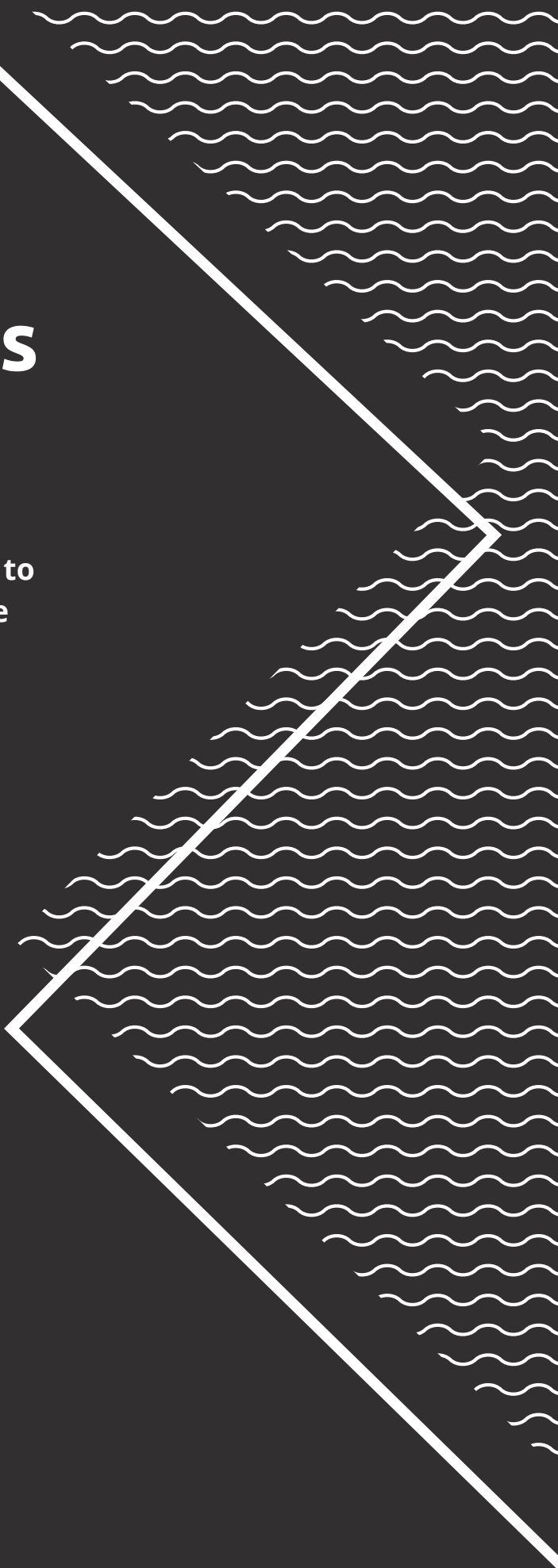
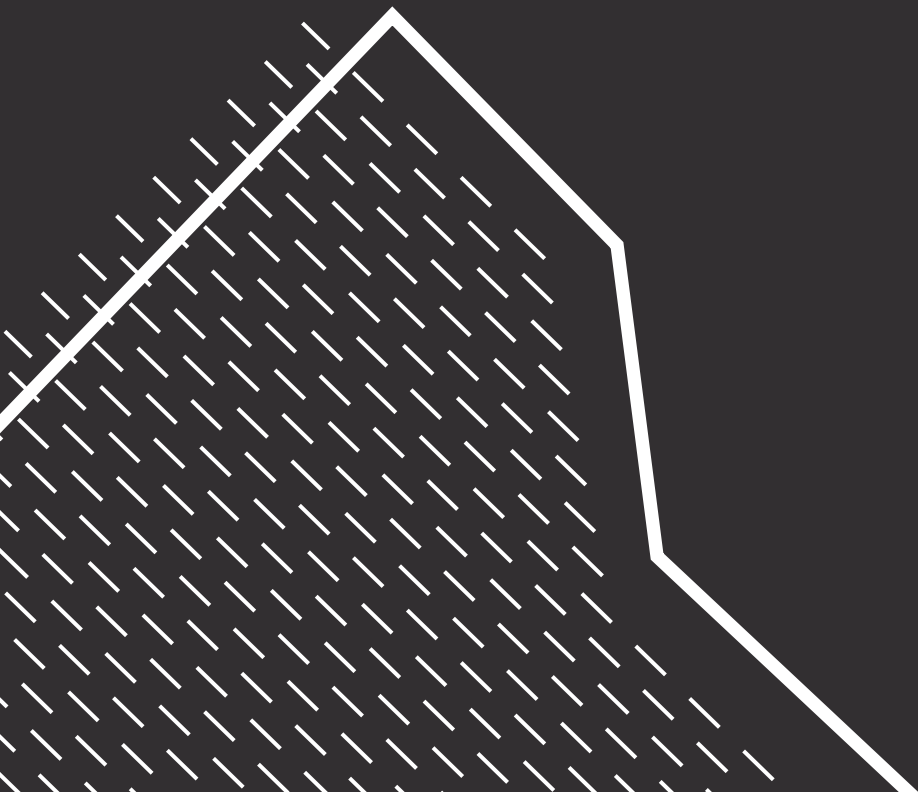


Version 1.3

Data Commons Blueprint

A high-trust, lower-cost alternative to
enabling data integration and reuse



Aotearoa New Zealand

Data Commons
Project



datacommons.org.nz



*This work is licensed under the Creative Commons Attribution 3.0 New Zealand License.
To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/nz/>*

Contents

4	Preface
5	Acknowledgements
7	Executive Summary
10	Introduction

14 Reusing data is largely a matter of trust

15	The emerging challenge to integrate and reuse data
17	The value proposition of data reuse
20	Risks of greater data reuse
21	The current “Ownership Model” does not encourage reuse
24	Breaking open the data reuse dilemma

26 A commons-based approach to data reuse

27	Managing the commons: the writing of Elinor Ostrom
29	Background to the commons-based design principles
37	Forming a Data Commons

40 Co-designing the commons protocols

41	Forming a Data Commons
42	Forming a Community of Interest around assets data
44	Understanding diverse interests in the commons
48	Evolving the commons: top-down and bottom-up
49	How do you develop a set of “rules of the road”
50	Technology can be protocol-based or a point solution
52	Social and institutional protocols
53	The Protocol Stack
58	Protocol-enabled relationships
60	Spanning institutions, relationships, and capability

62 Kick-starting the Data Commons

63	Incentives and disincentives
65	Starting up, curating, and growing the network effect
66	Appendix: Two case studies
67	Biosphere Data Commons
75	Person Data Commons

Preface

In June 2016 the New Zealand Data Futures Partnership, the NEXT Foundation, the Bioheritage Science Challenge, and Inflection agreed to co-fund the development of a blueprint for an alternative model to enable data sharing in New Zealand.

The project had two deliverables:

Firstly, the formation of a network of interested volunteers in New Zealand who have a range of technical expertise and a deep interest in data integration and reuse. EXP Ltd (part of the Enspirial collective) was contracted to convene an open conversation, using online tools such as Loomio and Gitbooks, a series of workshops and interviews, and a two-day retreat with this group of volunteers.

The outcome of those conversations was the second deliverable, this Blueprint document. Here we outline the conceptual thinking behind the idea of a “Data Commons” approach. We describe the high-level design features that will make it work, and outline the steps required to build one.

In preparing this document, the working group’s goal was to provide what we believe is a safer, lower-cost and higher-value alternative to the current approaches to the challenge of data integration and reuse.

While there is still much work to be done, we believe this document establishes a model that is worth further investigation. The technical experts involved consider it to be relatively low-cost and technically feasible to prototype.

The New Zealand Data Commons Blueprint is published under an open Creative Commons license so that other people can extend the conversation. We are not the only group trying to solve the dilemma of how to use integrated data for public, economic, scientific and environmental good while at the same time managing the significant risks inherent in doing so.

Acknowledgements

Project Sponsors

- The New Zealand Data Futures Partnership
- The NEXT Foundation
- The New Zealand Bioheritage Science Challenge
- Inflection

We would like to acknowledge our project sponsors, particularly Devon McLean, Diane Robertson, and Andrea Byrom for their support and advice. Not only were they able to provide essential financial backing, they also gave the project credibility, contributed expertise, and connected us with other individuals and organisations that made important contributions to the work.

Writing Crew

James Mansell (Lead Author), Rob Laking, Billy Matheson, Rohan Light

Auckland Retreat Working Group Members

Graham Scott, Hayden Glass, Haydn Read, James Gordon, Pieta Brown, Rachel Knight, Rob Laking, Robert O'Brien, Rohan Light, and Veronica Bennett.

Wellington Workshops (IR and Macs Function Centre)

Guy Kloss, Austen Ganley, Roger Pech, Miriam Lips, Rob Laking, James Mansell, Alanna Krause, Billy Matheson, Sarah Habershon, Robert O'Brien, Graham Scott, Rohan Light, Emily Mason, David Rutherford, Peter Newell, Tony Burton, Richard Poulton, Mike O'Neil, Veronica Bennett, Joshua Vial, Dominic Tarr, Jo Scothern, Sarah Greenaway, Anne-Marie Brook, Pia Waugh, Hayden Glass, Robert Guthrie, Jestlan Hopkins, Matt Muller, Nicola Brown, Murray Price.

We would like to personally thank the working group that participated in the Data Commons Retreat and helped us to resolve the Blueprint document. We would also like to thank the many people who generously gave their time to participate in focus group meetings, online discussions, case study workshops, and one-to-one interviews. Without this support the New Zealand Data Commons project would not have been possible.

A special call-out for Robert O'Brien and Haydn Read, whose experience and expertise guided the group towards making this a scalable and efficient alternative.

As part of the programme, we interviewed a range of people, and presented, attended or observed meetings with stakeholder groups who shared their challenges and hopes for doing data reuse better.

Landcare Research Workshop (PFNZ) Rachelle Binny, Bruce Warburton, Aaron McGlinchy, Nick Spencer

eDNA workshop (Auckland University and Landcare Research);
Robert Holdaway, Thomas Buckley, Austen Ganley

Platform Trust workshop Marion Blake, James Gordon, John Cook, Karla Bergquist, Kelly Pope, Naomi Cowan, Richard Woodcock, Robyn Shearer, Phillipa Gaines

Manaiaikalani Trust Visit and Q&A in regard to data challenges

Predator Free NZ Scientists' planning meeting at DOC in Wellington

Presented at Tahatu Rangi, October, Te Papa, Wellington

We would also like to thank Sarah Habershon from OptimalBI, and Alanna Krause and Rob Guthrie from the Loomio team, for their work helping us set up the project.




Executive Summary

The central challenge is trust. Data integration and reuse at scale can create significant value for all parties – data contributors, and data reusers – but only if people can create and maintain a high-trust relationship in regard to the transactions they are participating in.

The Data Commons working group has concluded that existing models for enabling data integration and reuse fail because they do not address this central challenge. The dominant approach tends to build technically focused point solutions that are highly specific to the particular context they are operating in. Moreover, data reuse interests tend to address only their own needs – frequently overlooking the interests of the data contributor. At best there is lip service to consent, minimal personal control for the contributor, or at worst coercive harvesting of data. Because these attempts fail at trust, they become costly and hard to scale.

The alternative proposed here is to establish a Data Commons. A commons-based approach builds trust and scalability into the DNA of the solution. This is achieved by adhering to a set of principles and goals which embed an inclusive and open approach to data for everyone who is participating in the commons. In addition, by setting up a “protocol-based approach” the Data Commons is scalable and lower-cost.

The Data Commons exists primarily to maximise the value of the participants’ data *for* the participants, and it is co-designed and co-governed by them. Moreover, the design aims at creating a data reuse ecosystem that rewards and encourages data reuse, rather than the on-selling or trading of data. The benefits of specific data reuse are valued (and in some cases sold for profit), but the data itself is not traded or owned.



In our proposal, data is treated as a common-pool resource. This is quite different from existing models that seek to either control or trade data based on agencies with exclusive monopolising interests in data reuse. The principle of universality encourages a protocol-based approach to the rules and technology, such that the solution becomes low-cost and easy to scale.

The work of building a Data Commons approach will involve two parallel processes:

Co-designing the Commons Protocols:

Community-forming and alignment around the Data Commons principles, and then co-design of data reuse protocols – from technology protocols through to social protocols.

Kick-starting the Commons:

Deploying specific high-value data reuse solutions that use the Data Commons protocols as the basis for their relationship with the commons community.

A Data Commons approach requires forming a Community of Interest around the high-level Data Commons design principles, and then facilitating more detailed conversations about how that community wants to manage data sharing and reuse through developing the community standards, institutions, and protocols to make high-trust sharing easy.

The outcome of these conversations about “how we do things around here” is a set or “stack” of protocols that participating organisations and individuals can commit to. The Data Commons Blueprint outlines seven challenges (or layers) that make up the “Protocol Stack” that underpins the Data Commons. This is how we enable high-trust and high-value data reuse transactions to take place across the community and between its various interests.

At the same time, there is another kind of work that needs to take place which involves building value in the commons. This is done by identifying, inviting, and supporting innovators and entrepreneurs to kick-start specific data reuse solutions that are based on these commons protocols. We need to build some data reuse opportunities that are valuable for members of the community, so that they will use them. This involves recruiting people and organisations who have pressing data integration and reuse challenges, and supporting them to use the Data Commons protocols to build their data solutions. This adds both data and users to the Data Commons and makes it more valuable for the next innovator, who now has even more data to work with, and so grows the value of the commons.

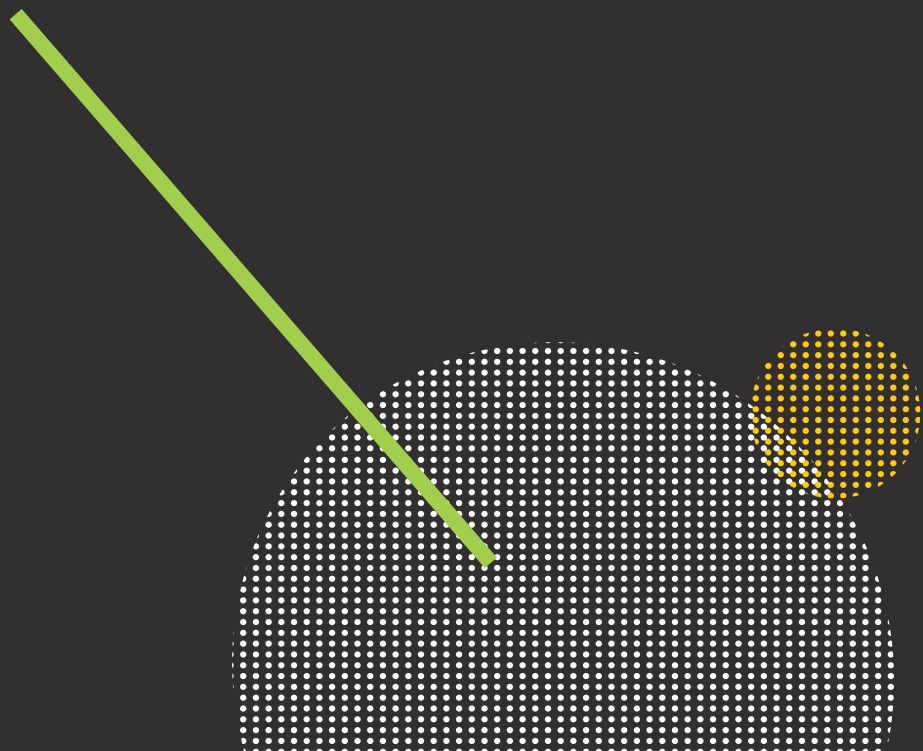
Key questions here are how and where to start to create early value, and how to enable the commons-based approach to be attractive to other participants. How best to add value to the community to establish a network effect to grow the value of their shared asset? How to convince potential data reusers that it is more valuable to build a commons-based approach rather than a one-off point solution? This will be harder in the first instance and will take a leap of faith, since there is initially no valuable data on the commons. It will get easier over time as the perceived risk of being on the commons is seen to be outweighed by having access to a wide variety of integrated data to develop high-value products and services.

