Auckland and Wellington through OAI-PMH. The only requirement is that the metadata elements be mapped to the appropriate Dublin Core elements. For example, 11 metadata elements from the DMP were successfully mapped to 11 Dublin Core metadata elements (see Figure 2). Other metadata standards can be used, but OAI-PMH specifies that Dublin Core be implemented.

The code is available to any interested parties but it may be simpler to code and develop any APIs from scratch. Given the small effort required to code this PoC API, this was not a significant overhead.

WP2 Platform

There are a number of technical solutions to metadata registries - which can be defined as catalogues of resources hosted at distributed locations. National aggregations or registers of metadata relating to research data exist in a number of countries and several of these are detailed in the CONZUL RDM Framework Report (2015)[8].

In the New Zealand context, the establishment of a preferred technology to harvest and present collected research data metadata requires an awareness of related national infrastructure and services which may inform, overlap and even compliment any solution. These include **data.govt.nz** which is a directory of publicly-available New Zealand government datasets (including Crown Research Institutes), and **NZ Research** which harvests research publication (theses, articles, working papers) metadata from university and polytechnic repositories.

[8] http://www.universitiesnz.ac.nz/files/CONZUL-RDM%20Framework%20Report%202015%20FINAL.pdf

At the same time, it is important to acknowledge that research ecosystems at NZ universities are in a development phase –all have Research Information Systems, but few universities have a research data repository in production as of June 2016, or an institutional data asset register (metadata store). This makes it difficult to demonstrate the harvesting of research data metadata, as established harvest end points do not yet exist, for the most part.

Platform Functional Requirements

At the macro-level the requirements of any National Data Registry are likely to be discovery, interoperability and facilitating research assessment (although this third function could be delivered from elsewhere). The functional requirements listed here are not a full needs analysis, but a lightweight review, based on project use cases and the UK Research Data Discovery Service Statement of Requirements. [9] It is understood that any selected platform may not be able to deliver all functional requirements.

- Ensure that no duplicate records show in the registry, should they be harvested from different sources (compare unique identifiers).
- Ensure different versions of a dataset are uniquely identified.
- Available, robust and documented APIs for presenting standardised data (The format made available needs to be defined: HTML, RDF, JSON, CERIF…)
- All metadata to be indexed by Google and other search engines
- Persistent URLS for individual records (and search results)
- A responsive, easy to configure front-end that supports standard web development (css) and the ability to promote as well as discover research.
- Pre- and post-search filtering options to narrow search on defined metadata fields (e.g. datatype, institution, licence types, published status
- The ability to and manage relationships between data objects
- Support for an extensible (and customisable) data schema
- Collection/harvesting of metadata from institutional data repositories and other data aggregator services across a range of data format including OAI-PMH, XML and RDF/XML.
- Geo-searching via map interface for georeferenced datasets

[9] https://www.jisc.ac.uk/rd/projects/uk-research-data-discovery

- Ability to interrogate and find data using semantic web technologies
- Authentication/ authorisation for administrator roles
- Authentication of users to create custom metadata sets – favourites etc.
- Publishing of metadata via a OAI-PMH stream
- Harvesting logs available for analysis
- Broken link monitoring
- Suppression of records or collections from the search index
- Crosswalk functionality for mapping metadata from a range of data sources
- Metrics / usage data should be captured for the purposes of analytics.

Note:

- Digital Object Identifier (DOI) creation at this point is the responsibility of the data owners rather than a national data registry. (There is certainly a case to be made for a national DOI minting agency).
- Crosswalks of metadata to an agreed schema could either take place at the institution level pre-harvest, or as part of the ingest process.
- The verification and assessment of research data value and impact (including PBRF reporting) are currently provisioned by institutional Research Information Systems (e.g. Symplectic Elements). While robust APIs for a national data registry should be a requirement, it is recommended that reporting for funders etc. should remain the responsibility of each institution. (To centralise this function would require either considerable functional enhancement to existing data registry applications, or to provision a national research information system – e.g. a national Symplectic Elements instance.

Analysis

A full analysis of all potential registry solutions was beyond the scope of this feasibility study. Six platforms were assessed against the functional requirements, as well as for potential synergies within the national research landscape. Other research repository software such as Eprints http://www.eprints.org/uk/, Hydra http://projecthydra.org, and Dataverse http://dataverse.org/ could potentially be repurposed as registry services, but were considered to be primarily repository applications. Nor were commercial services (such as Figshare https://figshare.com/) or hosted, third-party offerings like Zenodo https://zenodo.org/dev considered. Two options stood out as fitting closest to the functional requirements:

Comprehensive Knowledge Archive Network (CKAN)

CKAN is an open-source data portal platform providing tools to streamline publishing, sharing, finding and using data.[10] It offers a powerful API that allows third-party applications and services to be built around it. There is also potential to federate with other CKAN nodes (for example data.govt.nz, or Landcare Research). It holds potential for distributed administration of harvesting and publishing.

NZ Research/ Digital NZ

A platform and service developed and hosted by Digital NZ, a part of the National Library of New Zealand. Supplejack is Digital NZ's open source tool for aggregating, searching and sharing metadata records, supports HTML, RSS, XML, OAI-PMH and RDF/XML. NZ Research also manages metadata transformation into a unified search index and provides an open API data service.

The NZ Research service currently harvests research publications metadata from university and polytechnic institutional repositories (IRs).

Platform Solution

APPLICATION	TECHNOLOGY	LICENCE	ADVANTAGES	COMMENTS
CKAN http://ckan.org/	Python, Javascript PostgreSQL, Solr	AGPL	Customisable UI, version control, role based permissions, harvester tool. Dataset relationships. Persistent URIs. Extensible with rich API. Geospatial features	"CKAN powers more than 40 data hubs around the globe, including government data catalogues for UK's data.gov.uk, USA's catalog.data.gov, the European Union's publicdata.eu. Also used by Landcare NZ https://datastore.landcareresearch.co.nz/ and being considered as platform for data.govt.nz
Digital NZ http://www.digitalnz.org/	Ruby on Rails, MongoDB, Solr, Tomcat	GPL	Supplejack open source harvester, customisable UI, scalable architecture.	Already used for http://nzresearch.org.nz/ Development and support of hosted service may require funding. A hosted solution which would require costing development & support by Digital NZ

[10] http://ckan.org/features/

APPLICATION	TECHNOLOGY	LICENCE	ADVANTAGES	COMMENTS
RDA (ANDS) https://github.com/au-research	MySQL, Apache, Linux, Solr, Tomcat	ASL 2.0	Core ANDS codebase includes a metadata registry, front-end portal and access management system. Collections Registry, Harvester, XML Crosswalks, PIDs service.	While the Research Data Australia software is available under an open-source licence* it has been developed by ANDS for their own requirements (e.g. RIF-CS Schema). JISC trialled then rejected in favour of CKAN because of "considerable development effort" required. http://www.dcc.ac.uk/sites/default/files/documents/registry/UKResearchDataRegistryPilot_reportWP2_v04.pdf
DSpace http://www.dspace.org/	Java, Apache, PostgreSQL, Tomcat	BSD licence	Customisable UI, version control, role based permissions, harvester tool plugins available	Local expertise available. Could use the Skylight UI to improve the interface. Example of DSpace as a metadata catalogue at http://www.hauhake.auckland.ac.nz/
VIVO http://vivoweb.org/	Java, Apache, MySQL, Tomcat, Solr	BSD licence	Produces Linked Open Data available via SPARQL queries. Provides network analysis and visualization tools	Limitations as a metadata manager, and issue of complexity of transforming institutional data into RDF. Works best as a researcher profile and collaboration service
Islandora http://islandora.ca/	Drupal, Fedora, Solr.	GPL	Customisable UI, version control, role based permissions, harvester tool.	Available as a hosted solution or local install, geolocation tools available as add-on service. Already in use by the University of Otago as a digital repository http://marsdenarchive.otago.ac.nz/

Proof of Concept

Solution	Installation	Harvesting	Metadata	Search	Implementation Costs
CKAN	The CKAN application was installed by the University of Auckland Centre for eResearch as part of its data repository investigation, with minimal configuration and a small number of metadata records ingested via Figshare API. A full pilot would require a reinstall on a dedicated VM and further configuration of the harvesting tool. Landcare Research Datastore is an example of an effective local installation.	Customisable harvesting tool collects from a range of sources (Untested, however JISC CKAN project list a range of harvested sources including 12 OAI-PMH targets[11]	A minimal default set of metadata was created in the test instance. CKAN allows for additional metadata (beyond the default) for a dataset by storing arbitrary key/value pairs against a dataset when creating or updating the dataset. CKAN also includes tools to import geo-coded metadata in a number of formats and make it queriable ('discoverable') according to the INSPIRE standard. It can import major metadata schemas such as ISO19139, GEMINI 2.1[12].	By default, the CKAN test instance provided post-search filters for Organisation, Groups, Tags, Formats and Licences. Landcare Research have implemented a map widget based on MapQuest tiles and Mapbox.[13]	Staff: 1 FTE 6 months (initial)[14](not including local resource allocation any participating members, see work package 1) Hardware: 2 x Servers with 8GB of RAM (One for Web and one for the Database/solr) 160GB hard drive on both. Quad core processors.

