MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

Scientific Collections and Databases Review

Update report

New Zealand Government

Contents

Executive summary

Scientific collections and databases (C&Ds) are a type of research infrastructure comprising systematically collected groups of specimens or data and their associated metadata. The C&Ds covered by this review are public assets that represent information generated from years of public investment in scientific research and related activities. It is important these assets are used and managed in a way that maximises (net) benefits to New Zealanders and contributes to international knowledge.

The last substantive review of government investment in C&Ds was in 1996. Since then the science system has seen significant changes. These changes have led some stakeholders to question whether the existing C&Ds funding arrangements are fit for purpose.

The scope of this review is broad - it seeks to:

- assess the strengths and weaknesses of the current approach to managing and funding New Zealand's C&Ds
- provide recommendations aimed at increasing value generated from New Zealand's C&Ds
- provide a framework and principles to help guide public investment in C&Ds.

This report provides an update on the findings of the review to date, and proposes possible options for improving the management and funding of C&Ds. This update does not provide firm recommendations; rather, it aims to stimulate discussion on potential improvements to the system.

This update has been informed by extensive consultation with stakeholders, including semi-structured interviews and a survey of users and custodians of C&Ds. The update also draws on academic literature and previous government reports - both from New Zealand and abroad.

As a starting point, some key characteristics of an efficient, well-functioning C&Ds system are presented. These include:

- that C&Ds are used to enable excellent, high-impact research, and its application for achieving public-good outcomes
- strong strategic oversight and direction
- mechanisms to collect timely and reliable information on who is accessing C&Ds and for what benefit
- processes to periodically review the value of C&Ds and return on investment
- funding arrangements that are adequate, dedicated and flexible
- good data management practices as standard operating procedure
- processes for generating transparent information on the cost of curation and enhancement of C&Ds, and for adjusting funding accordingly
- a high level of formal, enduring cooperation and coordination across related C&Ds
- that key decisions are made by those with the incentive, information and capability to drive value from C&Ds
- that Māori data sovereignty perspectives are adequately addressed through co-governance arrangements.

Parts of New Zealand’s C&Ds system have some of these characteristics but there is room for improvement. Decision-makers across the system need a richer understanding of how C&Ds are used and what for. There are opportunities to improve coordination and efficiency, there needs to be better management of data from publicly funded research, and more clarity around who has responsibility for managing that data. There is a need to ensure New Zealand

is keeping up with technologies and international best practice for data management. Funding needs more flexibility to allow resources to flow to new C&Ds as their value becomes apparent.

There are lessons to be learnt from other countries, but they do not offer a panacea for addressing the challenges mentioned above. What works in one country may not work in another. Lessons from overseas include that:

- incorporating C&Ds into broader science infrastructure plans and roadmaps can help promote a more strategic and considered approach to C&Ds management, maintenance, expansion and funding
- in considering the 'value' of C&Ds it can be helpful to clearly differentiate between the 'cultural value' of a C&D and their 'scientific or educative value'
- while the use of concepts akin to 'national significance' can be helpful for prioritising funding, it is useful to apply them in a flexible and adaptive manner
- forecasting (and adequately) funding operational costs is important for ensuring the sustainability of C&Ds, however accurate forecasts appear to be difficult to develop
- to promote the use of publicly-funded data, many countries have adopted open data policies as a precondition of research funding.

What changes could improve the management of New Zealand's C&Ds?

A starting point for improving New Zealand's C&Ds is to actively promote (and incentivise) the Government's principles for managing data and information: data should be open, readily available, well managed, reasonably priced and re-usable, unless there are necessary reasons for its protection. Personal and classified information must remain protected, and Government data and information should be trusted and authoritative.

In addition to these overarching principles, investments in *scientific* C&Ds could be guided by the following principles:

- *Focus on enabling excellent, high-impact science and science-based activities*: The scientific value of a C&D should be the primary driver of funding (as opposed to the cultural or heritage value).
- *Focused on public good outcomes*: Use of C&Ds should be primarily focused on the maximisation of public good outcomes (as opposed to private interests).
- *Policy and strategy relevant*: C&Ds should be relevant to the long-term, strategic needs of New Zealand.
- *Responsive to change*: Investment in C&Ds should be agile rather than static.
- *Commitment to Māori data sovereignty*: Custodians should make measurable shifts towards best practice by partnering with Māori and actively protecting their interests. This should be achieved by involving Māori in decision-making about Māori data and appropriately funding such activities.
- *Promote open data principles*: Investment should be used to incentivise commitment to New Zealand's existing open data principles.
- *Transparent and accountable*: The outcomes sought by funders should be clearly articulated and curators should be held accountable for the outcomes they achieve.
- *Well-informed*: Funding decisions should be based on sound information about the current and potential future uses of C&Ds. The consequences of funding decisions should be understood and acknowledged by investors.
- *Sustainable*: Funding arrangements should be sufficient to allow curators to undertake long-term planning for curation, maintenance and expansion activities, and to avoid

deterioration or destruction. Funding should also be sufficient for custodians (acting efficiently) to achieve the outcomes articulated by funders.

This review suggests the C&D system could be improved by:

- more clearly articulating cross-government expectations around the management and funding of C&Ds
- increasing the flexibility of long-term funding to respond to opportunities and the changing relevance of specific C&Ds over time
- strengthening incentives for collaboration, connectivity and cohesion across similar organisations or disciplines
- custodians adopting a more strategic approach to managing their C&Ds (enabled by more flexible funding)
- funders requiring (and paying for) the collection of information on who is accessing C&Ds and for what benefit
- conducting periodic reviews of the funding system to ensure expected outcomes are being achieved
- establishing a mechanism under the Strategic Science Investment Fund and infrastructure whereby custodians can seek funding for discrete projects that add value to existing C&Ds.

1. About the review

Why the review is needed

Scientific collections and databases (C&Ds) represent the collective knowledge generated from years of public investment in scientific research[1]. They are a type of research infrastructure comprising systematically collected or compiled groups of items or data that underpin or contribute to a broad range of scientific research and science-based services. As with other public assets, C&Ds need to be managed and used in a way that generates the largest possible value for the people of New Zealand.

Each year around $19 million from the Strategic Science Investment Fund (SSIF), operated by the Research, Science and Innovation (RSI) portfolio, is spent on enhancing and maintaining 25 Nationally Significant Collections and Databases (NSCDs)[2]. The NSCDs are only a subset of New Zealand's C&Ds. Funding has been relatively static for many years meaning it has decreased in real terms. Many other C&Ds receive public funding through various sources.

For some time, stakeholders have suggested that New Zealand's systems for funding and managing C&Ds may be limiting the value extracted from them. However, a system-level review of how the government invests in C&Ds has not been undertaken since 1996. The science system has changed considerably since then, for instance:

- there have been significant technological advances, such as in genomics where the increasing volume and use of data have unlocked uses of specimens and data not initially envisaged
- there are now more opportunities for strategic science investment and research is generally becoming more data-intensive, meaning there is more data being generated from publicly funded research
- collaboration between research institutes and countries has increased significantly and data storage and sharing has become easier meaning there are opportunities to form a cohesive approach and for C&Ds to become more interoperable and enable data-intensive research
- there have been advances in best-practice data management, including adding a digital component to physical collections and increased ease of storing, accessing, combining and sharing data from a range of sources.

The Government has set a target of raising economy-wide research and development (R&D) investment to two per cent of GDP by 2027. While some of this investment will come from the private sector, publicly funded research is also increasing, and with it the generation of data.

With these issues in mind, it is an opportune time to assess New Zealand's system for funding and managing C&Ds.