We need to build some data reuse opportunities that are valuable for members of the community

To summarise: any group of organisations with an interest in data integration and reuse can form a Community of Interest to co-design the set of objectives and commons protocols (the “Protocol Stack”) that they think will best meet their needs. At the same time, that community needs to kick-start the commons by identifying specific high-value data reuse opportunities and building these on the Data Commons.

The building of generalisable commons protocols (top-down work), and the business of encouraging people to meet their data reuse needs via the Data Commons (bottom-up work), are ongoing and iterative.

As more data becomes available on the Data Commons, it becomes more valuable to the community of scientists, ecologists, entrepreneurs, and social investors and activists who will seek to use it and in turn add their data to a high-trust data ecosystem.



Introduction

Most data in today's world is controlled by large private and public organisations who in practice regard the data they collect as their own private property. They share this data only within narrowly defined parameters where the value of reusing data is only available to themselves.

Our sponsors and a community of experts and data stakeholders commissioned this project to develop an alternative model for scaling up safe reuse of data.

Over a six month period we explored a wide range of ideas, issues, and use cases around the Data Commons concept and summarised the results of that conversation in this blueprint. This blueprint details the new proposed "Data Commons" model. It is available now as a snapshot of where we have landed at this point in time and for potential data reuse interests to take forwards into prototype and testing.

Our sponsors agreed at the outset that this paper should be available under Creative Commons license to further the general open data movement, and to support others with similar ideas or facing similar challenges, and to aid transparency and public scrutiny of the proposal as it is developed. We hope this conversation, and the network of interest around this project, continue to develop and build on the approaches suggested here.

A Data Commons, simply put, is a way that communities can agree on how to share their data, add to the value of their data over time, and manage the risks of its integration and reuse. Through the establishment of a Data Commons, a wider group of potential data reusers can realise more of the value for themselves and their communities safely and in a way that is high-trust and mitigates the risk of misuse.

The document describes how a data integration and reuse solution founded on commons principles can enable individuals and organisations to work together to more effectively share, reuse, and integrate data in a high-value and safe way.

Any commons is formed by a Community of Interest for mutual benefit. Overall, our conclusion is that the work of enabling high-value, high-trust data

sharing is largely community-forming work (rather than technology work). The challenge of data reuse is the challenge of managing interests, and that is a relationship challenge, not a technology challenge.

Too often we see people leaping to technical point solutions without laying a solid foundation about how to establish the social protocols for how and when data will be used. There is some great new technology (such as the blockchain) that may unlock exciting new potential. But all of that is pointless unless communities of practice, often with diverse and divergent interests, can work together to establish collective rules for a shared common-pool resource. Unless all parties feel good about sharing their data, they will be unlikely to do so. Attempts at coercion lead to poor data or no data. A model where data is fenced off as private property reinforces silos of competing interests rather than data integration or sharing.

This blueprint is our first attempt at a fundamental rethink of how data reuse might be enabled. This is a rapidly emerging field of practice and we are still very much at the stage of feeling our way forwards, but there is a lot of great work being done that we can build on. Besides the community that was formed to explore this blueprint, we refer both to other communities internationally who are embarking on a similar journey for similar reasons, and to a wealth of examples of proto-commons-based ways of enabling data sharing. Here we present our first version of the blueprint – with, we hope, enough substance to get people interested in taking this from theory into practice in the coming year.

In Section One we examine the risks and benefits of data reuse and conclude that the central challenge is building trust into the system. In doing so, we explore exiting practice around data integration and examine why it is hard to scale or fails to lead to comprehensive data integration.

Section Two introduces the six design principles and objectives of an alternative model for enabling data integration and reuse. Here we introduce the core ideas behind the Data Commons approach.

Section Three introduces the first part of the work of building a Data Commons: community-forming and co-design of the community protocols that underpin the social contract governing the commons. This section introduces a framework for thinking about data interests – who needs to be at the table? – then introduces the notion of a “stack of protocols” that needs to be co-designed by those interests to allow the commons to function effectively.

Section Four outlines the second part of building the commons: kick-starting use of the commons to drive value.

The appendix contains further notes about two Data Commons case studies: one for person data and one for biosphere data. These are not fully developed case studies but reflect some of the Data Commons aligned thinking applied to particular classes of data.

**Enabling
high-value,
high-trust
data sharing
is largely
community-
forming work**

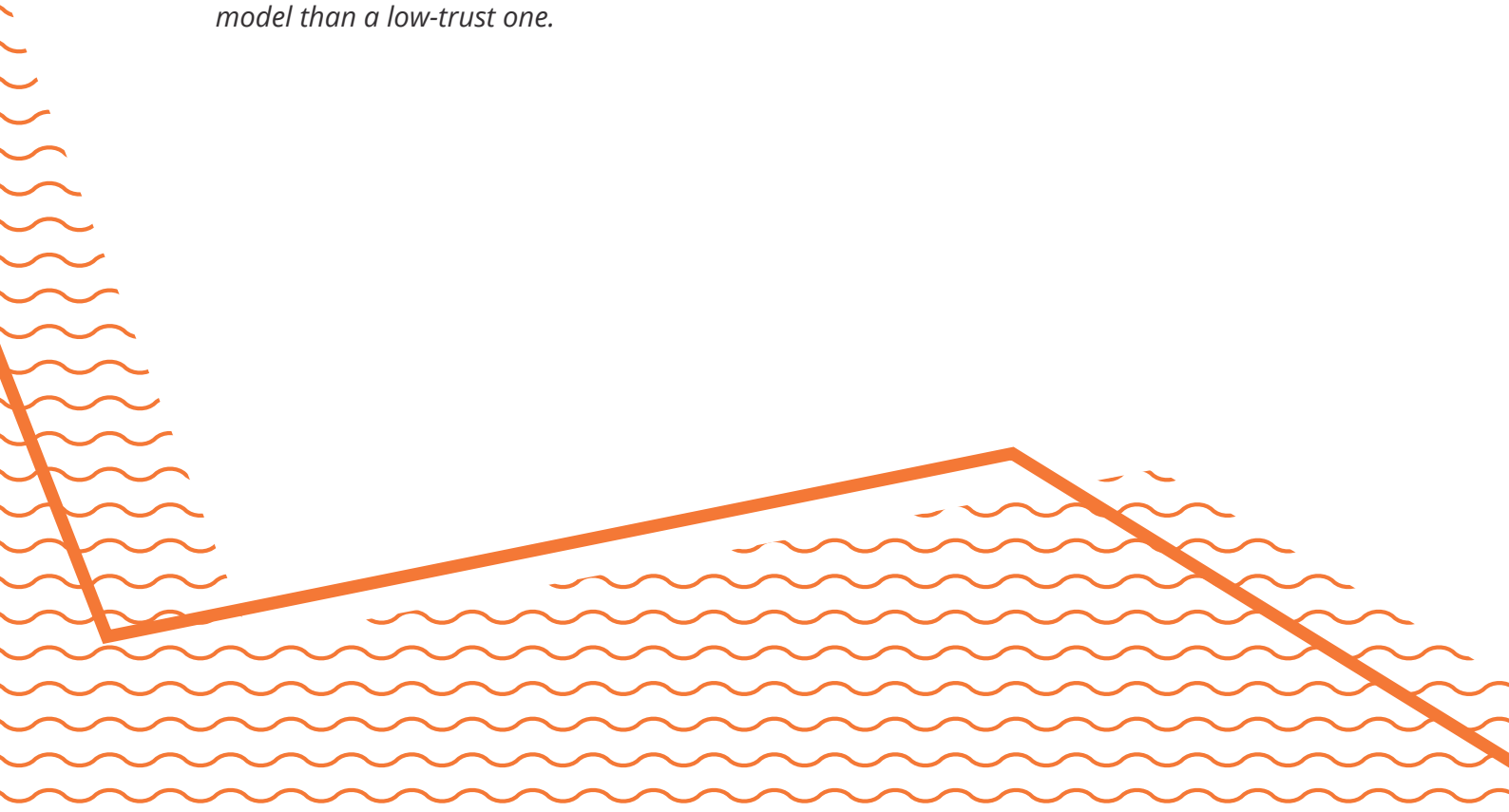
We hope that by proposing this alternative model for enabling data reuse we can reset the debate. It is a false dilemma that we must either be coercive to harvest value from data, or give up on the value proposition because the risk is too great. That only appears to be a dilemma when you don't directly address trust. There will be higher value for a wider Community of Interest where there is higher trust.



Section one

Reusing data is largely a matter of trust

Data integration and reuse is widely recognised as a way to drive social, economic, and scientific outcomes that improve people's lives. However, it is not without risk. We believe that current models that focus on data ownership, minimal contributor consent or control, and short-term commercial benefit do not work as a means of encouraging data integration or reuse. The fundamental challenge is enabling trust through forming relationships that enable (or disable) data reuse. There is more value to be gained in a high-trust model than a low-trust one.



The emerging challenge to integrate and reuse data

Data integration and reuse is widely recognised as a way to drive social, economic, and scientific outcomes that improve people's lives. However, it is not without risk.

We believe that current models that focus on data ownership, minimal contributor consent or control, and short-term commercial benefit do not work as a means of encouraging data integration or reuse. The fundamental challenge is enabling trust through forming relationships that enable (or disable) data reuse. There is more value to be gained in a high-trust model than a low-trust one.

Technology and digital media are transforming the world we live in, and offering us a potentially far more responsive, effective, transparent, and accountable approach to business, civil society, and government.

This transformation is fuelled by unprecedented amounts of data and information. While 'big data' promises a more prosperous, just, and equitable society, as with any innovation there are both risks and benefits. Big data can just as easily be used to steal IPs, erode commercial interests, and steal hard-won academic research data. State sector use of big data can intimidate citizens and unintentionally or intentionally target marginalised communities.

Practices for the safe management of personal and commercial, creative and scientific knowledge have built up during the last hundred years. These practices around privacy, commercial sensitivity and secrecy, and intellectual property balance personal and public interests, commercial and public goods. For example, in health research, bioethics practices inform how to obtain consent and undertake research safely to make scientific progress. Intellectual property law seeks (not always successfully) to balance openness and shared discovery with the need to permit commercial rewards in reaping benefit from investment in discovery or artistic creation.

These regimens work for the most part as an evolving system that enables data production and use, but they are woefully unfit for what is happening now. In today's world, data is networked, easily transmitted, and copied at almost zero marginal cost. What is more, with the advent of digitisation and wireless sensing, the cost of initially capturing data has also dropped to near zero.



In the past, when thinking about information sharing, we were largely referring to finished knowledge products such as scientific papers, patents, songs, or movies. But what happens when people want to integrate and reuse the low-level raw data? There may be value to be found, not just in the scientific paper but also in reusing the individual genome that was gathered as part of the research; or not just in information about your overall income, but in each financial transaction you make.

To understand how value arises from sharing data, and also the risk and how to manage it, it's helpful to distinguish clearly between data sharing, data reuse, and data integration.

Data sharing is simply the transfer of data between actors. A doctor may share your data with the surgeon in hospital to co-manage your health condition. The additional value from sharing arises because the data is then reused or integrated with other data. But reuse or integration may also carry risks – and this is where the challenge in sharing arises.

Data reuse is what happens when shared data is used for another purpose – for something that was not intended when the data was first collected. Sometimes data is not shared with another person, but is repurposed by the original holder of the data. I may have collected your email address for the purpose of providing an email service, and now I want to target advertising of other products to you using that email. Reuse of data also includes sharing it with somebody else who repurposes it. If a doctor shares your medical information with the government to help understand benefit liability, then that is repurposing it. Data reuse is what drives both potential for value and many of the concerns about risk of sharing.

Data integration is what happens when we link bits of data together to understand the relationship between them. An example might be to join your personal health data to information about your lifestyle to better understand your health risks. Another example might be a scientist linking environmental DNA samples with data about pest numbers in a national park to get a more complete picture of the future biodiversity of the park. Integrating data can let us answer questions that we previously couldn't answer – it creates a bigger picture of what is going on.

Reuse, repurposing, integration: they are all aspects of the basic idea that the value of data lies in its use and that further value can be added by its reuse. The value and risks are all based on the insight gained from the reusing or integrating. I can do new things, make different decisions, automate decisions by integrating and reusing data. In this document the focus is on data integration and reuse, hereafter termed Reuse.

In this
document
the focus is
on data
integration
and reuse,
hereafter
termed
Reuse.

The value proposition of data reuse

Data is already being Reused every day by individuals, businesses, scientists, researchers, and government organisations that have an interest in using it to make better decisions and add value to our lives.

After all, knowledge is central to all social, scientific, environmental, and economic activity. Data is at the heart of informed decision-making. For this reason alone, the potential benefits of increased data reuse and integration are likely to be enormous, and will affect all aspects of our lives.

Recent advances in computing and digital technology, wireless networking technology, and the miniaturisation of electronic sensing technology have led to an unprecedented ability to collect and process data almost instantaneously. These new technologies are already helping us to learn faster, better manage many kinds of risk, connect, collaborate, and form communities of shared interest. The evidence is all around us of the potential benefits for New Zealand of using data to drive innovation and economic growth, to provide better commercial and public services, to protect the environment, and to promote democratic participation and engagement.