[11] http://ckan.data.alpha.jisc.ac.uk/hr/harvest
[12] http://ckan.org/features-1/geospatial/
[13] https://github.com/ckan/ckanext-spatial/blob/master/doc/map-widgets.rst
[14]Initial resource required for installation and configuration; interface customisation; harvester configuration; initial harvest; crosswalks; API testing

Solution	Installation	Harvesting	Metadata	Search	Implementation Costs
DigitalNZ	Hosted by Digital NZ.	Digital NZ[15] currently harvests from: • Figshare – 1840 records • data.govt – 4,338 records • Landcare (CKAN)– 184 records with minimal metadata: • NZ Research[16] currently harvests from • 12 Institutional DSpace instances (OAI-PMH)	Metadata captured from **data.govt** into Digital NZ is a subset of available content - five fields only: • *By*, *Date, Description, Usage, & Category* (e.g. Dataset). Other metadata fields not captured: • *Reuse rights, Contact, Email, Last updated*, and *Keywords* not captured. Nor *Coverage* (coordinates), *Format* (raster) or *Source* available from **Landcare**. However, NZ Research was able to accommodate the metadata schema defined by contributors, including 7 mandatory and 4 optional (rights, coverage, relation & source) fields.[17]	Both NZ Research and Digital NZ offers a range of pre and post-search filters including: • *format, usage, content provider* and *date*.	Redevelopment of NZ Research interface & ongoing support: By negotiation

[15] http://www.digitalnz.org/
[16] http://nzresearch.org.nz/
[17] http://nzresearch.org.nz/system/resources/BAhbBlsHOgZmSSJAMjAxMi8xMC8zMS8wOV80M180OV8xMTJfQ29udHJpYnV0b3JfTWV0YWRhdGFfR3VpZGVsaW5lcy5wZGYGOgZFVA/Contributor_Metadata_Guidelines.pdf

Summary

There is a lack of infrastructure and services at the tertiary institutional level at this point to support metadata creation and management, making it difficult to create a national aggregation. Research Data Management services are beginning to develop, but institutional metadata stores (whether data repositories or data registries) are yet to appear.

However, being late to the game means that we are able to learn from others. One question the Working Group has raised is the rationale for creating separate metadata silos for data and publications. There is an opportunity to develop a single national research (outputs) registry which would include publications, lab notebooks, workflows, methodologies, correspondence, grant applications, simulations etc.

For these reasons the working group proposes a phased approach where a short term (2-3 years) solution could be established quickly with low risk (as infrastructure already exists) followed by medium term solution which would permit participating members to prepare for local data and metadata management and federation technologies.

1. **NZ RESEARCH (1-3 years).** A lightweight solution for an interim national *research* registry / discovery service for tertiary institutions.
- Digital NZ already contains metadata for 1840 datasets for New Zealand researchers. These could be reharvested into NZ Research to form the basis of a metadata set for research data.
- The existing schema would be extended to include data elements.
- The web interface and platform would need to be upgraded. (NB: Digital NZ is currently being upgraded. There is potential for this work to be leveraged for NZ Research).
- Supplejack would be utilised to harvest from developing institutional metadata stores.
- An API is available for interrogating and downloading metadata

2. **CKAN (Year 3 +).** A national research registry that incorporates metadata from CRIs and other research institutions.
- A pilot could be undertaken in Year 2 while NZ Research was still in production, with a test instance and harvested metadata.
- data.govt.nz has launched a beta CKAN site with the intention of porting production to the new platform. http://beta.data.govt.nz/ . CKAN instances can be federated i.e. a tertiary sector CKAN could potentially pull in CRI metadata from data.govt CKAN
- The metadata schema could be extended to include other research outputs.
- Geospatial search would be tested and implemented
- An API would be available for extracting metadata for assessment purposes.

WP3 Metadata

Metadata for a Proposed National Research Data Registry

The Framework report noted that any national aggregation of metadata must be underpinned by an agreed and appropriate metadata standard. The creation of a New Zealand standard requires evidence-backed guidance for the description of research data together with evidence on which metadata elements make the greatest contribution to discovery and reuse. The selection of metadata standards is crucial to the success of any repository so that metadata from several sources (in this case New Zealand universities) can be combined and recalled. There is no widely used standard for how research data should be catalogued in a non-disciplinary context, but there are existing standards and tools developed with research data in mind. The Digital Curation Centre provides a useful list of cross-disciplinary metadata standards.[18]

To enable a feasibility study for a New Zealand Data Repository it was decided that identifying and agreeing on mandatory minimum core metadata elements would be a useful way to start a discussion on metadata. The aim is not necessarily to create a new schema but to make sure that any future or existing schemas used include these mandatory fields.

Supplemental information including lite review of other implementations of repository metadata is provided in Appendix 3.

Key principles for metadata

In considering the minimum metadata elements for the feasibility study the following principles have been kept in mind:

1. Metadata must provide sufficient information for discovery and reuse as well as data citation
2. Metadata standards should be discipline agnostic
3. Focus on the minimum. “Less is more” - simplicity is important for a small scale feasibility study and a small set of core metadata elements will also lessen any additional burden on researchers
4. Metadata should be machine readable where possible with a minimum of free text
5. Metadata should be appropriate for use in a New Zealand context
6. Metadata standards should be based on best international practice

[18] Digital Curation Centre (2016). General research data. http://www.dcc.ac.uk/resources/subject-areas/general-research-data

Proposed metadata fields

Based on the key principles and the findings of the JISC UK Discovery Service Project[19,20], the Working Group identified the following minimum mandatory metadata fields for a New Zealand National Research Data Registry.

Dublin Core (DC) has been used as a basis, as cross-walking most library-centric systems will be trivial, and offering harvesting from the registry will be straightforward. Dublin Core is not perfect, but it is better to use an established and understood metadata schema than to invent a new one (note that the UK experience indicated that common metadata elements bore a close resemblance to Dublin Core and Datacite). Dublin Core is also used by existing New Zealand university repositories. Reusing metadata models from either Datacite or ORCiD is easily achieved as these were also based on Dublin Core.

Proposed Scheme

Field	Expected	Note	Dublin Core	ORCiD
Creator	Free Text (UTF 8)	Author's name and/or ORCiD ID	dc: reator	orcid: work-contributors
Title	Free Text (UTF 8)	Title	dc: title	orcid: work- title
Date	W3CDTF profile of ISO 8601	Machine readable date	dc: date	orcid: publication-date
Subject	Controlled vocabulary	Could be discipline specific, (MESH) rather than geographical or possibly ANZSRC	dc: subject	-
Description	Free Text (UTF 8)	Should include appropriate keywords in the abstract	dc: description	possibly orcid: keywords
Identifier	URI (for example) DOI: ISBN: ORCID: HANDLE: RINGGOLD:	DC is fuzzy about this. Accept machine readable URIs prefixed with the scheme.	dc: identifier	orcid:work-external-identifier
Format	MIME		dc: format	orcid:media-type
Rights	URI	URL of rights statement	dc: rights	
Institution			dc: publisher	orcid: organisation
Type	DCMITYPE	Mostly 'dataset', but could be 'software' or 'text'	dc:type	orcid:work-type
isReferencedBy	URI	Pointers towards already published outputs describing/analysing the data.	dc: IsReferencedBy	

[19] Dom Fripp. "Developing a Core Metadata Profile for the UK Research Discovery Service". March 11, 2016. https://rdds.jiscinvolve.org/wp/2016/03/11/core_metadata_profile/

[20] Dom Fripp. "Research Data Discovery: How much Metadata is enough?" March 18, 2016. https://rdds.jiscinvolve.org/wp/2016/03/18/how-much-metadata-is-enough/

Possible Ambiguities

Creating ambiguity in metadata schemes is not good practice. For example, 'dc:Identifier' can refer to an author identifier like ORCiD, an institution in RINGGOLD, or another URI. This is understood to not be optimal, but has a useful, and understood work-around. Controlled vocabularies are recommended for property values, but the choice of those thesauri are difficult in a multi-disciplinary environment - ANZSRC, MeSH, for subjects are all valid. MIME types and Rights indications should be machine readable, to encourage reuse and harvesting.