[1] This review includes social science within the definition of science, and recognises that the current Research, Science and Innovation investment is heavily weighted towards the natural and physical sciences.

[2] The list can be found here.

Objectives of the review

Appendix 1 provides a copy of the terms of reference (ToR) for the review. In essence, the review seeks to provide decision-makers with a:

- clear understanding of the strengths and weaknesses of the current approach to managing and funding New Zealand's C&Ds
- set of recommendations aimed at increasing the value derived from New Zealand's C&Ds
- framework and principles to help guide public investment in C&Ds, including guidance on when to invest further or to discontinue public funding of a specific C&D.

This report provides an update on the findings of the review to date, and proposes options for improving the management and funding of C&Ds. This update does not provide firm recommendations; rather, it aims to stimulate discussion on potential improvements to the system. Further policy development, including stakeholder consultation, is required to implement any changes raised in this report.

What collections and databases are covered by the review?

This review focuses on C&Ds used predominantly to achieve public good outcomes, and that are likely to receive (or have received) public funding for their establishment, maintenance or enhancement. As such, the review specifically excludes C&Ds held by commercial entities which are not in-principle publicly accessible.

In addition to the 25 NCSDs held by the Crown Research Institutes (CRIs) and the Cawthron Institute, an independent research organisation, there are a variety of other publicly-funded C&Ds held by research institutions such as museums, universities, and local and central government.

Different C&Ds have their own funding and access models, and, although many do not currently receive direct RSI funding, they may receive it indirectly through full-cost research funding.

What won't the review cover?

The review does not:

- provide a comprehensive stocktake of all C&Ds in New Zealand
- attempt a Cost Benefit Analysis of the returns to government investment in C&Ds
- review how well individual institutes are managing their C&Ds
- cover C&Ds held primarily overseas and funded by overseas organisations
- cover C&Ds primarily used for non-scientific purposes (such as arts and law).

At this stage, the review will not attempt to further define public-good or other definitions, such as when a collection of datasets becomes a database, or distinguish between a database and a model. Further definition of characteristics and concepts may occur should it be required in any subsequent phases of the review.

Our approach

This review is informed by:

- a survey of a wide range of stakeholders, for which 72 responses of varying completeness were received

- 35 semi-structured interviews with key researchers and decision-makers from a range of user and custodian organisations
- consultation with officials from relevant government agencies
- previous reviews of New Zealand's science system and of specific C&Ds
- relevant academic and 'grey' literature from New Zealand and overseas
- attendance at national and international conferences relevant to data management
- discussions with our Australian counterparts.

It is important to recognise that this review is not starting with a blank canvas. Many of the challenges found in this review result from path dependencies and legacy issues. This means careful thought is required about whether the benefits resulting from any changes are worth the effort and investment (ie we want to avoid change for change's sake).

2. The collections and databases system

The terms 'collections', 'databases' and 'data' can mean different things to different people. This report used the following interpretations:

- **Collections** are systematically collected physical specimens - such as animals, plants, fossils, rocks, soil, ice cores, and microorganisms, and their associated metadata. Many collections have a digital component as well as physical specimens.
- **Databases** are stored repositories of information on particular subjects of interest - such as climate observations, levels of toxins or pollutants, genetic sequencing, a particular cohort of specimens (ie longitudinal studies), and natural hazards. Both collections and databases may expand as more specimens or data are collected.
- **Data** are facts, observations or numbers collected for the purpose of answering questions or informing decision-making.

The **C&Ds system** includes C&Ds themselves, but also the people, infrastructure, processes and standards required to generate, store, manage and share specimens and data. More broadly, this can also include other infrastructure that enables the use and sharing of data such as high-performance computing and high-speed networks.

The role of collections and databases

C&Ds can be established for a variety of reasons, including for research, commercial, policy or regulatory purposes. They can contribute to many scientific disciplines and support research-based services and activities. Examples of current uses of C&Ds include:

- scientific research
- education (all levels)
- planning, policy and response (eg for biosecurity, environmental, health and social interventions)
- regulation (eg for issuing permits and setting environmental standards)
- reporting to contribute to policy fora and meet obligations (domestic and international)
- commercial products or service development.

Importantly, the value of C&Ds can become much broader than their original purpose. Well-managed C&Ds provide extensive option value through their unpredictable future uses. Technological advances can unlock new uses for existing C&Ds; genomics and data analytics capability are two well-known examples. In many instances, the longer the time over which the data has been collected, the more valuable a collection or database becomes.

System participants and their interactions

Broadly speaking, key participants in the C&Ds system are investors, custodians and users.

Investors provide resource to the system. Resources can take different forms but are generally provided in the form of money or access to skilled people[3], infrastructure and equipment. Investors make resourcing decisions based on an expectation they will receive a return on investment. In the private sector, the return is generally profit; in the public sector (the focus of this review), return can be conceptualised as the net economic, environmental, cultural or social benefit derived from C&Ds.

Custodian organisations have responsibility for maintaining and enhancing C&Ds. They use resource from investors for the costs and activities set out in the table below. Custodians are the key decision-makers regarding access, growth, preservation and retirement of C&Ds. However, flexibility in these areas can be constrained by investment conditions including available resources, investor policies, and contractual arrangements with investors or users.

Curatorial oversight; skilled people			
Maintenance	**Growth**	**Enhancement**	**Services**
- Upkeep of existing specimens, data, and infrastructure (ie preventing deterioration) - Storage infrastructure and consumables for C&Ds, such as buildings with controlled environments for collections, and rapidly changing hardware and software needs for data - May be more or less automated depending on collection type - Quality control and adhering to relevant standards[4]	- The initial cost of collecting the specimens and data - Ongoing addition of specimens and data, such as amalgamations or acquisitions of physical collections, outputs from specific projects, or the upkeep of monitoring equipment networks and other routine data collection processes to ensure new datasets are added to in a consistent way	- Activities such as digitisation, removing barriers to access, adding or improving metadata or images, or collaborative projects to increase cohesion across related C&Ds - May increase costs (eg adding a digital component to a physical collection)	- Response to requests for specimens or data, routine or bespoke - Providing data in specific formats - Interpretation, such as developing models or writing reports - Verification (eg identification of a specimen) - Hold, own and administer registered IP

Users are the 'customers' of the system. They utilise their intellectual capital to interact with C&Ds to produce outputs – such as academic papers, reports and further data or specimens.

[3] Can also include 'brokers' who facilitate relationships between users and custodians.

[4] An oft-cited example in wide use in the natural history collections community is Darwin Core

The outputs generated by users support a wide range of operational services – such as adding specimens or data to the C&Ds they utilise, monitoring compliance with regulations, evidence for policy, the management of natural resources and the development of new products and services. These outputs can generate outcomes and eventually impacts – society-wide increases in productivity or wellbeing[5].

Measuring the impact of C&Ds is difficult. However, measuring outputs or outcomes is important as it provides feedback to custodians, users and investors on the utility of a collection or database.

Nationally Significant Collections and Databases

Government has historically provided selected C&Ds with science funding. In 1996, the following selection criteria were developed by the Foundation of Research, Science and Technology (FoRST) and used to designate 25 C&Ds as "Nationally Significant".

Criterion 1: Is the science asset funded in whole or in part from the Public Good Science Fund?

Criterion 2: Is the science asset nationally important?

National importance was assessed against the following sub-criteria:

- Sub-criterion 2.1: Does the asset make a substantial contribution to the goals set out in the Statement of Science Priorities?
- Sub-criterion 2.2: Is the asset important to a wide range of stakeholders?
- Sub-criterion 2.3: Does the asset deliver substantial benefits to users?
- Sub-criterion 2.4: Is the asset unique nationally and/or internationally?
- Sub-criterion 2.5: Is the asset irreplaceable?