Here are a few examples:

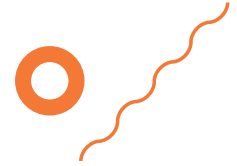
- 1** If researchers can use your lifestyle data (such as what you eat), link this to sensors you wear (to monitor blood chemistry, temperature, and heart rate), and link your personal data to government health data (the genome and the medical history of people and their families), they will likely find patterns that can potentially predict heart attacks in advance, or find early signs of diabetes or cancer, or track and reduce the spread of influenza or bird flu. If that integrated view can be used at a personalised level and made available to your GP, you then have deeply personalised evidence-based health care. Your Apple watch would be more than just a toy on your wrist.
- 2** The government wants to use integrated data to learn more accurately which of its social services achieve better outcomes. An integrated profile of a citizen allows the government for the first time to be accountable for outcomes because, over time, it can measure the actual effects of its services on individual lives. Access to citizen information by local services also probably means that more of the service design as well as delivery can be done outside of Wellington. As a result, the use of integrated data in government has the potential to create better outcomes for citizens, and a leaner, more effective, customer-focused service response.
- 3** If social sector groups have access to data collected by government, then policy will be better informed and more widely debated and not just left in the hands of a few Wellington officials and politicians. Mobilising support on local issues becomes easier with easy access to integrated data that can be reused by a more diverse community of interpreters and interest groups.
- 4** Wireless sensors put into concrete slabs in new buildings will be shared with engineers (indeed quite possibly with anyone possessing a smartphone) who can check in real time on structural damage from an earthquake and thus determine whether a building is safe to re-enter. If this data is widely available, then employees might be able to better plan for and respond to emergencies. Data from motion sensors can already be integrated geospatially with Geonet 'quake monitoring results so that scientists and engineers can better model the effects of 'quakes and understand how to manage the risks they pose for structures.
- 5** Where utilities companies, other businesses, and local government can share data on assets like pipes and roads, they can save money and time on maintenance and replacement. As a nation we can save on energy if businesses and utilities can share data and use smart meters and other smart appliances to make consumption smarter.

6 Shared data could make business and government more accountable for environmental impacts. By remote sensing water quality and making the data available to anybody, people can know if it is safe to swim in a river and close the loop back to those responsible for protecting water standards. Indeed, if sensors are placed at different points in a river or a storm water system, this will help identify polluters such as dirty dairying or people flushing toxic chemicals.

7 As one of the case studies in this report suggests, area pest eradication can be supported by communities with a common interest – DOC, regional authorities, local environmental groups, and individual householders – that collect and share data on pests and local ecologies for reuse by environmental entrepreneurs and scientists to plan and manage eradication campaigns.

8 The concept of a personal data wallet is that the data generated by all your personal transactions – with your bank, supermarket, energy company, electrician, doctor – is held securely by you and shared only with your consent. You can then agree to share it on your own terms and subscribe to services that help you understand and manage your personal budget and lifestyle choices. We are already seeing that Virtual Agents (like Suri) work better for you the more data they can refer to in your personal profile – which can be kept securely for you in your data wallet.

9 Duolingo crowdsources and reuses data from over 12 million people learning languages to constantly improve how it teaches them. “People with a profile like John’s like this kind of question, it keeps them engaged, so ask John to translate this.” This kind of self-learning integrated service is going to increase in value and sophistication. But much of this relies on mitigating the real commercial, personal, and professional risks that can arise if reuse is misuse.



Risks of greater data reuse

What all of these examples have in common is that to realise their benefits we have to share data.

Data can be integrated and reused in ways that are outside of your control and ways that can be personally, socially, or commercially harmful. The risks of this sharing are obvious in the case of deeply personal data. Personal data can be used by the people we know and by complete strangers to bully and blackmail us or steal our identity. It can also be used by government agencies or private companies to target us by using our data for enforcement or advertising. This constitutes manipulation – either psychological (if you are a marketer with no coercive powers engaged in customer behavioural engineering) or coercive (if you are a policy analyst in government pursuing social engineering).

But there may be other reasons, besides protecting personal privacy, why the producers of data see risks in sharing it. My data may be misused, affecting my reputation, reducing the value I get from exclusive possession, or breaching commitments I have made to others who supplied me with it. If you have commercial interests, these can be eroded by others with access to your data. Scientists spend a lot of time collecting reliable data and rightly have interests in reaping the value from publishing their research before rivals. Artists struggle to retain creative control over their work.

Clearly most of the current systems for management of data are dedicated to controlling these risks – but they rarely do so in the context of a commons. How – and why – access to and reuse of data is controlled goes a long way to explaining why creation of a commons faces real challenges.

The current “Ownership Model” does not encourage reuse

In the *private sector* a large and profitable industry has grown up around the collection, integration, and monetisation of data for marketing purposes.

This data reuse industry is based on an inherently extractive model, realising value from exclusive ownership. The data is obtained in exchange for “free services” such as email or search services or social media platforms. The same approach is also employed by non-digital services, such as buying shoes at your local shop and using a loyalty card. Digital footprints and fingerprints are traced and tracked as you shop and go about your daily life. The companies doing this want to build up the most complete profile of you that they can. Whoever does this best can realise a financial dividend by on-selling your integrated profile to marketing companies and other commercial interests for a profit.

This is a competitive and commercial approach to building up an integrated behavioural profile of our lives. The trouble is (apart from lack of control and privacy) that at the heart of the model is the fragmentation and siloing of data rather than sharing and integration.

The commercial imperative is to monopolise and own more data. Google extracts value from data for its shareholders by on-selling integrated profiles of its users to direct marketing companies. The same is true of Facebook and Loyalty New Zealand, and other integrators of personal data. These companies are competing with each other to monetise your profile, so they don't integrate their respective data assets (unless they get bought out). E.g. Microsoft just bought LinkedIn to add to its inventory of social network data.

The service user has little meaningful control over how that data is used and whether and to whom it is on-sold.

The *public sector* in New Zealand and elsewhere is falling into the same pattern.

Many government agencies have been, and remain, reluctant to share data with each other, partly because they are stewards of data acquired using the coercive powers of the state, but also from a bureaucratic instinct to stick to hierarchical lines and resist lateral collaboration. More recently, in an era of joined-up government and under significant pressure from the centre to work across organisational boundaries, at least some parts of the state sector have a newly developed interest in reuse and integration, particularly of personal data. There is genuine value to society in having a better understanding of life pathways, of what works, of how to invest better socially and economically. But the reuse of data in the public sector comes with risks. While sharing of personal information has to meet the requirements of the Privacy Act, there is still potential for misuse and increased marginalisation of individuals and communities.

For example, in New Zealand, non-governmental organisations (NGOs) are increasingly being caught in cross fire between citizens and the state. Agencies such as the Ministry of Social Development are increasingly demanding that third party providers (who rely on MSD contracts for their survival) hand over identifiable and deeply personal data about the people who use their service. So MSD is meeting its interests in obtaining and integrating data from third parties. But in doing so it may be overriding the interests of those third parties and their relationships with their clients.

The coercive approach to appropriating citizen data may represent high value to government in the short term, but it erodes confidence and trust between NGO providers and their service users who are the source of the data that government agencies value so highly. NGOs are already expressing concern that marginalised and 'hard to reach' communities may go further underground because they can no longer expect a confidential and high-trust relationship with the NGOs that were set up to support them.

In education, it is now possible to "tear down the classroom wall" (as one Ministry of Education official described it) through the use of an integrated profile of every learner. The information could be used for performance-based pay linked to how well a teacher does for each kind of student. But micro-level indicator-based control of teachers and students from central government is likely to reduce the quality of data and eventually stifle innovation and the quality of outcomes. People removed from engagement are poor decision-makers about data reuse and value since very often their interests do not align with the users of government services or the people who serve them directly. Getting elected or meeting the sector KPIs for career advancement is very different work from improving other people's lives.

Unfortunately, because of the power that public servants can have over people's lives, many citizens simply do not trust them to make good decisions



for other people about how their data should and should not be used. That decision is best placed back in the hands of the affected people who own their own data who can then share it exclusively with the people they trust. This is called consent – it is a basic principle, for good reasons, in privacy legislation but it can be expanded by embedding it in data management.

The reuse of data by business and government based on an Ownership Model has two fundamental problems that need addressing:

- You, as the user of the service in which your data is generated, don't get to use your integrated profile for your own benefit. The company or agency to which you surrendered your personal data has no interest in your realising the value of your own integrated personal profile to help you better manage your wellbeing, or your time, or your finances. Somebody else gets all the value from this fine-grained knowledge of your life.
- Society doesn't get to realise any benefit from your integrated profile either. People working to solve complex social challenges such as diabetes, homelessness, and child abuse cannot benefit from society's data about itself. Your data may have become the private property of a technology company and have to be bought back, or it may be locked up in different government sites by officials using it as an asset to further their own careers or by ministers who would rather not have the political risk often attached to transparency.

Whilst these kinds of concern are acute when applied to the reuse of personal data, similar kinds of concern emerge around data reuse for the scientific community, for professional interests, and for commercial interests. Scientific and commercial interests also have trust challenges with the reuse of data they (co)produce. How do I integrate and share my data in an environment of "publish or perish"? How can I be the first to market if everyone can use my hard-earned data?

In a nutshell, the current Ownership Model for data reuse is destroying value and frustrating collaboration. Entrepreneurs, politicians, public servants, community leaders, scientists, and citizens who can see the value in greater data reuse face significant institutional and systemic resistance, or they themselves feel the professional, personal, or commercial risks are too great. "How can I allow reuse without losing control over something I have an ongoing personal, professional, or commercial interest in?"

One of the fundamental barriers to greater reuse is lack of trust or control over reuse.

Should we acknowledge the risks and abandon the effort? Or perhaps further data reuse and integration for greater individual and community benefit is only possible with more private ownership and coercive appropriation? We're optimistic that the future for use of data for our collective benefit as a society is bright, but it will be based on the development of a new model of data sharing, reuse, and integration.

One of the
fundamental
barriers to
greater reuse
is lack of trust
or control
over reuse.

Breaking open the data reuse dilemma: trust, community, and commons

Most data solutions talk technology – the IT language of data acquisition, definition, platforms, and applications. That’s the wrong place to start.

The central challenge of enabling data integration and reuse is the challenge of trust and how to build that trust. Our basic assumption is that trust is embedded in a community with common values and institutions, so that building trust requires building that community and defining the relationships within it, in a contract for all community members.

The Data Commons is the way that a community shares its data. The commons principle implies a resource belonging to the entire community and shared equally by everybody in the community, subject to the community’s rules. It has a very different set of principles and objectives from the dominant Ownership Model.

What does a Data Commons approach to data reuse look like, and how is this different from the standard model?



Section two

A commons-based approach to data reuse

The “commons”-based approach to data reuse focuses on solving risk across competing interests through improving trust and increasing control by the participants. The value of data reuse is the focus, not ownership or trading of data. And it is governed as a shared common-pool resource but with licensed reuse according to a set of community-designed protocols that form the “community contract”.

Managing the commons: the writing of Elinor Ostrom

Throughout history, communities have shared common resources such as grazing land or local fisheries.

Land, forests, and fish are depletable resources, and the challenge for any community is to ensure that they are not overused. Nearly thirty years ago, Nobel prize-winning economist Elinor Ostrom argued, with many case studies, that most communities can evolve rules themselves for restricting access to commons to share the resource fairly and husband its use. An example close to home is the rules that iwi and hapu have developed for controlling access to their fisheries – including the rahui, or closed season, to allow the fishery time to regenerate. Ostrom argued that this form of self-regulation by communities was a viable alternative to both privatisation of the resource or rules laid down and enforced by outside authority – the lord of the manor or the government. She proposed some principles for community management of the commons, starting with agreement in the community on who would have access, when, and under what conditions, and how the rules would be collectively monitored and enforced.

More recently, Ostrom turned her attention to the application of the commons principle to the management of data or information. Here the problem is not fundamentally one of resource depletion. Data, information, and knowledge are different from natural resources: they are not depleted if used and their value to the community increases the more widely they are used. But with the growth of the Internet, people began to recognise that data shares some basic attributes with other resources. As Ostrom and her colleague Charlotte Hess wrote ten years ago:

There appears to have been a spontaneous explosion of “ah ha” moments when multiple users on the Internet one day sat up, probably in frustration, and said, “Hey! This is a shared resource!” People started to notice behaviors and conditions on the web – congestion, free riding, conflict, overuse, and “pollution” – that had long been identified with other types of commons. They began to notice that this new conduit of distributing information was neither a private nor strictly a public resource.

So data sharing and reuse have both value and risks for its producers. The community therefore needs rules to manage the risks and harvest the value – and they turn out to be quite similar to those Ostrom originally proposed for a natural resource commons.

Here we introduce several core principles that we think are likely to be fundamental to building a commons-based model of data integration and reuse.

Background to the commons-based design principles

The community design principles here build upon those first developed in the New Zealand Data Futures Forum (NZDFF) public consultation process.

The first NZDFF paper illustrated the balance between risk (fear of adverse consequences of increased data sharing) and value (desire to do more to improve lives). Too often this was seen as a dilemma: that protecting privacy or being risk averse necessarily squanders significant opportunities to realise sometimes life-saving value; or that to grab the value leads necessarily to trampling on human rights (including privacy) or commercial sensitivities. The NZDFF challenged New Zealand to hold both of those principles together at the same time. It argued that doing so was the only way towards the kind of data sharing ecosystem that we were aiming for – one that both managed risk and realised value.

With this in mind, the NZDFF came up with four design principles for a safe and high-value data sharing ecosystem. Two of these principles were focused on keeping the value side of the equation at the table; that (1) the data sharing ecosystem needs to direct value from data sharing back to its participants; and (2) it needs to be inclusive, thus providing shared value, not monopolised by partisan interests. Two principles were focused on keeping risk-control at the table: (3) the data ecosystem has to be high trust (I have confidence that the system will protect my interests); and (4) the system must be in the control of its participants (data sharing is something you do, it isn't done to you).

It was also concluded that if you achieve these four design principles for data sharing, that, far from being a dilemma, the result is a positive feedback loop. The more you enable value and inclusion, the greater risks people are likely to take to do more data sharing, because they themselves get that value from it. By the same token, the more you improve control and trust, the more people will be willing to try sharing to realise some potential value, because they are still in control and can reverse their decision if trust is eroded.

But this feedback loop can also spiral downwards. If you erode trust, people will cease sharing, meaning loss of value, making people more sceptical and so feeding back into increased mistrust – because the perceived risk outweighs value. That is basically the system dynamic that underpins the “Ownership Model” and why it tends towards fragmentation and mistrust.

These four principles are at the heart of creating (or destroying) a thriving data sharing ecosystem. The important point to grasp here is that it is not a dilemma, it's a feedback loop. You have to consider both sides of this equation together to build a thriving data sharing ecosystem. If you just think of value without trust, this will unravel. If you just focus on risk, you end up unable to realise value and so remain sceptical.

These principles were widely applauded by the likes of the UN Global Pulse and Privacy Commissioner who thought they added significant new thinking to working past the old confrontational debate between risk and value. However, the NZDFF did not have the time to figure out the next step: how to actually apply these in practice.

Further thinking was done in "Handing Back the Social Commons". Here it was argued that to apply the NZDFF principles required thinking about the way data could be managed as a common-pool resource rather than owned: additional data sharing activity could learn from the notion of common-pool interests and how these are managed.

You have
.....
to consider
.....
both sides of
.....
this equation
.....
together to
.....
build a thriving
.....
data sharing
.....
ecosystem.
.....

Principle 1

Data is a common-pool resource

The data in the commons is assumed to be a common-pool resource for the common good.

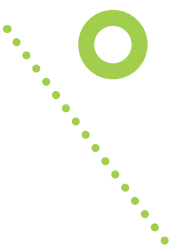
Data will be added to the commons where I get some personal, professional, or commercial value back by being able to meet my own purposes as a result of integrating the data. However, there is a common good in enabling that data to be freely available to other members of the commons – if they can meet my needs for safe reuse.