Additional elements

There are, of course, additional elements that could be included in a New Zealand national data registry to further enhance the discovery and reuse of research data. These could include Language and Geospatial data. The usefulness of a contact statement and version control was also discussed but in the interests of simplicity they were excluded for now. Funder was also excluded since the proof of concept was not intended as a reporting tool; it was envisaged that Elements would fulfil that purpose for now. The next steps for this work could be to formulate a formal specification and construct a data model that each participating institution could code against to harvest, submit to a central authority or to make available for a distributed search tool.

WP4 Use Cases

The working group engaged a variety of possible end users to understand the direct benefits and outcomes they might expect from an NRDR. This understanding was used to:

- Validate the findings of each work package
- Enable continue benefit monitoring throughout the study
- Provide a baseline of expectations to refer back to should aspects of the study go off track and interventions become necessary.

Core findings are provided here. Supplementary information including the methods is provided in Appendix 3.

Interview Approach

Each CONZUL member was asked to select and interview three stakeholders including:

a. A mature researcher
b. An emerging researcher/PhD student
c. A high-level administrator (i.e. someone with external reporting responsibilities).

CONZUL members were provided with an introduction email template with information on the study for participants.

At the start of the interview, participants were shown an example of a data registry http://ckan.data.alpha.jisc.ac.uk/dataset . They were then asked a series of four questions:

1. Do you think an NDR would be useful in NZ? Can you say why/why not?
2. Would you use an NDR?
3. If yes, what you use it for?
4. Would you like to make other comments about an NDR?

Most interviews were recorded and transcribed, or answers provided to the CONZUL members in written form.

Responses received

Interview responses were received from Canterbury University, University of Auckland, Massey University, University of Otago and University of Waikato. The breakdown per stakeholder group is:

- 10 established/mature researchers;
- 5 early career researchers and;
- 5 research administrators with external reporting obligations.

***Note that one interviewee held dual roles as a mature researcher and an administrator. They have been treated as separate roles in the analysis.*

***The larger number of responses received from the established/mature researcher group made it possible to carry out a content analysis of the information shared.*

Results

Questions	Early career researchers	Research Administrators
1) Do you think an NDR would be useful in NZ? Can you say why/why not?	Yes, useful Visibility of data increased (published and other data that has not been published/gone anywhere)	All agreed it would be useful. Reasons include: Increased value from publically-funded data; transparency and accountability To solve problem of identifying researchers for collaborative efforts (collaborative partners).
2) Would you use an NDR?	Yes, would use	All said yes. It would be something to advise researchers about. To get an overview of research being done.
3) If yes, what would you use it for?	Finding out what is out there/what's been done, especially cross-disciplinary. Searching as an information source and a way to find people outside their existing networks who are doing similar work.	University of Otago respondent said they would use it to pull together research teams at the start of the research process Visibility of data Overview of data Discovery of data.

4) Would you like to make any other comments about an NDR?	Positive about this project.	Good idea Generally supportive.

Content analysis of established researcher responses

The following shows the overarching themes that emerged from the content analysis.

Q - Would you use a National Data Registry?	
Yes, depends (70%)	***Yes, definitely (30%)***
"Depends on what project I am working on…if we were to move into a new area it might be a good way for us to identify existing data in that area, making sure we didn't replicate…existing data". **-Research Officer, Psychology**	*"I'd definitely use it, but I'd also be pointing my research students to use it"* **Mid-career researcher, Political Science**

Q - Why would a data registry be useful?	**Q - What would they use it for?**
Single point of access *"The idea of a single place to provide verification and assessment of NZ research would be attractive to funders (MBIE etc.).* – Academic, Geography	Single point of access • To data and metadata • To national and international research *"...because of what I'm working on now...there may be other things, such as current government material that may not be easily accessible, that could get on that..."* - Researcher (Humanities)
Storage vs Registry (confusion between the two) *"We were…looking for somewhere to host that data publically so something like this would be ideal".* – Mature Researcher	Teaching *"If I have this data registry, this platform, these datasets, I would definitely use it for both research and teaching...for teaching it would be really good. I've been thinking about all the possible cases to talk about in the class".* - Researcher (Marketing)
Networks • Collaboration • Mechanism for making contacts	Collaboration • Multiple cohort studies • Make contacts *"I'd use it to connect with other researchers working in my field. My field is so small...I've got no idea if there are any others".* - Post-Doctoral Research Fellow (Fundamental Sciences)
Discovery and Access • Published & unpublished data *"The great virtue of this…is that it's a place to put stuff that might later be published, or that isn't really publishable, but is nonetheless useful".* – Philosophy Academic	Search and discovery • Published and unpublished data • Efficient, effective search *"One thing that you might do, is look for data when you can't get funding, to do something that's very large that you'd like to know about...And they might have done it, but not managed to get it published, or for whatever reason it might not otherwise be in the public domain, so I imagine this being useful".* – Philosophy Academic

Support research • Meta-analysis • Analyse & re-analyse data *"You would have the potential for meta-analysis of different sets of data which would otherwise be a bit harder to access possibly"* – Mature Researcher	Support research • Meta-analysis • Secondary analysis *"It's the kind of meta-analysis stuff...and being able to access research findings without having to do the research. Research is expensive, we don't get a lot of time to do it, if other researchers are open to people using their data in slightly different ways...that is a potentially good use of the resources that have been put into research...we don't...have to continually reinvent the wheel".* - Mature Researcher

Q – Any other comments about a data registry or this [feasibility] project	
Access and use statements (ethics, protocols, caveats)	"I do wonder about the ethical implications for participants, you know I mean if there's material there it can be used in different ways to what they were led to believe, so it might mean that there have to be some caveats" - Mature researcher
Governance	"It gets a bit tricky once it's sort of out in the public domain, you know, who polices that compliance, but I guess there's ways of doing it". - Mature researcher
Administration	"One thing a National Data Registry will need, if done correctly, is a good and dedicated team of curators. The team will need to maintain the different datasets, ensure they are consistently formatted and discoverable". - Established researcher (Biological Sciences)
Funding/costs	"...Need to be assured of the longevity of this. As a researcher, not interested in committing to something that has only short term funding, isn't going to be maintained..." - Research Officer, (Psychology).
Standards (i.e. metadata)	"It is difficult to incentivise researchers to publish and create metadata. So data collection and description would have to have a value-case for the researcher inside the University, and a national discovery service would be an added, no-pain extra". -Established researcher
Single point of access	"Why should data be separate from publications for a discovery service - why don't we upgrade nzresearch.org.nz to expand to data and later to other types. To become a research discovery service, not just a data one. If you just want data, then limit the search." - Established researcher

Discussion

These results are indicative only and a wider study might be of value, prior to embarking on a national data registry project. Studies exist overseas, however we need an understanding of the New Zealand context.

There was general confusion across all groups about what a registry is, with a number of participants confusing it with a data storage facility. Although each respondent was shown a data registry, they may have assumed they could access the data being described. The comments across all groups indicate that people

want to know data exists; however, they also want to know where to find it, and be able to access the data itself from a single point of access.

Everyone liked the idea of a registry. They identified benefits to efficiency in carrying out research, collaboration and ease of search across published and unpublished data.