Criterion 3: Is funding of the science asset on a priority basis consistent with the Foundation for Research Science and Technology Act and with the Statement of Science Priorities?

The current NSCDs were funded on the basis that:

- they are being held on behalf of New Zealand, where continued provision, maintenance and utilisation are critical for New Zealand science to deliver public benefit
- the benefits accrue to many, varied users and third party beneficiaries while the costs of provision belong to the custodian.

Funding was to allow custodians to at least maintain the NSCD, and there was an expectation that the data would be freely and publicly available where it is inappropriate for the end-user to pay (otherwise, access could be cost-recovered).

In 1996, FoRST committed to give at least two years' notice to NSCD custodians if a funding cut was coming. This was to allow time for custodians to find alternative funding sources.

The relatively static level of funding over recent years has meant funding has declined in real terms. With the continued growth of some NSCDs, custodians have been using funding from other sources (such as SSIF Programmes or commercial revenue) to ensure NSCDs are maintained to a workable level, but they report that this is becoming increasingly difficult. For example, some of the maintenance work is done by retired volunteers.

[5] For more on these concepts see the Ministry of Business, Innovation and Employment's 2017 Impact of Science discussion paper.

To alleviate critical funding pressure, Budget 2016 allocated an extra $2 million per year to the SSIF Infrastructure appropriation to increase funding for the NSCDs. This funding was allocated on a pro-rated basis to NIWA, Manaaki Whenua, and GNS, which are custodians of the greatest number of NSCDs.

Other collections and databases

In addition to the NSCDs, there are a variety of other C&Ds held by organisations such as museums, universities, and government departments.

Different C&Ds have their own funding and access models. Some currently receive direct RSI funding, but others receive it indirectly through full-cost research funding or from other RSI sources such as SSIF Programmes at the discretion of custodians. Some C&Ds also receive direct public investment from, or are held by, other government agencies. Public funding includes funding from local government (ie rates).

Given the focus of this review is on public-good C&Ds, almost all C&Ds mentioned in the survey and interviews receive (or had received) government funding in some form[6]. Mechanisms for government funding included:

- **Direct funding** to create, maintain or enhance a specific collection or database (eg the NSCDs, Genomics Aotearoa, and specific disease databases)
- **Devolved organisational funding** (eg some museums are bulk-funded with rates via local government and they maintain collections as part of their core business)
- **Indirect funding** (there are many examples of project funding for research that utilises C&Ds and is sometimes used to access or enhance them as part of the project).

Other sources of funding include direct funding from the custodian organisation, lotteries grants, commercial users, and access fees or levies. These are often contestable and uncertain in the long term.

Approach to data management, and creation and growth of C&Ds

An important means by which new C&Ds are established or existing ones grow[7] is through RSI-funded research. This is not currently done in a coordinated way. Research contracts contain obligations for researchers to make their data available unless it would compromise privacy, commercial interests, ethics or legal obligations.

Data or specimens from a research project could be usefully added to an existing collection or database (and therefore would become more valuable and increase the impact of the research/activities), and some larger RSI investments result in the establishment of new C&Ds, such as the National Science Challenges, or Genomics Aotearoa. However, stakeholders report that, in general, ongoing maintenance and provision of data or specimens arising from research projects once the project funding ends is not well thought out at the start.

[6] For this review, the source of funding is more important than whether the custodian entity is public or private. There are a number of publicly-funded C&Ds held by privately-owned, independent organisations.

[7] A well-known example of a growing database is the Integrated Data Infrastructure (IDI), held by Statistics NZ.

3. Observations of current arrangements

This section summarises MBIE's observations of New Zealand's C&Ds system. The information presented is a collation of evidence from the stakeholder survey, interviews, and other personal communication with stakeholders. The chapter adopts a system view, and begins by articulating the characteristics of an efficient, well-functioning system for managing C&Ds. Drawing on these characteristics, the chapter then discuss the strengths and challenges of the current system.

What would an efficient, well-functioning system look like?

Testing the performance of the C&Ds system is difficult to achieve empirically. However, it is possible to identify characteristics of an efficient, well-functioning system, and to compare our observations of the current system against these characteristics.

The characteristics of an efficient, well-functioning system include:

- ***C&Ds are used to enable excellent, high-impact research and its application for public-good outcomes*.** In a well-functioning system, C&Ds would be widely used (and valued) as an enabler of excellent, high-impact research. Investors and users would look for opportunities to extract additional scientific value from existing C&Ds.

- ***Strong strategic oversight and direction***. In a well-functioning system, investment is guided by clear principles and investment signals that are as consistent as possible across public investments. Investors would articulate their expectations for C&Ds, and custodians, in turn, would use the strategic signals of investors to develop their own C&Ds strategy and management plans that would include (as appropriate):
  - governance arrangements
  - value assessments for their major C&Ds
  - preservation/maintenance plans and priority growth and enhancement activities for each C&D (noting that the challenges may be different)
  - access model and policies (openness, points of access)
  - any standards they will adhere to
  - how they will partner with Māori in decision-making about Māori data
  - mitigation of risk of unintentional loss
  - exit strategy including how they will determine how to disinvest without compromising potential future value
  - international activities, current and planned
  - approach to recording and reporting use in order to assess impact.

  Some organisations may be investors, custodians and users. Further, some custodian organisations may gather their own specimens and data to grow collections while others only store and manage other people's data. This means the management of C&Ds at the level of the individual organisation is important and any overarching strategy or principles need to be flexible enough to adapt to organisations' different circumstances.

- ***Mechanisms to collect timely and reliable information on who is accessing C&Ds and for what benefit.*** This information is important for understanding the needs of different users, and the value generated from individual C&Ds. Use information is particularly valuable for people making decisions about how best to use resources. This could be investors deciding which C&Ds or organisations to provide resource to (eg NSCD approach), custodians who

have flexible organisational funding (eg Te Papa), or users (eg a researcher with project funding looking for the best C&Ds to use for that project).

- ***Processes to periodically review the value of C&D investments.*** To ensure the greatest return from public investment, an efficient, well-functioning system would periodically review the value derived from each C&D and adjust investment strategies accordingly. Such reviews, provide an important 'feedback loop' for investors, allowing the system to be more dynamic and adaptive than it otherwise would. Value is broad in this context and includes economic, environmental, uniqueness, reputational, or cultural and heritage value. It also includes option value associated with currently unknown future uses.

- ***Sufficient, dedicated and flexible funding.*** The value and relevance of C&Ds can change through time. For instance, new technology can facilitate novel ways of extracting value from existing data, or, conversely, can make existing methods or data obsolete. A well-functioning system would have dedicated baseline funding that is flexible enough to adapt to such changes (while acknowledging the uncertain nature of future benefits) and be sufficient to maintain and enhance C&Ds as need or opportunity arise. Funding should be dedicated to reduce the risk of reallocation within custodian organisations (given the range of views on the value of C&Ds). Flexibility would also allow custodians to respond to opportunities such as a need to create a collection or database in a new area, or a user with funding for a large project that would grow a specific C&D.

- ***Good data management practices would be the norm***. Data management is the application of skills, tools and technologies to allow value to be extracted from the original investment in data collection (MoRST, 2008). A well-functioning system would include incentives to adopt good management practices, including planning for ongoing data management at the outset of projects. In many instances this will mean adhering to 'FAIR' data management principles (see box below[8]). Having data that is accessible and easy to find would allow researchers to explore opportunities to use data in new ways.