For reuses of data that don't intrude on my personal, professional, or commercial interests, data is held in common and available to all participants. Data integration and reuse may be exploited in various ways to improve lives and typically have many spillover benefits. It is intended that one of the primary purposes served by the Data Commons is enabling data to be managed as a common-pool resource for the common good.

A Data Commons is designed to be generative, in the sense that it is designed for the benefit of the community that owns and administers the platform. We assume that I and my community will get many direct and indirect benefits from improvements to science, society, health, the environment, and improved commercial and entrepreneurial activity from the reuse of data I have (co)produced.

For example, I indirectly get value back from the commons when a scientist uses the data to cure a disease. Similarly, science is the winner when scientists collaborate. We all win economically, through enhanced innovation and invention, when there is shared access to non-rival resources such as data.

The common good may be further supported by ways of building the commons that allow the value of reuse to be redistributed back to data donors directly. For example, reusers who make money from their innovations (e.g. sell "apps" that rely on commons data) may be taxed a portion of sale that is then redistributed back to the participants of the commons whose data has been reused. This allows net donors of data to also receive redistributed commons value, even though they don't develop their own on-selling reuses.



Principle 2

Value from reuse

Value is added to data by its reuse or integration with other data. The value to members of a commons flows from access to and reuse of other data besides their own, and also from the ability to do new things for themselves with their own data. The Data Commons is based on managing the value of reuse, not on profit from trading in owned data. Trading owned data for financial reward incentivises ownership, not integration and reuse. It is part of the current problem that the Data Commons seeks to overcome. The value of data reuse and integration should be an opportunity available equally to all participants of the Data Commons.

Too often data is surrendered (e.g. to marketers or government) in return for access to services (like email or income support respectively) who then capture the value of reuse and integration for themselves or on-sell it. The other participant in the exchange gets something else – but not access to the integrated use of their co-produced data. The value of reuse and integration of data should be available to all co-producers.

This principle of value will drive people to want to be included in the commons because they will receive the value of data integration and reuse for themselves. Google would get to use my profile, but so would I. As a scientist I can add my data to a shared pool and get more back in return. As a citizen I can add my profile and get new kinds of integrated data-enabled services.

In addition, what is rewarded is use of data, not ownership of data. Supporting the value of ownership (and trading data) encourages fragmentation and non-sharing. Rewarding the use of data encourages accessibility and innovation, driving better personal, commercial, and scientific outcomes. It discourages profit-taking from merely owning and on-selling.

This does not mean there is no opportunity to create financial reward from joining and using the commons. Quite the opposite. The fees paid for specific reuse applications will likely be higher than those relying on owned and still fragmented data. Trading a bunch of incomplete/fragmented data is not as valuable to the end user as integrated data. Integrated data offers improved insight and so improved opportunity for personal, scientific, economic or environmental benefits.

If you want to sell or buy access to data, go somewhere else.

Principle 3

Embedding trust

The Data Commons needs to build high trust into its DNA. Since the main ingredient to encourage and sustain data reuse and integration is the level of trust that the community's members feel in how their data will be used, then the Data Commons needs to build "trust-by-design" into all levels of ownership, protocol building, redress, technology, etc. Trust is the crucial test of the health of the community. Other principles follow from this basic requirement.

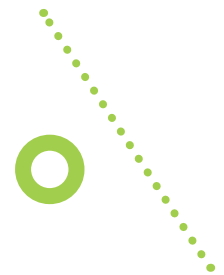
Principle 4

Participant design

Ostrom's first basic principle of a commons is inclusivity: its rules ought to be designed and agreed by all its members, not monopolised by single interests. The majority of interests should be represented at all levels of governance and protocol setting. Inclusivity is essential to trust by the members in their relationships with each other.

Failure to be inclusive or to eradicate partisan or monopolising interests will undermine the essential ingredients of the Data Commons: that it is high trust and for the common good. If state sector operational interests or big business interests were to design the rules of the road for data reuse and sharing, they would be very different than if NGOs or citizens were represented. We are already seeing the effects of that. By the same token, if research and scientific interests were to hold the pen, then operational or entrepreneurial interests would likely be squashed or subsumed in ways which were counterproductive.

Since the main challenge is forming high-trust relationships across multiple different interests, there is a need to enable co-stewardship to allow those interests to have a voice, to have access, and to receive value from the commons.



Principle 5

Participant governance and control

Data reuse in the commons should be deeply democratic: participant-controlled at all levels. People can vote on the board or with their feet. Participant control is central to realising value and managing and maintaining trust. The locus of control is placed in the hands of those most affected by the decision on sharing and reusing the data in question.

Data providers and co-producers have a higher level of interest in the way their data is used and what it is used for than third-party reusers. They are typically more affected by misuse and should receive more benefit from allowing more reuse. They are well placed to make the best decisions about reuse that affects mostly them.

The locus of control also mediates trust. Since trust is built and eroded between parties over many data reuse transactions, control over the nature of that relationship, and whether to increase or decrease the level of access, needs to remain with those who have most to gain or lose from data reuse. In short, I need to be able to terminate my relationship with you if you start acting in untrustworthy ways. Alternatively, I might want to dip my toes in the water to try something out and am more likely to do that if I am free to continue or opt out once I see what level of value I get.

Participant control is also central to allowing the Data Commons community to resist predation by monopolising interests. If big business or government seeks to monopolise the Data Commons, people can vote with their feet and leave, taking their data with them to form a community elsewhere.

Essentially the NZDFF recognised that data reuse, integration, and sharing is all about the kind of relationship that is possible between the provider of the data and re-user or integrator of the data. The form of that transaction matters. The NZDFF found that if the relationship was high-trust and in the control of the data provider, then it would be high-value, and that inclusivity, rather than exclusivity, would also improve the value obtained from data sharing and integration. The principles are in fact a recipe for the nature of that transaction, which is the core of the relationship. Control is an important input into rebalancing the relationship between centralising interests and democratic interests. It puts governance on notice that trust is contingent and needs to be earned and maintained.

Principle 6

Universality

One choice for a Data Commons is between a “point solution” – one tailored to the specific requirements of a narrow community and specific reuse application – and a “protocol solution” – one employing rules and specifications with more general applicability.

Many solutions today involve hardwired attempts to share data. Examples include the Statistics New Zealand IDI and most of the data integration projects within government, or the Loyalty New Zealand Fly Buys programme integrating personal data across fifty or more New Zealand companies. The same focus on bespoke solutions applies to Google or Facebook.

Programming and technology can hardwire virtually anything. So there is always the option to just build the specific Commons solution: buy a big computer and tailor individual data gathering and integration solutions to the available data, then build the high-trust, inclusive, participant-controlled, data reuse system in this “box” or “cloud”. But this costs more, is inflexible and is fragile. A point solution might be easier to prototype, but is often harder to scale. It also has a built-in centralising and so controlling tendency (though this can arguably be mitigated).

We think the better opportunity is to create a protocol-based generalisable approach. Technically, a protocol-based approach (as opposed to point solutions) usually builds more innovation, lower costs, scalability, and inclusion into the solution from the ground up. A protocol-based approach to the commons will be a slightly slower start, but far cheaper, more innovative and flexible, and more scalable in the longer run. It is a better solution for distributing control (mitigating some central actor tendencies). It also makes the solution scalable at low cost, since new interests are merely adding themselves to the network of other actors who adhere to the protocol. But there are more basic reasons for looking for generalisable solutions. The most basic has to do with the nature of data: the more opportunities for reuse and integration of data we can create, the better we understand the world we live in and the more value we are likely to add to our society. Each specific instance of a Community of Interest sharing and reusing data sits within a much larger data ecosystem, made up of all the ways in which the data we generate reflects the complex interconnections of the social, economic, and physical world in which we live. As we link that data together, we will find more connections – expected and unexpected.

So the rules for a Data Commons ought to be designed to facilitate scalability – widening the parameters of an existing Community of Interest, and trading/sharing data between Communities of Interest – to increase the opportunities for making these connections.

A point solution
might be easier
to prototype,
but is often
harder to scale.

At a technical level the solutions are available. A transparently published protocol will include things like “Application Programming Interfaces” and “Metadata standards” for the technology and data layers respectively. Protocols for both are being developed or have already been adopted. But what are more important for our concept of the generalisable Data Commons are the higher layers of these protocols. The proposal here is to create a method for people to form data sharing arrangements: not “the point solution” that is the place where this happens but a “market” where high-trust data integration and reuse can easily emerge, prosper, and be terminated between parties. So it might also include higher-level constitutional protocols: how we respect, manage, and make decisions about data reuse and manage the community. The more general – and widely accepted – these rules are, the easier it will be to facilitate both scaling up of, and trade between, Communities of Interest.

In addition, being protocol-based builds “distribution-by-design” into the DNA of the data reuse ecosystem. It is a better solution for distributing control (mitigating some central actor tendencies). It also makes the solution scalable at low cost, since new interests are merely adding themselves to the network of other actors who adhere to the protocol.

Forming a Data Commons

Once a Community of Interest aligns around a commons-based approach, how do you then in fact build a Data Commons?

There are two main steps to building a Data Commons:

1 Co-designing the Commons Protocols Community-forming and alignment around the Data Commons principles and then co-design of data reuse protocols – from technology protocols through to social protocols.

2 Kick-starting the Commons Deploying specific high-value data reuse solutions that use the Data Commons protocols as the basis for relationships with the commons community.

Forming a Data Commons requires forming a Community of Interest around the high-level Data Commons design principles and then facilitating more detailed conversations about how that community wants to manage data sharing and reuse through developing the community standards, institutions, and protocols to make high-trust sharing easy. The outcome of these conversations about “how we do things around here” is a set or “stack” of protocols that participating organisations and individuals can commit to. The Data Commons Blueprint outlines seven challenges (or layers) that make up the “Protocol Stack” that underpins the Data Commons. This is how we enable high-trust and high-value data reuse transactions to take place across the community and between its various interests.

At the same time, there is another kind of work that needs to take place which involves building value in the commons. This is done by identifying, inviting, and supporting innovators and entrepreneurs to kick-start specific data reuse solutions that are based on these commons protocols. We need to build some data reuse opportunities that are valuable for members of the community, so that they will use them. This will involve recruiting people and organisations who have pressing data integration and reuse challenges, and supporting them to use the Data Commons protocols to build their data solutions. This adds both data and users to the Data Commons and makes it more valuable for the next innovator, who now has even more data to work with, and so grows the value of the commons.

To successfully build a Data Commons, we need to create early value and enable the commons-based approach to attract new participants, establishing a network effect to further grow the value of their shared asset. This will require us to convince potential data reusers that it is more valuable to build a commons-based approach than one-off point solutions that address their immediate, short-term problems. It will be a challenge in the first instance and will take a leap of faith, since there is initially no valuable data in the commons. However, it will get progressively easier over time as the benefits of having direct access to a wide variety of integrated data and the ability to develop high-value products and services become clearer.

The co-design of the protocols and the deployment of specific solutions form the ongoing practice of the commons-based approach to data reuse. Community-forming around the commons protocol and kick-starting the commons are addressed in the next two sections respectively.

Section three

Co-designing the commons protocols

The core of the problem is the challenge of building high-trust relationships between the parties who want or need to transfer, integrate, and reuse what is sometimes highly sensitive data. Therefore the challenge of sharing high-value, low-risk data for reuse requires establishing a new kind of relationship, one based on an understanding of data as a common-pool resource that needs collectively agreed protocols so that it can be shared across the Community of Interest.



Forming a Data Commons

The community needs to establish suitable standards, procedures, and protocols to enable a trust market.

These form the basis of the social contract that will underpin the Data Commons. You must be able to trust – and therefore share and reuse data with – people you have not met, and correctly interpret data you have not seen before. A set of protocols needs to make these relationships easy to form and terminate.

Trust-building begins by ensuring that principles are discussed and protocols are developed openly and transparently in a conversation amongst all those interested in establishing the Data Commons. This Community of Interest needs to have the right kinds of conversation and activity at the right level across interest groups. The way community protocols are developed is just as important as the end result.



Case study one

Forming a Community of Interest around assets data

The best way to see how community-forming works is through a real example.

In 2011 Wellington City Council identified both a challenge and an opportunity. One of the biggest costs in managing a city's development is the management of investments in assets such as roads, pipes, and buildings. Typically, however, the data they used to understand and analyse and manage these assets was stored and analysed in siloes. "Historically water asset managers around here analysed pipes using the data one way, whereas building asset managers stored building assets data and analysed it to answer their specific building management questions another way. Roading asset managers similarly have a further derivation and silo of information and its use."

But the City Council had a broader interest. How do you effectively manage the lifetime and maintenance and investment in a broad portfolio of assets? It was very difficult at the time to integrate the data by location or service when it was all organised in silos. But there was potential for a different kind of management: "If I have to dig up this piece of road to fix this pipe, are there other works I can do at the same time in that location since it usually costs at least as much to dig up the road as it does to fix the pipe?" "How can I coordinate and manage and analyse shared road corridor space across all utility providers? And, if I'm a utility provider, how can I manage my own work programmes if each council has its own way of storing pipe data? It is almost impossible to integrate pipe data across councils to co-manage or learn from each other."

To solve this challenge, Haydn Read (Strategic Asset Planning Manager) organised a community to co-design metadata standards for pipes, roads, and other built assets. Metadata standards are a set of protocols for how data can be created, collected, analysed, and visualised to help asset managers and councils make informed investment decisions. If there are standards for storing "this is a pipe" data, then this makes it very cheap, easy, and efficient to integrate that data with other asset data adhering to the same storage and analysis protocol.

Wellington City Council recently did this for all data about infrastructure assets. An independent review by the New Zealand Institute for Economic Research (NZIER) found that this approach would save local government

\$10 million annually in improved asset management efficiency – and an estimated \$100 million in reduced IT costs due to a common protocol for integrating and using data about water, pipes, and roads. Not only could these councils answer asset management questions better and invest better in assets, the cost of integrating the data itself would be far lower in IT and related technical costs: anybody adhering to these standards would be able to integrate their data with existing data at low cost.

What is important for our purposes here is to note how the standards were developed. It was largely completed using a community co-design approach, with a group of practitioners who were “bound by a shared vision”. The development of community protocols for how data can be integrated and reused is the business of relationship-forming and consensus-building and co-design around a common objective.

There were several reasons for the success of this community:

- The community formed around a common objective to make all their lives easier by establishing metadata standards.
- There was a high level of pragmatism. The working practice of the group was to assume that, for all standards, assumptions could be reversed if found to be incorrect. But decisions had to be made, even when some people initially disagreed. Pragmatically, assumptions about standards were made on the 80/20 principle: they were likely to be imperfect, but adequate until such time as they were tested in operations. Peer review of initial uses of the integrated data standards identified that the logic was sound, and operational learning also uncovered the need to revise assumptions as more information became available.
- The community of co-designers were by and large practitioners, not managers. They were people who both understood the technical detail and had the pragmatism, honed in frontline engagement, to make things work.
- The facilitator was non-aligned to any one particular interest group, had the “thick skin” to keep people in the room, and was pragmatic enough to “call it” when things got stuck, but willing to be proven wrong and revise decisions.