The **early career researchers group** agreed a data registry would be useful, and they would use it. They would use a registry as a literature search tool, and as a way to find other people working in their discipline and to build contacts. It would be useful to them if both published and un-published datasets were registered.

Research administrators all said a registry would be useful. They would use it to get an overview of the data being produced, and would use it as another tool they could refer researchers to. None of interviewees said they would use this tool for reporting. One administrator raised a concern about taxpayer investment in the collection of data that is not realised when information about that data (and data itself) is not captured.

Established researchers liked the idea of a NDR. They would use it for research and teaching purposes. 'It depends' was a common response to the question on whether or not they would use it. This hesitation is likely because they have built up their networks of other researchers working in the same field, so would not need to use a registry for connections in the same way the early career researcher group would.

This group expressed the strongest concern about the sustainability of a data registry. A number of respondents commented that any such registry be well-supported by a management framework that would include access and use statements, assurance of on-going funding, data management standards and people. They do not want to contribute to anything that does not have a long-term future.

Conclusion

Stakeholders want the following from a national data registry:

- A register that will not only identify that data exists, but shows where the data is located and, if possible, link to the data itself. They want a registry that connects national and international research, from a single point of access.
- Registration of published and unpublished data: acknowledging that a lot of unpublished work is still useful.
- Assurance that any registry has long-term viability before they contribute to it, and they want it to have clear governance, be well-supported by administrators and have policies and procedures for access and use.

- A tool that can be used to provide an overview of research undertaken; and something that can be used to enrich teaching as well as research.

Stakeholders expect the following direct benefits and outcomes:

- A service that will strengthen and streamline the research process. They want something that can be used to carry out meta-analyses of existing studies, analyses and re-analyses data.
- More efficient and effective search and discovery than presently available, because sources must be identified and checked individually with no way of verifying how comprehensive the search has been.
- Early career researchers in particular will have another avenue to identify others working in the same field, making it easier to form relationships,
- Identifying researchers to form collaborations with will become easier.

WP5 Governance

Stable and agreed governance is required to ensure the sustainability of a national research data registry involving multiple stakeholders. Any model may be further complicated if the intention of such a registry is to include research institutions that are themselves governed by separate structures, e.g. New Zealand Universities, Crown Research Institutes, Wānanga, Polytechnics or private research institutions. While it is beyond the scope of this Working Group to include institutions other than universities in any proposed solution, the governance model options covered a range of possibilities from a pan-university model to a national institute model through to an international community model.

Without further information from the relevant organisations themselves on the success or otherwise of their governance model, it is difficult for the Working Group to provide much critical analysis of the respective models. Analysis of the models has therefore been based solely on an assessment of theoretical governance and public information.

Existing model analysis

National Library Model. NZ Research/DigitalNZ

The National Library of New Zealand ran the nzresearch.org.nz website from 2011. This service provided access to research papers and theses produced in New Zealand institutions (eight universities, plus Unitec, CPIT, Whitireia and Open Polytech, (Archives NZ, and the Alexander Turnbull Library). The service harvested metadata from repositories around New Zealand.

This harvesting service was established in 2007 as the Kiwi Research Information Service (KRIS), the result of a collaborative project between the National Library and tertiary institutions. Its governance composed a group made up of representatives from the academic community and key government departments. CONZUL took the lead in sharing expertise on the development of institutional repositories, while the National Library was funded by the Tertiary Education Commission to develop the harvesting service[21]. The National Library was responsible for taking an active role in the governance group and leading the day-to-day management of the website, promoting KRIS and awareness of research discovery, and where appropriate taking a lead in developing those services[22].

In 2011 NZ Research was migrated to the DigitalNZ platform. DigitalNZ was established in 2008 as a Digital Content Strategy initiative, and now has nearly 200 partners led by the National Library of New Zealand. The DigitalNZ team are part of a larger DigitalNZ Group within the Department of Internal Affairs. The DigitalNZ Advisory Board provides advice and guidance on the work of DigitalNZ and comprises representatives from the National Library, the education and higher education sectors, DIA, the culture and heritage sector, and law[23].

Critical analysis of a national library model

As noted, the NZ Research infrastructure falls under the governance structure of DigitalNZ and through the National Library, ultimately the DIA. While it seems that the service was initially set up to provide access to higher education institutions'(HEIs) research output, a governance model which appears to include limited representation from HEIs may inhibit this goal. Although it may be useful to include representation from a wider cultural and heritage sector this may also lead to the service being (or remaining) siloed into a service run by the National Library without having broader applications for a national higher education community.

National Institute Model. Australian National Data Service (ANDS) /Research Data Australia

The Australian National Data Service[24] (ANDS) is a partnership led by Monash University, together with the Australian National University (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO)[25]. ANDS' task is to create an infrastructure which enables Australian researchers to easily publish, discover, access and re-use research data and accordingly build an Australian Research Data Commons[26]. Strategic directions, policy and milestones are set by the ANDS Steering Committee, and comprises

[21] NZVCC Electronic News Bulletin Vol. 7 No. 20 6 November 2007. Kiwi Research Information Service (KRIS). http://www.scoop.co.nz/stories/ED0711/S00039.htm

[22] KRIS and nzresearch.org.nz. Research and repositories in New Zealand. Matthew Oliver. National Library of New Zealand. CHRANZ Housing Research Workshop. 30 May 2008 (in Digital NZ Scoping Study p. 26

[23] http://www.digitalnz.org/

[24] http://www.ands.org.au/

[25] http://www.ands.org.au/about-us/governance

[26] http://www.ands.org.au/guides/discovery-ardc

representatives from universities, research institutes, the research council, government departments and ANDS[27].

Research Data Australia (RDA) is the data discovery service of ANDS Collections Registry (a free service available to all Australian research institutions and universities, and government agencies)[28]. RDA is a set of web pages describing data collections produced by or relevant to Australian researchers, and now provides access to research from over one hundred Australian research organisations, government agencies, and cultural institutions. Metadata and links to the data from data publishing partners or contributors are provided within RDA, while the actual data is stored within local institutional repositories etc.[29].

Critical analysis of national institute model

It could be expected that having representation from universities, research institutes, the research council, government departments and ANDS on the ANDS Steering Committee would be beneficial for the governance of ANDS/RDA. Advantages would include buy-in from all major stakeholders across Australia, the ability to consider all stakeholders' viewpoints before decision-making, and the ability for the steering committee to make recommendations and lobby for funding to the government in the knowledge that they represented a cross-section of relevant Australian institutions, i.e. strength in numbers.

Inclusion of all parties may however lead to an inability to act quickly on new developments or directions and a tendency to get bogged down in the details of how decisions might affect one party over another. Some stakeholders may still feel that their issues are not adequately addressed as they may be subsumed into those that affect larger or more influential institutions.

Representation from those sectors with multiple institutions i.e. universities and research institutes would need to be fairly shared over time so that smaller institutions received adequate representation too. Clear feedback channels and reporting structures would need to be in place to ensure that issues were moved along and decisions were made in a timely manner.

International Community Model. CKAN (Comprehensive Knowledge Archive Network)

CKAN (Comprehensive Knowledge Archive Network) is a data portal platform, i.e. software for building a catalogue and repository for datasets[30]. The system can store datasets, or hold metadata for datasets hosted externally. The CKAN website lists over 70 instances in use across the world, including national data portals for Australia, Austria, Canada, Italy, Norway, the Netherlands, Romania, Slovakia, the UK, and the USA.

---

[27] http://www.ands.org.au/about-us/governance
[28] http://www.ands.org.au/online-services/research-data-australia/collections-registry
[29] http://www.ands.org.au/online-services/research-data-australia
[30] http://www.dcc.ac.uk/resources/external/ckan

CKAN is provided by Open Knowledge Foundation International (previously known as Open Knowledge Foundation- pre-May 2014), a worldwide independent non-profit organisation[31]. It is incorporated in the UK and operates globally. The CEO manages Open Knowledge International in terms of general management and strategic direction, and reports to the Board of Directors, who are responsible for the financial and legal probity of Open Knowledge International.