- ***Processes for generating transparent information on the cost of curation and enhancement of C&Ds, and for adjusting funding accordingly***. In a well-functioning system, investors would have reliable information on the cost of curation and enhancement activities. This not only promotes transparency, but helps investors to better understand potential funding gaps – and the consequences of gaps.

- ***High level of formal, enduring cooperation and coordination across related C&Ds.*** In a well-functioning C&Ds system, custodians, investors and users would work together to improve the system. A system approach with a high level of collaboration and coordination across actors would help to facilitate data sharing, reduce duplication of effort, and improve the efficiency of the system. It would ensure that various investors (for example, different government agencies or private investors) take an outcomes-focused approach that is as consistent and transparent as possible, and that the broader system including project-based uses of C&Ds and other research infrastructures are considered. Such cooperation needs to be formal and enduring to ensure it continues through changes in staff and organisational structure.

[8] https://www.nature.com/articles/sdata201618

- ***Key decisions are made by those with the incentive, information and capability to drive value from C&Ds.*** As a general rule, a well-functioning system would allocate key decisions to those with the incentive to allocate resources to the highest value use, the information needed to judge what that use may be, and the capability to implement the decision.

- ***Māori data sovereignty perspectives are adequately addressed through co-governance arrangements.*** Ensuring culturally appropriate management of and access to C&Ds will unlock benefits that could be extracted from Māori data or information, including benefits for Māori. There is an opportunity for New Zealand to be a world leader in incorporating indigenous perspectives on data into the broader data management system.

FAIR data principles

A development in recent years is the emergence of the FAIR principles for data management. These principles provide guidance for scientific data management and stewardship and are increasing in use in data management circles. FAIR stands for:

- *Findable* – the existence and content of C&Ds should be easily discoverable by a wide range of potential users
- *Accessible* – once found, C&Ds users should be able to access them at minimal cost and effort, or there should be clear reasons why access is limited
- *Interoperable* – data or tools from non-cooperating resources should be able to integrate or work together with minimal cost and effort
- *Re-useable* – C&Ds should have detailed descriptions of provenance, limitations and attributes. They should have clear terms of access and meet relevant community standards.

The FAIR principles are generally built into data management practices over time rather than being an end-point that can be achieved through a single investment. Also, not all the principles are suitable for all types of data, but generally data should be as FAIR as possible.

Along similar lines, New Zealand's open data principles were developed to ensure there is high quality management of the information the government holds on behalf of the public. These principles are discussed further in Section 5.

What role should government play as an investor?

As a steward of the science system, MBIE has a strong interest in ensuring New Zealand's C&Ds support excellent science and provide taxpayers with value for money.

The role of government investment is a key question in this review due to the complexity, size and growth of the issues. The relationship between custodians and users is often closer than that between investors and users. This means that custodians can more fully understand and respond to users' needs, which will increase the investor's return on investment.

Government investment should not be overly prescriptive. Rather, a balance should be struck between providing sufficient freedom to custodians to manage C&Ds in a way that best meets users' needs, and ensuring that custodians are managing C&Ds in a way that is aligned with government policies such as open data and public good.

Therefore, any guidelines government (as an investor) sets should be outcome-based and flexible to allow custodians to apply them to their own circumstance and adapt to their users' changing needs.

It is important to note that there will always be fewer resources than opportunities. The C&Ds landscape is large, growing and dynamic, and government must invest in a way that provides sufficient resources to increase benefits from C&Ds in a financially sustainable way.

Opportunities in the current system

Throughout the course of the review, MBIE officials identified several opportunities that, while currently patchy or nascent, could be built upon to make improvements quickly through increased connectivity, coordination and dissemination of best practice. This would increase the efficiency and effectiveness of the system. Opportunities identified in the system include:

- ***There appears to be a strong desire to do things better.*** The vast majority of stakeholders interviewed (both users and custodians) agreed with the premise that, where appropriate, publicly-funded data should be freely and openly available in a useful format.

- ***Strong informal networks facilitate data sharing.*** New Zealand has a relatively small C&Ds system. There were many comments from users who simply emailed someone they know in a custodian organisation to access C&Ds and make specific requests.

- ***There is close international collaboration (in some areas).*** New Zealand's RSI system is already well-connected internationally and this is reflected in the C&Ds system. Many C&Ds have international users, and researchers and custodian organisations already submit their data to international journals or aggregators – especially in the areas of genomics and taxonomic research.

- ***Examples of good data management are emerging***. Some organisations already have a strategic plan to clear backlogs, move towards international best-practice, and make data more FAIR. These efforts were usually as a result of strong organisational culture and practice, and supported by adequate funding (either from the organisation itself or through other funding sources such as lotteries grants).

- ***Efforts to improve coordination are underway***. There are several initiatives already in place that demonstrate a desire to ensure data and metadata are collected consistently and presented in a unified format. For example, Land Air Water Aotearoa (LAWA) is used by many councils to make their data available through a single point of access.

- ***New Zealand has existing open data policies and Māori data frameworks.*** Examples include a set of data management principles that were agreed by Cabinet in 2011 (see Section 5), the Open Data Charter[9], and the New Zealand Government Open Access and Licencing (NZGOAL) framework[10]. Te Mana Raraunga, the Māori Data Sovereignty Network, has also developed a framework that may be useful for investors and custodians considering Māori data issues. Both of these could be built on and used to guide investments, strategies and enhancements.

[9] https://www.data.govt.nz/blog/open-data-charter/
[10] https://www.data.govt.nz/manage-data/policies/nzgoal

Challenges facing the current system

The system faces some significant challenges. We have grouped the challenges into broad categories but they are all inter-dependent and not mutually exclusive. These challenges are summarised in the table below.

Category	Description	Importance
Management and curatorial challenges	Enhancing C&Ds through improved coordination	high
	Avoiding deterioration and backlogs due to gaps in capability/capacity	high
	Developing richer information on the value and use of C&Ds	medium-high
	Involving Māori in decision-making for Māori data	high
Financial and budgetary challenges	Total funding level	high
	Accounting for the cost of ongoing data management during project development	high
	Improving cost recovery processes	medium
	Making funding more flexible	medium
Technological challenges	Keeping up with technological advances	high
	Adhering to standards	high
Policy and legal issues	Improving system oversight and strategic guidance	high
	Addressing barriers to openness	high

Management and curatorial challenges

Enhancing C&Ds through improved coordination

As noted in Section 2, custodians often undertake activities aimed at enhancing the value of C&Ds. Enhancement can occur, for example, through the addition of specimens or data, increasing findability and access, or consolidating similar collections or databases[11].

Enhancement can be costly, yet the current system is generally unresponsive to the increasing costs incurred by custodians. This is in part due to limited information on the size and nature of cost increases, and in part due to limited flexibility within current funding mechanisms

[11] As a collection or database grows, it obtains greater statistical power, and can be enhanced in a number of ways to increase its use.

(discussed further below). In any case, enhancement activities of many custodians face financial pressures[12].

Finding efficiencies in the system will be important if enhancement activities are to be undertaken within the current funding envelope (which may require redesign to enable further efficiencies). However, the cost of implementing any efficiency gains needs to be weighed carefully against the benefits. Efficiency gains appear most likely to occur through:

- *Coordinated service provision* such as providing a single user interface for accessing multiple databases
- *Shared capacity* such as sharing costs of skills or expensive capital, or consolidating C&Ds.