This process in the Wellington City Council for developing common metadata standards for assets is about to be applied across New Zealand. However, it is important to note that the metadata standards were not merely imposed upon the rest of the country. A larger group of the same kind of people has worked under the same kind of process to develop improved asset metadata standards for the whole of New Zealand. This scaled-up community-building and trust-building enabled two things. Firstly it socialised the stakeholders and enabled a much wider engagement and adoption across New Zealand on the basis of trust and inclusion: people knew what they were building and signing up for. Secondly, including a much wider range of interests at the table was effective in developing a more versatile and reliable set of standards. The second version of the standards was better than the first one developed by the Wellington community only.

**The community
of co-designers
were by
and large
practitioners,
not managers**

Understanding diverse interests in the commons

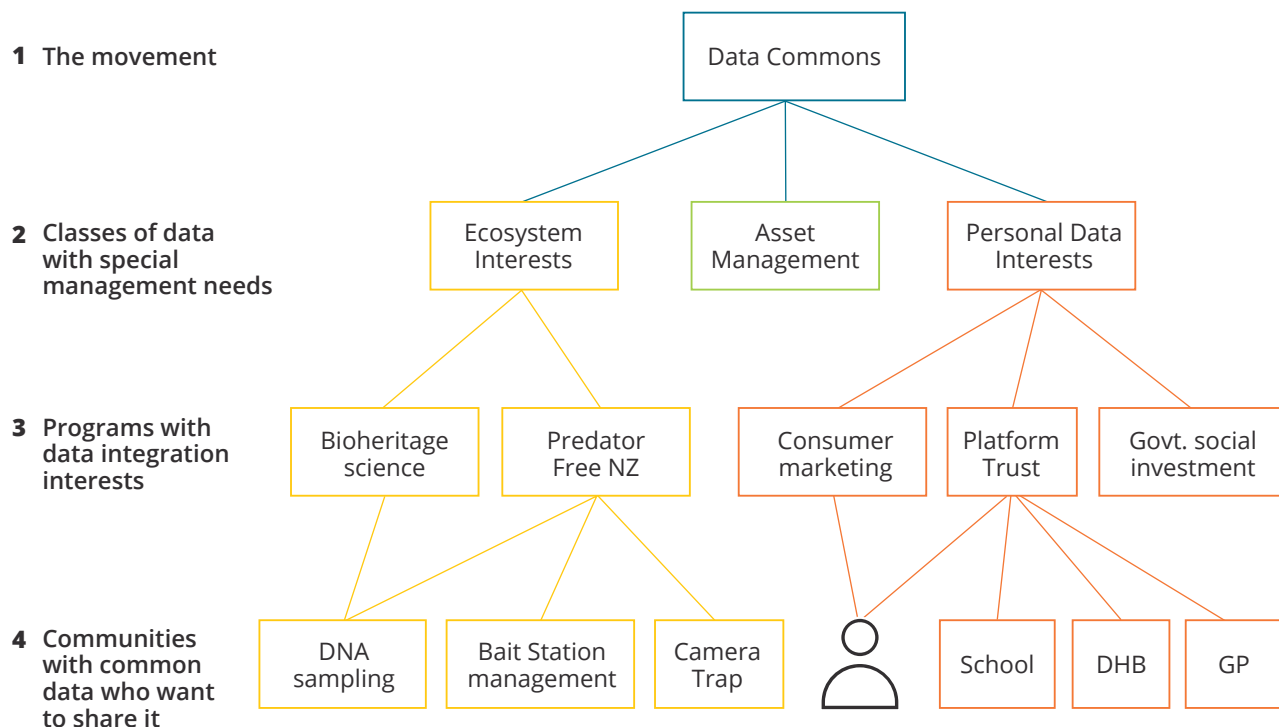
An important component of effective community-building and co-design is to understand the interests at the table.

Here we provide a brief introduction to some ways of thinking about data reuse interests that might be useful to help a community understand perspectives.

Having the right conversations at the right level

At this point it is useful to introduce a rough schema of the kinds of interest in data integration and reuse. This will be used for the remainder of the document to understand various roles and considerations in actually doing the work.

The chart below is an example of the different ways of cutting the cake and organising the conversations, based on the types of interest of the participants.



The overall movement (I) is concerned with supporting the commons as a generalised solution for how to integrate and reuse all data. This paper is interested in this level of question. Below this, there are subgroups with interests in classes of data (II): clarity in this area is necessary to contain the scope and size of the discussion about how to handle particular classes of data. Wellington City Council's interest was in asset data integration and didn't have much to say about predator-free NZ or personal data. Programme interests (III) have a particular integration and reuse challenge in mind that they want to solve, and typically a well-formed community of such interests (where relationships have been established). Data Interests (IV) are siloed interests in particular datasets: people who are sharing that data within their community but not thinking about integrating it with other kinds of dataset or reuse for purposes outside of those it was collected for.

With these levels of interest in mind, we consider the work that needs to be done to coordinate them to start up and sustain a commons-based approach to data sharing. This is defined in more detail below.

1 The Data Commons movement. There is a movement of people interested in building a commons-based approach to data integration and reuse. They have a top-down interest in helping build the relationships, high-level design principles, and institutional frameworks appropriate for all specific instances of the Commons. At the level of the whole commons community, there will need to be some top-down protocol-setting, including answers to some of the big questions such as "right to forget". The Commons community as a whole is also likely to have a big say in the general provisions for transactions and sharing. They will have a third role too: ratifying bottom-up standards from data experts to ensure they are generalisable across the commons.

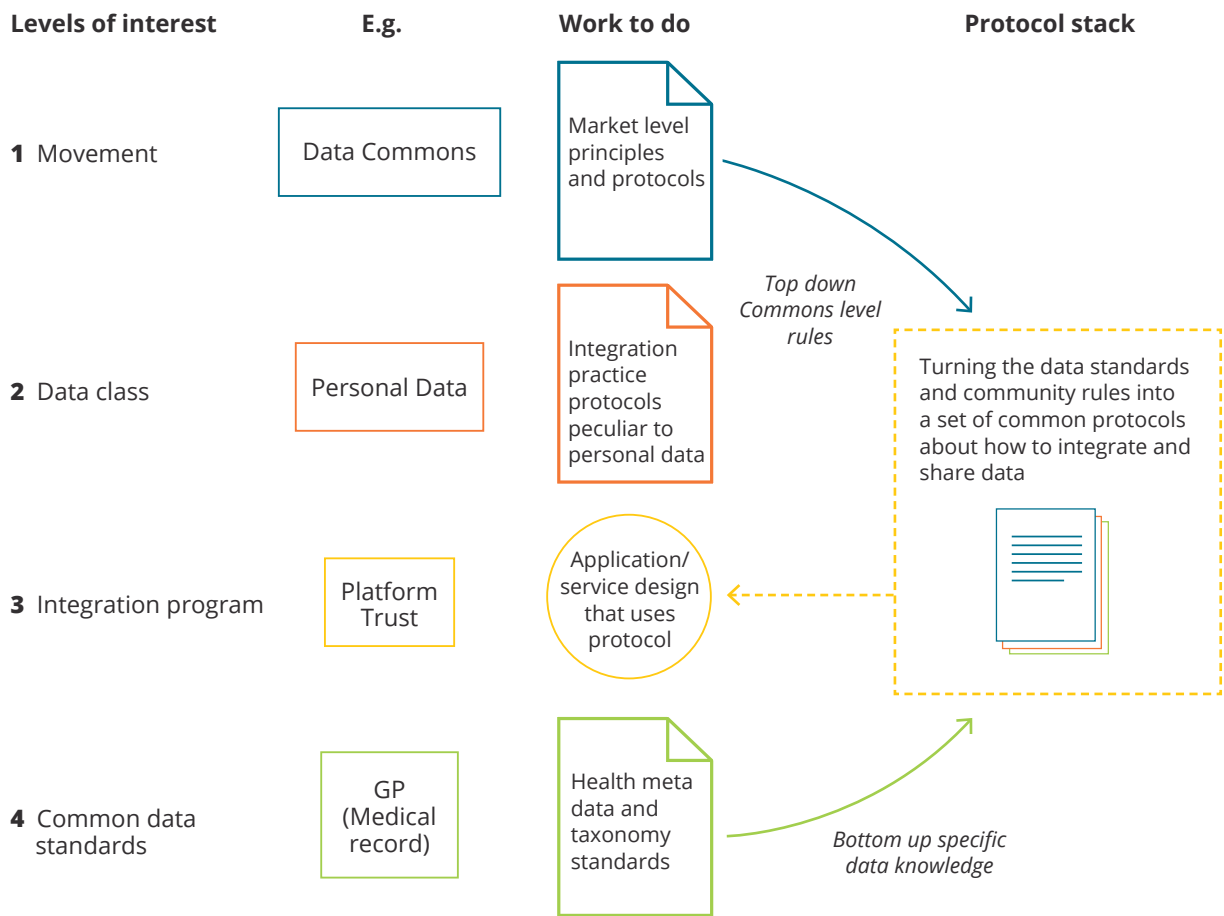
2 Data class interests. There are several large classes of data that have characteristic properties in terms of the way that data can be used or integrated. These include data about people (personal data); assets and the built environment (roads, pipes, buildings, networks – increasingly including the Internet of things, i.e. where people interact with the built environment); the economy (production, incomes, employment, tax, finance, etc.); and the biosphere (the natural environment, its organisms, and their movement). These definitions allow us to limit the scope of the conversation based on how these subcommunities of data interest identify themselves. But data classes are merely pragmatic subgroupings. They break down, for example, with social housing, which is the intersection of asset data and personal data. Then you have to start another conversation.

3 Integration programme interests. Communities of Interest are particular communities of people who may have a common interest in solving an integration and reuse challenge for that community's own purposes; or entrepreneurs with a particular data reuse

idea that they wish to build. Examples include Manaiakalani Trust (education achievement), the Platform Trust (better coordination and collaborative delivery of community mental health), Predator Free NZ (sharing data among scientists, trappers, and volunteers engaged in a mission to eradicate predators from the New Zealand ecosystem). Typically, such activity would begin without reference to a commons-based approach, leading to a point solution that is hard to scale or build upon. However, if they were to climb aboard the Data Commons, they would find a ready-made broader community allowing high-trust data reuse transactions and a potentially lower-cost solution with access to a wider range of data of interest. Connecting these Communities of Interest – which have a natural incentive to share information – to the Data Commons movement can help them scale and efficiently manage their integration challenge in a high-trust way for their constituents.

4 In common data interests. Within existing data silos, there are communities of common interest in a specific kind of data, who use it for the purpose for which it was co-produced – not necessarily repurposing it. At this level – which is where a lot of data sharing opportunities and challenges first emerge – we are primarily concerned with the potential for sharing data with people and organisations that have common interests and shared values. They are best placed to develop at least the technical protocols and standards for data capture, storage, and integration. Because they are more familiar than anyone else with their particular kind of data and its typical uses, they best understand the sensitivities and risks around reuse. Medical professionals, for example, are interested in medical records and have developed taxonomies (ICD10 codes, Read Codes) to allow that data to be transferred and interpreted easily among that community. Scientists are developing metadata standards for the management, sharing, and integration of DNA data obtained from soil and stream samples. Predator Free NZ will need to develop metadata standards and taxonomies for bait stations and camera traps data. There may be a low level of interest in repurposing or reusing. However, if data integration interests (I–III) wish to use this data, then they need to engage with the In Common Data Interest Groups (IV). These groups have a key role in developing the lower-level data standards (the bottom-up protocols) that allow their data to be used and interpreted by non-specialist interests – so connecting them to people with data integration interests is important.

The work of building and maintaining the commons protocols, according to the role of the interest group, is depicted below using an example applied to personal data.



This schema is useful for considering some of the (relationship and community-building) reasons why the status quo seems to be stuck at low-level data sharing. As already noted, most data in New Zealand is trapped in organisations operating at level IV – people, banks, government agencies that find it too costly, risky, or relatively low-value to engage in data sharing or integration any more widely than their immediate community of common interests. Or they have leapt to point solutions at level III and find them hard to scale – since no work of developing standards and more general protocols was done. The StatsNZ IDI point solution and the Social Investment Unit’s and MSD’s coercive practices make those solutions hard to manage or scale. The same is true in the private sector. It is difficult for business-to-business data integration solutions not to get bound up in red tape because they are largely extractive point solutions that have not been set up in conversation with other interests (such as the shoppers themselves). If the solutions are aimed at owning and monetising data, they will be designed specifically to limit rival access – and so data remains fragmented and only of limited general value.

Evolving the commons: top-down and bottom-up

A lot of our discussion over the last few months has been about how the Data Commons might evolve.

If there is an ambiguity in our presentation, it is in the different meanings of the term. We have a vision of a Data Commons for New Zealand – the Class I data interest, unified by a common set of principles and standards. But we also believe that Data Commons will naturally form first around Communities of Interest in Class III – for example, in asset management, predator control, or delivery of personal social services – who can see the value in development of specific data sharing, reuse, and integration applications.

We need both. The vision we have for a Data Commons is for sharing and reuse of data without limits – because nobody can predict for certain what we as scientists or as a society can discover about ourselves from data integration. So as not to foreclose on the possibilities of expansion, we need the development of a common set of principles and standards at level I, to which we hope the specific communities at level III will join themselves.

How do you develop a set of “rules of the road” across diverse interests and data sources?

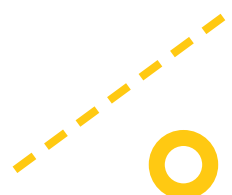
The Data Commons community needs to define the contract for data reuse across an often bewildering array of interest groups.

The challenge will be applying the principle of universality.

For example, there will likely be tension between bottom-up data interests (level IV) who tend to want highly specific rules and top-down commons interests (I, II) who will be aiming for generalisability and universality across a community of diverse interests. Success and the ability to scale and integrate across diverse communities will rest on the degree to which the commons community can solve this riddle.

One recommended practice is to be “protocol-based” rather than try to solve these relationship challenges with a “point solution” approach.

Here we use “protocol” in both the technology sense and the social sense of the term.



Technology can be protocol-based or a point solution

With information technology a designer always has to choose between building a data integration application as a hardwired “point solution” or as a more flexible “protocol-based solution”.

The Data Commons is a protocol-based solution because this is a universalisable solution based on standards.

The argument for the latter is rather like the reasoning for having standards around power plugs. Yes, everyone can build their own unique plug type, but there are real advantages to the community adopting a standard design. New Zealanders can buy any appliance (fridge, toaster) and plug it into any household power source because we know they fit. The opposite occurs where technology companies cannot agree on standards for computer connections – which is why we have endless numbers of adapter types that fit USB, USB-C, and any number of other Samsung-only or iPhone-only ports.

HTML and TCP/IP protocols (the backbone of the Internet) are successful protocols and show what can happen when technologists get it right.

The same choice is there for data sharing solutions. Code can be written for any particular solution. We could build one big integrated data store (on the cloud perhaps) that lapped up every specific kind of data people threw at it without any standards. But it is not hard to see how costly and fragile that would be. There are advantages to a standards-based approach: solutions then become more scalable, cheaper to implement, interoperable.

