Our Land and Water National Science Challenge – A Data Ecosystem for Land and Water Data to Achieve the Challenge Mission

Data · September 2016

7 authors, including:

David Medyckyj-Scott
Landcare Research
29 PUBLICATIONS   **297** CITATIONS

SEE PROFILE

Kristin Stock
Massey University
50 PUBLICATIONS   **167** CITATIONS

SEE PROFILE

Robert George Gibb
Landcare Research
23 PUBLICATIONS   **49** CITATIONS

SEE PROFILE

Mark Gahegan
University of Auckland
169 PUBLICATIONS   **3,742** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

"sense+know" From environmental monitoring to management: Extracting knowledge about environmental events from sensor data, ARC DP120100072 View project

GBS pipeline View project

Our Land and Water National Science Challenge

A Data Ecosystem for Land and Water Data to Achieve the Challenge Mission

Authors

Dr D Medyckyj-Scott

Technical Director, National Land Resource Centre, Landcare Research

Dr K Stock

Director, Geoinformatics Collaboratory, Massey University

R Gibb

Research Associate, ex Informatics Science Team Leader, Landcare Research

Professor M Gahegan

Director, Centre for e-Research, University of Auckland

Dr H Dzierzon

Team Leader Bioinformatics, Plant & Food Research

Dr J Schmidt

Chief Scientist Environmental Information, NIWA

Dr A Collins

Director, National Land Resource Centre, Landcare Research

Contributors

The following individuals and organisations have joined us in the spirit of co-innovation and generously contributed time and expertise to this project.

Alison Fordyce	Senior Investment Manager, Strategic Investments, Ministry for Business, Innovation and Employment
Andre Post	Manager Geospatial Management, Ministry for Primary Industries
Martin Workman	Manager Land and Water Policy, Ministry for Primary Industries
Robert Deakin	Chief Steward - National Spatial Data Infrastructure at Land Information New Zealand
James Turner	OL&W NSC Science Theme Leader, AgResearch

Prepared for:

Our Land and Water National Science Challenge

AgResearch
Ruakura Research Centre
10 Bisley Road
Private Bag 3123
Hamilton 3240
New Zealand

Reviewed by:	*Approved for release by:*
	Justine Daw
	General Manager
Landcare Research	Landcare Research

Landcare Research Contract Report: LC2664

Disclaimer

Version 1.0   23rd September 2016

Contents

Executive Summary

1. The growing diversity, complexity, and volume of data represent a rich source of opportunity to lift primary sector productivity, social license to operate, and value for premium product. Thus one of the greatest 'additionality' gains for the Our Land and Water National Science Challenge ('the Challenge') is gathering this amorphous collection of data into a dynamic, shared data ecosystem in which data can be widely used, and more easily understood, integrated, and analysed.

2. To address the question 'What are the best data structures for land and water information to achieve the Challenge Mission?' a small group of experts were commissioned to produce this 'think piece' focusing on what a data ecosystem that allows data to be used to its full potential for all stakeholders should look like.

3. Short surveys, discussions with MBIE, Pan-challenge consultation, insights from the wider national and international research data community as well as stakeholder contributions were used to canvas issues and needs and shape the think piece. Expert panel membership of data-related international programmes was also leveraged with respect to the selection of principles, best practice, and appropriate technologies.

4. Throughout the think piece the concept of 'data ecosystem' is used to describe a system made up of people, practices, values, and technologies designed to support particular communities of practice. In such an ecosystem 'data is valued as an enduring and managed asset with known quality'.

5. This ecosystem approach is critical as the Challenge involves a wide range of data generators and consumers, including individuals and groups from different disciplines and sectors and each with different capabilities, expertise and motivation around data.

6. The vision is an advanced data ecosystem which enables frictionless data access and sharing across New Zealand's land and water stakeholders to meet the Our Land and Water Challenge Mission efficiently and effectively. The vision has the potential to seed benefits that are far more wide reaching than the Challenge including support for communication, participation and collaboration that considers Pakeha and Māori values and worldviews.

7. To evolve a data ecosystem that is capable of supporting the Challenge and its mission as well as dealing with large data volumes, complexity and heterogeneity (often constrained by issues of privacy, IP and licensing), the use of a Data Management Maturity Model is recommended as a framework for thinking and action.

8. Under the Data Management Maturity Model we recommend a shift from ad hoc approaches to managing and exchanging data (Level 1) towards the development and adoption of community-wide practices and standards for data sharing and data governance (Level 3). Certain aspects of the data ecosystem will need to grow to levels 4 and 5 if the proposed vision is to be obtained.

9.   As the Challenge looks to the future, we identify the environment needed to enable the data ecosystem. This includes a 'Roadmap to Achieve the Our Land and Water Data Ecosystem' that highlights the steps required to build effective governance, mature individual and pan-institutional behaviours, and realise technical capacity (moving to Level 3 of the Data Management Maturity Model).

10.   Other critical pieces in the development of the enabling environment we recommend include:

    a.   Changing the data management culture.

    b.   Establishing a data analytical structure for the ecosystem (such as cloud-hosted)

    c.   Increasing interoperability

    d.   Building collaboration

11.   We also identify a set of research questions that must be addressed in order for the vision of a mature data ecosystem to support the Challenge to be realised.

12.   A lack of confidence and willingness to move towards significant transformation is often a barrier to change, we therefore recommend a series of 'first steps' to take place over the next 6 months including:

    a.   Engagement with senior management (CIOs) to initiate promotion within participating institutions and identify and name data ecosystem Champions

    b.   Endorsing the vision and principles as a Challenge, including stakeholders and collaborators in the endorsement process

    c.   Agreeing priorities for first steps and research

    d.   Establishing a cross-Challenge data management governance group

    e.   Establishing the role of data stewards for themes

    f.   And creating a collaborative space that allows forums to discuss data-related issues.

1   Introduction

'Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?' [1]
Where is the information we have lost in data?

from 'The Rock', T. S. Elliot, 1934

The unprecedented and growing diversity, complexity, and volume of data and derived information products represent a rich source of opportunity to lift primary sector productivity, social license to operate, and value for premium product. Thus one of the greatest 'additionality' gains for the Our Land and Water National Science Challenge ('the Challenge') is gathering this currently amorphous collection of data generated in science, practice, policy and society, into a dynamic, shared data ecosystem in which data can be widely used, and more easily understood, integrated, and analysed. With a data ecosystem that is increasingly populated with real-time, fine-scale data from a variety of sources (including data from sensors and citizen science contributions) and increased accessibility, this assemblage of 'big data' has the potential to power the Challenge. It will provide the primary sector with defensible sustainability credentials (theme 1) to participate in global value chains and markets, support complex decision-making about the way land and water is used (theme 2), and provide the means by which individual land and water users, communities, and iwi agree and implement co-developed solutions to guide communities' innovation aspirations for their land, water and people (theme 3).

Big data, the tools that leverage off that data and the analysis carried out using it, are fundamental to the Challenge having impact. But the Challenge must address a range of interwoven technical and social challenges caused by the fragmented and heterogeneous creation, management, supply and governance of data across the multiple stakeholders involved in the Challenge. At the same time a step change is required to keep pace with evolving needs and technological developments, respond to the complex requirements of the Challenge such as co-innovation, data generation and integration, analysis and modelling, as well as information delivery and presentation. This step change includes the way data are collected, managed, and made available, and the behaviour of the actors within a 'data ecosystem'.

To create an advanced data ecosystem requires significant stretch and the advancement of a unique science mix of business and systems analysis; elements of psychology; user profiling (Liaskos et al. 2010); data science, informatics and computer science, semantics and ontologies (Stock et al. 2013); related systems development and implementation, visual semiotics and uncertainty visualization (Maceachren et al. 2012); and software ergonomics and organisational practice. The level of effort and thus investment required to achieve an advanced data ecosystem is significant. However, there will be components of the underpinning infrastructure that will be required by all the Challenges and the science sector in general e.g. data storage, data curation and data citation services; collaboration with the other Challenges and the research institutions on these shared services will help lower the investment required by the OL&W Challenge.

There is significant potential for Māori knowledge, resources, and data to make a valuable contribution to the Challenge. However, this knowledge and data will have very specific attributes that will require thoughtful incorporation into the wider data ecosystem, and collaboration throughout the Challenge to ensure solutions also meet the needs of Māori next

---

[1] T S Elliot, extract from one of the Choruses in the play The Rock.

and end-users. The approach required needs to incorporate Māori values and worldviews, through providing support for collaborative decision-making processes among kin groups and other parties. It needs to consider different worldviews in the design of access mechanisms; and explore approaches to incorporate diverse data including qualitative and oral data in the mature data ecosystem that we envisage.

Collaboration is a key Challenge value. Again we envisage a move towards a mature data ecosystem that supports, enables and develops collaboration and co-design among data creators and data consumers.[2] A large range of different parties are involved in the issues that the Challenge is addressing, each with their own objectives, perspectives, and disciplinary background. The Challenge's success relies on an effective data ecosystem that can effectively support collaboration among such a diverse set of participants.

1.1   Project Background

At the end of June 2016, the Our Land and Water National Science Challenge funded four 'think piece' projects. These small projects of 3 month's duration were intended to inform larger pieces of research addressing gaps in the overall Challenge programme. Landcare Research was awarded funding to work with a small group of experts to write a 'think piece' addressing the question *"What are the best data structures for land and water information to achieve the Challenge Mission?"* Answering this question was identified by the Challenge Board, Directorate (Ken Taylor and Rich McDowell) and Science Advisory Panel as critical to the success of the Challenge. It is intended that the think piece produced by the expert panel will be used by the Challenge leaders to refine a research question that will be discussed at a workshop in October 2016 from which a collaborative research proposal will be created. Many of the recommendations made in this think piece are more about governance and operational aspects of the data ecosystem than research-oriented and Challenge leaders may choose to implement these without invoking a research question.

1.2   Scope and operating context

The term 'data' can encompass a huge variety of types of objects (e.g. word processing documents, spreadsheets, database files, charts, graphs, electronic mail, logs, photographs, programming notes, etc.). For the purpose of this white paper we view data as *opinions, reports, observations, facts, and statistics collected or created for a specific purpose of studying or analysing, gaining understanding and communicating*. The scope of this document white paper also encompasses data as it relates to modelling and data-related publications. Administrative data such as those arising from project management and contracts, and data in the form of documents such as reports and research publications, are however, considered out of scope.

The Challenge research and business plans[3] mention tools in some detail, and closely relate them to data. While we recognise the importance and relevance of tools in meeting the Challenge objectives, we do not consider tools directly in this document other than a cursory review (Section 3.5). We do, however, consider the need to ensure that any developments

---

[2] We use the term data consumers in preference to next and end users. As noted in section 3.3, data consumers can be direct or indirect users of data.

[3] Our Land And Water - Toitū Te Whenua, Toiora Te Wai National Science Challenge Revised Research And Business Plans September 2015

towards a more mature data ecosystem consider the requirement for tools to be able to interact with data in the ecosystem, with tool interoperability an important consideration.

It is important to understand the context for this white paper, including the operating constraints, such as the limited timeframe for delivery (less than three months) and funding. This, for example, dictated the level of engagement the expert panel could have both with those working in the Challenge and with stakeholders. A secondary issue was that the first phase programmes of research in the Challenge were also in development, and still working on gaining greater clarity on end-user need.

1.3   Consultation

Given the operating constraints, it was essential the expert panel undertook cost-effective consultations to canvas issues and needs.

Short surveys

To better understand the data requirements of those working within the Challenge, and key stakeholders/end-users two short surveys were conducted. The findings are reported in this white paper[4]. The first survey was an internal one of Challenge Theme and Programme Leaders. The second was with 15 key stakeholders who were selected by the Challenge Leadership. The people approached in the stakeholder organisations were senior: CEOs, Principal Advisors, and senior managers. Unfortunately, unlike the internal Challenge survey, the response rate from stakeholders was very poor. This was despite the Director of the Challenge contacting those surveyed and follow-up telephone calls by project staff.

At the request of the Challenge leadership, selected government agencies contributed content and reviewed the report including Andre Post (MPI), Rod Deakin (GSO, LINZ), and Martin Workman (MfE/MPI). Finally, for specific topics, experts were consulted. For example, Kevin Ashley, Director of the Digital Curation Centre, UK, and his colleague Angus White regarding data management in virtual organisations and Garth Harmsworth, Landcare Research, regarding the use of data by Māori.

MBIE discussions

A Senior Investment Manager involved in the National Science Challenges at the Ministry of Business, Innovation and Employment was approached with a small set of questions about MBIE's expectations with respect to data management, access, and use. As a result the following direction came from MBIE[5] on data structures to achieve the Challenge mission.

1. Reiterating views in the Challenge research and business plans,[3] that the Government's preference is for open access to data and outputs generated with public funding such as the OL&W Challenge.

2. That New Zealand's approach to science data management, curation, and access would move towards international best practice.

3. That the primary objective of Challenge activity should be to deliver on its objective. MBIE do not see it as the responsibility of any individual Challenge to create new analytical

---

[4] The results have been made available to the Challenge leadership in a separate annex, Annex 1.

[5] Email communication from Dr Alison Fordyce, Senior Investment Manager, Science System Investment and Performance, MBIE. 31 August 2016. The email noted that MBIE does not currently have a formalised position on data management or sharing but it has been identified as an area of focus for future policy.

infrastructure or services. MBIE encouraged Challenges to, where possible, link to existing infrastructure/service providers.

MBIE's position mirrors those of the expert panel and is reflected in the approach and recommendations subsequently made in this white paper. However, as we identify in this paper, the lack of key infrastructure and services will be a problem for all the Challenges and affect their ability to deliver.

Pan-challenge consultation

We approached the Directors of a number of the national science challenges[6] seeking information on what they were doing with respect to data structures in their Challenge. We also asked what other organisations and data infrastructures their challenge needed to interoperate with and how. In some cases the Challenge Director's redirected our enquiry to others working in their challenge. From these we were able to gather some useful information. However, the majority of the challenges reported that it was too early for them to have answers to our questions although the importance of activities like data management was being considered and specific partner institutions that had relevant expertise in this area were being consulted.

Towards the end of the project we became aware that a new project had been funded through the 'Science for Technological Innovation National Science Challenge' entitled *'Te Tāhū o te Pātaka Whakairinga Kōrero: Next Generation Indigenous Knowledge'*. This project will develop, in consultation with Vision Mātauranga (VM) teams, a platform for digitally managing and distributing Indigenous Knowledge (IK) within and across each of the National Science Challenges (NSC) that uses spatial hypermedia. We have initiated discussion with the project lead (Dr Hēmi Whaanga), and consider collaboration with this seed project key to the development of the data ecosystem to fulfil the objectives of the Challenge, particularly those that relate to iwi. Such collaboration would benefit from the data management considerations discussed in this white paper, with the potential for co-design benefiting both endeavours.

Wider national and international data community

We spoke directly with individuals working on related national initiatives, e.g. the Regional Councils Land, Air, Water Aotearoa (LAWA). Intelligence from a number of national initiatives such as the *Geospatial Senior Officials Group (GSOG), the Natural Resource Sector (NRS) Information Directors Group, the Biodata Services Stack (BSS)/New Zealand Organisms Register (NZOR) Steering Committee, the National Environmental Monitoring Standards (NEMS) Steering Group, National e-Science Infrastructure (NeSI)*, was also incorporated within the white paper, in part via expert panel membership on these national governance groups. Similarly, expert panel membership of data-related international programmes, including the *Open Geospatial Consortium (OGC), the World Meteorological Organisation (WMO), the Geosciences Network (GEON), the NSF Earthcube Programme[7], Group on Earth Observations (GEO), the Global Biodiversity Information Facility (GBIF)*, the *Research Data Alliance*, was leveraged for the report.

---

[6] The NSCs we consulted were those we considered as conducting research in the environmental data space, namely, The Deep South, Resilience to Nature's Challenges, New Zealand's Biological Heritage, and High-Value Nutrition.

[7] EarthCube was initiated by the National Science Foundation (NSF) in 2011 to transform geoscience research by developing cyberinfrastructure to improve access, sharing, visualization, and analysis of all forms of geosciences data and related resources. In 2015, NSF awarded 14 new EarthCube activities totalling approximately US$35 million. http://www.earthcube.org/

2   A Data Ecosystem

Common terms used when talking about data in the context of scientific research are 'data life-cycle'[8] and 'data infrastructure'.[9] In the context of the Challenge we believe these terms, and also the term 'data structures' used by the Challenge, are deficient with respect to the breadth of the Challenge's concerns. We have instead adopted the concept of 'data ecosystem' (Pollock 2011; Vanschoren et al. 2015). A mature data ecosystem is a distributed, adaptive, open socio-technical system with properties of self-organisation, scalability, and sustainability that turns data into information and knowledge[10]. It comprises a system made up of people, practices, values, and technologies designed to support particular communities of practice. It is our view, which is elaborated in Section 4, that the current Challenge data ecosystem is highly immature and chaotic and that it needs to evolve to a much more mature model if the Challenge is to deliver on its mission.

The term data ecosystem is thus used in this paper to emphasize the need to look at the wider context in which the Challenge operates when examining data structures issues. It highlights the fact that all the pieces of a data ecosystem are necessary and must work together to build a longer-term viable solution with respect to data collection, management, curation, sharing, use, and re-use.