Providing shared capabilities or sector-wide services requires a coordination function, and the provision of capability and/or equipment to multiple users with similar needs. Existing examples of this model are the National eScience Infrastructure (NeSI) and the Research and Education Advanced Network of New Zealand (REANNZ).

Yet, seemingly similar C&Ds can in fact be very different, and the importance of metadata in determining the value and interoperability of C&Ds cannot be overstated. The use of different collection methodologies, infrastructure and standards can reduce the ease with which services can be coordinated, capacity shared or C&Ds consolidated. These are often legacy issues which traditionally have been difficult and costly to solve.

Avoiding deterioration and backlogs due to gaps in capability/capacity

The risk of degradation of specimens or data was almost universally raised by stakeholders. Yet, substantiating the extent and scale of the risk proved difficult. The review found many instances of large cataloguing backlogs, and orphan datasets with poor findability.

The consequences of degradation or loss are more significant for physical specimens that hold greater historical and cultural value, and for data that cannot be re-generated (observational and/or longitudinal).

For physical collections the primary mechanisms to mitigate risks are controlled storage conditions and access, regular inspection for deterioration, and a level of redundancy that balances specimen rarity with the ability to replace it. Similarly, database preservation mechanisms include controlled access to hardware, replication across multiple sites and regular inspection for format redundancy, 'bit-rot' and software versioning.

Deterioration and backlogs were commonly attributed to shortfalls in capability and capacity, which could possibly be alleviated through enabling greater efficiency in the system. While most C&Ds currently have a curator, they may not be a dedicated FTE and curation capacity was very commonly reported as insufficient. Skills in curation, data management, and technology were considered paramount to the maintenance of C&Ds, however, custodians often find it difficult to recruit individuals with these skills[13].

[12] For custodians of physical collections, cost pressures mainly impact the enhancement of management and curation activities. For commercial and government collections, cost pressures impact the ability of custodians to make data more accessible and useable.

[13] Although it wasn't entirely clear whether the current recruitment difficulties were caused by lack of funding or lack of skills available (or both), a repeated concern across almost all stakeholders was the

Developing richer information on the value and use of C&Ds

Measuring the use of C&Ds is a helpful way to assess their impact and value. There are several common ways to measure and record use of C&Ds, but some stakeholders did not assess usage as they did not think it was a priority within their already limited resources. Perceived lack of value in gathering use statistics is in part driven by the fact that it is difficult, or even impossible, to attribute impact from user activity without associated metrics that document and evidence public benefit, and which link it back to the use of C&Ds. Stakeholders generally reported a need to infer public good impact from user metrics.

The most commonly reported metrics to record use were counting website hits and recording requests, inter-loans and physical visits. Tracking user type and/or purpose was much less common where it was not already obvious (ie a researcher is probably accessing data for research purposes). Getting a true measure of use through attribution was difficult, especially outside of research uses and over time (eg in taxonomy, users do not attribute original work once the scientific name is accepted). It appears that the more freely and widely available data is the more it is used, but also the more difficult it becomes to measure that use and subsequent impact.

There are emerging services, such as digital identifiers and aggregation services, that can assist with tracking use and impact but they are currently only available to research users. For example, the Open Research and Contributor ID (ORCiD) disambiguates individuals and can be implemented as a primary key to aggregate other identifiers associated with those individual researchers using C&Ds. Also, DataCite provides the Digital Object ID (DOI) system, which can be used to trace and measure the use of data across an integrated digital system. Together with any existing ID systems implemented across collections or databases (eg the herbaria registry ID), a rich network of system-wide research activity can be created and measured.

Partnering with Māori in decision-making for Māori data

Māori data sovereignty is an emerging consideration in the management and use of C&Ds, particularly in international data sharing and use, and benefit sharing.

Stakeholders' knowledge and consideration of Māori perspectives and data sovereignty principles was varied. However, there was general agreement across stakeholder types that, much like the non- Māori population, there is no single Māori opinion, and acknowledgement that collections of native specimens, mātauranga Māori, or samples or information from Māori people needed to be managed in a culturally appropriate manner with sufficient funding to do so.

There was a strongly-voiced intent to include Māori representatives in the development and use of C&Ds, particularly those relevant to Māori health outcomes or intellectual property. While some organisations are more advanced than others in partnering with Māori in decision-making, there was agreement that it is not done well on the whole.

There was a mixture of opinions on whether Māori perspectives on data management and sovereignty were a barrier or an opportunity. Collectively, Māori hold nationally and regionally

---

lack of curation capability coming through the pipeline. The capability pipeline is out of scope for this review but it appears to be a pressing issue, particularly in taxonomy.

significant information about themselves and New Zealand's native biodiversity and ecosystems. Unlocking the potential of this information and ensuring beneficial returns to Māori through their active involvement in culturally-appropriate data management practices and intellectual property rights was generally seen as a significant opportunity. Māori provision of expertise in relation to Māori and engagement with other indigenous people is consistent with the practices of Tikanga Māori.

However, some stakeholders consider Māori reticence around information sharing and open access to be a barrier to research, including, or in some cases especially, research that could benefit Māori. These stakeholders were concerned that inequities in health and economic outcomes would continue to widen without Māori involvement in decision-making about the use of Māori data.

Additional funding for systemic improvement in this area and partnering with Māori in decision-making where appropriate may be required. This could include increasing the number of current and future experts in this area as demand on the few existing experts is high.

Financial and budgetary challenges

Total funding level

Almost all interviewees considered a general lack of funding to be the main reason behind limited access to C&Ds, through either lack of curatorial capacity, degradation/obsolescence of specimens, data or data infrastructure/formats, or inability to distribute physical specimens or respond to data requests. Many reported an internal mechanism to fund C&D management but, increasingly, maintenance had to compete for operational funds with other internal services. Some stakeholders used a portion of the research funding for projects that utilise C&Ds to support routine maintenance or curation.

Stakeholders report that lack of funding at least partially drives many of the other challenges. However, it is not clear that overall funding is insufficient. Particularly with funding that is devolved to custodian organisations, the perceived lack of funding may be due to organisational culture and inefficiencies created by the current structure and limitations of the funding rather than an overall lack of funding from the investor. If funding is insufficient, it is unclear by how much, where additional funding is most needed, and how organisational culture issues might be addressed to ensure efficiency. The current funding level may be sufficient, but redesign is required so that existing inefficiencies can be overcome.

Accounting for the cost of ongoing data management during project development

Project-based funding was frequently reported as a key issue. Research, educational/citizen science, or monitoring projects can create data (including sizable databases such as in the National Science Challenges) that will likely have ongoing value if they are stored and curated in a way that makes them findable, accessible and able to be combined with related data. However, stakeholders reported that ongoing data funding, access and management was rarely considered at the outset of a project, meaning that data often became orphaned, lost, or otherwise not utilised beyond the life of the project.

Improving cost recovery processes

There are a variety of different modes of cost recovery: some organisations provide free access; some cost-recover only for large or complex access requests; some must cost recover everything to maintain financial viability; and some charge beyond simple cost recovery. All

stakeholders charge for bespoke, value-add activities such as model development and data interpretation.

Charging policies were generally dependent on whether a collection or database had commercial potential (ie could create a beneficial return for the custodian) rather than how the data collection was funded in the first place (public versus private). There were reports of some custodians charging for access to data collected wholly through public funding, which was thought to be inappropriate.

Another situation reported by several custodians was that they had to charge for access in order to make ends meet and keep the collection or database available to users for whom it was critical, even though they knew this reduced the overall use and impact.

From a user perspective, some stakeholders reported frustration at the variability of access arrangements, both across organisations with similar C&Ds the user wanted to access, and within organisations over time.