The Integrated Data Infrastructure (IDI) solution run by Statistics New Zealand is costly because there are no metadata standards across the social and productive sectors that would allow Statistics NZ easy and low-cost integration of that data. When things change at source, this can break the integration point solution. The same problem arose for a budget bid to integrate education sector data. Not much thought was given to first developing metadata and API standards which would make the solution less costly by tens of millions, and less fragile once built.

Emerging practice for smarter companies and institutions is the use of metadata standards and other protocol-based approaches to reduce cost and increase value.

Xero uses “APIs” (Application Program Interfaces) that allow a thriving ecosystem of third party-providers to develop bespoke solutions for specific accounting needs. There are “Farming Accounting” add-ons, for example, that hook into Xero’s data sharing ecosystem. IRD in turn is now publishing APIs (rather than a hardwired point solution) to allow third parties like Xero and MYOB to hook into tax data to develop integrated tax-accounting solutions.

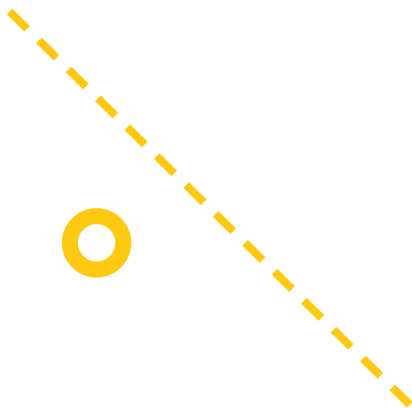
Social and institutional protocols

But the commons contract is not just about technical protocols.

It is also about human “social” protocols – the protocols that govern the way relationships between people are managed.

The use of a protocol-based approach determines how members of the community should interact with each other and what should be expected. In this sense the term “protocol” is closer to “diplomatic protocol” than to “technology protocol”. But it does the same work. A technology protocol is all about how data and technology relate to each other. A social protocol is about how communities relate to each other. (No doubt we are doing violence to both technologists and diplomats!)

The basic work of the commons is to develop and steward a protocol-based approach to allow a low-cost (“frictionless”) market for high-trust data sharing solutions. This includes both the human and the technological protocols.



The Protocol Stack

To define these relationships, we have borrowed the concept of a protocol stack from the IT world of network management.

A protocol is a set of rules or transaction relationships. In a technology sense, a protocol stack is a layered set of protocols defining all the transaction relationships in a network: between applications sending and receiving data and all the layers of software in between required to transport the data across the physical network.

The Data Commons Protocol Stack goes further still to include the social protocols necessary to enable the design, implementation, and management of a high-trust system for data sharing and reuse.

The Protocol Stack is a set of agreements. Some of these agreements will be instantiated in technology and some will be embedded in ethical standards and institutional arrangements. Together, this stack of agreements constitutes the contract made between members of the community that underpins the particular institutional frameworks and technology protocols necessary to create and manage the Data Commons.

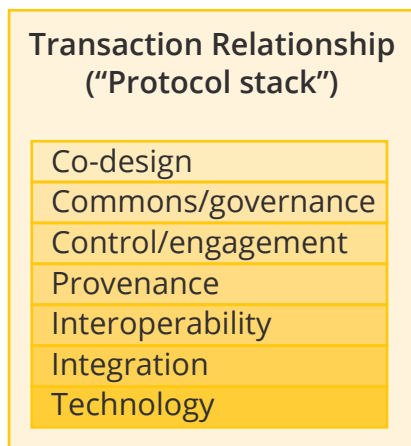
We envisage the co-designed commons Protocol Stack as a collective asset: anyone who wants to create a trusted data sharing initiative can do so with the knowledge that they are drawing on current best practice and community standards.

Co-design of the commons protocols needs to solve seven specific challenges. We've developed an initial stack of seven protocols, shown in the diagram, that each references one of these challenges and together form the contract for a Data Commons.

We see the process of developing this contract as an ongoing discussion amongst the people interested in a Data Commons. The discussions are unlikely to be a linear progression through each of the layers in the stack. As members talk through arrangements for co-design, governance, and management of the commons and work through the more technical requirements of defining and controlling transactions, they are likely to come back to higher-level questions of value and risk and revisit earlier agreements. So the stack should be seen as a set of challenges, not as a project plan or critical path.

Emerging practice for smarter companies and institutions is the use of metadata standards and other protocol-based approaches to reduce cost and increase value.

Xero uses “APIs” (Application Program Interfaces) that allow a thriving ecosystem of third party-providers to develop bespoke solutions for specific accounting needs. There are “Farming Accounting” add-ons, for example, that hook into Xero’s data sharing ecosystem. IRD in turn is now publishing APIs (rather than a hardwired point solution) to allow third parties like Xero and MYOB to hook into tax data to develop integrated tax-accounting solutions.



1 Collaboration and co-design protocol

Our concept of a Data Commons rests firmly on the concepts of community, collaboration, and co-design. We built this philosophy into our own work on developing this blueprint. We chose a co-design approach that involved potential users of the Data Commons. We could have raised venture capital to build an extractive business and kept this strategy to ourselves. Instead we invited inclusion, maintained transparency, and involved as many people as we could with the resources we had available. The people who were attracted to the project tended to reinforce this approach. It has been an interesting conversation amongst technologists, financial cryptographers, scientists and environmentalists, extollers of free markets and frontline social sector NGOs.

Out of this process we have largely agreed on the nature of the challenge and the thrust of the solution. In a word, high trust and inclusion were at the heart of the project. We hope our approach serves as a model for the collaborative and co-design approach which we believe is essential to the continued development of the Data Commons.

So the first step in forming a Data Commons will be to identify a shared set of interests in data and forming a group which represents all those interests. The discussion above should also help provide a checklist of not only who should be included in this collaboration, but at least broadly what their shared interests are which will form the basis for the Commons. Based on our own experience, the early discussions need to identify these interests, how they can be supported by a Data Commons, what each member of the group can bring to the discussion, and what value and risks group members can identify. The next step will be for this group to agree on how the Data Commons project will proceed and how they will work together on it. As indicated above, it's likely that this and other steps will be revisited in the course of the discussions.

An important idea that has emerged from the project to date is that of a “community of practice.”

2 Community governance protocol

The Data Commons will need a constitution in which its members agree how decisions will be made about the operation of the commons and by whom. The constitution should reflect the principles we have set out in this paper: collective ownership, participant governance, and participant design. An important idea that has emerged from the project to date is that of a “community of practice”. This suggests that leadership and governance for the commons should come from the people who are actively involved in developing the commons.

This group provides the decision-making and community boundaries for the commons. They are responsible for making rules, managing community assets (e.g. standards), and ongoing co-design. We think that this approach has the best chance of avoiding the project becoming extractive over time. We are proposing a commons-based co-ownership model that supports better self-regulation, including agreed ethical standards and clear sanctions.

We suspect that the governance model that will emerge from discussions will be one that is common in clubs or incorporated societies: a general membership that votes for an executive.

3 Engagement and control protocol

Standards need to be developed that allow people to upload, control, and license use of their data.

For personal data we predict the need to build in a Personal Information Management System (PIMS) – including consent, right to forget, exclusions, data provenance, and the ability to license other users of data. We will also need a licensing layer in the metadata too. There are many useful lessons emerging from existing PIMS such as Xero and MyWave that can support this process.

This is also true for commercial and scientific assets and other forms of data not about people. The methods of uploading and controlling access to data need to be developed as community-level expectations and protocols and technical standards.

4 Transaction provenance protocol

To manage data on the basis of how it is used, the community must be able to see who is using it for what. Standards around tracking reuse allow a couple of important community-level assets:

- Transparency about where value is being generated, which then allows redistribution back to data contributors of any value the commons has generated.
- That same feature of the commons also enables control and enforcement: it informs decisions about the forming and terminating of relationships, allows misuse to be detected, and supports sanctions.

Technology such as the blockchain and distributed ledger systems in general may be one way to enable this kind of traceability.

5 Transaction interoperability protocol

To share or reuse data we need a way of describing it. A clearly-set-out range of taxonomies and classification systems will allow sensible data usage and data sharing through a set of metadata standards, agreed and implemented by all members of a Data Commons. Metadata standards solve problems of interoperability: I will know that this piece of data is a genome, who it is associated with, when it was uploaded, what I am allowed to do with it, and other “meta” facts about the data I am sharing, integrating, reusing. These standards will be set at both level III (for a specific data-sharing community) and level I (for the entire Data Commons). For example, the genomics community might agree on how genome data needs to be described so it can be shared within the community. Then this specific metadata needs to be joined with supra-community standards, to answer questions like how genome data can talk to social sector data, or how face recognition data can talk to identity data.

6 Integration protocol (2+2=5)

Data integration is the main purpose of the Data Commons. When data can be meaningfully integrated, its value increase can be non-linear. To do this requires being able to integrate diverse piece of data together to form new insights. Data integration needs points of joining. At the most atomic level (for the management of the commons) there are probably several main integration protocols: (a) who – joining data by personal identity, (b) What – joining data by asset identity, (c) where – joining data by geospatial location, and (d) when – joining data by time stamp. Developing standard approaches to this allows data to be integrated at low cost and effort.

Rules about what kinds and purposes of integration are also needed. Integrating data to re-identify somebody who wishes to remain anonymous, for example, might not be sanctioned.

7 Technology protocol

The technology has to allow safe and secure transfer across a network. What is the locus of control and storage? How federated is it? Is data owned and housed within individual bitcoin wallets? By service providers? Centralised? How do all these elements talk to each other so that we can have remote deletion (right to forget) or copying and integration?

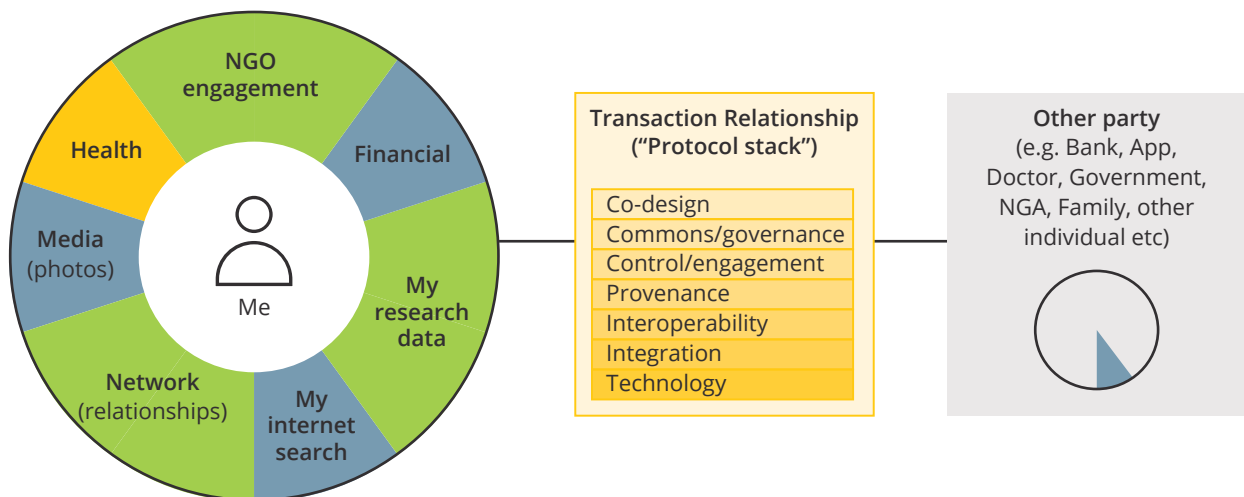
The Data Commons community needs to set the bar on how data transactions are handled technologically by referring back to the high-level principles of the commons, in particular embedding control and trust into the technology.

Protocol-enabled relationships

Deploying the Protocol Stack is a systems-level intervention. It describes the factors or preconditions that we need in place in order for a Data Commons to function.

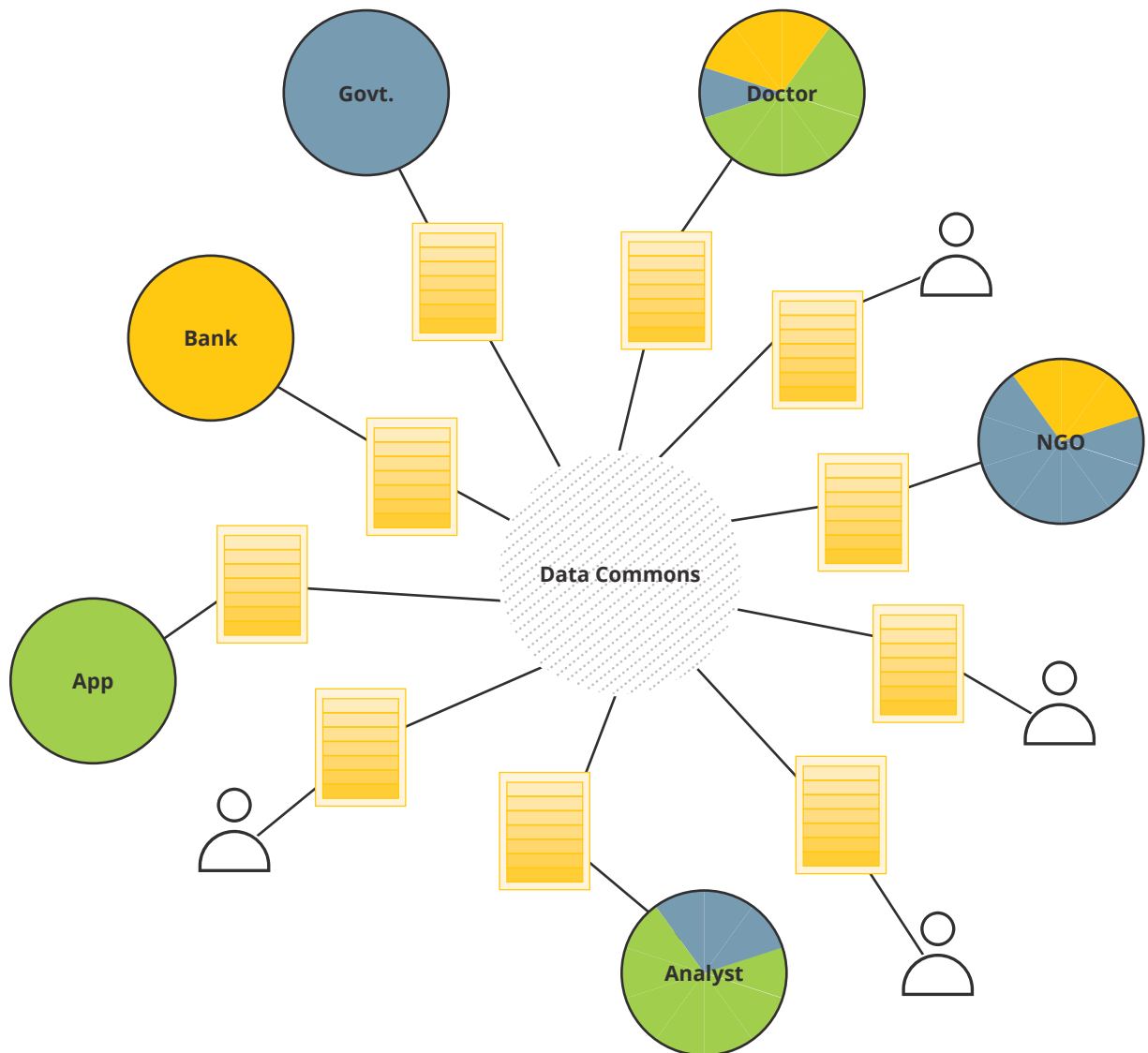
We think this is true at the scale of a small local data sharing initiative involving a handful of organisations, right up to the national and global scales.