Local groups, working groups and other activities such as Open Knowledge Labs are community-run, and supported by the Open Knowledge International team. There is also a distinguished Advisory Council who provide specialist expertise and guidance[32].

Critical analysis of international community model

As a data portal platform currently with over 70 instances internationally, it would be expected that CKAN would benefit from the governance structure of a worldwide independent non-profit organisation. The structure appears to incorporate international and local expertise, in the form of local groups, a Board of Directors, and an Advisory Council. It could be expected that such a cross-section of communities would engender a certain amount of buy-in from major stakeholders, with the related ability to include a variety of viewpoints before making any decisions.

As with the ANDS governance model, inclusion of all parties may however lead to an inability to act quickly on new developments or directions and a tendency to get bogged down in the details of how decisions might affect one party vs another. Some stakeholders may still feel that their issues are not adequately addressed as they may be subsumed into those that affect larger or more influential institutions.

Again, representation from those sectors with multiple institutions such as universities, research institutes and government departments would need to be fairly shared over time so that smaller institutions received adequate representation too. Clear feedback channels and reporting structures would need to be in place to ensure that issues were moved along and decisions were made in a timely manner.

Governance Model recommendation

While the Framework envisages a truly national meta/data registry for New Zealand, it is largely restricted to offering suggestions for national universities as this is its stakeholder group. An international governance model is unrealistic for the purpose of this study or for consideration other than as a comprehensive community support network. Comprehensive national governance with representation across all research capable national institutions would be

[31] https://okfn.org/about/
[32] https://okfn.org/about/governance/

difficult to manage with as many competing interests would stifle decision making, a critical task at the early stages of a service.

This work package recommends, in the first instance a 'universities only' Governance Board that would comprise representatives from CONZUL and/or other university directorates, ensuring that a range of university size and geographical location is covered. The Governance Board would be responsible for leadership, strategic input, technical capability and sector consultation with clear guidelines around establishment, tenure, resignation and a proposed schedule of key deliverables.

Funding could come from memberships or agreed in-kind resource, e.g.CONZUL or Universities NZ or another pan-university structure. A technical/operations lead would be selected on behalf of the Board and would most likely be a member institution. If it was decided to take advantage of existing infrastructures like DigitalNZ then technical implementation could leveraged.

Each University would require a dedicated point of contact to manage the local metadata and present it to the registry for harvest or search according to defined harvest/search protocols. Further governance of local data management conditions would be provided within individual University structures.

Subsequent intent to expand to non-HEI could be considered as a Term of Reference of the Board as it sees fit.

Future options

If a pan-institutional national meta/data registry is created, a national Governance Board would require further representation from Polytechnics, Wānanga, CRIs, government departments, local council bodies, Research Institutes, NeSI, REANNZ, NZGL, MBIE. As suggested in the eResearch2020 report, cross-sector research programmes such as the National Science Challenges could also be incorporated[33] (2016, p. 42). Such a governance model could be along the lines of what exists for ANDS/RDA.

The governance board would set up working groups as needed to be tasked with specific actions i.e. metadata schema management, as per the eResearch2020 report (2016, p 43). Funding would need to reflect intellectual investment and could be local resource allocation or at a national level i.e. MBIE or the DIA (DigitalNZ/ Research NZ). There should also be an associated network of "data management practitioners" from each institution, to offer feedback and advice to working groups or the governance board.

The expanded comprehensive governance model can be regarded as an extension of the 'universities only' model. A NRDR can be piloted across universities more easily than attempting a truly national registry with over-powering

33 http://www.eresearch2020.org.nz/eresearch2020_nationalresearchdataprograme_f_single-2/

representation and stifling cross-governance negotiation. Once the university registry matures and the technology stabilises, expanding the governance to include other university-related or non-university institutions will be less effort and have a greater likelihood of success.

Project Findings

Institutional effort in collecting research metadata

New Zealand universities use a variety of reporting systems which house metadata that could inform a NRDR. Extracting metadata from HR, grant reporting, financial reporting or publication systems would be challenging. Many internal systems are not well integrated, if at all, and are controlled by different departments within institutions making it difficult and time consuming to obtain useful metadata. In addition, some of the systems have security concerns that preclude external access, or even internal access from outside controlling directorates. Once negotiated and collected the metadata would need to be cross walked into an agreed metadata schema. Neither of these challenges is trivial and particular challenges change between universities.

In short, significant effort would be required to make the metadata about the research data available and (where appropriate) provide services to ensure the researchers are able to make this information accessible for sharing and reuse.

Technology Solutions

Deciding a technology solution is often focused on a clear set of requirements and features at the expense of pragmatic opportunity. The technology space for distributed repositories and registries is changing rapidly and provides many solutions, some already established in the NZ University and Library space. Taken together such a situation will risk adding burden of new technology to fulfill all requirements rather than fulfilling most requirements with established technology.

Given the findings of work package 1, that most institutions need to invest in local management of metadata and underlying data, it would seem pragmatic to reuse existing and implemented infrastructures rather than implement more. However, it was also determined that implementing a universities-only federated registry is a low overhead and quite achievable provided local metadata is harvestable.

This work package considered two technology solutions, an established service run but he National Library, DigitalNZ, and a university-only service that would need implementing, configuring and managing. There were benefits to either solution but no real clear preference. The result is a recommendation of a phased approach that made use of existing services of DigitalNZ which gave universities to focus on local metadata management and registries. Once local

registries are sufficiently stable, a decision can be made to move the federation back to university as either a dedicated registry with a CKAN service, or newer technology like distributed search or federation on the fly.

Managing metadata

Symplectic Elements is used in most member institutions and initially reusing the metadata model associated with this product seemed a logical proposition. However, it was determined that even reaching agreement on a minimum set of metadata would still require significant cross-walking between different versions or implementations of Elements. While this effort was not considered overly burdensome, it was a newly defined effort and because control of institutional implementations was not generally in the same organisational unit across all institutions, there was a significant risk of error or lengthy negotiation.

The alternative solution was to adopt various metadata elements from existing schemas and to normalise these to Dublin Core. This would be straightforward and require modest effort as most of the necessary elements are present in DataCite and ORCiD data models, which are based on Dublin Core. However, in doing so the effort of managing the metadata schema would shift from the institution, to the governance of the NRDR.

User requirements and expectations

Work package 4 was included in this study to ensure that the aims of the feasibility study reflected the user benefits as defined in the Framework Report. From recorded semi-structured interviews, the conclusion that the benefit to many different users of an NRDR remains positive. Semi-structured interviews were run with a number of key stakeholders and the findings from these interviews suggest that a NRDR would be beneficial to a range of different users. Early career researchers considered an NRDR an extra avenue for search and discovery in addition to traditional information tools. Mid to late career researchers often added an 'it depends' comment that queried the sustainability of such a resource; this appeared to be key to getting further buy-in from them.

Respondents were also confused as to the content and purpose of a NRDR. Many researchers believed that the presence of a metadata entry indicated the availability of the underlying research data, while for administrators (who were interested in meta-analysis) the presence of the underlying data was not as relevant. Evidently, there needs to be a clear message about any NRDR to accurately communicate context and also to manage expectation.

Governance and sustainability

Several different governance models were considered in work package 5 including National Library, comprehensive stakeholder, university-only and international community. Models based on comprehensive representation and international communities were discounted as either being un-manageable in the first instance, or unrealistic for the purposes of the study. The remaining models namely, National Library and university-only were considered more appropriate.

With a university-only model of governance the core concerns of a NRDR could be controlled to greater effect with all New Zealand university parties involved in much the same type of academic activity and with many of the same constraints and/or concerns. In this sense the most useful governance could be built from an existing CONZUL-type committee supplemented with technical leads and cultural/training programmes tailored by individual institutions.

If there was interest in utilizing the existing DigitalNZ infrastructures, then it would be important to seek greater influence on the governance of this service for CONZUL benefit. However, the result may be less control over collection, presentation and other operational decisions.

Funding any registry service would be complex. For a university only governance model the support could exist as 'in-kind' resource contribution from participating members, although this may result in sustainability issues as member priorities change or personnel leave. In contrast established services like DigitalNZ are exposed to politically derived risks, particularly if governments change or financial efficiencies are imposed on departments. A subscription model to members would be premature at this stage. The working group was unable to identify a preferred funding solution but agreed that either of the preferred governance models would be sufficient for any pilot activity without significant extra financial support.