Applying the concept of data ecosystem in the context of science data is not new to New Zealand; it first appeared in MoRST's Environmental Data Management Policy Statement (April 2010)[11]. The term is also currently used within the Government, for example, Statistics New Zealand's mention the concept within their data delivery plan *"Unleashing the power of data to change lives"* (January 2016)[12] and James Mansell, member New Zealand Data Futures Forum, talked about the need for a national data ecosystem in his lecture to Treasury on 5 September 2016.[13]

The data ecosystem term encompasses:

- policies regarding data management planning, data custodianship and curation, legal frameworks, and the use of externally sourced data;

- procedures and processes to execute those policies and manage data;

- a data governance framework and organisational structures;

- engagement with data consumers and stakeholders; and

- technology platforms that will support data collection, storage, description, analysis, linking, delivery and curation.

Critically, data ecosystems can be nested with community and regional ecosystems at the micro-level, national ecosystems at the meso-level, and a global ecosystem at the macro-level. Each ecosystem primarily deals with data from its own level, but may intersect with others at times (Heimstädt & Saunderson 2014). Thus aspects of the Challenge data ecosystem will overlap

---

[8] https://www.dataone.org/data-life-cycle

[9] https://www.epsrc.ac.uk/research/ourportfolio/themes/researchinfrastructure/subthemes/einfrastructure/strategy/roadmap/data/

[10] This definition is a modification of that of Marinos and Briscoe (2009).

[11] http://www.mbie.govt.nz/publications-research/publications/science-and-innovation/MBIE_Environmental-Data-Management-Policy-Statement.pdf/at_download/file

[12] http://www.stats.govt.nz/~/media/Statistics/about-us/corporate-publications/he-hui-tatauranga-aotearoa/data-delivery-plan-Jan-2016.pdf

[13] http://www.treasury.govt.nz/publications/media-speeches/guestlectures/mansell-sep16

with the data ecosystems of those of partner institutions, other Challenges, regional and central government, and those operating in businesses and business sectors.

3   Setting and Contexts

3.1   Challenge Themes and Programmes

The vision for the Challenge is that New Zealand is '*world-renowned for integrated and successful land-based primary production systems, supported by healthy land and water and capable people'*. This vision will be achieved through a research programme that has the following impacts:

- *Individual land and water users, communities, and iwi will have the social processes, data, tools and increased capacity to agree and implement co-developed solutions*. These solutions will produce mutual benefits to meet their aspirations and achieve sustainable outcomes by operating within agreed resource limits.

- *New Zealand land users and regulators will have a menu of tested technologies coupled with new innovative land-use options and land- and water-use practices* that achieve primary production growth targets within community and regulatory limits.

- *The New Zealand primary sector will sustain higher economic growth through participation in global value chains that are generating new products, services and market segments* that are aligned with and validated against stakeholder environmental, social and cultural values.

The Challenge is currently structured into a series of themes and related programmes of research (see Table 1).

Table 1 Challenge Themes and Programmes

Theme	Programme	Summary
Theme 1: Generating greater value from global markets	**Market access, value chains**	A nexus project to identify the best types of value chain and value chain factors most likely to effect changes to land management practices and land uses at relevant scales is currently underway. This work will help shape the design and focus a detailed research programme that addresses the theme's objective.
Theme 2: Innovative and resilient land and water use	**Sources and Flows**	Managing contaminant pathways and attenuation to create headroom for productive land use.
	Suitability	Using resilience in receiving water bodies and soils to guide land use suitability decisions and meet community objectives.
	Next generation systems	Next generation primary production systems: opportunities to change the face of production.
Theme 3: Collaborative capacity	**Mauri Whenua Ora**	Developing a mātauranga-centred framework to aid land and water utilisation and community innovation.
	The Collaboration Lab	The collaboration lab: the transformative role of collaboration in managing our land and water.

Nexus	**Performance Indicators**	For monitoring and integrated assessment at different scales, establish boundaries of land suitability (trade-offs/impact), and what metrics are best to use. Indicators of productivity, at various scales.

3.2   Challenge Data Goals

Access to data is fundamental to achieving the Challenge mission, and the following goals important for realisation of a data ecosystem that is in turn able to support the Challenge in achieving its mission. These requirements have been extracted from Challenge documents and consultation outlined in Section 1.3 (further details on requirements and sources can be found in Appendix 1).

1. Collaboration is crucial. The data ecosystem must support and enable collaboration and must be created through co-design, co-innovation, co-development and co-production.

2. The data ecosystem must support different world views (e.g. Māori and Pākehā) and diverse kinds of data, including qualitative, quantitative, written and oral.

3. The data ecosystem must support data-driven science, enabling data to be used and produced at all research stages.

4. The data ecosystem must integrate and support analysis of diverse data types, sources and domains; with different structures and semantics (or meaning), different geographical and temporal scales; data collected from official, business and citizen science sources, by methods including field collection, sensors, imagery, and terrestrial survey instruments.

5. The data ecosystem must enable intelligent search and discovery of data from the 'user' perspective, supporting the data → information → knowledge → wisdom transition.

6. The data ecosystem must make data available in a form that enables interoperation with tools and visualisation techniques to present both data and results.

7. The data ecosystem must operate a policy of open access to data and information, having due regard for the rights of third parties, cultural sensitivities and the appropriate protection and management of Intellectual Property.

As mentioned in Section 1.3, the stakeholder survey on data needs had a poor response rate. We would therefore recommend opening the survey for a longer period and to a wider group to establish greater clarity. The same or a similar survey could be used as an ongoing tool to gauge conceptual and cultural shifts as the Challenge progresses and the data ecosystem matures.

3.3   Challenge People and their Interactions with the Data Ecosystem

The Challenge involves a wide range of different types of users, including individuals and groups from different disciplines and sectors. Each user brings his or her own capabilities, expertise and previous experience, which will influence their perspective, terminology, and motivation around data.

There are also a number of roles users may play in the Challenge data ecosystem, and different users may play different roles at different times, and in different situations, for different data sets. The same user may be a provider for one data set and a consumer for another. The same data set may be used by a range of user groups at different stages during its life-cycle. Table 2 attempts to clarify the complexity around user and consumer, as well roles within the data

ecosystem. The values in brackets indicate where we believe users could be playing a more involved role in the Challenge.

The roles in the data ecosystem (top row) are defined as follows.

- **Data providers** create a data product which they provide to others. They may have collected or created the data themselves, or sourced it from elsewhere. A data provider is an accepted supplier of data into the Challenge ecosystem.
- **Data collectors** collect data directly, e.g. through surveys or interpretation of imagery, or indirectly, e.g. using real-time sensors such as water quality samplers and sensors, and make them available through the data ecosystem. This may include data collated via citizen science projects, or through volunteered geographic information (VGI) applications.
- **Data creators** create data through combining data with other data, or deriving new data as a result of analysis or modelling.
- **Data managers** are responsible for setting policies, QA of the data and ensuring correct and adequate governance of the data. They manage the data using good practice across its entire lifecycle. Data managers are often referred to as "data custodians".
- **Data owners** (of Challenge data) own the data set and grants rights of use to others via a license.
- **Data customers** are those who actually need the data to achieve some purpose, although they may not directly use it. Another agent may access and use the data on their behalf.
  - **Data consumers** (direct) consume data directly from the data source.
  - **Data consumers** (indirect) consume data, most likely as information, through an application, phone app, web site, etc.
- **Data reviewers** provide feedback on data quality or requirements to the original data collector or creator. Often data reviewers will also be users.
- **Value adders** take a data set, add value and then provide it to another party. That other party could be businesses, consumers or even scientists.

We have not used the terms next and end-user which appears in the Challenge documentation as we do not believe these terms provide much clarity on the nature of the use or their role. Rather we have used the terms direct and indirect data consumers.

We have excluded the role of data steward in this categorisation. A steward works at a systems and strategic level, taking a holistic and aspirational view in the national interest across a set of related data (theme). The agency acting as a steward promotes good practice, and coordinates activities to ensure best outcomes, bringing the views of the users to the table (LINZ, Steward and Custodian Framework, 2014[14]).

---

[14] http://www.linz.govt.nz/system/files_force/media/file-attachments/steward-and-custodian-framework-image.pdf

Table 2 Data roles in the context of OL&W Challenge

		Provider to others (product)	Collector	Creator	Manager / Governance	Owner (Challenge data)	Customer (need)	Consumer (direct)	Consumer (indirect)	Reviewer (validation, feedback)	Value Adder
Science	**OL&W Challenge leadership**	-	-	-	*****	-	-	-	-	-	-
	Scientists – in OL&W Challenge	****	**	****	***	*****	*****	*****	***	***	****
	Scientists – in Other Challenge teams	***	**	****	***	*****	*****	*****	***	***	****
	Scientists – non-OL&W Challenge *(Govt., universities, CRIs, other Challenges, business)*	***	**	***	-	-	*****	*****	***	***	****
	Challenge participant organisations	*****	-	-	*****	-	-	-	-	-	-
Govt. (central, local)	**Policy**	-	-	-	***	-	*****	*****	*****	**	-
	Senior decision makers	-	-	-	**	-	***	*	*****	-	-
	Regulatory - DOC, MfE, MPI, RC *(e.g. compliance officers, environmental reporting)*	***	***	**	**		*****	***	***	***	*
	Operational	*****	*****	*****	***	**	*****	*****	*****	****	***
Business	**Agribusiness** *(large producers, e.g. Fonterra, Ravensdown)*	**	***	***	*	**	*****	**	***		*
	Primary sector organisation *(marketing, regulatory, policy, DairyNZ)*	**	***	*	*	**	*****	***	***	**	**
	Farm consultants	-	-	-	-	-	*****	****	***	***	-
	Land owners *(e.g. owns farm & pays someone else to manage it)*	-	-	-	-	-	***	**	****	-	-
	Managers – land not in productive/residential use *(e.g. conservation land, national parks)*	*	***				****	***	*****	*	-
	Farmers/farm managers *(incl. forestry and other kinds of primary production)*	** (****)	****	**	*	-	****	****	*****	* (****)	-
	Tourism *(businesses & regional tourism organisations)*	*	**	**			**	***	***		
	Inshore fisheries *(e.g. salmon farming)*	*	**	**			**	***	****		
	Māori *(may also fall into one or more of the above categories)*	** (****)	***	**	*	**	*****	****	*****	* (****)	**
Consumer	Local NZ consumers	-	-	-	-	-	-	-	*	-	-
	Off shore consumers	-	-	-	-	-	-	-	**	-	-
Vested interests	**Iwi / hapū / whanau / kaitiaki** *(with an interest in/ connection to land, often nationally / internationally dispersed, often not running agri-businesses as covered above)*	** (****)	**	**	*	***	*****	**	****	*(****)	****
	Community / interest groups *(groups with an agenda, catchment, recreational, friends of …)*	**	**	*	-	*	**	**	***	**	-
	General public	-	- (*)	-	-	-	-	-	***	-	-
	Local communities	*	*	*	-	-	*	*	***	*	-
	NGOs *(e.g. Forest and Bird, Landcare Trust)*	**	**	**	*	*	****	**	***	**	**

3.4   Types of Data

The Challenge[15] describes an ambitious programme of research that relies on and/or creates a large number of social, economic and environmental data sets together with tools to discover, intelligently integrate, and derive new knowledge.

Existing data sets will be important for the data ecosystem, as background or as input into the calculation of other data sets or models. It is anticipated that much of the data in this category is already held in long-term databases maintained by CRIs, Central and Regional Government or the private sector. However, and as noted in the Challenge documents, it is likely that some of these databases will not have sufficient quality, currency, resolution or completeness to meet the requirements of the Challenge or may not use the correct classification systems or semantics[16].

Lack of data of appropriate quality and fitness-for-purpose has the potential to hinder achievement of the Challenge objectives if not addressed. In addition, it is likely that some of the data required to achieve the Challenge mission has not yet been captured in coherent data sets. There may be multiple sources that hold fragments of data in different forms and using different representations. The data ecosystem will need to address these issues of heterogeneity and evolve to support scientists dealing with data constrained by privacy, IP and licensing. As the number of sources increases, overcoming the diversity of organisational approaches to privacy, IP and licensing hinder progress. This effort should not be underestimated.

New data sets may also be created by the Challenge through the integration of a number of other data sets, citizen science or Volunteered Geographic Information (VGI) approaches or the creation of data resulting from research activities (e.g. land suitability, contaminant pathways, primary sector performance, and performance indicators).

Appendix 6 lists some of the key data themes that are likely to be required to meet the Challenge mission. These were identified from Challenge documentation and our survey of Challenge leaders. However, a detailed analysis of goals, requirements, and existing data resources is necessary to fully specify the data needs of the Challenge. This activity relies on the detailed development of the Challenge programmes over the coming months.

3.5   Challenge Tools

Tools are mentioned in many places throughout the Challenge documents and are a core part of the way the Challenge will deliver impact (e.g. *"Land and water managers and hapū/iwi are using the solutions and tools developed within the Challenge to increase production and profitability…"*[15] p. 20). The term 'tool' is used broadly, encompassing classification systems, modelling tools, cost benefit analysis and GIS applications (see Table 3).

While the data ecosystem is focussed around data, tools must also be considered as many ingest or produce data sets or form part of a data workflow in which the output data from one tool becomes an input for another tool. The fitness-for-purpose of data used by tools is critical. Data requirements and interpretation of output for some tools are specific to the New Zealand environment, and *"the Challenge can act to ensure the right data are used with/for the right tools*

---

[15] Our Land And Water – Toitū Te Whenua, Toiora Te Wai National Science Challenge Revised Research And Business Plans September 2015.

[16] The 2013 Environmental Domain review conducted by Statistics New Zealand identified similar serious shortcoming in the land and water related datasets that are used to provide insights on the state of our natural environment.

and that tools are able to talk to one another i.e. tool interoperability",[15] pp.4–5). Data interoperability is thus an important pre-requisite for tool interoperability.

The Challenge will interact with tools in a number of ways, including extension or enhancement of existing tools as a result of research outcomes (e.g. GIS decision support tools in Theme 3) and co-development of new tools (e.g. classification-based land use suitability tool in Theme 2).

Table 3 Categories of Challenge Tools

Category of Tool	Examples
Data management tools	ArcCatalogue, data repositories, MS Excel, MySQL, Dropbox, MS Access
Generic data manipulation and analysis tools	GIS, e.g. Arcgis, R, ETL and Python scripts, Excel. NVivo (qualitative data analysis)
Scientific tools specific to a particular domain or problem space	SEDNET, CLUES, OVERSEER, FARMAX, IFM, GAMS (General Algebraic Modeling System), LTEM (Lincoln Trade and Environment Model)
Tools targeted at data consumers	Land/farm management tools (e.g. OVERSEER, ApSIM, MyLand), new databases, web sites, custom GIS applications.

Our investigations (Annex 1[17], and through a collaborative workshop on land management tools[18]) found the following with respect to end-user tools:

- Outputs from tools are often NOT interoperable due to the use of different source data.

- Many existing tools require a high level of competency by the end user.

- Outputs from tools are difficult to evaluate due to lack of provenance information about their input data.

- Data sharing from tools is hampered by different licences used by multiple source data.

- There are too many tools and users need guidance on which to use and how.

- There are barriers between landscape-scale tools and property-scale tools.

- Tools need to be developed that will utilise real-time data where appropriate.

3.6   National and International Initiatives

A data ecosystem developed for the purpose of the Challenge needs to be built within the context of relevant national policies and initiatives, international obligations and programmes, and application in other/related science challenges. Appendix 3 describes a selection of relevant national initiatives that: (1) describe best practice data management; (2) provide guidelines for the development of data management infrastructures like the Challenge data ecosystem described in this document; and (3) show examples of guidance and policy

---

[17] Annex 1: Surveys & Responses for 'A Data Ecosystem for Land and Water Data to Achieve the Challenge Mission' is available separately on request.

[18] New Developments in Land Management Tools and Their Application at the Farm Scale. Workshop run by the National Land Resource Centre workshop in April 2015.

documents developed for other similar/related national initiatives. Appendix 4 summarises a selection of relevant international initiatives, and is focussed on examples that have similar collaborative data sharing efforts to the Challenge. These documents have provided valuable background material for this white paper. In particular the following guidelines are identified that are relevant to the context in which the Challenge operates.

- *The Open Government Information and Data Programme* includes an open data licensing framework (NZGOAL[19]) and principles for managing data and information, and recommendations for data release.

- *The New Zealand government ICT strategy*[20] as well as the Land Information New Zealand (LINZ) Our Location Strategy[21] includes support for data standards and data release.

- Ministry of Business, Innovation and Employment (MBIE) requires data generated as part of its funded research project (including Science Challenges) to be made available consistent to the *Open Government principles*.

- A cabinet paper put forward by the Minister for Land Information in December 2010[22] recommends that government agencies be directed to support and be involved with the development of *Spatial Data Infrastructure.*

- The *Declaration on Open and Transparent Government*, approved by Cabinet in August 2011 stating: "Building on New Zealand's democratic tradition, the [New Zealand] government commits to actively releasing high value public data."[23] Supporting this declaration and operating under the New Zealand Data and Information Management Principles, is essential in ensuring "high quality management of the information the government holds on behalf of the public."[24]

- The International Council for Science international accord *Open Data in a Big Data World*[25], now endorsed by many international science partners, identifies the opportunities and challenges of the data revolution as today's predominant issue for global science policy. It proposes a set of fundamental principles for open data arguing that open data is a fundamental pre-requisite in maintaining the rigour of scientific inquiry and for maximising public benefit from the data.