Making funding more flexible

Direct funding was seen as desirable but, when it is linked to specific C&Ds, it is inflexible in the face of shifts in the value of C&Ds, variable user priorities and needs, and specific opportunities.

This review suggests C&Ds operate in a sophisticated, complex and dynamic ecosystem, and that the value of C&Ds can change rapidly in response to unpredictable events (such as earthquakes, biosecurity incursions or the emergence of new technologies). It is as yet unclear whether the investor selecting specific C&Ds on the basis of national significance is the most efficient and sustainable approach.

A collection or database that is currently assessed as low value because it has deteriorated, fallen out of fashion, or has been superseded by more advanced versions does not mean it has no future value. Also, many C&Ds are irreplaceable because they contain observational data that cannot be repeated or specimens of extinct species.

Furthermore, the current approach is not financially sustainable as costs will increase with each new collection or database, and definitions and scope to inform the current investment approach are difficult to establish and may change over time.

Technological challenges

Keeping up with technological advances

Findability, access, interoperability and maintenance are often sub-optimal due to dated and rapidly changing data infrastructure and user interfaces. Risks to infrastructure and the capability required to continue growing and extracting value from longitudinal C&Ds (eg regular monitoring programmes or networks of monitoring equipment that feed into an existing database) were frequently reported.

Rapid technological advancement is enabling enhancements such as digitisation and computational algorithms to help interpret and use data (eg artificial intelligence, deep learning). However, technology obsolescence is another risk to ensuring that data remains available in a format that is useful to users. Ongoing service provision requires skilled people and sometimes significant investments each time new technology is required.

Adhering to standards

There is a lack of standards and quality controls (domestic or international) in many disciplines or across organisations that collect and hold similar types of data. For example, various organisations may collect data on air quality slightly differently and have different metadata[14]. Some stakeholders thought that legacy issues would be difficult to resolve. However, standards could be applied to data collection going forward, provided they are sufficiently similar to existing norms so as not to diminish the value of the data already collected (and therefore the database as a whole).

Policy and legal issues

Improving system oversight and strategic guidance

Many custodians currently receive little guidance from investors on how to prioritise or make trade-offs, for example between maintenance and enhancement. Some custodians report organisational culture issues because there are no incentives to improve practice (given that it is not currently linked to funding). Stronger signals and system-level oversight from government as an investor could create incentives for best-practice, and greater consistency and efficiency in managing the range of C&Ds.

Addressing barriers to openness

Openness was generally considered good practice but there are a number of barriers. C&Ds may not be readily accessible for valid reasons, such as: security and privacy concerns for ethically sensitive data; contractual or commercial arrangements with third parties; endangered or potentially dangerous (eg poisonous or hallucinogenic) species or specimens; and management of specimen degradation or quarantine. Such concerns were often reported as a trade-off; custodians had to balance competing agendas, and the level of buy-in regarding the benefits of data being as open as possible was mixed across various stakeholders.

Privacy and security were almost always a concern for health and social C&Ds. It is a legal requirement that people cannot be identified from their data, the trade-off being that making it more open to researchers or policy-makers could result in better health outcomes for those people or groups. Anonymised data was made available where possible (often only when a researcher asked for it) but that required skilled people's time, which is where lack of resource sometimes became a barrier.

The Māori view on open access is often cautious, requiring careful management of Māori data sovereignty in the context of wider Māori aspirations. There will be some information that Māori do not want to share and this could be addressed on a case-by-case basis. However, mutually beneficial trading relationships are a part of Māori culture so in order to derive benefits from Māori-owned information, Māori would need to be partners in decision-making about sharing that information.

The trade-off between openness and commerce was reported frequently. Where commercial interests were concerned, both the databases and any specimens considered a valuable asset would only be shared under certain conditions. There was evidence of this from CRIs, and it was suggested several times there was a possible conflict arising from pressure to generate revenue streams independent of CRI funding and providing access to publicly-funded data.

[14] Although technical solutions such as data dictionaries mean this is less of a problem now than it was.

Commercial reasons why access to C&Ds might be restricted include that: it may be jointly funded and enact confidence or embargo contracts to protect investments; there may be legal ambiguities with native title indigenous native flora and fauna; or there may be legislative responsibilities for data protection (eg drill cores). There was some anecdotal evidence that having overly open data policies could discourage commercial users from interacting with C&Ds or even the custodian organisation in general if they were not assured their investment would be protected.

For privately held and funded C&Ds, this is not a problem, however tensions arise when a collection or database contains data that was collected with public funding or that receives public funding for its maintenance and provision. Where there was a mixture of public and private funding for C&Ds (and the data they contain) a range of bespoke access models were used. For example, selective or delayed data availability was common, or charging commercial users (eg for product or service development, or privately funded research) but not charging other users (such as publicly funded researchers, schools, government departments).

For C&Ds that do not have a public access mandate, access and licensing decisions were often based on the protection of intellectual property rights. The default position is generally to control access to the content rather than licence use. This is because 'potential' innovations cannot be protected, only defined innovations, so it is easier to protect the asset and potential intellectual property by controlling (or preventing) access than it is to define a complex use or shared ownership license. This approach is recognised to stifle innovation, as all potential innovation is only available to a small group of owners or users, which was seen as inappropriate for publicly funded C&Ds.

A common and more innovation-friendly approach to publicly funded C&Ds was to encourage commercial opportunity via 'value-add' activity (ie not restricting access to the data, but permitting commercialisation and subsequent protection of the products developed by use of the collection or database). Developing a licensing framework able to permit use without stifling innovation can be a complex task, but NZGOAL and Statistics New Zealand's Data Ventures work[15] may be useful starting points should this work proceed.

4. What can we learn from overseas?

Many countries are facing similar C&Ds policy, technical and funding issues as New Zealand. Countries have responded to these challenges in different ways, with their approaches influenced by history, funding, political environment and other factors. The uniqueness of each country means care must be taken when looking overseas for 'off-the-self solutions' to the issues raised in Section 3. Nevertheless, exploring how different countries tackle common issues provides useful insights for New Zealand. This section highlights some of these lessons.

Policy lessons

Plans and roadmaps promote a strategic approach to C&Ds

Several countries (such as Australia, Canada and the United States) include C&Ds in research infrastructure roadmaps and investment strategies. In doing so, these countries promote a strategic, long-term approach to C&D investment and management. Further, discussing C&Ds

[15] https://dataventures.nz/

alongside more 'traditional' types of research infrastructure, such as scientific equipment, brings the importance of C&Ds into focus, making their value to the science system seemingly more widely appreciated.

For example, in 2018 the European Strategy Forum on Research Infrastructure updated its Science Infrastructure Roadmap. Among other things, the roadmap lists projects of strategic importance to the European Union science system. Of the five new projects added to the roadmap in the 2018, two relate to the strategic management of C&Ds taking a more centralised approach. These are:

- the transformation of dispersed and fragmented natural science collections into a single integrated pan-European resources[16]
- the development of a unique access point for historical documents and resources relevant for social and cultural research into the Holocaust.

National significance is used, but generally more flexibly than in New Zealand

All countries have finite resources to allocate to C&Ds. These resources need to be allocated in a way that creates as much value as possible for the country in question. Generally, this requires a way of differentiating between high and low value C&Ds. Assigning them 'national significance' (or equivalent) is one way to make this distinction.

The Australian Department of Environment and Energy (2010) notes that "*Significance is a proven persuader. Whether it's making the case for a new acquisition, substantiating a funding application, or lobbying for education and online resources, significance goes to the heart of why collections are important and why they should be supported*".