Options:

Option 1. Do nothing

- Business as usual
- No benefit to discovery, reuse or meta analyses other than individual bespoke approaches currently used
- No additional effort or resource required

Anticipated outcome

The most significant outcome is the lost opportunity for national co-ordination around managing and sharing research data; individual approaches may continue but without co-ordination this may lower any national benefit to all as a consequence. Doing nothing means there is no basis to extend to other non-university research bodies.

Option 2. Establish a universities-managed metadata only registry

- Federated harvest or distributed search is owned and managed by lead member
- Burden will be agreeing and managing crosswalks and standardised national schema
- Each institution would locate, collect and maintain a registry of agreed standard metadata

- Main benefit to administrators and those interested in meta analyses
- Little benefit to researchers who expect data access implied by a metadata register
- Can be extended to provide data access as institutional capability matures

Anticipated outcome

The primary benefit lies with those stakeholders interested in meta-analyses. While individual members may still need to locate and collect metadata but without any benefit to researchers (either perceived of real), the metadata is more likely to come from existing administrative systems; researchers are unlikely to provide additional metadata without an incentive to do so. Federation via a harvest protocol or distributed search protocol is a low overhead provided agreements on metadata schema and crosswalks are established. Sustainability depends on the administrative need to maintain a registry as it offers little extra to traditional academic discovery pathways. This option could be undertaken rapidly and could form the basis of the development of a richer NRDR which focuses on researcher benefit where underlying research data are also accessible.

Option 3. Use existing service with metadata from local repositories

- Request DigitalNZ to maintain a service on behalf of all NZ research without the need to agree to metadata data standards
- Individual institutions would contribute only those metadata records where the underlying data could be accessed as DigitalNZ concern content rather than simply metadata.
- This would enable a slow but stepwise approach to national metadata surfacing and encourage good practice for open data (licencing or not)
- Will still require institutions to manage metadata collection and data access

Anticipated outcome

The reduced effort required to agreeing and crosswalk metadata schemas would enable a faster federation of national research data records but individual institutions would still be required to manage their own metadata. Content could be limited to metadata which are linked to the research data (i.e. the data must also be available), however this condition may result in extremely low representation of research data until universities are able to manage the underlying research data. The risk of expecting data availability may be offset by a visibility issue within Digital NZ, where metadata associated with research data are co-located with traditional published articles. This is an extremely attractive outcome and could be exploited by targeting those data that are deposited in community repositories rather than waiting for institutional solutions for data storage and access.

Option 4. Support individual approach to discovery and seek harmonization over time

- No federation service but support a community activity, e.g. existing repositories community, for harmonisation of registries, metadata standards and user interaction
- Will require ongoing and active participation/hosting of community events like conference attendance, workshops, symposia etc.
- Would favour a lower establishment and buy in threshold
- Slower time to benefit for all stakeholders

Anticipated outcome

A less proactive approach to a NRDR that would seek the most inclusive strategy of community participation in favour of direct action with a smaller set of stakeholders. The main risk would be the time required to benefit all stakeholders, as agreements are harder to reach in large communities of diverse discipline and governance. However, a more inclusive approach would be more effective once agreements were reached. This approach could ultimately prove costlier as it would require attendance at conferences, workshops and community symposia over many years to gather support and discuss benefit and solution.

Options Analysis

Option	Quality	Time	Risk	Cost	Comments
1	L	H	L	L	Current BaU not altered. Individual repositories evolve independently as time and budgets permit. Federation benefit is lost. No real risk except to lost opportunity and low cost as no significant new effort required.
2	L	M	H	L	Full control over infrastructure but even with comprehensive UNZ stakeholder involvement the quality remains low as access to data currently out of scope. Time to benefit tied to local collection of metadata. High risk due to unmet user expectation of data access but can be mitigated with communication and expectation management. All costs remain at individual institutions with one lead taking trivial cost in federation. Federation trivial and if fully implemented could provide greatest benefit.
3	H	M	L	M	Co-location of publication and data through same access portal increases quality significantly but relies on institutional access features or collection of external pointers. Main time constraint in institutional readiness. Lower risk but limited take up for foreseeable future. Primary burden remains with institution in providing content and metadata. Can be established rapidly but requires co-governance. Potential for strong benefit.
4	H	H	L	M	Community driven conventions are generally the most fit for purpose but reaching agreement can take time. Low risk as current behaviour not altered significantly but the continued support of community participation would require continued contributions from CONZUL members, whether in kind or through active participation in conferences and meetings

Recommendation

The Working Group believes the most successful strategy would be a phased approach to a federated NRDR that begins by leveraging existing NZ Research infrastructures to represent research data records alongside more traditional articles. This first stage would require a partnership with DigitalNZ rather than a service fully controlled by NZ Universities but this was a small governance concession to establishing a NRDR rapidly. In the longer term the Working Group believed a university-led service could become part of a wider network of metadata registries similar to data.govt.nz, NZ research, the CRIs and other research institutions. That said, universities could maintain their independence by managing metadata and access to the underlying research data for their own concerns while still being part of a wider landscape. The need for a comprehensive registry may not be a technically demanding or may be superseded by a 'distributed search' or 'federation on the fly' protocol rather than a dedicated registry. The Working group recognised the main effort lay in collecting metadata and accessing research data and this remains and institutional responsibility.

Either approach would be low to medium risk but by starting small and using the existing services of DigitalNZ to provide those metadata for which the underlying data are available through institutional repositories or external repositories would realise more benefit to a greater section of stakeholders than providing a metadata-only NRDR. In supporting an open access agenda for appropriate research data this approach also drives key research data licencing

issues and engages rights owners in addressing the current barriers to reuse of research data[34]. Should access to the data increase and the open data culture spreads (as anticipated) then the presence of stakeholder metadata in DigitalNZ that points to research data content will grow. In undertaking this approach the expectation of users will be constantly met, albeit with modest entries to begin. If the existing services fail or are discontinued then federating the existing individual institutions research data registries and repositories is a relatively small effort to NZ Universities stakeholders; initially for those stakeholders interested in meta-analyses but also researchers, as the access to underlying research data grows. The substantive efforts of local metadata registries and data repositories are critical. If established it will also be feasible to implement other technologies, including federated search. A stable and useful service would be attractive to other institutions, who would experience benefit earlier with less investment and facilitate a truly national research data registry more rapidly.

This Working Group recommends Option 2 but recognises that benefit would be realised earlier with Option 3, which could be undertaken as a phased approach to Option 2.

[34] see accompanying discussion paper "Ownership and Licensing of Research Data"

Appendices

Appendix 1: Risk and Issues

Risk ID	Date	Risk Name	Description	Owner	Risk Analysis	Mitigation
NRDR001	26th Feb 2016	Scope Creep: *Quality* *Cost* *Time*	Ill-defined scope for this study will risk a delay in project delivery and increase the likelihood of poor solutions. Both these risk increasing cost effective project management	Max WILKINSON (Programme Manager)	MED: Benefits are not immediately apparent to all stakeholders. There are different incentives or utility to each.	*Agree scope and work packages, then monitor frequently reporting any creep back to sponsor where appropriate.* *Ensure all working group members work to Project Scope. Ensure regular reporting to Sponsor and CONZUL* *>20160318. Feasibility agreed to progress as a throw-away product. To inform and quantify further development without committing to long term sustainability* *>20160420. Harvest API built at University of Otago and established as harvested endpoint at Auckland. Feasibility demonstrated. Documentation-lite but de-risked.* *>20160610. Scope of project considered of lower benefit than expected. Consistent feedback from user stories were negative about 'only metadata' option. Options used to mitigate possible scope creep.*
NRDR002	26th Feb 2016	Stakeholder Engagement *Quality*	Staff engagement at individual institutions is minimised	Max WILKINSON (Programme Manager)	LOW: Difficult to adjust to changing priorities,	*Maintain realistic timelines/commitments* *Conduct face-to-face meeting early in the project to maximise engagement*