- Various international programmes support the implementation of data management and federation practices. Notably the *Open Geospatial Consortium (OGC)*[26] and the *Global Bio Information Facility (GBIF)*[27] are developing and maintaining a range of standards for data federations. These standards are applied and supported in various programmes, for example within the *Group on Earth Observations (GEO), World Meteorological Organisation Information System (WIS), INSPIRE* (see Appendix 4).

---

[19] New Zealand Government Open Access and Licensing framework (NZGOAL) https://www.ict.govt.nz/guidance-and-resources/open-government/new-zealand-government-open-access-and-licensing-nzgoal-framework/

[20] https://www.ict.govt.nz/strategy-and-action-plan/strategy/

[21] http://www.linz.govt.nz/about-linz/our-location-strategy

[22] http://www.linz.govt.nz/system/files_force/media/file-attachments/cabinet-minute-capturing-benefits-of-location-based-information.pdf?download=1

[23] https://www.ict.govt.nz/guidance-and-resources/open-government/declaration-open-and-transparent-government/

[24] https://www.ict.govt.nz/guidance-and-resources/open-government/new-zealand-data-and-information-management-principles/

[25] Open Data in a Big Data World – An international accord - http://www.science-international.org/

[26] http://www.opengeospatial.org/

[27] http://www.gbif.org/

3.7   Technological Context

New ways of acquiring data are emerging (animal, on-vehicle and environmental sensors, RFIDs, UAVs, even social media), many providing data in real-time. The agricultural sector in many countries is taking the large volumes of data collected using such technologies, combining it with third-party data, such as weather forecasts or food prices, and feeding it into algorithms and models being created by scientists. This derived data is then used to gain insights into system behaviour; optimise land use; predict and improve crop yields; inform precision farming; monitor the state of the environment; and provide oversight across the value chain.

With data capture and analytical technologies changing, and changing faster than ever before, this will have radical impact on both the science opportunities open to the Challenge and constraints in operation. Traditional data collection for science is predominantly the domain of and under the control of the scientist, and engagement with farmers and Māori has been one of seeking permission to operate on their land. The convergence of sensor miniaturisation and price reduction as well as the increasing reach of the internet into the rural environment are driving the emergence of what is being called the Internet of Agricultural Things (IoAT)[28] and a shift of responsibility for gathering data from the scientist to individual farmers and farm equipment suppliers.

This convergence results in greater data *'variety'*: data from an increasing number of different sensors; greater *'velocity'*: data sampled at increasingly small time-steps and available in near real time; and most obviously, greater data *'volume'*, with all three experiencing exponential rates of change. These are the classic three *Vs of Big Data* (Laney 2001). With the loss of control over the context of each measurement come additional challenges of *'veracity'*: its origin, ownership, availability, and authenticity (Lukoianova & Rubin 2014) and of realising the emergent *'value'* for science and the Challenge's stakeholders. Extracting value requires a radical shift in analytical techniques, in understanding the value proposition of data streams that have not been designed for science, and in stakeholder engagement to negotiate access to data that is mostly not owned by the scientific community.

A mature data ecosystem with appropriate governance, coordination, data orchestration and management environments would allow these issues to be tackled much more effectively and efficiently.

3.8   Data Challenges

Global investigations into the re-usability of research data show the necessity for improvement (e.g. Roche et al. 2015). In New Zealand, data management and infrastructure are fragmented and data are sometimes located behind corporate firewalls. This creates significant friction when exchanging and sharing data, and poses an obstacle for releasing opportunities afforded by better data integration, including enhanced analysis and the resulting creation of new knowledge. Integrating data from disparate sources requires minimal friction and a good understanding of the acquired data. While a culture of sharing data is encouraged by the New Zealand government, particularly when tax money has been used (see *NZGOAL* and *NZGOAL-SE*[19] guidelines), little in the way of common data sharing policies currently exist. Underlining these policies is government support for centralised infrastructure initiatives like *New Zealand eScience Infrastructure (NeSI)* and *New Zealand Genomics Limited (NZGL)*.

---

[28] http://www.eletimes.com/technology-news/internet-of-things-technology-news/iot-technologies-in-agriculture-at-display-at-iot-tech-expo/

The move towards a mature data ecosystem to support achievement of the Challenge mission poses a number of specific challenges that must be addressed in order to move from the current situation, in which data are largely shared in an ad hoc manner or not at all, towards a research data management environment that will allow the data to realise its full potential in supporting the Challenge. These challenges include:

1. *Data variety is massive.* Different formats (e.g. NetCDF, HDF5, ESRI shapefile, CSV, Excel, XML), semantics, quality, source, media, and data-handling strategies are used and are often not coordinated. The effort required to share and integrate such heterogeneous data is significant[29].

2. *Data collection processes are not standardised*, and nor are the data that are recorded. Data collection is often duplicated by different agencies.

3. *Data supply is not coordinated* and often inefficient (e.g. via email, FTP, web browser).

4. An analysis of the supply of data identifies the (largely hidden) and *significant costs* borne by data consumers of modifying data products which are not fit for purpose. The onus is on the consumer to 'wrangle' data for use. Furthermore, data supply is often ungoverned with data providers able to arbitrarily change structure, format, frequency of delivery, condition of access and use.

5. *Multiple copies* of the same or what appears to be the same data exist.

6. *'Garbage in – garbage out'.* The analysis and development of models requires high quality data. A lot of potential input data is not fit-for-purpose (scale, currency, coverage, etc.), resulting in compromised output.[30]

7. Many of the questions being asked in the Challenge require atom level data (e.g. paddock or sub paddock) but that *data doesn't exist or is not uniformly available*.

8. *Lack of metadata and quality assurance* make it hard for data consumers to understand the limitations of the data they will gain access to, undermining user confidence in the data.

9. *Tensions between the requirement for open data and requirements for self-funding*, commercialisation and privacy and sensitivity of some data. For example, as well as sensitivities about sharing data on Waahi tapu, Waahi taonga and Mahinga kai sites, Māori business and agri-business/farming are cautious about sharing financial details and records of farms, their future land use plans, and nutrient budget plans.[31]

10. *Scaling up* from Challenge case studies to regional and national data sets and systems may not be trivial. For example, case study areas are often data-rich but and scaling up requires that much larger data volumes and processes are handled.

11. *New technologies* (e.g. sensors, UAVs) *provide opportunities* for gathering more timely, targeted and comprehensive data and for more advanced forms of data analysis and data analytics; however they provide *challenges* in building expertise to access, manage, or process the data.

---

[29] MfE and Stats NZ have estimated they spend about $140K on data collation and quality checks for each environmental domain report.

[30] Evidence for this can be found in the OL&W theme and programme leader survey, OL&W documentation, the sister think piece to this one on High Impact Indicators and the EMaR Scoping report of the Environmental Monitoring and Reporting Land topic (unpublished).

[31] Garth Harmsworth (Te Arawa, Ngati Tuwharetoa, Ngati Raukawa), senior environmental scientist, Landcare Research, personal communication.

4   Data Management Maturity

Capability Model Maturity Integration (CMMI[32]) is a standard approach to understand and assess the maturity of an organisation's processes and is used as a basis for continuous improvement. Developed by Carnegie Mellon University, it is heavily used worldwide in IT governance and software development.   Various versions of the Maturity Model exist, geared for different organisational challenges. Below, we introduce CMMI's *Data Management Maturity* (DMM) model that was specifically developed for organisations and communities of practice that "*…seek to evaluate and improve their data management practices*". It provides a common set of themes and language to describe, evaluate, plan, and improve data management activities. The themes: Data Governance, Data Quality, Data Management, Platforms and Architecture, Data Operations, and Supporting Processes,[33] as shown inFigure 1. We propose the model as a framework within which to contextualise the detailed issues that the Challenge community faces and a structured approach to advancing data management maturity.

The Data Management Maturity model addresses a number of specific data-related *themes* and describes five maturity *levels*, starting from an **initial** level, where the individual organisations and researchers have ad hoc approaches to managing and exchanging data. The second **development** level represents the emergence of some shared approaches based on specific projects or tasks; the third **defined** level represents the development and adoption of community-wide practices and standards for data sharing and data governance. The further two stages of **managed** and **optimising** represent movement towards a highly optimised and reactive state.
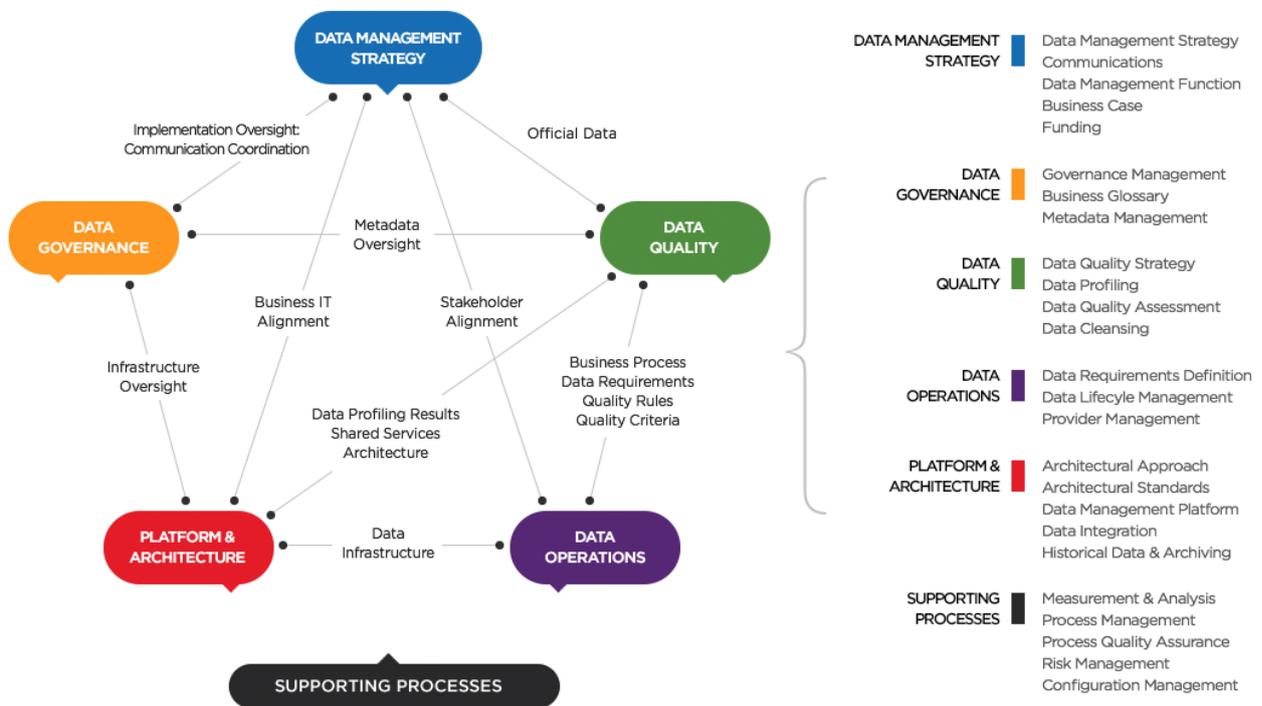

Figure 1 The six key themes of CMMI's Data Management Maturity (DMM) model [34]

---

[32] https://en.wikipedia.org/wiki/Capability_Maturity_Model_Integration

[33] http://rdm.ischool.syr.edu/xwiki/bin/view/CMM+for+RDM/Research+Data+Management+Maturity+Levels#H0.3ResearchDataManagementMaturityLevels

[34] Sourced from http://cmmiinstitute.com/data-management-maturity

4.1   Data maturity and the Challenge

Based on recent experience among some of the Challenge participants (including those in the Lincoln Hub), we expect a DMM assessment, while identifying pockets of activity at every level of maturity across all the themes, would find Level 1 operations are likely to be most commonplace. As an initial step we recommend the Challenge conducts a fuller assessment, to ascertain at what levels(s) it is currently functioning across the various themes. Using the DMM model as a framework partners in the Challenge can then work together to set shared expectations for how far they need to move from the current state towards an optimised state, including the technical and cultural steps needed to reach the desired level of maturity.

Greater benefits will accrue with the Challenge moving to a higher level of data management maturity, such as *Level 3*, where data can be more effectively used for discovery, sharing, publishing, and reuse. Level 3 is where open data starts to thrive although it represents the low-end of international best practice for institutions.

It is not always desirable or practical to strive for a maturity level of 4 or 5 across all aspects of data management, but it is desirable to bring communities up to the 'defined' third level across all aspects. However, for the Challenge to have impact we believe stretch to maturity *Levels 4* and *5* to meet specific data sharing and governing needs will be necessary. Note also that while the DMM model is a very useful framework within which to contextualise the many specific issues that the Challenge community faces, it does not enumerate all of these issues[35].

Table 4 below shows the full Data Maturity Model matrix, combining the 6 *themes* from Figure 1 above with the 5 *levels* of maturity. Some entries in the table text have been adapted from the Australian National Data Service (ANDS) DMM matrix,[36] which is based on earlier versions of the CMMI model. The row entries in the table are matched with the various *themes* in Figure 1 and use the same *theme* colours. We have added the italicised text in the table to help ground each level description in a practical issue likely to be experienced by participants.

4.2   Data Maturity and Community Maturity

While the Data Maturity Model was developed with institutions and communities equally in mind, non-institutional communities have their own social dynamic that will influence how they approach implementing the model. Critically, it is unlikely that maturity in handling data will emerge if in other ways participants lack a strong sense of community.

In preparing this white paper, Challenge participants reflected that the Challenge community is at present 'loosely-bound' with some expressing the view that they are disconnected from the Challenge community. As the Challenge is still young, this is not surprising, but it does point to both the need and potential to build 'community' and 'connectedness' to realise the Challenge mission.

The Challenge states that successful collaboration and co-innovation are at the heart of delivering on the mission. The community aspirations of the Challenge and the data ecosystem

---

[35] One of the weaknesses of the DMM and other data maturity models is the fact it is targeted at single institutions. Being a virtual organisation (VO), the Challenge has additional data related issues it needs to deal with. For example, institutions within the Challenge will have their own data ecosystems and research cultures, goals and capabilities, and their maturation rates may be faster or slow than the Challenge as a whole as a result. Also, since the Challenge is not a legal entity, it can neither own equipment nor hire staff, so it has to do everything with respect to physical assets through its partner agencies.

[36] http://www.ands.org.au/guides/capability-maturity

proposed are thus well-aligned; however it is important to recognise that some practices encourage while others impede a sense of community.

For example, increasingly diverse virtual research communities tend to adopt existing cloud-based services, rather than provisioning their own as an institute might. (Cloud-hosted alternatives to more traditional institutionally-hosted solutions include *zoom.us* for meetings; *github.io* for document hosting and version tracking; and *teamwork.com* for project planning.) Not only do these offerings specifically support open research collaboration they have the further advantage of being highly scalable, and their very openness says 'you are welcome to join our party anytime' as distinct from 'please go through these administrative hoops and we will give you restricted access to our firewalled private garden'.

Large science collaborations, on a scale similar to the Challenge, have evolved data and community maturity in new ways. An international example is the US National Science Foundation's EarthCube[37]. The NSF's Geoscience team noticed that as the science they funded became more complex, recipients of funding were spending an increasing proportion of their funds on getting started, finding and understanding diverse datasets, and negotiating access to data. The NSF's Geoscience team partnered with the NSF Cyberinfrastructure and engaged the entire Geoscience research community in a new collaborative initiative called EarthCube, and handed over the whole funding negotiation, governance, and development of solutions to the community – saying to them 'This is our collective problem. Join us on a journey to discover what the solution is, and we will commit to it'. In New Zealand the National e-Science Infrastructure (NeSI) collaboratively purchased and now operates High Performance Computing (HPC) infrastructure on behalf of New Zealand science. As part of the negotiation process for the second tranche of funding NeSI adopted a much more open and mature approach to its operations, service offerings, and community engagement. A small but significant part of this larger change includes use of both zoom.us and teamwork.com in preference to the platforms being operated by any of the respective academic and research institution partners. (Both examples are described more fully in Appendices 3 and 4.) Cloud-based modelling environments specifically for supporting science are also emerging, and these are discussed in Section 8.2.