The commons protocols enable low-friction, high-trust data integration and reuse. This is essentially having the rules of the road available for forming a quick and safe relationship between data providers and users. The diagram – an example using personal data – shows how I can share my story with another party, safe in the knowledge that I have at least equal influence in that relationship.



A protocol-enabled community

The resulting data ecosystem that is enabled by the Protocol Stack creates a community where data reuse and integration can happen at scale and at low cost due to the high trust and control the protocols afford.



The Internet protocol ushered in the ability for a wide range of solutions to increase value for users – email, web pages, file sharing, social media, etc. Financial regulation aids citizens and organisations to transact money safely with all of the social and institutional protocols that govern those kinds of transaction.

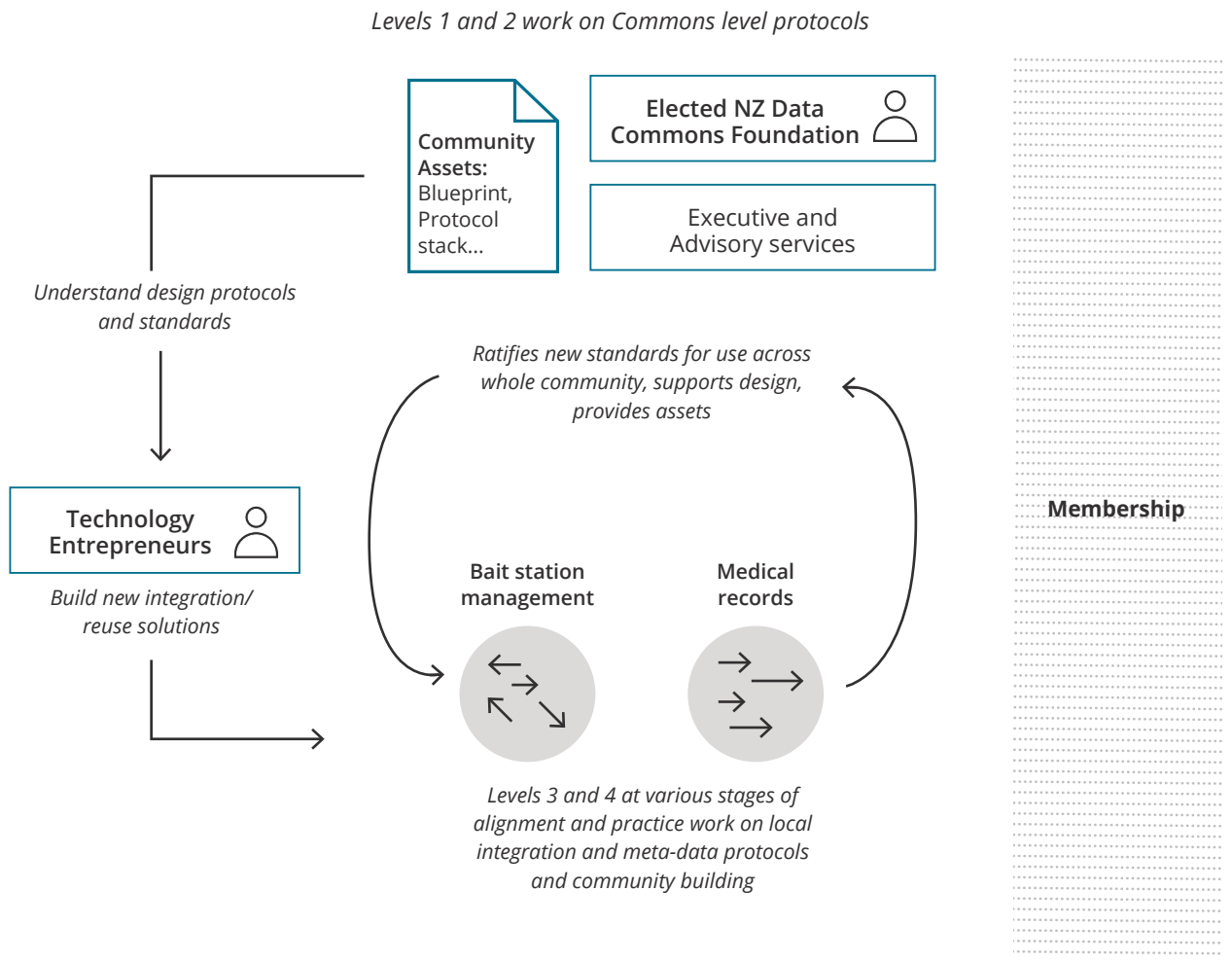
By the same token, data sharing protocol should allow a community that enables, not disables, innovation and engagement in the reuse and integration of data to drive value. It will do this because it solves the current big problem, that there are no standards for consent, control, licensed reuse, etc, so it is difficult to form data sharing relationships.

Spanning institutions, relationships, and capability: a national foundation?

We don't wish at this stage to get any more prescriptive that we already have.

The ideas here are presented to illustrate the kinds of institutional arrangement that are in keeping with the design principles of the commons. There will probably be some kind of "commons foundation" whose job is to support the commons community and build an inclusive process for curating and maintaining the standards and protocols. This will likely be a non-aligned not-for-profit that administers the community standards in the same way a market regulator would do – except we recommend that it is constituted by a wide range of interested parties and not governed or captured by narrow or extractive interests or interests non-aligned with the commons community – interests that would be coercive and exclusive.

This institution also needs to have expertise in facilitating bottom-up groups of interested parties to define their own local metadata standards, as well as being able to facilitate a broad bandwidth discussion at the community level about commons-level protocols.






Section four

Kick-starting the Data Commons

Compared to expensive centralised data repositories, the Data Commons is almost “virtual”. It is made up of a number of communities of common interests who are developing the set of protocols that allows them to share, integrate, and reuse data. They also enable other people to come up with valuable applications based on this integrated data. The Data Commons is in many ways a self-organising system that forms and regulates a “market” for data integration and reuse.

Like all markets, the value of the Data Commons will only grow with use. Entrepreneurs, social entrepreneurs, scientists, analysts, community groups, and others who can benefit from data reuse need to climb on board and build their ideas on the back of the Data Commons protocols. Supporting this growth becomes a matter of managing incentives to join and identifying high-value starting points to kick-start a network effect.

Note that this leads to a chicken and egg challenge: the need to grow a set of generalisable commons protocols whilst at the same time actually building value for people on a nascent Data Commons.



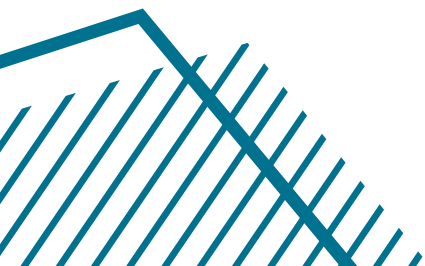
Incentives and disincentives to have a commons-based data integration solution

The advantage of building your own point solution is that you have complete control over that particular data integration challenge.

This will be the correct solution for some kinds of interest – those with more interest in controlling and owning data as a business model, for example. So a Data Commons solution is not for everyone.

For people and organisations who want data to deliver value (rather than owning it to generate a dividend by on-selling it), and who want to enable the benefits of reuse, the advantages of doing this via the commons protocol are several:

- They will be able to illustrate to the people they want to sign up that they are building their special project in a high-trust way which gives a lot of control back to the signee.
- Once there is data already on the Data Commons, it becomes very easy and low-cost to add some extra data and integrate your solution with the existing data on the commons. So this is a low-cost and efficient platform for your particular use of integrated data. Leveraging off the market rather than building a bespoke solution will be cheaper and more scalable.
- There will be a real market of participants needing high-trust solutions whom you can do business with, and who value a commons-based approach to forming a data sharing relationship with you. They can trust that people won't on-sell or misuse their science data or personal data.
- Your solution won't be fragmented: so if Samsung stays off the commons, its heartbeat sensor will never be more than just a toy. But if a New Zealand entrepreneur's heartbeat sensor is hooked into the commons, it can be joined with people's medical records and what they eat from their supermarket shopping basket and directly shared with their general practitioner to become a complete personal health solution. The heartbeat sensor that can integrate data using the commons protocols moves from toy to service enabler. These kinds of "economies of scope" mean that you'll be able to design better services for your customers through the high-trust use of their integrated data rather than staying fragmented and in control of siloed data that is of lower value to everybody.



- Growth of value with scope is non-linear: as the network builds in more kinds of data and a larger number of data providers and users, the commons becomes exponentially more valuable for its members as the innovative uses of more data integration grow geometrically as well.

In short, we think that the availability of a commons-based approach will generate increased value for the community as it grows and will likely put pressure on siloed “ownership” interests to integrate their data for fear of being able to provide only a poor service by comparison. For the individual entrepreneur who is trying to do some social good, manage social investment, or build a commercial service, the benefits of using a commons-based approach will far outweigh any value obtained from a point solution.

The Data Commons enables a competition in ideas and innovation to drive value, not extraction of value from ownership. It’s a very different beast to what is now possible.

This is why people don’t go off and design their own Internet protocol. The value of the network is in everyone else who is on it. We wait patiently for an update to the protocol as the community of common interest refines it. Then there is mutual value in hopping on board to drive the value we seek.

Starting up, curating, and growing the network effect

So the second part of the work becomes signing up and encouraging participants and trying to get the network effect going.

And this is where we recommend having catalyst projects that solve high-value data integration challenges for particular Communities of Interest. Incentives might need to be provided at the start to attract high-value data. Once high-value data is on the commons, a network effect should start to kick in and attract further data, and so begin a snowballing effect.

In effect a Data Commons is really just a market for high-trust transactions in the reuse and integration of data. That market needs some transactions and some data to begin. Key tactical objectives for anybody building a Data Commons are to attract high-value data early, and to ensure that early communities can obtain high value immediately. Effective community management becomes key to scaling and attracting participation.



Appendix: Two case studies

This appendix has two purposes. The first is to apply the theory of the Data Commons to some real-world cases and illustrate what those high-level community principles might look like when put into practice. If we could form a Biosphere Data Commons or a Person Data Commons, what will that enable? The second (meta) purpose is to rehearse the kinds of question that any such Community of Interest would need to answer before they could begin building their own Data Commons.

Two case studies informed this investigation of designing and building a Data Commons. Using these real-world challenges provided a useful foil to test thinking and develop the ideas in this report.

These two data integration and reuse case studies illustrate some of the concepts, community questions, and value propositions of a Data Commons approach.

The first case is the example of New Zealand biosphere data integration and reuse. This is an example of the kind of commons where there are highly aligned interests in integrating and reusing data, and where their data can be relatively open access: there are fewer personal risks or commercial sensitivities. It also illustrates integrating data by place. There are several shared goals around which the community can align to scale up data reuse, such as the massive national undertaking to make NZ predator-free by 2050, the need to manage our nation's unique indigenous bioheritage, and the need to stop new invasive pests entering New Zealand.

The second case study, about the integration and reuse of personal data, is at the other end of the spectrum, where licensing semi-open access to highly personal data becomes a key consideration. How do we design a Data Commons that allows us to license some kinds of open use of personal data whilst limiting other kinds of use to consent only? The same kinds of consideration might also apply as lessons for the management of commercial sensitivities around data integration and sharing. This case study illustrates integrating data by person. One burning issue currently in New Zealand is the state sector's poorly realized desire to integrate and reuse government data to support social investment. On a more positive note, there are emerging opportunities for personalised health, big-data-based science, and accelerating innovation if only New Zealand could unlock the potential for low-friction access to integrated personal data by keeping it high-trust and safe for people.

Case study two

Biosphere Data Commons

This is the case brought to us by Predator Free New Zealand and the New Zealand Biological Heritage Science Challenge.

These two overlapping Communities of Interest have a shared aim to build a national data integration and reuse capability that enables scientists, ecologists, and all groups with an interest in New Zealand's ecosystem to be able to collaborate, coordinate, and build intellectual and social capital that helps New Zealanders support the environment. They have a data reuse and integration challenge.

In particular there is renewed interest in data integration due to a national level-interest in supporting the effort to make NZ predator-free by 2050.

Interests in biosphere data

The New Zealand Biological Heritage Science Challenge and the NEXT Foundation have co-funded our Data Commons work due to their interest in the potential for improved data reuse and integration to support the coordination and mobilisation of efforts to sustain New Zealand's ecosystem. There are two principal areas of endeavour for these two communities. At the superset level is the interest of the scientific and conservation community in monitoring the health of New Zealand's bioheritage and biosecurity. Within this is a Community of Interest in Predator Free New Zealand who have a restoration focus. We think that there is enough overlapping interest to treat these both as a single case study. This is because much of the data will be of interest to both parties if it can be shared and is likely to be collected and used for a range of ecology, science, and conservation projects.

Note also that the Predator Free and bioheritage interests may overlap with biosecurity and border security interests and primary sector and economic interests in detecting and stopping invasive pests.

The co-producers of value and insight and consumers of biosphere data include:

- Scientists, conservationists, philanthropists, the Department of Conservation, the Ministry for the Environment, citizen volunteers, and NGO groups with a professional interest in ecology and pest eradication.
- Recreational users of the natural environment such as tourists, trampers, hunters, and watersports enthusiasts.

- The primary sector, as it interacts with ecology, bioheritage, pest eradication, and farming and agriculture interests.
- Government in general: the benefit of success in the 2050 objective will be to achieve a long-run reduction in pest eradication liability. There are also national-level economic, health, and environmental goods to be gained through effective management of the ecosystem.

The value of a high-trust low-cost data integration and reuse commons for ecosystem data

The value propositions for ecosystem data interests include:

- Enhanced detection of what is going on, leading to improved insight for decision-making and strategic targeting of resources.
- Improved ability to coordinate and collaborate at the local level, reducing the cost of operations and leading to efficient targeting of action.
- The ability to recruit, motivate, enable, and retain a large and diverse range of participants around a shared objective (particularly for Predator Free New Zealand).
- Opening new sources of investment through improved measurement and tracking of success.
- Far lower costs to adding new ideas, communities, solutions, and operational processes. They can hook into a highly efficient data sharing ecosystem (and don't have to hardwire bespoke solutions and replicate work every time something new comes along).
- Improved capacity to learn what works.

An ecology Data Commons for bioheritage and Predator Free NZ might look something like this (next page):

What can a fully functioning data commons and coordinated approach to data sensing, integration, use, can add to stewardship of NZ eco-system

Draft reverse brief for discussion 1.0

Use

Prediction

What is going to happen next? Mast year?
Forecast diversity? Forecast incursion rate?