Risk ID	Date	Risk Name	Description	Owner	Risk Analysis	Mitigation
					especially when staff line management changes or institutional strategies change.	*Ensure project remains visible to CONZUL members* *>20160318. CONZUL RDM framework mentioned numerous times in the eResearch 2020 NRDP case for support. May need to craft a consistent message about role and longevity to removed risk that CONZUL are committing to this long-term.* *>20160520. Membership of WG unstable due to changed priorities at some member institutions. Some work package activity is transferred Sponsor notified and risk accepted.* *>20160527. Reports of over allocation of resource at some member's institutes. Risk managed by refocusing effort to key tasks rather than full analyses* *>20160610 Risk remains*
NRDR003	26th Feb 2016	Project under resourced *Cost* *Time*	Project under resourced	Howard AMOS (Sponsor)	LOW: minimal effort required of working group members and considered resource contributions to CONZUL	*Maintain realistic timelines/commitments* *Ensure project remains visible to CONZUL members* *Report on un-recognised effort and resource needs.* *>20160520. Some members consider effort beyond resource allocation. No immediate risk to project deliverables* *>20160527. See above. Resource re-focused on key tasks rather than full analyses* *>20160610 Risk remains*

Risk ID	Date	Risk Name	Description	Owner	Risk Analysis	Mitigation
NRDR004	26th Feb 2016	Commitment to change *Quality*	CONZUL members lack commitment to adopt recommendations	Howard AMOS (Sponsor)	HIGH> Commitment to change is a local business decision that is influenced by budget and operational factors the working group have little influence over.	*Ensure project outcomes are well documented and visible to CONZUL members by reporting on progress from each meeting*
NRDR005	26th Feb 2016	Feasibility target group *Time* *Quality*	Most stakeholders have different capability in presenting harvest points to external parties, mainly because WG members do not have control over the necessary internal processes. Waiting for a full partnership risks significant and critical delay in feasibility study and poorer quality	Max WILKINSON (Programme Manager)	MED. Not all stakeholders are in control of their institutional metadata solutions and so cannot confirm external harvest.	*A smaller group of member institutions will be self-selected based on WG member's ability to provide technical control over the necessary processes and local infrastructures. Participation in G ensures rapid delivery of solution should this move to a wider pilot activity.* *>20160311: Pilot group identified as University of Otago, University of Auckland and Victoria University of Wellington.* *20160317: Victoria University of Wellington has removed itself from target group due to technical politico-technical barriers. May be resolvable with the University of Otago solution. Mike and Gillian to speak.* *>20160520. De risked. Harvest demonstrated between the University*

Risk ID	Date	Risk Name	Description	Owner	Risk Analysis	Mitigation
						of Otago and the University of Auckland. *>20160610. Re-risked. Consistent user feedback suggests metadata only efforts of much lower benefit*
NRDR006	4th March 2016	Ethical authority to collect and publish user survey *Quality* *Time*	As part of WP4 WG members will conduct semi-structured interviews with stakeholders. If we go on to publish this activity, then ethical review may be necessary.	Natalie DEWSON (Massey University)	MED. Ethical authority to conduct surveys can sometimes be difficult. As it turned out this was de-risked as ethical authority achieved without many complications	*Seek advice from local ethic board members or advisors and decide strategy to collect subject consent where necessary.* *>20160308. De-risked as Natalie and Gillian have established effective process to work with ethics committee at the University of Otago should it be necessary but advice from current committee is consideration is not yet required. May be in future should we wish to publish results.*
NRDR007	10th March 2016	Common term: *Quality* *Time*	Concept ambiguity: There are presently several different names used for referring to this feasibility study, e.g. metadata register, data catalogue, research data registry. Shared understanding required a common language and so lack of and agreed name risks misunderstanding.	Gillian ELLIOT (University of Otago)	LOW. Ongoing risk as a relatively recent activity. Emerging vocabularies are often complicated by semantic ambiguity of mis-use	*In consultation with the working group agree a common set of terms to use for this feasibility study.* *>201603018. De-risked, we are likely to call this feasibility goal a 'National Research Data Register'. Will confirm at RDMWG5 in April* *>20160527. On-going risk but not confined to this project. Accept and continue to minimise* *>20160610. No change*
NRDR008	10th March 2016	Configuration Management *Quality* *Time*	Document control can get confusing and risks quality and time in project delivery	Max WILKINSON (Programme Manager)	LOW. Difficulty in using shared documents is	*Work package owners will own work package documentation and all revisions will proceed through them. Contributors will submit directly to*

Risk ID	Date	Risk Name	Description	Owner	Risk Analysis	Mitigation
					the responsibility of the PM and their conventions which are communicated at each working group.	*work package owners and the owners will regularly re-publish version of the work package document to the shared drive. >20160313. Guidance on filename conventions and contributor process provided. >20160520. Google docs not working for everyone. Not a stable platform for distributed config management that this project seeks. >20160527. Ongoing risk but limited deliverables means documents managed effectively despite increased effort. Minimal risk. 20160610. Ongoing risk*
NRDR009	1st April 2016	Duplicated effort *Cost* *Time*	Wasted effort if service already exists or can be subcontracted with minimal effort, e.g. DigitalNZ	Max WILKINSON (Programme Manager)	LOW. Primary output from this working group is to investigate the most appropriate approach given the landscape.	*Need to establish if Digital NZ is able to undertake this activity as part of their remit rather than CONZUL. Risk may be multifaceted with sustainability and governance issues between university libraries and the national library. Discussion on 8th April will establish the mitigation. >20160523. Efforts made to communicate 'proof of concept' purpose of this project. Part of the outcome will be an estimate of effort required to establish and maintain any tangible outcome. >20160527. Will re-enforce feasibility of this project to any members seeking to speak publically about the project as part of the valuable dissemination activities.*

Risk ID	Date	Risk Name	Description	Owner	Risk Analysis	Mitigation
						>20160604. External services promoted as opportunity to exploit. Institutions focus on local management of research data while Digital NZ is utilised to present federated catalogue.

Appendix 2: Work Package Supplemental Reports

Work Package 2: Platform

HARVEST ENDPOINTS. Requirements:

1) Each institution will need to provide either (a) an OAI-PMH endpoint for harvesting metadata or (b) a (documented) API interface for extracting metadata

INSTITUTION	METADATA STORE	STATUS	OAI-PMH END-POINT	API
The University of Auckland	Figshare https://auckland.figshare.com/	Production	On the product roadmap in 2016	Yes https://docs.figshare.com/
Auckland University of Technology				
The University of Waikato				
Massey University				
Victoria University of Wellington				
University of Canterbury				
Lincoln University				
University of Otago*	DMP Metadata store	PoC	PoC	No

* See Work package 1 Collection

APPLICATION	OAI-PMH END-POINT	API	COMMENTS
Figshare	On the product roadmap in 2016	Yes https://docs.figshare.com/	Serves both as a repository metadata store
DSpace	Yes	Yes https://jspace.atlassian.net/wiki/display/DSPACEAPI/API+Documentation	Already used as Institution Repositories
Fedora/Islandora	Yes https://github.com/Islandora/islandora_oai	Yes https://github.com/Islandora/islandora/wiki/Working-With-Fedora-Objects-Programmatically-Via-Tuque#fedoraapia	Simon Fraser University have an Islandora repository in production. http://researchdata.sfu.ca/
Symplectic Elements	Not (currently)available	Not publically available, but Elements Reporting API could be used.	Advantages: (1) easy to provide a consistent metadata schema. (2) Would also reduce any government funders reporting requirement from a National Data Registry. Issues: (1) requires a metadata workflow from institutional Data Repositories. (2) University of Otago does not (yet) use. (3) Concern expressed about giving external access to the primary data via API. One potential solution would be for someone (The University of Auckland have the capability) to create an OAI endpoint from Symplectic Reporting database and make this code available to other sites.
CKAN	Yes https://github.com/kata-csc/ckanext-oaipmh	Yes http://docs.ckan.org/en/latest/api/	Most popular open source data portal. A number of examples available http://ckan.org/instances/# including Landcare https://datastore.landcareresearch.co.nz/ Review: http://www.dcc.ac.uk/resources/external/ckan
Data Management Plan Store. (Bespoke application)	Yes.	No	University of Otago tested a proof of concept DMP metadata mapping to minimum Dublin Core

			for OAI-PMH harvesting. This was harvested by The University of Auckland DSpace OAI harvester.