---

[37] http://www.earthcube.org/sites/default/files/doc-repository/Caron - EarthCube Governance Whitepaper _ Realizing expectable returns on EarthCube investments in community building and democratic governance.pdf

Table 4 Full Data Management Maturity matrix

Planning & action horizon	Individual	Project	Challenge (Institutional)	Challenge+ (National)	International
	Level 1 INITIAL *Process is disorganised and ad hoc*	**Level 2 DEVELOPMENT** *Process is under development*	**Level 3 DEFINED** *Process is standardised, communicated*	**Level 4 MANAGED** *Process is managed, measured*	**Level 5 OPTIMISED** *Focus is on continuous improvement*
DATA MANAGEMENT STRATEGY	Metadata management is chaotic & understood by only a few. *"When I share data I discuss exchange formats with my end user."*	• Responsibilities are defined & skills are developed. • Management and exchange processes are established, defined, & documented as needed. • Metadata applied to key datasets & shared externally. *"My project has agreed to standardise on Dropbox for data sharing, but I don't know what other projects use."*	• Processes are standardised & integrated. • All data are assigned a globally unique and persistent identifier (DOI). • Metadata, including recording provenance, is applied to new datasets & shared externally. *"I am aware the OL&W Challenge Management has written a SOP describing how I should manage data and I endeavour to use it."*	• Metadata quality metrics are collected. • All datasets described & metadata shared. *"Having learnt how to use the OL&W Challenge's data management guidelines I find sharing data is a lot easier and it is satisfying to see the monthly statistics as adoption spreads."*	Continuous improvement applied to processes & capabilities. *"I find I can easily discover, access, use, and publish data that the Challenge holds common, using the latest tools and standards."*
DATA GOVERNANCE	Policies & procedures may be underdeveloped, not up to date, and/or inconsistent across the Challenge Community. *"I didn't know there was any specific OL&W data governance."*	Policies & procedures are developed & harmonised for specific tasks. *"I was pleased when my manager heard that two of my projects were using the same technology, and asked me whether it would be helpful to share the ideas more widely across the OL&W Challenge Community."*	Policies & procedures are defined community-wide and absorbed into behaviours. *"I have helped write a data governance document for the OL&W Challenge Community covering policies and procedures. This feels like a really promising step as we move towards a more coordinated approach."*	Policies & procedures are accepted as part of culture & subject to audit. *"It is great to see the way people are just using the recommended practices – and all the old issues that used to waste our time with constant discussion and data wrangling have now become easier."*	Policies and procedures are periodically reviewed, improved and aligned with current best practice. *"I'm doing work in multiple challenges and its now so much easier that each of our agencies have aligned thinking across the governance groups and I can use the same practices for all the challenges I work in."*
DATA QUALITY	Data quality measures are ad hoc or absent. *"I keep notes on how my data is collected, but I wouldn't trust anyone else to understand these."*	Some quality metrics are used for specific tasks or projects. *"My project is using a template we found for recording data quality and metadata."*	Data quality strategy is developed: quality metrics are used in a consistent manner across the Community. *"I have adopted the OL&W Challenge Community's metadata template and find some parts of it are better suited to what I do than others."*	Quality metrics are refined to be fit-for-purpose. *"I provided feedback for a revised version of a metadata template that will be released for the Challenge, better suiting our collective needs."*	Quality metrics are refined periodically, based on feedback from data consumers. *"The OL&W Challenge has supported my participation in a global standards community for my discipline and we expect only a few changes will need to be made to the Challenge's systems so that we can become a reference implementation."*

DATA OPERATIONS	• Simple data sharing can be a challenge. • Curation & preservation services are absent or disorganised. *"I believe my organisation's storage is backed up regularly, but I still keep copies of my data on USB drives as a precaution."*	• Project-based data sharing services become available. • Data preservation organised around shared projects. *"My project coordinates versioned data releases that we all use and we are confident that old versions and releases will remain available to be retrieved as well."*	• Community-wide data management strategy and values developed. • Community-wide data sharing and preservation becomes straightforward. • Widespread availability and uptake of data services. *"My project has developed automated processing workflows for our data and models and we are being encouraged by the OL&W Challenge Community to share these across the community."*	• Curation & preservation understood as critical to the ongoing Community mission. • Data sharing becomes commonplace and embedded in practice. *"The OL&W Challenge has recognised that the effort we put into managing our data and model workflow benefits the whole community and leads to productivity gains and more repeatable outcomes."*	Customer feedback is used regularly to update & improve data operations & services. *"We use the OL&W Challenge Community suite of workflows, and regularly receive suggestions for improvement from others in the challenge and then contribute improvements to the published workflows we are responsible for."*
PLATFORMS & ARCHITECTURE	IT infrastructure is patchy, disorganised & poorly understood. *"We find that different members of our project experience different levels of hardware and systems support in their organisations so we tend to have to constantly take that into account in the way we do things and who does what."*	• Funds are invested in Community-wide technology & skills. • Responsibilities are defined. • Documentation & training developed. *"We still operate quite disjoint infrastructures across our various partner organisations, but we have established some shared access points and as a consequence, it is becoming easier to collaborate."*	• Widespread availability of data platforms. Facilities are well defined and communicated, standardised and integrated. • Management shows active support for shared platforms. *"It's good that the OL&W Challenge has endorsed best of breed solutions from each stakeholder and endorsed them as preferred solutions for the challenge, with stakeholders managing those resources for the challenge participants and not just for their own staff."*	• Architecture is managed as a Community resource. • Funding adapts to platform needs. • Documentation and training are up to date. *"Our IT departments have worked together to design a system that supports multi-agency islands of trust so that services can be set up for the OL&W Challenge that we all have seamless access to (and I only need to use my institutional username & password)."*	Concerted efforts to optimise platforms and architecture to fit emerging needs. *"Developers of commercial on-farm tools participate in our architectural design workshops and implement on-farm solutions that contribute the famer's raw sensor data to the Challenge Community's data repository and use the repository as a source of data and to deliver actionable on-farm intelligence."*
SUPPORTING PROCESSES	• Data management planning is unsupported. • Training is ad hoc or missing. • QA is ad-hoc or absent. *"Over the years I've developed a set of processes that I find helpful and have been using them ever since. I'm reluctant to change because they are all working for me."*	Investment in skills and processes: Data management planning is used on projects, documentation & training developed. *"We have begun to use project-wide data management plans (DMPs) that give clear guidance to all of the project team."*	• Widespread availability and uptake of training and skills development in data management. • QA becomes feasible on training, processes to share and curate data. *"The Challenge now has a set of processes to guide us with making decisions on how to store, publish, describe and access our shared data assets."*	QA is routinely applied to processes, results feed into future planning. *"We have begun to meet as teams to better understand how our supporting processes work and whether they can be improved."*	Processes are optimised and periodically refined. *"I'm confident that we have highly efficient processes that put time and effort into important activities. They help me to focus on what is important."*

5   Vision and Mission

'We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely.'

E. O. Wilson, Entomologist, Author, Pulitzer Prize Winner
from his book 1998 book Consilience: The Unity of Knowledge

Our recommended vision and mission statements for a mature Our Land and Water Challenge data ecosystem are contained in Table 5.

Table 5 Vision and Mission Statements for the Data Ecosystem

Vision	**Data is valued as an enduring and managed asset with known quality. The vision for the data ecosystem is to enable frictionless data access and sharing across New Zealand's land and water stakeholders to meet the Our Land and Water Challenge Mission efficiently and effectively.**	
Mission	Social and Institutional	The mature data ecosystem will be achieved by a best practice approach to the management of data in which **agreed policies, principles and practices are adopted by all participating stakeholders throughout the data and research lifecycle** in order to facilitate sharing and access to quality assured data by Challenge stakeholders.
	Technical	The data ecosystem will be supported, enabled and facilitated by a federated infrastructure in which data may be collected from traditional sources and new technologies, curated, published, analysed, modelled, linked, used and reused but accessed through a single point of access, from its authoritative point of origin, with discovery and visualisation tools.

This vision has been developed for the wider community of the Challenge and its stakeholders, which includes science sector participants, but also government, regional councils, and the primary business sector. Members in each these groups will find it difficult to fulfil the vision and its mission without their institutions commitment; however the vision has the potential to seed benefits that are far more wide reaching than the Challenge.

Realisation of the vision and mission for the Challenge data ecosystem is to advance data management maturity, thereby creating a step change in data collection, management, sharing and use and related user engagement, offering a number of benefits including:

- Faster and more convenient access to quality assured data by data consumers.

- Improved transparency; the data, and claims derived from it can be scrutinised, can be reanalysed and tested for their validity.

- Increased opportunity to access existing data sets for value adding, rather than having to create new data sets, through better awareness of available data resources.

- Increased opportunity for collaboration around data collection and reuse so that data duplication and data management are decreased and awareness of work in particular themes across the Challenge is increased.

- Better awareness and participation around environmental and agricultural decision making by a broader range of stakeholders (for example, dispersed kin groups with interest in a piece of land) enabled by tools for virtual communication around data.

- Opportunity for users to be contributors by providing feedback through the ecosystem on the need for and utility of data created by the Challenge scientists and through integration of citizen science and crowd-sourced data collection.

- Access to an infrastructure that may be used to move towards mechanisms for scientists to be acknowledged for their data (data citation).

- Access to data to support environmental reporting through interoperation with other data infrastructures.

- Ability to scale up scientific models and tools developed by the Challenge for national use.

- Ability to quantify and connect to a broader range of data resources of different types (quantitative, qualitative, oral), allowing data quality and quantity to be determined to assist in justification for data collection activities to better support land and water science.

- Improved support for communication, participation and collaboration that considers Māori values and worldviews.

The following role-based scenarios illustrate the potential benefits for two users operating with a data ecosystem that meets the stated vision.

A scientist has the task of developing a model to predict the effects of increased nutrient concentration in a particular type of streams on the ecosystem function, as measured by invertebrate and fish communities. As underlying data, streamflow, nutrient concentrations, invertebrate and fish counts are required for the stream type in question. Scientists can now identify all stream reaches in question through thematically interrogating a nationally-maintained and quality-assured GIS database. The data is provided in a standard format using an agreed vocabulary of reach types and attribute definitions, representing New Zealand's best knowledge about its river systems from various sources.

The scientist can query all available observational data for these reaches using a geospatial search of observation data provided by many organisations in a standard geospatial format using agreed metadata and provenance schemes. The data can be accessed from multiple providers in a common format and seamlessly combined into one dataset, as individual datasets using the same agreed vocabularies or are mapped to each other through agreed and managed ontologies.

Finally, the data can be evaluated as a common data quality coding scheme is being used as part of the data generation. The scientist is able to develop the model directly on the obtained dataset instead of going through a consolidation process, which typically takes weeks of time.

A land manager wants to evaluate potential impacts of nutrient input/application on groundwater for his land. He can interrogate data on irrigation application, nutrient application, and stocking rates from all farmers in the area, as the data is provided in a common scheme, using shared vocabularies, with stated levels of quality and to a common geospatial standard.

The data can be instantly aggregated and fed into a model predicting maximum nutrient concentration contribution to groundwater and evaluated against groundwater concentration data provided in real time. The process is then automated and provided in a daily updated web portal to all relevant stakeholders.

While the benefits of moving towards a mature data ecosystem are significant, the effort involved is not trivial. However, the risks in continuing with the status quo include the following.

- As complex scientific and social challenges problems require greater data volume and diversity the knock on effect will be *significant demands on Challenge funding for marshalling and managing data* and liaising with colleagues and the wider participants regarding access to, and right to use data [38, 39].

- *Lost opportunity across the National Science Challenges to collaboratively create step change.* There is a limited window of opportunity to create an integrated and interoperable infrastructure in which data can be shared and accessed efficiently and effectively across challenges. Economies of scale can be achieved by a collaborative approach to reduce duplication both within and across the challenges and by leveraging off other current initiatives (e.g. the Environmental Monitoring and Reporting (EMaR)/LAWA data federation, LINZ work on a national Spatial Data Infrastructure (SDI)). This is particularly important for a small country like New Zealand with limited resources and which is thus unable to invest in research data infrastructure on the scale being seen in other countries.

- *Greater cost associated with a delayed approach to implementing best practice data management approaches* with programmes and projects established without incorporating the necessary data management, quality control and governance considerations.

- *Continuing reliance on non-interoperable solutions*, as currently occurs. This brings greater expense and increases the barriers to data sharing and integration by different users and sectors.

- *Problems adapting to new technological developments* and their data management requirements. Modern, best practice, interoperable infrastructure best positions the Challenge to meet the task of dealing with the new advent of data collection technologies such as sensors and UAVs, etc. More broadly, any delay in implementation of data management projects in the modern climate is risky due to the fast pace of technological development.

- *A risk that others will capture the value* instead and the Challenge will be left having to fund its own very expensive parallel data collection.

6 Principles and Expected Practices

The Challenge has an ambitious vision of bringing together a wide range of data sets from disparate sources, scales and quality. This integration will be necessary to support a wide variety of tools, to assist with the development of performance indicators, economic and environmental modelling and the creation of a land suitability map.

In order to empower researchers in the Challenge and release a high level of innovation to create impact, an overarching data management practice and a culture of collaboration are

---

[38] MfE and Stats NZ have estimated they spend about $140K on data collation and quality checks for each environmental domain report.

[39] In launching EarthCube in 2011, NSF estimated upwards of 65% of funds awarded to long tail research groups (i.e. research groups similar in size to NZ research groups), went towards preliminary marshalling and managing data and liaising with colleagues, prior to commencing the meat of the research.

essential. Wilkinson et al. (2016) have developed four guiding principles of data management that have been widely adopted internationally that will facilitate such an empowered and credible research community. According to these principles, data have to be:

Findable, Accessible, Interoperable and *Reusable* (FAIR)

The implicit aim of these principles is to generate a research data ecosystem. In an ecosystem, the elements must interact, either as a result of human actions or automation processes, which requires both collaboration and agreements (e.g. standards). Interaction is an active process that can only be guided by principles, but these principles must be filled with life by practical guidelines, such as those outlined in Table 6, to enable the ecosystem to help researchers to meet the Challenge mission. We have extended these practical instructions to a more detailed set of expected practices to be adopted by Challenge participants in order to provide a mature data ecosystem that meets the Challenge mission. These expected practices are described Appendix 2.

Table 6 The FAIR Guiding Principles for scientific data management and stewardship

Data should be **Findable**	F1. (meta)data are assigned a globally unique and persistent identifier (DOI)
	F2. data are described with rich metadata
	F3. metadata clearly and explicitly include the identifier of the data it describes
	F4. (meta)data are registered or indexed in a searchable resource
Data should be **Accessible**	A1. (meta)data are retrievable by their identifier using a standardized communications protocol
	A1.1 the protocol is open, free, and universally implementable
	A1.2 the protocol allows for an authentication and authorization procedure, where necessary
	A2. metadata are accessible, even when the data are no longer available
Data should be **Interoperable**	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
	I2. (meta)data use vocabularies that follow FAIR principles
	I3. (meta)data include qualified references to other (meta)data
Data should be **Reusable**	R1. meta(data) are richly described with a plurality of accurate and relevant attributes
	R1.1. (meta)data are released with a clear and accessible data usage license
	R1.2. (meta)data are associated with detailed provenance
	R1.3. (meta)data meet domain-relevant community standards

Following these guidelines will enhance innovation and the impact of the research the Challenge will produce. Sharing data will also generate a high level of credibility for the Challenge as a whole by providing a high level of transparency.

7   A Roadmap to Achieve the Our Land and Water Data Ecosystem

Building a data ecosystem is a significant undertaking requiring effective governance, maturing individual and pan-institutional behaviours, and realising technical capacity. The level of thinking needed to address these issues first emerges at Level 3 of the maturity model and in particular Level 3 in governance. For this reason, we recommend adopting an iterative approach

designed to demonstrate early benefits, minimise risk as well as encourage further development and cultural change.

Specifically, we recommend an agile development style, in which:

- data ecosystem development is *iterative, incremental* and *evolutionary*;
- Challenge stakeholders receive *early* and *continuing increases in value and benefit*;
- each iteration *produces something specific that can be used*;
- there is *active user participation* (co-design, co-production, implementation), and
- each iteration *builds on earlier developments* in a way that responds to changes within the Challenge contexts and settings.

We also advocate a pragmatic approach balancing investment and operational costs against incremental benefits. The potential costs, operational risks, and time frames associated with implementation of the roadmap can be substantially reduced by leveraging solutions developed nationally and overseas, by studying the lessons others have learnt, and by adopting best practice.

The two most important factors that will dictate the success of the data ecosystem to support the Challenge are:

Management buy-in. It is essential that the data ecosystem is championed by senior management from the very beginning. Some resistance may be expected due to the cultural change in data practices required, and the project will not succeed unless management are committed and require their staff to be as well. Cultural shifts require change management, and "the CEO helps a transformation succeed by communicating its significance, modelling the desired changes, building a strong top team, and getting personally involved", (Aiken & Keller 2007).

Stakeholder engagement. The Challenge has recognised the importance of stakeholder engagement in the early stages, and this is no less important for implementation of the data ecosystem. The sustainability of the research infrastructure hinges on the active involvement of the Challenge community in building and using the infrastructure.

Table 7 describes the activities that should be completed at each phase of the roadmap. While the phases should proceed sequentially, this is not a requirement, and work could begin on the next phase in some areas while an earlier phase is still being achieved in others. In particular, thinking about the Governance requirements of Level 3 early, and putting governance in place at the start of each phase will set the right *tone* and *environment* for the success in the iterations within that phase. The phases correspond to the maturity levels, except that the first phase involves planning for the later phases, performing requirements analysis and identifying two key case studies to use during the second phase. We recommend selecting two case studies with different properties (e.g. different data types, sources, media). Appendix 5 contains some examples of projects that may be considered as potential case studies, extracted from the OL&W Research and Business Plans[40].

As shown in Table 2 above and Table 7 below, achievement of maturity Level 3 would result in data sharing across the Challenge using agreed standards, and would enable many of the within-Challenge objectives to be achieved. This maturity level would also provide discovery tools to

---

[40] Our Land And Water - Toitū Te Whenua, Toiora Te Wai National Science Challenge Revised Research And Business Plans September 2015

data across the Challenge, and a platform on which to build applications accessing and integrating multiple data sources.

We include Levels 4 and 5 in Table 7 for completeness. They would see the Challenge data ecosystem integrated with other national and international data ecosystems, and initiate a process of continuous improvement. While these may be worthwhile goals to consider in the long term, particularly in specific areas, e.g. data analytics and data interoperability, achievement of a data infrastructure at Level 3 is far from trivial, and would support the kinds of Challenge objectives described in the proposal.