While several countries use concepts akin to national significance, the concept tends to be applied more flexibly than it is in New Zealand. Periodic reviews are common, with C&Ds 'graduating' to national significance as their value becomes apparent or as circumstances change.

Significance is most commonly used in connection to large nationally-controlled collections – such as the Smithsonian Institute's Air and Space Museum or the socio-cultural collections in the Scottish Museum.

In the UK, Research Councils maintain C&Ds that, while national and significant, are not classified as being of 'national significance'. These C&Ds are driven by the scientific community to participate in international networks (climate change), or specialised disciplines (UK Data Archive or archaeological data service). Their enduring value is recognised and this appears the most notable criteria for support.

It is common to differentiate between historic, cultural and scientific values of C&Ds

C&Ds can be valued for different reasons. Understanding the 'type' of value derived from a specific collection or database provides useful context for decisions about how it should be managed and funded.

[16] The 'Distributed System of Scientific Collections' project (DiSSCo), and the 'Atlas of Living Australia' was developed for a similar purpose.

While countries use different criteria to assess the significance of C&Ds, it is common to make a distinction between the *historical, cultural* and *scientific* value[17]:

- *Historic value* recognises some C&Ds (typically collections) are irreplaceable heritage assets and part of the historical record of a country. Criteria for assessing historic value can include whether the collection is associated with a particular person, group or event of recognised historic importance.

- *Cultural value* recognises some C&Ds (again, typically collections) have cultural or spiritual value to particular communities or groups in society. Criteria for assessing cultural value can include whether the collection embodies the beliefs, ideas, customs and traditions of a particular group, and whether there is demonstrated contemporary attachment to the collection.

- *Scientific value* recognises C&Ds are often used as an input into further research. Scientific value covers both the value derived from *existing* (or known) uses of C&Ds and *potential* (future) uses that may arise.

Ultimately, investors must make judgements about the magnitude of cultural and historic value provided by individual C&Ds. Investors must also make judgements around the likelihood that a collection or database will provide scientific value in the future.

Devolved decision-making is common

There are many examples of the management of collections being devolved to operational organisations. For example, in the UK the government devolves operations of C&Ds to various Research Councils to make decisions based on their strategic visions and operational budgets. This may entail them retiring or archiving a collection or database that is no longer used but remains part of the scholarly record.

Similarly, collections owned by the US National Institute of Health (NIH) are managed using available and applicable appropriations through the NIH Institutes and Centres or the Office of the Director's administration budget or programs.

Funding lessons

Forecasting operational costs is important but difficult

The NIH recommends investors and custodians develop realistic long-term projections of operational costs, and use these estimates to inform C&D management and funding strategies. This practice, however, appears not to be widespread or well implemented.

Custodians in many countries are concerned about the sustainability of funding when the majority of support is capital rather than long-term operational funding. For instance, custodians in Australia have noted that the absence of operational funding has significant long-term implications for research infrastructure. The University of Texas also commented that a lack of long-term funding is a significant challenge for the survival of both collections and databases.

[17] See for example *'Significance 2.0: a guide to assessing the significance of collections'*

The 'best' funding model depends on the circumstance

There are different ways to fund C&Ds. The most appropriate model to use depends on the circumstances and characteristics of the collection or database in question.

Examples of funding models used overseas include:

- *No Charge, user registration*. Under this model C&Ds are publicly funded. Users are not charged for access but are required to register. Registration allows use of the database to be tracked (albeit at a high level). Registration also helps differentiate users according to their data requirements and risk profiles.

  The UK Data Service (UKDS), for example, is funded by renewable grants from the UK Economic and Social Research Council (ESRC) together with contributions from the Joint Information Systems Committee, and universities (currently the Universities of Essex and Manchester). Access to the data collections is available to anyone upon registration but there are different levels of access according to the security of the data and the conditions placed upon the registered user. All projects funded by the ESRC are obliged to present the collections and data they generate from the project to the UKDS for assessment and inclusion in the collection.

  This model is commonly used when there is a strong rationale for public funding and where custodians need to control access to sensitive data. The commitment of funders to promote the sharing and re-use of publicly funded data is also important.

- *Deposit fees*. Dryad is a community run, not-for-profit data repository service for the global bio-life science community[18]. Dryad aims to facilitate the re-use of scientific research data. Users can deposit their data and 'grey material' (project documentation) into the repository with one-time deposit costs. The Archaeological Data Service[19] operates a similar model to Dryad, although they also operate a commercial service for non-academic operations (eg industrial archaeology units).

  This model is best suited to situations where the research community strongly supports open access to data (or where they are required to deposit their data as a condition of funding).

- *Cost recovery for value-added requests (partial cost recovery)*. Under this model, publicly funded custodians are obliged to make their collections available for use. While access is mostly free, custodians are permitted to recover the costs of value-adding services that go beyond standard access requirements (ie additional short run marginal costs).

  This model works best when there are large public benefits from the provision of collections but where non-standard requests impose high costs on custodians (and where the public benefit of a non-standard request is low). For example, the UKs Natural and Environmental Research Council fund data collection from sophisticated instrumentation and operational infrastructures (such as remote monitoring stations). Data from completed projects are maintained in data repositories and made available on request.

[18] https://datadryad.org/
[19] http://archaeologydataservice.ac.uk/

Partial costs are recovered from making data available and sometimes a contribution to the preservation costs.

- *Full cost recovery*. Under this model, users pay all the costs of providing a collection or database, including a portion of all overheads (such as rent and machinery depreciation). Full cost models can apply a standard (flat) fee to all users, or be calculated on a case-by-case basis (variable charges).

  Full cost recovery models work best when the party accessing the collection or database is able to capture all the benefits of access. This is often, but not always, commercial entities. For instance, a number of academic organisations offer full commercial services to industry. These services are supplied in a competitive market and attract full recovery of cost. For example, the Australian Synchrotron has a spectrum of commercial services that range from access to beam time, through to managing samples and analysing data.

  Guidance on the application of full or partial cost recovery is provided in many countries. The guidance provided depends on the circumstance, for instance free access for educational purposes, nominal fees for funded research, market value for commercial access. Some countries also include access conditions within relevant legislation (eg the UKs Environmental Information Regulations 2004[20]).

- *Public-private partnerships (PPPs)*. Some C&Ds have both commercial value and broader public benefits. Public-private partnerships are long-term contracts for the delivery of a service, where the service requires the construction of a new asset or the enhancement of an existing asset. In the context of C&Ds, this may involve the private sector financing the construction of new data infrastructure, in return for the right to charge an access fee for a given period. Once the initial investment is recovered (and a rate of return delivered) ownership of the asset would be handed over to the public sector. In the US and UK, several biobanks are funded through PPPs.

Technological lessons

Many countries have open data or open science policies as a condition of funding

Most OECD governments are working towards open access data policies for C&Ds for which they are the custodians or that are publicly funded. Further, many major international funders have made open access to data a requirement of funding. For example, to receive funding from the National Science Foundation (NSF) it is a prerequisite that researchers make project data freely available. One instance of this occurring is the Natural Hazards Engineering Research Infrastructure, which is accessible to researchers worldwide free of charge and which is funded by the NSF.

Similarly, the National Academy of Finland mandates that publication outputs from projects receiving their funding are published in open access journals and that data is also stored and made available through relevant national or international data archives.

[20] http://www.legislation.gov.uk/uksi/2004/3391/contents/made

Open access from shared international research infrastructure poses unique challenges, such as the need to balance national interest against the interests of specific research groups. This introduces a greater level of complexity in international collaborations. Physical and digital access to scientific collections must be balanced against preservation, privacy, proprietary and security concerns, or budgetary and human resources consequences.