Some metadata store options (for institutions)

Work Package 3: Metadata

What do other repositories use?

As mentioned there are numerous potential metadata schemas available. The Australian National Data Service, for example, uses RIF-CS. This schema has many mandatory elements because it is providing a well-resourced and well developed national solution for data creation through to data preservation. The New Zealand NDR is envisaged as a discovery platform only and will be pointing to data in disciplinary repositories which may have more developed metadata schemas. At this stage RIF-CS is too complex for our needs.

In the UK, there have been two recent surveys examining metadata schemas from British institutional research data repositories in an attempt to discover common elements. Torsten Reimer from the Imperial College London compared the metadata schema in use at his institution with the schema used in the data catalogues at Cambridge University[35]. Similarly, as part of his work on the JISC Research Data Discovery Service, Dom Fripp[36] reviewed eight metadata schemas (*Datacite, EU Data Portal, ANDS, EUDAT, ETSIN, INSPIEm, ReCollect* and *DDI Lite*) to establish what metadata fields were most used. After analysis, Fripp found that the key fields a bore a not unexpected similarity to DataCite and Dublin Core. His research was then incorporated into a service profile for the planned UK Research Data Discovery Service.[37] The results of both investigations are recorded in Table 1 below and reveal "complete overlap".

Cambridge & Imperial Universities (Reimer)	UK Research Data Discovery Service (Fripp)
Title	Title
Author/Contributor	Creator
Author/Contributor ORCID id(s)	Creator Identifier
Abstract	Description
Keywords	Keywords
Licence	License
Identifier (DOI ideally)	Unique Resource ID
Publication Date	Date
Version	Relation type/related identifier*
Institutions (of authors)	Publisher/Creator Affiliation
Funders	Funder
Grant reference	Project Number

Table 1 – Metadata comparisons sourced from Fripp (2016) and Riemer (2016)

[35] Torsten Reimer. "Less is more? A metadata schema for discovery of research data". *Open Access and Digital Scholarship Blog*, February 19, 2016, http://wwwf.imperial.ac.uk/blog/openaccess/2016/02/19/less-is-more-metdata-schema-discovery-research-datary-of-research-data/

[36] Dom Fripp. "Developing a Core Metadata Profile for the UK Research Discovery Service". March 11, 2016. https://rdds.jiscinvolve.org/wp/2016/03/11/core_metadata_profile/

[37] Dom Fripp. "Research Data Discovery: How much Metadata is enough?" March 18, 2016. https://rdds.jiscinvolve.org/wp/2016/03/18/how-much-metadata-is-enough/

Work Package 4: Use Cases

Method

1. Firstly, the responses as a whole were examined to identify general themes. These initial observations were reported to the working party and are presented in tables in section 4 of this report.
2. There were enough responses received from established researchers to carry out a content analysis to identify concepts and themes.
3. The content analysis involved going through each questionnaire response one sentence at a time and assigning a short summary (1-3 words) to describe the meaning of the text (the code).
4. The codes were listed under each question, with groupings made for similar codes, and isolated (redundant) codes removed. Once codes had been identified, supporting quotes were selected.
5. The themes were discussed and conclusions drawn.

Results

This section identifies the main themes that emerged for each stakeholder group interviewed.

Early career researchers (5 Responses)

Representative comments:

> *"I wanted to see the performance of these two things, to compare…it would be useful for my study to comparison (sic) between these two banks, but unfortunately I didn't get this type of data. So that led me to change my study direction".* ***– PhD Candidate (Finance)***

> *"I think especially within New Zealand, at the local level there might be more collaborative opportunities in terms of sharing information, and it might be a way to access that instead of just relying on your own contacts you already know". –* ***PhD Candidate (Veterinary Science)***

> *"I would find it really useful to see what else is there or what other people have done before me…. using the knowledge that people got while building the methodologies and the data management side of things". –* ***PhD Candidate (Media and Communication)***

Research administrators (5 Responses)

Representative comments:

> *"If you can just make sure that my view is recorded about it needing to be truly national and not just something the Universities do by themselves, and address the issue of how we make it truly a national database".* ***– High-level administrator (Deputy Vice Chancellor Research – Director of Research & Enterprise Office).***

"Research Advisors and Enterprise Managers would find it especially useful when helping PIs and Directors for pulling together project teams e.g. a medical person working on an HRC application and might need some economics input. While they may know everyone within the subject area within Health they are floundering when it comes to identifying experts outside their research areas."

Operations Manager, (Research and Enterprise)

"I think Research and Innovation would generally – it would certainly be another arrow in the quiver, something that we would be able to advise researchers about…we would be able to encourage them to lodge their own research data and encourage them to use the registry when they are thinking about their own research". – ***PBRF Advisor, (Research and Innovation Office)***

Established researchers (10 responses)

Question 1: Do you think an NDR would be useful in NZ? Why? Or: why not?

Representative comments:

"The idea of a single place to provide verification and assessment of NZ research would be attractive to funders (MBIE etc.)" - Geography Academic.

"Would be useful having a one-stop-shop for understanding the breadth of data, would encourage collaboration". - Research officer, School of Psychology.

"We were…looking for somewhere to host that data publically so something like this would be ideal…" - Mature Researcher.

"The great virtue of this…is that it's a place to put stuff that might later be published, or that isn't really publishable, but is nonetheless useful". Mature Academic, Philosophy.

"You would have the potential for meta-analysis of different sets of data which would otherwise be a bit harder to access possibly". - Mature Researcher.

Question 2: Would you use a National Data Registry?

Representative comments:

"Depending on the project, yes. There aren't many of me in the country" - ***Post-Doctoral Research Fellow, Fundamental Sciences.***

*"Yes, on occasion, for discovery purposes"–**Established Researcher, Geography.***

*"…I would be interested in using that, provided that the, the National Data Registry is already…full of useful data"- **Established Researcher, History.***

*"…I think I would, yes. I'm not working on those things quite as relevant to New Zealand as much now, but certainly there are things I can think of, such as imports, and things like that? Ah, people would have studied the different patterns over the decades and that might come up…on such a register" – **Mature Researcher (Humanities with an interest in Pacifika).***

Question 3: If 'yes', what would you use it for?

Representative comments:

*"If I have this data registry, this platform, these datasets, I would definitely use it for both research and teaching…for teaching it would be really good. I've been thinking about all the possible cases to talk to about in the class". - **Established researcher (Marketing).***

*"I'd use it to connect with other researchers working in my field. My field is so small…I've got no idea if there are any others". – **Post Doctoral Research Fellow, (Fundamental Sciences).***

*"It's the kind of meta-analysis stuff probably and being able to access research findings without having to do the research. You know research is expensive, we don't get a lot of time to do it, if other researchers are open to people using their data in slightly different ways I think that is a potentially good use of the resources that have already been put into research, you know we don't have to continually reinvent the wheel and sometimes it's not entirely necessary to start from scratch". – **Mature Researcher***

*"I think this is open ended, I mean, one thing that you might do, is look for data when you can't get funding, to do something that's very large that you'd like to know about…And they might have done it, but not managed to get it published, or for whatever reason it might not otherwise be in the public domain, so I imagine this being useful". **– Mature Academic (Philosophy).***

Q4: Would you like to make any other comments about a National Data Registry or this project?

Representative comments:

"…Need to be assured of the longevity of this. As a researcher, not interested in committing to something that has only short term funding, isn't going to be maintained, needs to include statements on intellectual property & who should have access, methodology used".

Research Officer (Psychology).

*"One thing a National Data Registry will need, if done correctly, is a good and dedicated team of curators. The team will need to maintain the different datasets, ensure they are consistently formatted and discoverable". –**Established Researcher (Biological Sciences).***

*"I do wonder about the ethical implications for participants, you know I mean if there's material there it can be used in different ways to what they were led to believe, so it might mean that there have to be some caveats…It gets a bit tricky once it's…out in the public domain, you know who polices the compliance". - **Mature Researcher.***

*The registry/discovery service should not just be thinking about data – in future it will need to consider methods, workflows, ontologies, theories and people". – **Established Researcher (Geography).***