Table 7 Activities in each Roadmap Phase

Phase	Data Management Strategy	Data Governance	Data Quality	Data Operations	Platforms and Infrastructure	Supporting Processing	Outcome
1	Establish leadership buy in. Conduct landscape review of existing data, requirements, standards in use and potential technologies across the whole Challenge, including input from all stakeholders. Select case studies with different properties and from different parts of the Challenge for Phase 2.						A detailed analysis and description of data requirements and the existing situation exists. Projects selected as case studies for phase 2. OL&W leadership fully supportive of maturing the data ecosystem to level 3 and providing necessary (financial) resources to achieve this. A data champion has been identified to drive the changes required.
2	Encourage the use of appropriate discipline standards for organising data and a common metadata discovery standard across the whole community. Explore options for partial automation of metadata population.	Establish a cross-Challenge governance group with a focus on achieving the second phase.	Select and adopt a shared approach to documenting data quality consistent with the selected metadata standard. Design and develop tools to assist users in the interpretation of data quality and lineage information.	Select & negotiate access to an existing repository instance (e.g. CKAN) to describe and reference (citation) data resources. Decide on backup, data archiving and publication policy. Define a stack of community standards to be used for data exchange. Design and implement mechanisms for communication around data resources in the ecosystem.	Acquire access to suitable hardware and software. Allocate responsibilities and resources for running the service.	Educate community in use and (organisational) implementation of data management principles and discipline standards through a set of workshops (part of this will be a maturity assessment). Train participants in the use of the repository. Promote uptake.	**Maturity Level 2:** A shared, managed, operational repository exists, populated with data from one or two representative projects (case studies). Policies and standards regarding data formats, QC and licences have been produced. Staff fully adopt the data management principles and discipline standards and understand their responsibilities within the data ecosystem. Data can be cited in publications using a permanent digital identifier.
3	Standardise processes for describing, publishing and managing data across the entire community. Require that all newly created community data sets are made available using these standards. Publicise the required processes and standards via a web site.	Require that data sharing occurs through the data ecosystem. Engage with all stakeholders to champion the role of data sharing as a core value of the Challenge community (Challenge leadership).	Develop a Challenge wide data quality strategy. Define quality metrics (e.g. nulls and zeros are carefully disambiguated), establish mechanisms for measuring them and ensure their uptake. Extend data quality interpretation tools to consider integrated data sets.	Design, implement and make available data discovery and access tools, considering the needs of different user groups. Explore options to develop tools to automate integration of data from multiple sources. Design and implement a set of minimally required vocabularies / ontologies (based on requirements analyses) to support semantic interoperability.	Define levels of service. Put measures in place to ensure that levels of service are met. Document the facilities that host the data. Move towards planning of an architecture that is optimal for data sharing across the Challenge using existing infrastructure and services, where those exist.	Train all Challenge participants in good data management practices at the level that is appropriate for their role in the Challenge through an ongoing set of well organised workshops.	**Maturity Level 3:** A shared, managed repository that is being used by all participants in the Challenge populated with data from all Challenge data related activities. Policies and standards regarding documenting data, data formats, QC and licences are being followed. Ways to semantically integrate data researched and implementation pathways defined for key projects. QC data is being shared with key data consumers in a controlled and managed way.

Phase	Data Management Strategy	Data Governance	Data Quality	Data Operations	Platforms and Infrastructure	Supporting Processing	Outcome
4	Improve data description and employ metadata quality metrics, use these as incentives to improve descriptions.	Conduct data audit to ascertain how well data policies are being used, and how effective they are.	Use feedback from data quality reporting to improve the reporting content and presentation, making it more readily consumable to users.	Measure the impact that changes to data sharing have had on the Challenge community, to see how much difference they are making in practice. Design and implement mechanisms to allow users to notify others of their planned data collection activities.	Move to a seamless data infrastructure where the Challenge is completely supported as a single virtual organisation, with no visible aspects of local data ownership or systems.	Use feedback from participants to improve the Data Management Planning process and other related data processes.	**Maturity Level 4:** A data ecosystem is operating in which all information is captured once, as close to source as possible, as close to real time as possible, available to direct data consumers, and the cost/benefit of the information captured is reviewed regularly. Data improvements are being made thanks to a process for getting feedback from data consumers.
5	Continuously improve descriptive metadata and metadata quality metrics to align with best practice.	Extend governance to connect with national, and international collaborations and directions.	Become involved in appropriate data quality standards communities to ensure alignment ongoing needs.	Measure impact that Challenge data is having outside of the Challenge community, and gather feedback on user experiences in using this data. Adjust systems as a result. Design and implement mechanisms to support collaboration around planned data collection activities.	Use the emerging data management and analysis needs of the Challenge to define an integrated computation and data platform to support future needs.	Lead reference implementations of appropriate data standards as they are developed.	**Maturity Level 5:** The Challenge data ecosystem is interoperating with data ecosystems operating in other Challenges, national ecosystems run by Government and businesses and internationally, and processes are in place for continuous improvement.

8   Looking to the Future: Creating an Environment to Enable the Data Ecosystem

The maturity model and the roadmap provide the framework in which the Challenge data ecosystem can evolve, as well as the activities that are necessary to achieve it. In the following sections we highlight a number of specific areas that are particularly important to the success of a mature data ecosystem in order to meet the objectives of the Challenge.

8.1   Changing the Data Management Culture

The data ecosystem described in this document relies on a more coordinated and thereby more mature approach to data management than is prevalent across the stakeholders at present. The maturity model also points towards the necessity of fostering ongoing cultural change and education for many Challenge participants. In a classic open data publishing model, there may be tenuous involvement of data consumers in the publishing process. In the Challenge, however, end-users, producers, and others participants in the data life-cycle are all brought together under the Challenge umbrella. This gives the Challenge a rare opportunity to fully weigh up the costs and benefits of all aspects of their data ecosystem irrespective of who benefits, who pays, and what necessary changes in behaviour can be anticipated, so that the Challenge's requirements are met throughout the data's life-cycle and across all stakeholders at appropriate phases of the Challenge's duration. The recommended iterative approach to the data ecosystem's evolution, coupled with appropriate governance and communication, should ensure that a balance is maintained between provider push and consumer pull for change, and that issues can be addressed while they are still small and manageable.

Currently, while some data are managed very well, a significant amount of data is stored in ad hoc, individual structures and shared in proprietary formats (e.g. Excel spreadsheets), resulting in lack of awareness of data that has already been collected, duplication of data collection efforts, and/or lack of access to data. In order to realise the open access vision of a data ecosystem, yet still ensure sensitive management of copyright, IP, privacy, as well as correct and valid application of data, changes to the culture of data collection are required. Some of the most significant areas of change are:

1. Data will be published in *open formats* to enable access and integration with other data.

2. When data are published, it will be described with a set of *standardised metadata, which includes data provenance and quality information,* that will enable other users to evaluate, interpret and use it, and with licensing, copyright and access controls.

3. When (meta)data are published, *agreed/standardized vocabularies* (for example, units of measurement, parameters, methods, quality codes, taxonomies, etc.) will be used that will enable other users to evaluate, interpret and use it.

4. When data collection is planned and data is being collected, data ecosystem users will *register the details of that collection activity* in the ecosystem to promote co-design and coordinated collection.

Experience shows that one of the major obstacles in the cultural change is the view that data belongs to "me" and that it is not treated as an asset. A pervasive cultural shift can only happen through on-going coordinated appropriate processes at both institutional and individual levels. This is an example of where operating at a Level of maturity of 4 in the data governance and supporting practice themes can pay dividends in leading the momentum going towards increasing maturity in the other themes. Overseas experience shows re-education is a vital part

of the mix to ensure that data ecosystem users are aware of the benefits of properly describing and publishing data in open formats, in accordance with data management best practice. Issues of data confidence and trust may also be addressed through an education programme, helping users to develop skills in evaluation and interpretation of data quality and in examination and understanding of data set lineage. Tools may be developed to help users interpret data quality, and mechanisms put in place in which well-trained staff take ownership of the data management process and support researchers in their effort to follow good data management practice (e.g. DataUp[41, 42] excel plugin that can be linked to a data repository such as CKAN[43]).

As noted in section 3.5, the Challenge proposes a large number of tools that rely on access to a wide range of data sets and easy integration of data and work processes. From a data management perspective, the implied goal is very ambitious, and will take time to fully mature. Governance and supporting processes operating at the level of maturity discussed above will ensure data consumers have realistic expectations, and remain engaged as the ecosystem evolves and matures.

The definition of clear roles and responsibilities in data governance across the Challenge is also important for effective implementation of the data ecosystem. For example, the distinctions between *data creator, data manager/custodian, data owner,* and *data steward*.

Through good governance the Challenge has the opportunity to establish a multi-party Memoranda of Understanding (MoU) for levels of service, and terms and conditions for Challenge data publishers and consumers, to ensure the overarching data integration aspirations are met. Custodians publishing data should be given the opportunity to negotiate terms and conditions of use of their data, and consumers and publishers should come to agreement on appropriate levels of service. Terms and conditions may include aspects such as: acknowledgement or attribution by data consumers; adherence to licensing and copyright conditions; adoption of any data security measures; and agreement on apportionment of future revenue and disclaimers. Selectable levels of service appropriate for real-time delivery vs. occasional file downloads that are sensitive to consumer's needs and the provider's costs and their capacity to deliver should also be negotiated. The Challenge might consider that having a single multi-party MoU providing access to the full diversity of Challenge data could be a market differentiator between those who have and have not yet signed up to the Challenge.

8.2   A Data Analytical Structure for the Ecosystem

At present there are four different approaches to infrastructures for data processing and analysis available to New Zealand scientists:

1. **Desktop systems** – laptops, desktops and workstations provided by individual institutions – primarily designed for everyday non-intensive data analysis tasks, and characteristic of Maturity level 1 though some scientists may have bigger workstations that they leave running for longer running more 'serious' analyses.

2. **Institutional servers** – typically providing shared storage for the desktops within the institution rather than science compute services.

---

[41] Strasser C, Kunze J, Abrams S, Cruse P 2014. DataUp: A tool to help researchers describe and share tabular data [version 2; referees: 2 approved]. F1000Research 2014, 3: 6 (doi: 10.12688/f1000research.3-6.v2)

[42] Latest DataUP version available through https://datastore.landcareresearch.co.nz/dataset/dataup (http://doi.org/10.7931/J26D5QXF)

[43] CKAN, http://ckan.org/about/, created by the Open Knowledge Foundation (http://okfn.org/)

3. **Cloud-hosted** – typically hosted off-shore and only rarely tailored specifically for NZ scientists. If they are tailored for a particular scientific analyses either individual 'Virtual Machines' (VMs) or collections of VMs configured as 'Virtual Labs' (VLs) are established to support standardised processing. These systems are often closely associated with particularly large datasets or datasets that are shared across a number of scientists in different institutions. VLs can be thought of as the digital equivalent of a traditional Telarc[44] registered physics or chemistry lab full of machines or equipment of known quality and staffed by skilled technicians who can be trusted to produce quality results.

4. **High Performance Computing** (HPC) – primarily provided in New Zealand by NeSI. NeSI is currently planning to start providing Virtual Machine hosting services to a small initial group of NeSI users that have some well-defined non-HPC analytical needs as an adjunct to their HPC needs. This is a first step in exploring the wider demand for such services in NZ.

There are three differentiators that distinguish the suitability of these infrastructures for particular purposes relevant to the Challenge (these are described below and through Figure 2).

Problem Scale: Desktops on their own are most appropriate for smaller scale analysis and smaller scale data volumes. Institutional, Cloud, and HPC support progressively larger and or more complex computational problems, but there is no reason why Cloud solutions should not be used for smaller analyses, and in fact they offer some advantages for small analyses because the environment is available to the scientist even when they are away from their desk, and they offer workstation capability even if the personal device is a small portable computer. Supporting growth in problem scale will be critical for the Challenge to meet their need to scale solutions across paddock to plate, irrigator to estuary, and farm to international reporting.

Collaboration: Desktop solutions supported by institutional servers are limited to individual and small institutional teams. Cloud is the best environment for multi-institutional virtual teams and widely shared functionality. Further, as discussed earlier in section 4.2, Cloud solutions allow collaboration to scale beyond a single institution and for data volumes to scale very significantly. Critically, Cloud solutions provide the Challenge with the opportunity to establish their own collaborative culture, independently of contributing institutions to support their goals of increasing collaborative capacity and delivering on co-design, co-innovation, co-development, and co-production. Finally HPC supports very large datasets, and/or large scale analysis but the increased specialisation typically means a less direct support for deep collaborative teams.

Best Practice Repeatability and Transparency: There is an emerging international principle that scientists and scientific publishers must provide access to the data, metadata and code used in their research and that not doing so is tantamount to scientific malpractice[45]. Cloud-based solutions provide the easiest pathway to support adherence to these principles.

We propose the best fit for the Challenge are cloud-hosted solutions, based on the three differentiators. The difficulty for scientists is that the culture of using shared systems, whether Cloud-based Virtual Labs or HPC can be significantly different from that of desktops and there may be resistance to changing. Those brought up on desktop systems will need specific training and support through the *transition*. Different ICT support structures are also needed to fully

---

[44] Telarc are New Zealand's certifier of quality, environmental and occupational health and safety management systems http://www.telarc.co.nz/

[45] http://www.science-international.org/

realise the benefits of collaborative environments so the change will be best made as a deliberate managed change. We recommend that the Challenge should establish a roadmap for adopting a set of cloud-based systems to support data analytics. A significant part of the early roadmap will be identifying partners, e.g. NeSI and/or other Challenges for provisioning the generic infrastructure and service support.
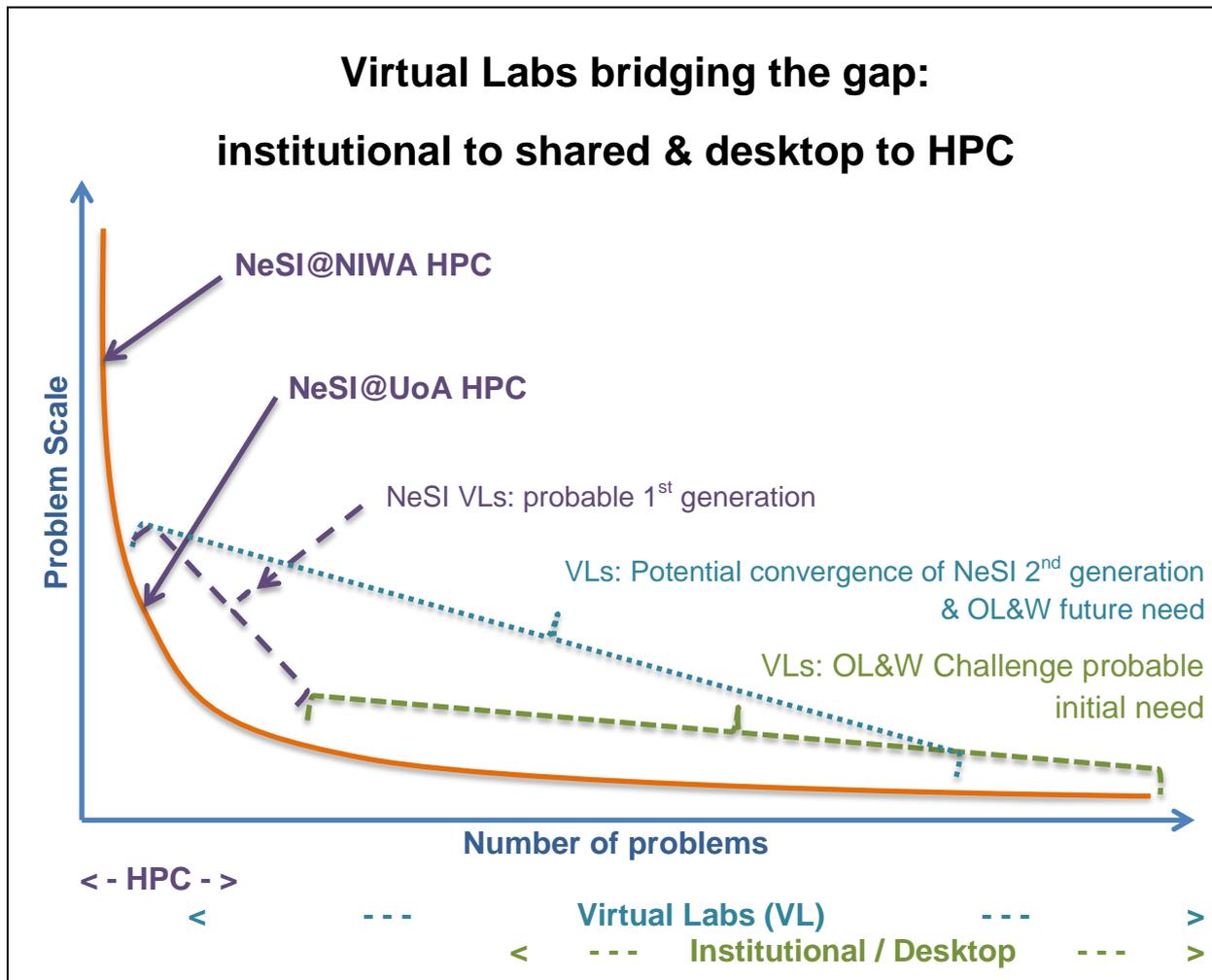

Figure 2 Relative scalability of different system solutions [46]

8.3   Interoperability and the Data Ecosystem

Interoperability is "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" (Geraci 1991). Interoperability also requires data integration; synthesizing data from different data sources – usually independent of each other – into a unified "view" according to a "global" schema (Lenzerini 2002). Thus successful data interoperability requires achieving data integration and data exchange as well as enabling effective use of the data that is being shared.