Some countries, such as China, have clear open science policies and require or encourage papers arising from publicly funded research to be made publicly available (for example published in open-access journals). However, not all countries considered have clear policies on the accessibility of the underlying data.

Other countries are grappling with indigenous data sovereignty

Access policy around C&Ds of significant value to indigenous peoples varies widely internationally. Many museums and other institutions that hold cultural valuable physical collections work alongside relevant indigenous peoples to apply culturally appropriate practices to their custodian efforts. However, this is not often the case with databases or scientific collections.

Indigenous data sovereignty is an emerging policy area and is yet to be addressed in any systematic manner, although New Zealand, Australia, Canada and the US are making some progress in this area. Groups such as the US Indigenous Data Sovereignty Network have made good progress in promoting the application of indigenous data principles. However, to date indigenous issues are largely unaddressed in C&Ds investment plans or strategies.

5. Towards a better approach

The current approach of allocating RSI funding to specific C&Ds based on a concept of national significance creates a number of challenges, and it may no longer be the most appropriate or efficient approach. This section outlines principles that can apply to all public investments in C&Ds, and highlights areas where change could be most useful. This section will form the basis of policy development for a new approach, including consideration of whether linking funding to specific C&Ds and the concept of national significance is still useful.

Principles to guide investment and management of C&Ds

General data principles

A starting point for all C&Ds is adherence to the government's overarching principles for managing data and information (approved by Cabinet, August 2011 (CAB Min (11) 29/12)).

These principles sit nicely alongside the FAIR and NZGOAL principles. However, for simplicity, it is proposed to apply the **New Zealand Data and Information Management Principles**[21] (details are provided in the box below.) supplemented by principles specific to public investments in C&Ds that are outlined below.

[21] Recognising that these may need to be updated or applied in a nuanced way to C&Ds.

New Zealand Data and Information Management Principles

Principle: open

Data and information held by government should be open for public access unless grounds for refusal or limitations exist under the Official Information Act or other government policy. In such cases they should be protected.

Principle: protected

Personal, confidential and classified data and information are protected.

Principle: readily available

Open data and information are released proactively and without discrimination. They are discoverable and accessible and released online.

Principle: trusted and authoritative

Data and information support the purposes for which they were collected and are accurate, relevant, timely, consistent and without bias in that context. Where possible there is an identified authoritative single source.

Principle: well managed

Data and information held and owned by government:

- effectively belong to the New Zealand public
- are a core strategic asset held by government as a steward on behalf of the public
- should only be collected or generated for specified public policy, operational business, or legislative purposes.

Agencies are stewards of government-held data and information, and must provide and require good practices which manage the data and information over their life-cycle, including catering for technological obsolescence and long-term preservation and access. Good practices also include collaborating with other agencies and the public, facilitating access, strengthening awareness, and supporting international cooperation.

Agency custodians must implement these practices on a day-to-day basis.

Principle: reasonably priced

Use and re-use of government held data and information is expected to be free. Charging for access is discouraged.

Pricing to cover the costs of dissemination is only appropriate where it can be clearly demonstrated that this pricing will not act as a barrier to the use or re-use of the data. If a charge is applied for access to data, it should be transparent, consistent, reasonable, and the same cost to all requesters.

Principle: reusable

Data and information released can be discovered, shared, used and re-used over time and through technological change. Copyright works are licensed for re-use and open access to and re-use of non-copyright materials is enabled, in accordance with the New Zealand Government Open Access and Licensing framework.

Data and information are released:

- at source, with the highest possible level of granularity
- in re-usable, machine-readable format
- with appropriate metadata
- in aggregate or modified forms if they cannot be released in their original state.

Data and information released in proprietary formats are also released in open, non-proprietary formats.

Digital rights technologies are not imposed on materials made available for re-use.

Additional principles for public investments in *scientific* C&Ds

In addition to these overarching principles, the following principles could be applied to public investments in scientific C&Ds:

Principle	Description
Focus on enabling excellent, high-impact science	The scientific value of a C&D should be the primary driver of funding (as opposed to the cultural or heritage value[22]).
Focus on public good outcomes	C&Ds should primarily focus on the production of public good outcomes (as opposed to private interests) through research and its application.
Relevant to long-term, strategic needs of New Zealand	C&Ds should be of relevance to the long-term, strategic needs of New Zealand, including supporting relevant legislative and international obligations.
Commitment to Māori data sovereignty	Custodians should make measurable shifts towards best practice for Māori-relevant material by partnering with Māori and actively protecting their interests. This should be achieved by involving Māori in decision-making for Māori data and appropriately funding such activities.
Sustainability of funding	Baseline funding arrangements should be secure enough to allow curators to develop and implement long-term curation and maintenance plans. Funding levels should be sufficient to cover an efficient level of enhancement activity.
Responsive to change	Investment in C&Ds should not be static. Rather, baseline funding should adapt to changes in the external environment. This implies the need for systematic review of the investment approach.

[22]Noting that some C&Ds will have a broad range of value propositions, scientific value needs to be clear in order to be classified as a *scientific* collection or database and to be guided by these principles.

Incentivise good practice	Investment should be used to incentivise commitment to FAIR data principles, NZGOAL and New Zealand's Data and Information Management Principles. This implies the need for up-to-date technology and management practices, the ability to integrate specimens and data from multiple sources, and protection from destruction (accidental or otherwise).
Transparent and accountable	The outcomes sought by funders should be clearly articulated and curators should be held accountable for the outcomes they planned to achieve.
Well-informed	Funding decisions should be based on sound information about the current and potential future users of C&Ds. The consequences of funding decisions should be understood and acknowledged.

Possible changes to the current model

This review suggests C&Ds system could be improved by:

- **clearly articulating cross-government expectations** around the management and outcomes of publicly-funded C&Ds and linking these expectations to funding
- **increasing the flexibility of funding** to respond to opportunities and the changing value of specific C&Ds over time.
- **strengthening incentives for collaboration**, connectivity and cohesion across similar organisations or disciplines (led by organisations with a strong C&D management culture)
- **custodians adopting a more strategic approach** to managing their C&Ds, enabled by more flexible funding and strong guidance from government as the investor
- **improved collection of information on users** (ie who is accessing C&Ds and for what benefit)
- **conducting periodic reviews** of the funding system to ensure expected outcomes are being achieved
- **establishing a mechanism** whereby custodians can seek funding for discrete projects that add value to existing C&Ds thereby enabling greater use and impact
- **increasing international collaboration**, particularly with Australia, such as through exploring the possibility of shared infrastructure, capability, and standards, and increasing New Zealand's contribution to international C&Ds and standards for data collection and management
- **where appropriate, unlocking the potential of Māori data** by partnering with Māori in decision-making on Māori data use and management.

Where possible and appropriate, work and consultation led by other organisations, such as Genomics Aotearoa and Statistics New Zealand's work on Māori data principles and management, should be leveraged to reduce duplication of effort.

Complementary, system-wide actions

Other actions within MBIE's mandate are system-wide rather than through SSIF. Implementing stronger data management practices at the research proposal stage for all MBIE-managed funds would ensure that data generated from research projects are well-managed and available beyond the life of the project. Proposals could be required to include information about the data the project is expected to generate and the preferred option for what happens to it at the conclusion of the project. This could involve incorporating them into an existing publicly-funded database.

Next steps

The principles and changes in this report are preliminary. **We will now move into a more detailed policy development phase** to further explore the appropriateness of the principles and how best to shape investment in C&Ds in the future.