To develop interoperability from this initial state, the roadmap suggests selecting a couple of case studies with differing candidate datasets, providers, and consumers to develop their level

---

[46] Adapted from http://jetstream-cloud.org/files/Jetstream_XSEDE15-2015_jul_28.pptx

of maturity. Since some areas are already operating at a high level of maturity, one of the use cases could be designed to incrementally bring all the data ecosystem themes together to achieve effective end-to-end interoperability, albeit within a narrow scope. This would provide the challenge with an early fully operational system that could be incrementally built on. Critically, it would also allow the Challenge to start gaining experience in *end-to-end data governance.*

Land and water data are, of course, already interoperable. Data in a CSV file or Excel spreadsheet passed from one scientist to another is an act of interoperability, albeit a highly constrained one. Agreeing a set of standards data formats and interpretations of that data and then providing a facility for discovering and accessing that data is a first step to improving data interoperability (University of Auckland's data repository '*Figshare*' or Landcare Research's *'DataStore'* are candidate facilities). This would be the first step in establishing an architectural pattern for a Challenge data ecosystem that delivers a higher level of maturity for simple data interoperability and increased efficiency.

The same repositories can be used to publish data to data consumers. Alternatively, more advanced data publication tools can be used such as Landcare Research's *LRIS Portal* and NIWA's *Environmental Information Browser,* which have been customised for publishing land and water data respectively. As the Governance matures we would expect end-users within the Challenge to take a greater role as maturing data consumers, which will allow the Challenge to work towards aligning the needs of data publishers and consumers, incrementally lifting the game of the entire Challenge community. Right from the outset consumers using a data repository will start to benefit from the standardisation it encourages not only for search and discovery, but also for data formats and for more detailed metadata describing the data and its quality.

With increasing maturity over time providers and consumers will expect and demand more advanced levels of interoperability. The implementation of more advanced forms of data interoperability by others, such as the regional council's SOS-based water quality and quantity services, has the potential to make it easier for scientists and others working in the Challenge to access that data. Realising that potential will be easiest if the Challenge's governance has reached a level of maturity where it is routinely engaging directly with the governance across the regional councils to ensure the widest possible benefit to the sector.

Those providing tools to end users, whether businesses or those working in the Challenge, will desire that the data created in the Challenge are accessible via methods that are fully automated using well established protocols (e.g. web services based on Application Programme Interfaces (APIs)) so that they can pull data into the tools they provide for their end users (Tayyebi 2016). Increasing connectivity with the end-user systems will enable them to make better informed decisions and improve product quality or productivity using the latest versions of data created by Challenge scientists. At the same time, scientists in the Challenge will have access to real-time data and authoritative data sets provided by government and businesses using open APIs and access to an analytical infrastructure where data moves between systems using similar protocols.

The drive for greater data interoperability is already being seen in the sectors with which the Challenge will interact. The New Zealand Government's desire to increase the sharing and use of data will be underpinned by advanced forms of data interoperability. For example, the NRS Environmental Integrated Data Infrastructure (e-IDI) has the goal of making environmental data more discoverable, shareable, accessible, traceable, aggregable and interoperable by using semantic web technologies to provide federated access to a harmonised and collated view of environmental data.

The activities in the NRS and the regional councils are all good examples of the emergence of shifting levels of maturity from Levels 2 through 4. This indicates an increasing readiness for pan-sector governance and is a role the Challenge should consider engaging with, as it would be

an excellent way of earning the credibility needed for it to lead the environmental sector. Increasingly, agricultural and farm information systems will be able to exchange data with scientists through the use of web services. These web services will provide not only data but also computation and models for decision making and operational activities in agriculture e.g. land use planning and precision agriculture[47] (Han 2012; Řezník 2016).

In order to foster a successful data ecosystem the Challenge needs to come to a consensus on consistent standard data formats and interpretations of that data. The choice of standards should consider the requirements of data consumers and primary and secondary data use. The *National Environmental Monitoring Standards* (NEMS), the *Government Enterprise Architecture for New Zealand* (GEA-NZ) standards reference document, and the *NZGOAL* Guidance Note 2: File formats (August 2015) are good starting points regarding the data standards that could be used but international initiatives should be looked at also for more domain specific standards, e.g. the *FAO Agricultural Information Management Standards*,[48] and the *INSPIRE Data Specifications.*[49]

Standards for achieving more advanced levels of data interoperability exist for both APIs and data encoding (the *Geospatial Offices Spatial Data Infrastructure Cookbook* is a good starting point). However, internationally accepted domain specific standards are still maturing; there are data encoding standards for hydrology, e.g. WaterML (OGC 2015) and geoscience data, e.g. GeoSciML[50] but standards are still evolving for soil and agricultural data. There is an opportunity for the Challenge to participate in the international developments and further improvements of such standards.

Implementing mechanisms to support high levels of data interoperability is quite complicated requiring specialised knowledge representation (semantics, domain models) (NITDA 2014) and technologies (xml schemas, RDF, vocabularies and ontologies, feature catalogues, registries) (Tóth 2014). Given this, one of our key recommendations is the need for the Challenge to use and build on the existing infrastructures that are being developed by partner organisations or evolving in other sectors rather than building a Challenge-specific infrastructure. Examples of the former are the technologies being used to provide Landcare Research's *S-map Overseer WFS data supply service, NIWA's Sensor Station, River Water Quality Data and Lake Quality Indicator WFS data supply services* (feeding into LAWA), and the systems being developed in the MBIE funded Innovative Data Analysis project being run by Landcare Research. Examples of national-scale infrastructures to provide data interoperability include the data federation underpinning LAWA and the NRS e-IDI shared service. With the exception of NeSI there exist no National (Research) Data Services that the Challenge can make use of. (Compare this to, for example, Australia, the USA, Canada, Finland, the UK, where large-scale science programmes can benefit from access to nationally provided services for data storage, data preservation, data certification and related training and support.) The Challenge leadership should look for opportunities to co-invest in the development of components of the infrastructure or areas where there is mutual concern, e.g. addressing issues of privacy and security in order to advance interoperability.

There is a risk to the Challenge in relying on third parties to provide exchange data services in that the infrastructures/services mentioned may cease to exist because they are unsustainable or do not come become operational due to a lack of investment. Looking at opportunities to use

---

[47] http://www.cgiar.org/press-releases/cgiar-opens-agricultural-data-to-the-world-using-amazon-web-services/

[48] http://aims.fao.org/

[49] http://inspire.ec.europa.eu/index.cfm/pageid/2

[50] http://www.onegeology.org/technical_progress/geosciml.html

overseas research data infrastructures, for example the Australian Terrestrial Ecosystem Research Network (TERN), may be required to mitigate such risks.

8.4 Collaboration and the Data Ecosystem

Collaboration is a key theme of the Challenge[51], as recognised in the requirements for the data ecosystem (Appendix 1) and the expected practices for ecosystem participants (Appendix 2). The following examples illustrate the kinds of collaboration for which the maturity model and roadmap provide a framework.

First, participating in a common and effective governance model, which is able to demonstrate the ability to deliver concrete outcomes, is an excellent way of establishing a strong sense of collaboration across the Challenge sector.

Second, the ecosystem has the potential to enable virtual communication and collaboration with data resources in the ecosystem, enabling ecosystem participants to discuss data sets (for example, consideration of different options for a piece of land). This would allow kin groups who are spread around the globe to communicate and use data to support and inform that communication in a collective decision-making process. For example, tools such as *CKAN* provide options to create groups in the style of social media, and thereby encourage communication and collaboration. This mechanism is included in Phase 2 of the Roadmap (Section 7).

Third, the data ecosystem as described supports collaboration on data collection and use of data in an effort to minimise duplication, create data sets for multiple uses, and enable economies of scale to be realised. The data ecosystem infrastructure provides a foundation to support tools to enable data collectors to advertise their planned data collection activities so that others can become aware of them, in case they are planning a similar activity and thus allowing parties to work together to reduce effort. This would require achievement of a relatively high level of data ecosystem maturity, particularly in terms of the cultural change required to trigger such collaboration.

Finally, collaboration on the design of tools and scientific models is supported by the data access and awareness enabled by the described data ecosystem. In such an ecosystem, users are more aware of other people's activities and data resources, the methods they use to create them, and the tools provided by the repository (via metadata). This paves the way for better collaboration on tool design, enabling co-design, co-innovation, co-development, and co-production as per the Challenge objectives.

9   Research Required to Achieve the Vision

Parts of the vision for a relatively mature data ecosystem may be realised through the application of policies and best practice that have already been developed, and the use of existing tools, infrastructure and services. However, there are also a number of research questions that require attention in order for the data needs of the Challenge to be properly achieved in the way they are described in the Challenge proposal. These are summarised in Table 8.

---

[51] Our Land And Water - Toitū Te Whenua, Toiora Te Wai National Science Challenge Revised Research And Business Plans September 2015

Table 8 Summary of Research Questions required to support the Challenge's data needs

Category	No	Research Question / Topic	Research Field / Discipline Expertise Required	Themes / Programmes involved
Data–related Research Questions	1.1	What data is currently available, and what data is needed to achieve the Challenge mission? What gaps are there? Can these be prioritised around Challenge goals? This research question requires study of each of the themes and programmes as they develop to define a set of data themes and groups of themes (by priority), and where there are gaps between what is required and what is available, identification and design of approaches to data capture, potentially linking with Q1.2.	Geoinformatics, land and water thematic experts	All themes and programmes.
	1.2	How can data be collected in a cost effective manner? In order to achieve the Challenge mission, a large number of data sets are needed, many of which either don't exist or only partially exist with poor quality or at the wrong scale. This research question will investigate whether we can create new approaches to creating/synthesising data sets that are more cost effective, including new technologies for data collection; citizen science and crowdsourced data collection and intelligent methods to infer likely data patterns for semi-manual validation (e.g. through sampling, rule-based approaches). The suitability of approaches to meet the requirements of data sets for each theme and programme will be evaluated. This questions is also identified as a research priority in the Land Information New Zealand set of research priorities.[52]	Geoinformatics, land and water thematic experts	All themes and programmes.
	1.3	How can data be integrated semantically, so that different classification systems, different terminology and different world views are supported? This research area will investigate the development of dynamic data integration and data conflict resolution methods suitable for the data themes relevant to the Challenge, including the consideration of how to integrate qualitative, quantitative and oral data while maintaining the contribution of each one yet creating a synthesised picture. This work will connect and synergise with the Science for Technological Innovation Te Tāhū o te Pātaka Whakairinga Kōrero: Next Generation Indigenous Knowledge seed project (funded September 2016, and contact already established with the project leader). This research area is consistent with the research question 'How can developments in technologies, such as the semantic web, improve the usability of geospatial information?' This is also identified as a research priority in the Land Information New Zealand set of research priorities.[53]	Geoinformatics, semantics, land and water thematic experts	All themes and programmes. Particularly strong link with Theme 3: Mauri Whenua Ora

---

[52] New Zealand Geospatial Research and Development Priorities and Opportunities 2016 – 2020, Overview

[53] New Zealand Geospatial Research and Development Priorities and Opportunities 2016 – 2020, Overview

	1.4	How can qualitative, quantitative, oral, etc. data as well as data from new technologies (UAVs, sensors, RFIDs, GPS tracking etc.) and crowd-sourced/citizen science data be effectively combined and displayed, while respecting the need for privacy and cultural sensitivity, in a way that is meaningful and understandable by Challenge participants? This question is linked with Q1.3, but focusses on the visualisation and reporting of combined data to provide information about the sources from which it was created (whether oral, qualitative or quantitative), rather than issues of data conflict resolution and merging at a semantic level which is the focus of Q1.3. This area has not been researched in depth previously, and the Challenge requirements present an exciting Challenge to develop some new work in this area that can provide benefits for Challenge participants in carrying out the research programmes. This question is also identified as a research priority in the Land Information New Zealand set of research priorities.[54]	Geoinformatics, user interface design, land and water thematic experts, Matauranga Māori	Theme 3: Mauri Whenua Ora, Theme 2: Next Generation Systems
	1.5	How can data with different levels of quality (accuracy, etc.) be combined and the resulting quality be calculated, represented and effectively visualised in a way that is meaningful and understandable by Challenge participants? This includes data sets that have been directly collected as well as those that are integrated from multiple sources (how can the integrated quality be displayed?), and that are both raw and processed/converted into different formats. This is linked to Q1.3 & Q1.4 in being an aspect of data integration, but focussing on the data quality issues. This is identified as a particular issue for the Challenge in which data sets from multiple sources will be used to determine land suitability, identify potential enterprises in areas that are not currently ideally utilised, and calculate performance indicators. The research will require development of new methods of data quality visualisation and evaluation with Challenge users.	Geoinformatics, land and water thematic experts	All themes and programmes.
	1.6	Can the data ecosystem enable collaborative, data-driven science (Science 2.0) using the new data gathering technologies like UAVs and sensors, and thereby assist Challenge participants in meeting the Challenge objectives more effectively and provide additional benefits for the primary production sector? The research area would work with stakeholders to identify new opportunities and determine how and whether such an approach will lead to substantial benefits for the primary production sector.	Informatics and computer science, land and water thematic experts	Theme 3: The Collaboration Lab, Theme 2: Next Generation Systems

---

[54] New Zealand Geospatial Research and Development Priorities and Opportunities 2016 – 2020, Overview

Category	No	Research Question / Topic	Research Field / Discipline Expertise Required	Themes / Programmes involved
Ecosystem-related Research Questions	2.1	What ecosystem (technical) functionality must be provided in order to enable Challenge participants to effectively and efficiently access the data needed to ensure that NZ has the right enterprise in the right place at the right time to deliver the best outcome for property owners, the environment and NZ? What options are already available that could be used to reduce effort in achieving the data ecosystem that can allow the Challenge goals to be achieved? This includes requirements for ecosystem management and maintenance tools, tools for data publishing to the ecosystem, tools to support metadata entry, etc, and is closely linked with Q3.4, but more focused on specific software requirements to support ecosystem development and operation.	Geoinformatics, business and systems analysis.	All themes and programmes.
	2.2	How can interaction with the ecosystem respect/reflect and support effective communication between those with different worldviews and values (Pākehā, Māori), enabling Māori values to be brought to bear on the effective management of the land to maximise productivity while protecting water resources and supporting all Challenge participants in collaborating and communicating effectively? This research area will connect and synergise with the Science for Technological Innovation Te Tāhū o te Pātaka Whakairinga Kōrero: Next Generation Indigenous Knowledge seed project (funded September 2016, and contact already established with the project leader).	Geoinformatics, user interface design, Matauranga Māori	Theme 3: Mauri Whenua Ora
	2.3	How can interactions among data ecosystem users be modelled and tracked to determine whether the ecosystem is improving collaboration, citations, etc., to study what kinds of interactions are happening and what has been successful, and how can this information best be used to support the Challenge in promoting collaboration, co-design and co-development?	Informatics and computer science	Theme 3: The Collaboration Lab
	2.4	How can we design and develop discovery tools that enable Challenge participants to achieve the Challenge mission effectively and efficiently? What form should these discovery tools take, and how many tools should be added to discovery (e.g. integration, translation), to create the best balance between cost-effectiveness, data usability and flexibility to data consumers, and to allow Challenge participants to achieve the kinds of cross-institution, cross-disciplines, cross-scale and cross-geography knowledge envisaged by the Challenge in order to meet its objectives (for example, to find other enterprises that share commonalities in some areas, but not others, and that would not be found using traditional keyword discovery)?	Geoinformatics, UX designers, land and water thematic experts	All themes and programmes.

Category	No	Research Question / Topic	Research Field / Discipline Expertise Required	Themes / Programmes involved
Social and Institutional Research Questions *The 'research questions' in this category could be addressed directly by Challenge management as part of an agile project execution process*	3.1	What is a suitable manifesto/set of agreements/principles that can be agreed by ecosystem participants, to which all participants can and will commit, and that will support/enable the kinds of outcomes envisaged by the Challenge and that can trigger and support the kinds of changes to data management culture described in Section 8.1, in order to bring about the data access and sharing environment that will allow Challenge participants to answer their own research questions? This will involve investigation of existing principles adopted by data ecosystems around the world with similar objectives, and development of an approach to apply those principles in the OL&W context. This question is also identified as a research priority in the Land Information New Zealand set of research priorities.[55]	Information systems, organisational psychology	All themes and programmes.
	3.2	What support do participants need to implement the manifesto (above)? For example, data management guidelines, data management culture, data management policies, data management plans, executive support, training, education, audits. This will involve data collection with Challenge participants to determine their current position with respect to data management practices and analysis of the required support to move Challenge participants from their current position to one that would enable a mature data ecosystem to be realised, and thus enable Challenge participants to access the data they need to realise the objectives of the individual Challenge programmes.	Information systems, organisational psychology	All themes and programmes.
	3.3	How can an environment be created in which businesses are willing to share their data with other scientists, and vice versa? This research question addresses one of the common points of failure in data sharing efforts, and will aim to develop strategies, incentives and processes to encourage data sharing beyond its current sphere of influence. Qualitative research methods will be used to identify existing obstacles to data sharing.	Organisational psychology, business / management, geoinformatics, land and water thematic experts	All themes and programmes.
	3.4	Can existing efforts to create spatial data infrastructures by government departments, CRIs, Universities etc., be leveraged, shared and/or coordinated to support the data interoperability needs of the OL&W Challenge? This research area will investigate existing infrastructures, tools, mechanisms, cloud services, etc., to identify potential areas in which common elements can be shared or adopted, and identify components that are not currently covered and may need to be created in order for the Challenge programmes to meet their objectives.	Geoinformatics, organisational psychology	All themes and programmes.
	3.5	How can threats to data sovereignty that result from new agricultural technologies (e.g. data collection from sensors by farm machinery manufacturers, data from UAVs, human and animal GPS tracking) be evaluated, monitored and managed in New Zealand's best interests? This involves the investigation of approaches to ensure that New Zealand businesses and government do not lose control and access to the data they create using these new technologies.	Geospatial Science, law	Theme 2: Next Generation Systems

[55] New Zealand Geospatial Research and Development Priorities and Opportunities 2016 – 2020, Overview

	3.6	What methods can be developed to manage the protection of data privacy and sensitivity (whether cultural or commercial), including that of data created new technologies that provide large volumes of data to a high level of detail? This involves balancing protection of privacy and sensitivity with the benefits to be achieved from access to the data by Challenge participants, and innovative approaches will be explored to allow data to be accessed in a form that can be used to achieve the goals of the Challenge (for example, geographic anonymization, theme-specific aggregation).	Geoinformatics, law	Theme 3: Mauri Whenua Ora, Theme 2: Next Generation Systems

10   Starting on the journey: First Steps and Recommendations

'Information is the seed for an idea, and only grows when it's watered.'

Heinz V. Bergen

The Roadmap (Section 7) provides a high level plan that moves the current data management environment within Challenge participants to a more mature data ecosystem as described in the maturity model summarised in Section 4. The realisation of Phase 3 within the Roadmap might be expected to take several years (depending on the resources applied).

The strategies, policies, and actions outlined in this paper will ensure the maturing of the Challenge data ecosystem. To initiate this process, we have identified a set of first steps to begin the journey. Taking the first steps is often the biggest obstacle, and therefore we recommend the following actions be initiated in the first 6 months of the larger agenda described in the Roadmap.

First 6 months (providing a foundation for all subsequent activity)

- Engage with senior management (CIOs) to initiate promotion within participating institutions (senior management buy in) and identify and name data ecosystem champions.
- Endorse the vision and principles as a Challenge, including stakeholders and collaborators in the endorsement process (governance).
- Identify key priorities for first steps and research (all).
- Establish a cross Challenge data management governance group (governance).
- Establish the role of data stewards for themes including a descriptor for that role (governance).
- Create a collaborative space that allows forums to discuss data-related issues (platforms).

First 12 months (initiating Phase 1 and parts of phase 2)

- Initiate an extensive analysis of data requirements including quality criteria (quality).
- Initiate an audit of systems and services that are already operational within partner organisations that the Challenge could leverage (governance).
- Consider candidate case studies for Phase 2. Appendix 5 lists potential areas that may be suitable, as extracted from the OL&W Research and Business Plans[56]
- Define minimum metadata and quality standards and incorporate a data management plan into project proposals (quality, governance).

---

[56] Our Land And Water - Toitū Te Whenua, Toiora Te Wai National Science Challenge Revised Research And Business Plans September 2015

- Ensure appropriate financial structures are in place to support and sustain the development, implementation and operation of the data ecosystem (governance).

- Setup a data sharing portal (Landcare Research's DataStore (CKAN), Geonetwork, or the University of Auckland's FIGSHARE are possible candidates that should be considered) (platforms).

- Engage with related work programmes operating outside of the Challenge to socialise its identified tasks and to make linkages to / leverage from any related tasks within these where collaboration would be beneficial and feasible (governance).

11 References

Aiken CB, Keller SP Feb 2007. The CEO's role in leading transformation. http://www.mckinsey.com/business-functions/organization/our-insights/the-ceos-role-in-leading-transformation

Geraci A 1991. IEEE standard computer dictionary: compilation of ieee standard computer glossaries. Piscataway, NJ, IEEE Press.

Heimstädt M, Saunderson F 2014. Toddler to teen: growth of an open data ecosystem – a longitudinal analysis of open data developments in the UK. JeDEM - eJournal of eDemocracy and Open Government 6(2): 123–135.

Laney D 2001. 3D Data management: controlling data volume, velocity, and variety. Stamford, CT, META Group.

Lenzerini M 2002. Data integration: a theoretical perspective. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS '02). New York, ACM. Pp. 233–246.

Liaskos S, McIlraith S, Sohrabi S, Mylopoulos J 2010. Integrating preferences into goal models for requirements engineering. 2010 18th IEEE International Requirements Engineering Conference. Pp. 135–144.

Lukoianova T, Rubin VL 2014. Veracity roadmap: Is big data objective, truthful and credible? Advances in Classification Research Online 24(1): 4–15.

Han W, Yang Z, Di L, Mueller R 2012. CropScape: a Web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support. Computers and Electronics in Agriculture 84: 111–123.

Marinos A, Briscoe G 2009. Community Cloud Computing. Proceedings First International Conference, CloudCom 2009, Beijing, China, 1–4 December 2009. Lecture Notes in Computer Science 5931. Berlin, Springer. Pp. 472–484.

Maceachren AM, Swingley D, Roth RE, O'Brien J, Li B, Gahegan M 2012. Visual semiotics & uncertainty visualization: an empirical study. IEEE Transactions on Visualization and Computer Graphics 18(12). Pp. 2496–2505.

NITDA 2014. Standards for data interoperability – achieving data interoperability through standards. National Information Technology Development Agency (NITDA), Version 1.2, June 2014. http://nitdademo.azurewebsites.net/wp-content/uploads/2016/06/Standards-for-Data-Interoperability-Final-Draft.pdf

OGC 2015. WaterML2.0: Part 2 – ratings, gaugings and sections. Doc # 15-018r2 http://docs.opengeospatial.org/is/15-018r2/15-018r2.html

Pollock R 2011. Building the (Open) Data Ecosystem. Open Knowledge Foundation Blog. Retrieved 10 July 2013, from http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/

Řezník T, Lukas V, Charvát K, Charvát Jr, K, Horáková Š, Křivánek Z, Herman L 2016. Monitoring of in-field variability for site specific crop management through open geospatial information. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLI-B8, XXIII ISPRS Congress, 12–19 July 2016, Prague, Czech Republic. Pp 1023–1028.

Roche DG, Kruuk LEB, Lanfear R, Binning SA 2015. Public data archiving in ecology and evolution: how well are we doing? PLOS Biol. 13, e1002295.

Stock K, Karasova V, Robertson A, Roger G, Small M, Bishr M, Ortmann J, Stojanovic T, Reitsma F, Korczynski L, Brodaric B, Gardner Z 2013. Finding science with science: evaluating the use of scientific knowledge and semantics to enhance discovery of scientific resources. Transactions in GIS 17(4): 612–639.

Tóth K, Portele C, Illert A, Lutz M, de Lima MN 2014. A conceptual model for developing interoperability specifications in spatial data infrastructures, European Commission Joint Research Centre. doi:10.2788/21003

Tayyebi A, Meehan TD, Dischler J, Radloff G, Ferris M, Gratton C 2016. SmartScape™: A web-based decision support system for assessing the tradeoffs among multiple ecosystem services under crop-change scenarios, Computers and Electronics in Agriculture 121: 108–121.

Vanschoren, J,Bischl B, Hutter F, Seba M, Kegl, B, Schmid M, et al.2015. Towards a data science collaboratory. Advances in Intelligent Data Analysis XIV. Proceedings 14th International Symposium, IDA 2015, Saint Etienne, France, 22–24 October 2015. New York, Springer. Pp. xix–xxi.

Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3. Article number: 160018. URL: http://www.nature.com/articles/sdata201618.

Appendix 1: Key (known) Requirements and Sources

The following requirements have been identified from the OL&W Challenge documents (e.g. research and business plans, reports from stakeholder workshops), the surveys conducted by the project team and requirements inferred from those documents based on the experience and knowledge of the expert panel.

Number	Description	Source
1	Data sets may be geographically, temporally or thematically partial, either due to practicalities of data collection or privacy. For example, data on details of enterprise profitability is likely to be sensitive, so may only be accessible by certain users.	Panel
2	Project management tools across the NSC.	OLW Challenge Leaders Questionnaire
3	Strategies for data anonymization.	OLW Challenge Leaders Questionnaire
4	A collaborative document sharing space with some level of version control.	OLW Challenge Leaders Questionnaire
5	Providing an infrastructure that supports a transparent environment, including information that supported decision making processes and lead to particular decisions, information about sources (lineage) of integrated data and visualisation of accuracy and data integrity.	OLW Challenge Leaders Questionnaire Challenge Proposal
6	Monitoring and indicator reporting of the managed data (metrics), including who is downloading data, how much it is being used, etc.	OLW Challenge Leaders Questionnaire
7	Requirement for co-funders to agree to the data ecosystem data sharing policies (e.g. provide OL&W contract templates).	OLW Challenge Leaders Questionnaire
8	Management infrastructure across NSCs for RfP preps, proposal assessment, contracting etc.	OLW Challenge Leaders Questionnaire
9	Interoperability of data with existing and future tools, as the creation of tools is an important part of the Challenge. Tools may include environmental and social models, as well as algorithms for the definition of indicators based on a collection of data sets.	Panel
10	Support for collaborative decision making, including stakeholders residing internationally (for example, some Māori kingroups).	Challenge Proposal
11	Support for the Māori world view, including recognition of individual values and knowledge.	Challenge Proposal

12	Accommodation of official, volunteered (crowd-sourced), citizen science and sensor data sources.	Challenge Proposal
13	Dynamic (real time) data integration to bring together multiple data sets with different formats, semantics, etc.	Challenge Proposal
14	Intelligent search tools to identify similar situations in different locations, scales and times (e.g. who else is doing something similar at a different scale? who else has tried this approach in a different situation?).	Challenge Proposal
15	A multi-geographical and multi-temporal resolution (national, regional, sub-regional, catchment, enterprise level) view of New Zealand to allow different scenarios to be explored.	Challenge Proposal, Business Plan p4-5, OLW Challenge leaders questionnaire
16	Support for qualitative, quantitative and oral data, as well as data from new technologies (UAVs, sensors, RFIDs, GPS tracking etc.).	Challenge Proposal
17	Ease of use combined with flexibility/richness for a wide range of user levels.	Challenge Proposal
18	Publicly accessible data, taking into account existing IP and contract constraints.	Challenge Proposal
19	Coordinated with data approaches used by other Challenges.	Panel
20	Technical language needs to be translated into something more meaningful to users.	Panel
21	Atom level data (e.g. paddock or sub paddock) in order to allow data to be aggregated for higher level access.	Panel
22	Documentation of the origins, accuracy and precision of the data, along with other quality related attributes (including the level of confidence/uncertainty), and quality assurance mechanisms applied.	Panel
23	Easy discovery of data sets that are ready, fit for use and presented in a meaningful way.	Panel
24	Need to provide data, information and tools that can be used by next and end users with different abilities and knowledge.	OLW Challenge Leaders Questionnaire
25	The data ecosystem needs to interoperate with other ecosystems operating at different scales or in different problem spaces. e.g. other challenges, EMaR data federation/LAWA, eIDI etc.	Panel
26	Important that data and tools involves end users – co-design, co-development, co-implementation.	OL&W Challenge plan

27	The Challenge needs to be the "one source of truth".	OLW Challenge Leaders Questionnaire
28	The cultural context must be taken into account when designing the data sharing policy for data sets that may have sensitivities.	Panel
29	Where data is to be used from external sources, Challenge partners agree on a single authoritative source for a particular data set.	Social Architecture for NEII
30	Provide mechanisms to ensure secure access by appropriate individuals or groups where data is not open access.	Panel
31	Ensure that data in the data ecosystem is easily discoverable by a range of different users with different perspectives, world views and cultural frameworks (including Māori).	Panel
32	Provide mechanisms to combine, integrate and visualise data from multiple sources.	Panel
33	Provide mechanisms for easy import and export of existing data.	Panel
34	Design for the future by considering future system migration as the data ecosystem and user needs evolve.	Panel
35	Enable distributed access to data in existing mature, robust systems through open standards, as well as a repository for data that does not currently exist within a suitable system environment.	Panel
36	Provide mechanisms to ensure secure access by appropriate individuals or groups where data is not open access.	Panel
37	Ensure that data in the data ecosystem is easily discoverable by a range of different users with different perspectives, world views and cultural frameworks (including Māori).	Panel
38	Provide guidance on availability, suitability and currency of tools for particular purposes (e.g. https://researchit.cer.auckland.ac.nz)	Stakeholders Questionnaire
39	Accommodate data collected using new technologies (UAVs, sensors, trackers, etc.) in the data ecosystem.	Stakeholders Questionnaire
40	The ecosystem needs to provide mechanisms for interoperability with land and farm management systems.	Panel

Appendix 2: Expected Practices

The following provides a detailed specification of the practices that would be expected from users of a mature data ecosystem (level 3+) at each stage in the data lifecycle, in order to meet the Challenge objectives.

Lifecycle Stage	Expected Practices
1. Data Collection Planning	1.1 Work with users, stakeholders and other Challenge partners to analyse the requirements for a new data set, through co-design, co-innovation, co-development and co-production. 1.2 Identify data set requirements on the basis of demand from specified research projects. 1.3 Summarise key details describing the data collection activity that is planned or underway (prior to data publication) and publish in the data ecosystem. This enables other data ecosystem users to identify data collection activities that may be relevant to them and potentially reduces duplication of data collection. 1.4 Identify a data owner and data custodian for the data set. Each data set should have a custodian and an owner throughout its life. The same person or organisational unit may take both roles. 1.5 Identify the owner of the intellectual property that will be present in the data. 1.6 Investigate existing data sources that may be suitable for reuse, cleaning, restructuring or value adding and either reduce or eliminate the need for new data collection. 1.7 Ensure scope of ethical consent enables future re-use of data.
2. Data Creation/ Collation	2.1 Build user and stakeholder feedback mechanisms into the data collection process so that fitness for purpose can be assessed and refined. 2.2 Select appropriate data types and file formats. 2.3 Plan for data storage by estimating final data volume. 2.4 Use open, machine readable file formats. 2.5 Use logical file names and data organisation strategies. 2.6 Build in data checking and quality assurance during data collection. 2.7 Keep data in raw format whenever possible to facilitate future re-analysis and analytical reproducibility. 2.8 Use existing standard classification and coding schemes for data where possible. 2.9 Create shared data representations (data models/ontologies) or mapping mechanisms across information communities for commonly used data sets. 2.10 Engage with and reuse standards development activities outside the Challenge (e.g. National Environmental Monitoring Standards). 2.11 Select stable standards with widespread adoption (e.g. well

	established standards of the OGC, ISO, W3C) rather than creating new ones. 2.12 Fully document the provenance of the data.
3. Data Storage	3.1 Preserve raw data from instruments and associated metadata where it may be useful later on. 3.2 Have a systematic backup scheme. 3.3 Select a storage method based on size and nature of data, costs of storage, how the data will be used, time to transfer, who needs access, and privacy concerns. 3.4 Clean or reduce raw data as appropriate. 3.5 Store data securely.
4. Data Publishing	4.1 Give data a permanent, unique identifier and publish in the data ecosystem (obeying any data restrictions or privacy concerns). 4.2 Create discovery metadata along with documentation or links to provide the context needed to interpret the data and to ensure the authenticity, reliability and integrity of the data. 4.3 Specify the publisher's knowledge of the level of data quality (accuracy, correctness, completeness) as a key element of the metadata so that data consumers can be clearly aware of the limitations of the data. 4.4 Ensure that metadata descriptions support long term storage and reuse of the data, to support the Challenge's goals of temporal data analysis. 4.5 Cite and link data in publications. Develop mechanisms to support dynamic data citation where subsets of data sets are published. 4.6 Publish data at the highest possible level of granularity (without compromising privacy). 4.7 If privacy is a concern, publish data in anonymised form (e.g. through removal of identification, aggregation, geographic distortion). 4.8 Create linkages and data reuse beyond and the Challenge by publishing metadata in public repositories (e.g. data.govt.nz)
5. Copyright and Licensing	5.1 Establish and publish copyright of data. 5.2 Use the NZGOAL framework for copyright and licensing. 5.3 Publish data using a license that allows open access to the public (e.g. Creative Commons) except where this would compromise privacy issues. 5.4 Enable re-use by choosing open formats, structures, classification, coding schemes and terminology to facilitate easy combination with other data.
6. Data Analysis and Modelling	6.1 Decide whether data produced during intermediate and final steps in the analysis should be persisted or regenerated on demand.

	6.2 Track versions of data and any processes used to generate them.
	6.3 Record descriptive and technical metadata that will later be published with final data for reuse.
	6.4 Ensure data is suitable for reuse by enabling easy interpretation by third parties.
	6.5 Develop software tools that will result in data sets that are suitable for reuse using robust software engineering practices.
	6.6 Use tools that are fit for purpose, rather than adopting tools for another purpose that are not a good fit.
7. Reuse and Value Adding to Data	7.1 Ensure that consent or legal rights to reuse data have been obtained from custodian.
	7.2 Identify and raise any issues of ownership of the data product that result from the reuse or value adding as early as possible in order to deal with them efficiently.
	7.3 Identify and raise any potential privacy issues that may arise from the reuse or value adding as early as possible in order to deal with them efficiently.
	7.4 As far as practicable, use the best quality data available of known provenance.
	7.5 Promote good quality data collection and documentation by data suppliers.
8. Reporting	8.1 Ensure compliance with all funder, government, and institutional policies on how data will be managed and shared.
	8.2 Consider industry-specific and government reporting requirements for specific data themes.
9. Archiving and Long-term Preservation	9.1 Invest in data curation early in the project design.
	9.2 Consider preservation and curation issues, how and where the data will be stored or accessed, and the need for active migration of data to different formats or media through time.
	9.3 Preserve data for long term reuse to support temporal analysis.
	9.4 Consider when and how data should be deleted or destroyed.
10. Data Access and Discovery.	10.1 Data should be made accessible to other Challenge partners and other data consumer unless there is a compelling reason for access to be restricted (for example, due to legitimate privacy concerns that cannot be resolved by data anonymisation).
	10.2 Relevant metadata (particularly quality information) will be provided and easily accessible alongside the data itself.

	10.3 Data will be accessible through the data ecosystem using open standards.
	10.4 All Challenge data will be accessible through a single point of access and discovery.
11. Data Ecosystem Design	11.1 Include Challenge partners, data consumers in the process of design and governance of the ecosystem.
	11.2 Design the ecosystem with a view to potential interoperation with other data ecosystems.
12. Data Feedback	12.1 Where relevant and practicable, provide feedback to the data custodian on issues of data quality.

Sources:

- Centre for eResearch Research, Research Data Management.

- GNS Data and Collections Management Policy.

- Massey University Data Management Policy.

- Lincoln Hub Research Data Management Policy Template.

- New Zealand Data and Information Management Principles: Cabinet Minute CAB Min(11) 29/12.

- Research Data Management Framework Report by the Council of New Zealand University. Librarians (CONZUL) Working Group for New Zealand Vice-Chancellors' Committee (Feb 2016).

Appendix 3: A Selection of Relevant National Initiatives

- Bio Data Services Stack Project https://teamwork.niwa.co.nz/display/NZBSS - Completed national research project on building a demonstration on federating biological observation.

- Data Futures Partnership http://datafutures.co.nz/ - National (ly funded) initiative to help lead the development of a data-use system that will create value for all New Zealanders by making better use of our data. Includes a range of 'catalyst' projects to improve data availability.

- Environmental Monitoring and Reporting (EMaR). EMaR explores what needs to be done to standardise of methods and sharing of data collection, management and exchange protocols to allow national scale interpretation of regional data.

- Geospatial strategy for a spatial data infrastructure (SDI) http://www.linz.govt.nz/about-linz/our-location-strategy/geospatial-strategy-for-spatial-data-infrastructure - SDI is the technology, policies, standards, and human resources necessary to acquire, process, store, distribute and improve the usability of geospatial data. Essentially, an SDI is the full framework supporting the use of geospatial information.

- Government ICT strategy https://www.ict.govt.nz/strategy-and-action-plan/strategy/ - Led by Department of Internal Affairs (DIA) GCIO; works towards "a coherent, NZ information ecosystem". It talks about exploiting emerging technologies, unlocking the value of information, leveraging agency transformations (building on existing components of the ICT ecosystem) and partnering with the private sector. The related implementation plan includes: building and enabling data environment and policy settings; support development of data standards; improved data analyses; accelerating data release.

- Initiatives on building / providing institutional data infrastructures consistent with international standards are underway in Crown Research Institutes, Universities, and Government Agencies, as well as through industry (especially geospatial industries).

- Land, Air, Water Aotearoa (LAWA) https://www.lawa.org.nz/ - Regional Council led projects to improve data interoperability and data federation through an evolving data portal.

- Lincoln Hub Data and Information Architecture Project - Provides a set of policy elements and guidelines to ensure that data generated as a result of research are appropriately managed, stored, and made available. Each hub partner should adopt these common elements, and tailor them to their specific organisation's needs.

- LINZ Our Location Strategy http://www.linz.govt.nz/about-linz/our-location-strategy - LINZ is working with central and local government, and the private sector to improve the value and accessibility of the location information.

- Ministry of Business, Innovation and Employment MBIE requires data generated as part of its funded research project (including Science Challenges) to be made available consistent to the Open Government principles.

- National Environmental Monitoring Standards (NEMS) https://www.lawa.org.nz/learn/factsheets/%28nems%29-national-environmental-monitoring-standards/ - Programme championed by MfE for establishing a set of national standards for data collection and exchange.

- National Environmental Reporting http://www.mfe.govt.nz/more/environmental-reporting - a range of initiatives, mandated through the National Environmental

Reporting Act 2015, led my Ministry for Environment (MfE) and Statistics New Zealand to improve national environmental reporting.

- National Resource Sector (NRS) Information Programme - lead by Land Information New Zealand to coordinate initiatives in data sharing in the Natural Resources Sector agencies.

- NeSI is a collaboration to purchase and operate High Performance Computing (HPC) infrastructure on behalf of NZ Science. Since it is not a legal entity, it can neither own equipment nor hire staff, so it has to do everything through its investors. In the first 4yr MBIE contract NeSI's mode of operation was essentially a loose collaboration between the collaborating partners, with all the services coming from behind the contributing institution's respective firewalls, which put significant strain on its operations and engagement. As part of the negotiation process for the second tranche of funding NeSI adopted a much more open and mature approach to its operations, service offerings, and engagement. A small but significant part of this larger change includes use of both zoom.us and teamwork.com in preference to the platforms being operated by any of the respective partners.

- New Zealand Data and Information Management Principles https://www.ict.govt.nz/guidance-and-resources/open-government/new-zealand-data-and-information-management-principles/ - Principles for Managing Data and Information held by the New Zealand Government, approved by Cabinet on 8 August 2011 (CAB Min (11) 29/12 refers)[1].

- NZGOAL https://www.ict.govt.nz/guidance-and-resources/open-government/new-zealand-government-open-access-and-licensing-nzgoal-framework/ - guidance for agencies to follow when releasing copyright works and non-copyright material for re-use.

- Open Government Information and Data Programme. https://www.ict.govt.nz/programmes-and-initiatives/open-and-transparent-government/open-government-information-and-data-work-programm/ https://www.ict.govt.nz/guidance-and-resources/open-government/ Based on the Declaration on Open and Transparent Government 2011, a programme led by Land Information New Zealand, providing guidance and advice to support agencies manage and release their data and information for re-use.

- Proposed National Research Data Programme (NRDP), developed through eResearch 2020 http://www.eresearch2020.org.nz/ - eResearch 2020 is a national programme that develops strategy with thought leaders across the research sector and aims to assemble a comprehensive vision of researcher needs (eResearch 2020 is part of New Zealand eScience Infrastructure (NeSI) https://www.nesi.org.nz/)

- Proposed New Zealand Environmental Integrated Data Infrastructure (e-IDI) - Water Services Pilot Project – MfE (through EMaR) & LINZ (through NRS) Better Public Services proposal to develop a national data infrastructure, especially to support national environmental reporting (with Horizons Regional Council, Landcare Research, NIWA).

Appendix 4: A Selection of Relevant International Initiatives

- AgGateway http://www.aggateway.org/ - AgGateway is a non-profit consortium of businesses serving the agriculture industry, with the aim of promoting, enabling and expanding eBusiness in agriculture.

- Digital Curation Centre, University of Edinburgh http://www.dcc.ac.uk/ - The Digital Curation Centre (DCC) is an internationally-recognised centre of expertise in digital curation with a focus on building capability and skills for research data management.

- EarthCube: The NSF's Geoscience team noticed that as the science they funded became more complex, recipients of funding spent an increasing proportion of their funds on getting started, finding and understanding diverse datasets, and negotiating access to data. Consequently upwards of 65% of their funds were consumed before they embarked on the science problem. This was particularly prevalent in the many institutions with small, highly specialised research teams – who collectively received 80% of NSF Geoscience's funding. Realising this was neither sustainable nor efficient they partnered with the NSF Cyberinfrastructure and engaged the entire Geoscience research community in a new collaborative initiative called EarthCube, into which they threw the whole funding negotiation, governance, and development of solutions to the community – and said to them 'This is our collective problem. Join us on a journey to discover what the solution is, and we will commit to it'. The whole project exists on an open wiki that anybody from any country can join. The early week-long workshops involving hundreds of people were run entirely using google docs and google calendar with every breakout session having an associated virtual meeting and all meeting note taking done in google docs that all participants could see and contribute to. As part of the early work on Earthcube Governance, a white paper was written that explores the different benefits that participants in a mature collaboration receive from donating their sometimes considerable time and energy to EarthCube. NSF also acknowledged that the style of funding contract that the US Government mandated was potentially part of the problem, and they asked the community to advise them on that to assist them in creating change in that area too, just to underscore the breadth and depth of change they were open to.

- eFoodChain http://www.efoodchain.eu/ - This demonstration action is expected to identify and remove the existing technical and organisational barriers, in order to increase the efficiency of the agro-food supply chain; It includes principles and rules for interoperability among business processes and data exchange models to allow for seamless information and data flows underpinning transactions along the food supply chain.

- FAO Agricultural Information Management Standards - http://aims.fao.org/openagris - A collaborative network of more than 150 institutions from 65 countries, maintained by FAO of the UN, promoting free access to agricultural information.

- Global Biodiversity Information Facility GBIF http://www.gbif.org/ (and Darwin Core http://rs.tdwg.org/dwc/) - International open data infrastructure, funded by governments, working by encouraging and helping institutions to publish data according to common standards, and providing a global access portal.

- Global Earth Observation System of Systems (GEOSS) http://www.earthobservations.org/geoss.php - GEOSS is a set of coordinated, independent Earth observation, information and processing systems that interact and provide access to diverse information for a broad range of users in both public and private sectors. (see especially the GEOSS Architecture Implementation Pilot (AIP) http://www.earthobservations.org/documents/cfp/201501_geoss_cfp_aip8_architecture.

pdf and GCI User Requirements
http://www.earthobservations.org/documents/portal/20160419_towards_gci_user_requirements.pdf)

- Global Open Data for Agriculture and Nutrition (GODAN) http://www.godan.info/ - GODAN supports the proactive sharing of open data to make information about agriculture and nutrition available, accessible and usable to deal with the urgent challenge of ensuring world food security. A number of partners are collaborating to work to the data infrastructures needed for the global agricultural sector.

- ICT-AGRI – ICT and robotics for sustainable agriculture http://www.ict-agri.eu/ - The overall goal of ICT-AGRI is to strengthen the European research within the diverse area of precision farming and develop a common European research agenda concerning ICT and robotics in agriculture.

- Infrastructure for Spatial Information in the European Community (INSPIRE) http://inspire.ec.europa.eu/ – The INSPIRE directive aims to create a European Union (EU) spatial data infrastructure. This will enable the sharing of environmental spatial information among public sector organisations and better facilitate public access to spatial information across Europe.

- Jetstream (NSF funded) – a cloud-based, on-demand system for 24/7 access to computing and data analysis tools that are integrated into the national research ecosystem. Jetstream will provide the following core capabilities: Jetstream will be attractive to communities who have not been users of traditional HPC systems, but who would benefit from advanced computational capabilities. http://jetstream-cloud.org/

- National Data Services – Given the increased value of research data as a national research asset, many countries are establishing national services to provide data services, coordination and support for research data storage, management, publication, reuse and sharing. Countries running or establishing National Data Services include Australia, the USA, Canada, UK and Finland. http://www.nationaldataservice.org/

- National Environmental Information Infrastructure (NEII) http://www.neii.gov.au/ - The Australian National Environmental Information Infrastructure (NEII) is an information platform designed to improve, discovery, access and re-use of nationally significant environmental data. It proposes a network of standards-based IT components, developed in collaboration with a number of technical and strategic partners.

- NISO, the National Information Standards Organization – data management http://www.niso.org/ - is a non-profit association accredited by the American National Standards Institute (ANSI), identifies, develops, maintains, and publishes technical standards to manage information in today's continually changing digital environment. NISO standards apply to both traditional and new technologies and to information across the whole lifecycle, from creation through documentation, use, repurposing, storage, metadata, and preservation.

- Open Agricultural Data Alliance http://openag.io/ - Aims at building an open source framework and a community of commercial vendors, farmers, academics, and developers upon which the emerging agricultural data market can rapidly grow.

- Open Geospatial Consortium (OGC) http://www.opengeospatial.org/ - The Open Geospatial Consortium (OGC) is an international industry consortium of over 525 companies, government agencies and universities participating in a consensus process to develop publicly available interface standards.

- Open Water Data Initiative – initiated by the US Department of the Interior's Assistant Secretary for Water and Science http://acwi.gov/spatial/owdi/ - Purpose is to improve water information for decision making about natural resources management and

environmental protection, through the U.S. Geological Survey (USGS), as the lead agency. Other federal organizations that fund, collect, or use water resources information work together with the USGS to implement program recommendations.

- Organisation for Economic Co-operation and Development (OECD) http://www.oecd.org/ - The aim is to provide information for data users and data providers on the OECD's direct and indirect data collection and statistics production practices. (see in particular OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information https://www.oecd.org/sti/ieconomy/40826024.pdf).

- Research Data Alliance Interest Group on Agricultural Data IGAD https://rd-alliance.org/groups/agriculture-data-interest-group-igad.html - IGAD is a domain-oriented group working on all issues related to global agriculture data. It represents stakeholders in managing data for agricultural research and innovation, including producing, aggregating and consuming data. IGAD promotes good practices in research with regard data sharing policies, data management plans, and data interoperability, and it is a forum for sharing experience and providing visibility to research and work in agricultural data.

- SageMathCloud (NSF funded) supports collaborative computational mathematics using any or all of the standard open source computational tools scientists use such as Python, R, SageMath, Julia, GAP, and Octave for instance. https://cloud.sagemath.com/ SageMathCloud is available as a ready to use hosted service, but it is also available as source code for installing on other clouds e.g. JetStream or Amazon or in NZ a NeSI VM for instance etc.

- United Nations Environment Programme (UNEP) GEMS/Water Global Data Centre http://www.bafg.de/GRDC/EN/05_cllbrtn/53_prtnrcntrs/gemswater.html - GEMS Water provides the world community with sound data on fresh water quality to support scientific assessments and decision-making.

- World Meteorological Organisation (WMO) Information System (WIS) http://www.wmo.int/pages/prog/www/WIS/ - Coordinated global infrastructure responsible for the telecommunications and data management functions for managing and moving weather, climate and water information in the 21st century.

Appendix 5: Example Use Cases

For full development of a mature data ecosystem, the definition of a set of example use cases is suggested. Potential areas in which use cases could be developed include the following, extracted from the Challenge research and business plan[2].

1. How far can I reduce my contaminant loss and still be profitable? Is this enough to meet the NPS-FM? If this loss is too much, what other land uses can I alter and where is this most profitable?

2. Do time lags exist between land use, the loss of phosphorous to groundwater and the re-emergence of phosphorous and in surface water (as for nitrates)? p.11

3. Can we identify the impacts (opportunities/threats) of enhanced removal of nitrates from groundwater denitrification zones? p.11

4. Development of a groundwater classification system that could align land use capability with the suitability and sensitivity of aquifers to changing land use.

5. The linking of long-term databases of the characteristics, state and trends in land and water resources (S-Map, River Environmental Classification, Land, Air and Water Aotearoa to climate forecasts and databases of crop and animal performance to inform where the best gains can be made.

6. How long can I continue to operate in my current fashion without impairing catchment indicators; is the most cost-effective placement of strategies to mitigate contamination at the source or sink; and if mitigation strategies are not enough then what is the most profitable land use for my property that will not impair catchment indicators and what is the level of associated investment risk?

7. Supporting Māori communities in the production of honey in Northland (also forestry). p.64

8. Pa to plate enterprise level study. p.64.

Appendix 6: Key Challenge Data Sets

A full analysis is required to determine the data needs of different projects and programmes within the Challenge. However, the following data themes are among those that are key to the realisation of the Challenge mission and that are either mentioned or implied in the Challenge documentation:

- Land tenure, including details of rights, responsibilities and obligations over land and customary title and holders of those rights, responsibilities and obligations.
- Details of enterprise activities, profitability etc.
- Topography.
- Soil types.
- Geology.
- Land use.
- Land cover.
- Land use practices.
- Water use.
- Census data (e.g. demographics and population location).
- Water body location.
- River networks
- State and trends in land and water resources (e.g. S-Map, River Environmental Classification, Land, Air and Water Aotearoa) – water quality and quantity.
- Data from microbial tracers (rivers)
- Contaminants (N, P, faecal microbes).
- Contaminant sources
- Climate including historical data.
- Crop productivity.
- Market analysis.
- Animal performance.
- Catchment location.
- Potential sources of contamination.
- Water quality indicators.
- Human knowledge and skills.
- Social and professional networks and kingroups.
- Data from Overseer
- Māori cultural values, cultural resources and practices
- Generation and transport of land use pressures (new) e.g. leaching

- Land use pressure response relationships (new)

- Transportation networks

- Location of facilities e.g. ports and processing plants.

- Agricultural yields and productivity

- Cost of inputs to agricultural production

- Genomics

- Data from new technologies (e.g. RFIDs, Animal traceability with RFIDs, on-vehicle data like Farm Angel and GPS tracking, UAVs, sensors, real time local weather stations)