Scale science

Pooled data to get ecosystem view, re-use of high cost data, track gene drive impact on other species (economy of scope)

Investment decisions

Learn what works at scale, early. Measurability provides opportunities: Profit/risk share reduced forward pest risk with government. E.g. Forward investment approach. Bio-bonds

Motivate/mobilize/enable/incentivize

Because results can be tracked could have rewards, prizes, rankings, ...have to be careful about perverse incentives! E.g. Dob-a-rabbit photo scanning for schools and volunteers. X-prize for science/technology development. Community rankings for land parcels and quality of sensing data

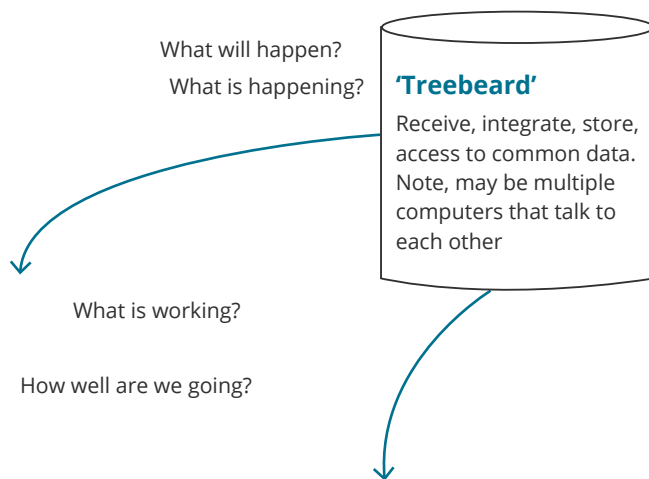
National investment and learning

National coordination/ collaboration across shared view and orienting KPIs

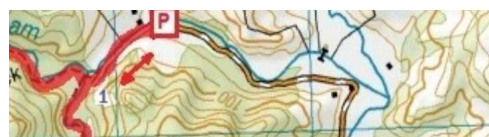


Mapping; predicted mast season, predictor burden, biodiversity indicator, which seasons doing well (by intervention), community engagement level, incursion directions, gene drive spread. Instead of one-off papers, put it on a shared self service dashboard = easy access, shared view, drive innovation and self awareness

Integration



Local coordination



"Our land parcel" analytics:

Localised community planning and monitoring, operations and data capture. Where have we been, what's happening on our borders, which traps need servicing, who has been where (micro-GIS lines of our troops)?

Includes standard dashboards of in-common metrics;

Estimated possums; 4,600

Rats 120,000 (down 12%)

Predicted mast year; high risk 2017

Soil bio-diversity quality; No data

Water bio-diversity quality; 23 organisms per gram

National rank: 3rd

Similar ecosystems; 1st

Bell-bird 200 (up 40%)

Interoperability

Data Commons Protocol Stack.
(Data commons "Blueprint")

Solving the 7 challenges to having a coordinated data community.

1. Design
2. Commons
3. Engagement and control
4. Transaction providence
5. Transaction interoperability
6. Integration
7. Technology

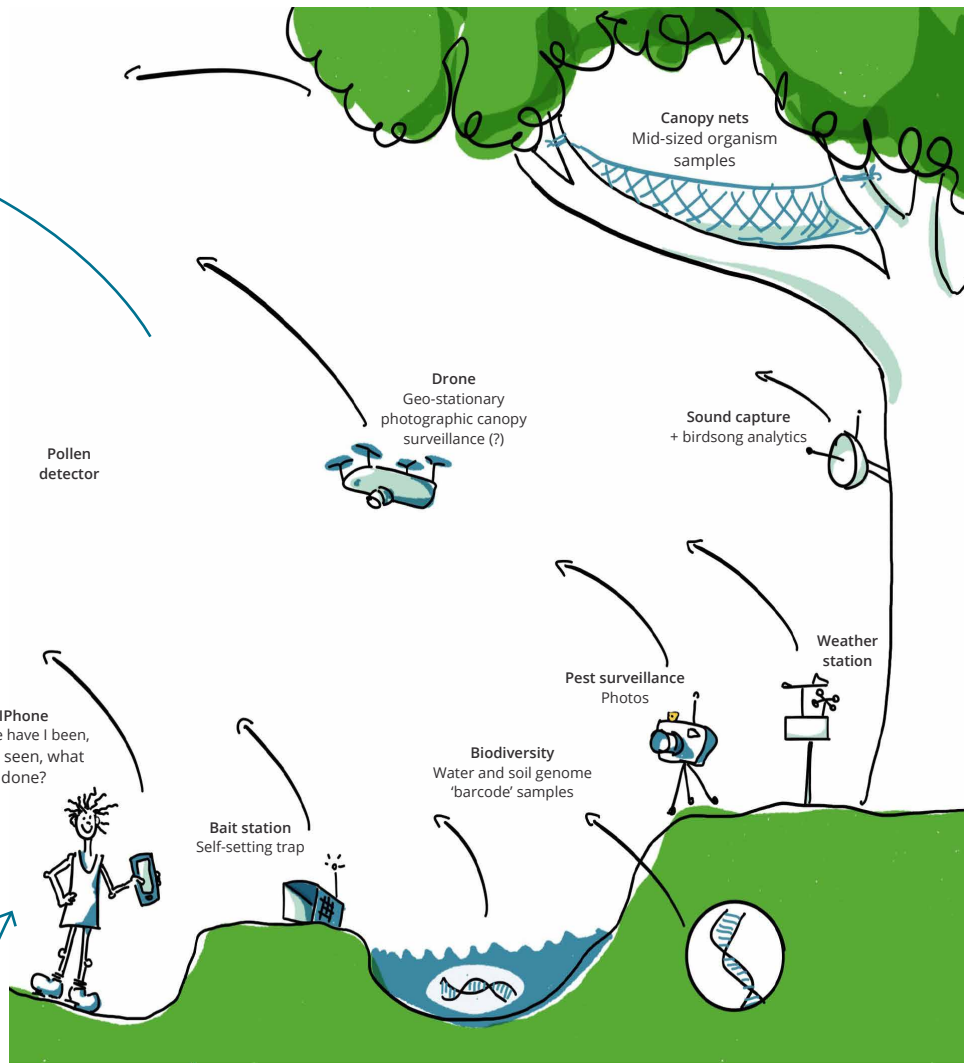
Sensing

Two things happening here. 1. Sensors are cheaper, smaller, remote, real time, connected, wireless, self charging, smarter (analytics at source) and can measure new things (e.g. Genome content (bio- diversity) of stream). 2. Organization and uploading the data itself is also easier; Use of QR codes, what3words, wireless networking and upload and GIS and photo capability of phones using standardized apps provides lower cost, more standardized data capture by (sometimes) less expert collectors, or automated collection

Feedback

Strategically: New initiatives, collection, investment. E.g. Do more of X here

Tactically: near real time operational feedback - e.g. photo snapped a ferret, batteries flat, trap needs clearing



Human 'sensors'/entrepreneurs/apps:

Can connect via a range of apps that can hook into the data commons via smartphones (field workers) or Internet/tablets for people interested from home/office/community. Tech Entrepreneurs have low cost way to integrate their data or use existing data in innovative ways.

Lizard Match (Photo ID)	"Dob-a-ferret" school photo prize	Farmer volunteer and social network	Timaru Rat Catchers Gang App	...
-------------------------	-----------------------------------	-------------------------------------	------------------------------	-----

Building a Data Commons for biosphere management

A community of exchange comprises producers and consumers. And although the two roles are not mutually exclusive, both are required for a functioning market. The core roles in the biosphere data market are sensing (gathering data on organisms and the health of the ecosystem and biosphere), and consuming that data to do science, manage pest eradication, understand and manage bioheritage, and stop invasive pests.

Sensing activity (data capture)

Sensing activity is the people and processes that gather new data. Sensors might be human or machines, including remote sensing devices, cameras, hunters, trappers, farmers, citizens, sample collectors, station managers, NGOs, and pest eradication groups. These all collect data about the environment, whether it is their stated intent or an unintended byproduct of their activity.

Data reusers

These are the analysts and consumers of data, the generators of insight, who sometimes need to integrate and reuse that data to do their work. This can include:

- Direct operational sharing of data. “Somebody else’s bait station just went off, I’d better go check it for them.”
- Adding value to data and returning it to the commons for somebody else to use. “I’ve been scanning your camera trap photos and this one contains a ferret” or “I have taken your soil sample and determined its biodiversity using DNA analysis.”
- Analysts add value to the data by integrating it and examining it and drawing conclusions that are useful for decision-makers. Progress against agreed metrics can be tracked over time, to see whether efforts to achieve a certain end are having any effect. Sophisticated analysis involving multiple controls can even attempt to attribute progress to a particular factor using integrated data.

There are consumers of data at all levels and across the entire breadth of the interest spectrum – from bait station managers who need to know which bait stations to check, to community groups monitoring their area, and Predator Free New Zealand’s monitoring of the New Zealand-wide situation. Funders investing in biodiversity and pest eradication need insight

into progress, scientists pursuing research and publication need access to quality data, and school and community groups who want to get involved will engage best when information is available. To develop a successful ecology data market, we must define the core engagement and have a strategy in place to leverage the network effect of the data market for greatest impact.

This data sharing capability could be built as a point solution which adds specific data sharing applications over time. But it would be less costly, scale faster, be less fragile, and solve a number of other challenges (trust, control, inclusion) if it were built as a Data Commons according to the principles outlined above.

Application of a commons-based approach to biosphere data

Looking at the communities of interest forming around Predator Free New Zealand in the Biological Heritage Science Challenge, it is clear that there are common purpose, common data, and potential large economies of scope from data reuse. Scaling up better integration would be of huge value in helping this community to mobilise for a national-level challenge. The application of a Data Commons approach to biosphere data integration will, we think, provide a lower-cost solution that could grow with the community and diversify as trust and community-forming grow around the shared challenge.

We think this community would be well-positioned to back the kind of Data Commons principles outlined in this report.

The scientists and volunteers who form the bedrock of this community are interested in public and environmental value, not commercial gain through trading data. The value of data reuse is in the ability to have a collective impact at a greater scale than any individual can have using fragmented and siloed data.

The Community of Interest in biosphere data has well-aligned, inclusive values, and various groups are already at the table. We have identified interests in reuse at the predator-free coal face (e.g. bait station management can be streamlined and made more efficient) and reuse cases for central planners and investors. The aligning values are broad-based. However, there are also specific niche interests, such as scientific interests in “publishing first from my data” that also need to inform any reuse protocols for the community. There may also be competing commitments between parties such as commercial sensitivities in the primary sector and the ecological sensitivities of volunteers. Creating a robust and generalisable set of community protocols requires that these interests are at the table.

This principle of inclusion, together with that of control, enables the development of community-based protocols that allow scientists, for example, to integrate data while retaining the right to publish first.

The value of
data reuse is
in the ability
to have a
collective
impact at a
greater scale

These trust issues, whilst still important, are not as severe as those involving highly personal data.

Data reuse is participant-controlled at all levels and thus deeply democratic: people can vote on the board or with their feet. If individuals can exercise control over specific data-sharing activity, such as timing it or withdrawing their data, then this control will translate into higher trust and value, due to people being able to participate with minimal risk.

Although there is probably a limited number (perhaps 20–50) of specific reuse cases for this Data Commons community, there is still value in developing a protocol-based approach rather than a series of discrete point solutions. It is likely to be lower-cost and far more scalable, and new applications will have a very fast time to market.

The biosphere commons is a market that licenses data reuse on a basis of trust, not ownership or on-selling of data; the trade here is the trust. If I put my data on the commons because I trust the commons, I can develop reuses that integrate other data sources to help me with my specific needs. Data reuse should be governed as a common-pool resource for the common good; clearly there is a common good for New Zealand and this Community of Interest in particular in allowing free and unfettered access to the data that forms and mobilises this community.

We think that biosphere data is ripe to be developed as a shared, common-pool asset where the protocols around reuse are collectively developed across the various interest groups within this community to reflect their common interests and their specific interests. If this is achieved, then each participant in the Biosphere Data Commons will be better off individually – they will be able to do better science, better community management, better mobilisation of community interest, and better investment in ecological outcomes. New Zealand will be better off collectively by developing a thriving data sharing ecosystem that informs all aspects of the way we manage our biosphere.

Case study three

Person Data Commons

Note: we do a lot of preparatory conceptual work here just because people can get confused about the distinctions between personal and impersonal uses of person data.

A Person Data Commons is capable of enabling citizens to engage with the various entities that use their co-produced data on a reciprocal basis of value and safety. This will deliver more and better quality data to the communities of interest such as NGOs, service providers, business, innovators, and government, whilst also affording unprecedented opportunities for individual citizens to manage and control the data they generate, and derive value from it for themselves.

With personal uses of data there are several underlying concepts that need to be considered.

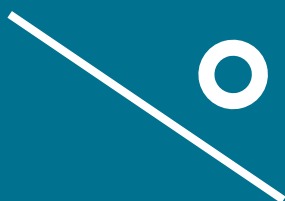
Personal vs impersonal uses of data about people

In describing the Person Data Commons we are talking about “data about people and their engagement with the world”.

Sometimes a distinction is attempted between “personal” and “non-personal” kinds of data, but this is artificial; it is more helpful to think of personal and impersonal uses of data about people. In other words, it’s better to think of data about people and their engagement with the world in terms of the purpose it is being used for, rather than the subject of the data itself. There are plenty of non-personal uses for data about people. I can analyse groups of people and not target them individually. This is a non-personal use of data about people. I could analyse that same data to target an individual with a service. That would be a personal use of data about people – including data about this particular identifiable person.

So, thinking about this data in terms of its use, rather than the kind of data:

- **Personal use** of data means using a person’s data to target them specifically with a product or service, or create a bespoke solution for them based upon what is known about them. Personal uses are where analysts want to know some aspect of My Story to do something with that.
- **Impersonal use:** when data about people is used non-personally, it is used in an anonymised way to learn about a population and make decisions at a higher level. So it is Our Story – the story is about a group of people and analysts want to understand their story and do something with that information.



Communities of interest in a Person Data Commons

A Data Commons that facilitates the sharing and integration of data about people has value for multiple communities of interest.

There are two important aspects to the interests in personal data reuse. One is the distance from the person whom the data is about: here we make distinctions between uses of personal data ranging from first-party to fourth-party use. The second is to look at the kinds of user: government, marketing, technologists, and their various interests in personal data.

I have a special “first-party” relationship to data about me, and it is for me to reuse too. My data is about me, it’s “my story” – or at least the digitally captured part of it.

Data about an individual, their home, their travel, their interaction with other people, the services they use, their health, their behaviors, and their preferences can be integrated to generate incredibly powerful insights and a more complete picture of their story. These insights can be used to develop personalised services, products, and experiences, which might be commercial or public services. I might want to share what’s going on for me (my heart rate) with another person (my GP). It’s important that individuals have control over whom they allow to use their data in a personal way, as data could otherwise be used coercively to identify and target a person against their wishes.

So people have a special relationship to data about them. Any Person Data Commons needs to respect that as a first principle.

Therefore the primary Community of Interest for a data market transacting data about people is, of course, the people whose data is it is. We want to maximise value for ourselves. I want to integrate my story and manage whom I share it with. Adding banking data to my tax data makes book-keeping easier. Adding it to my Flybuys data helps me with budgeting. I want to share my Fitbit with my GP or fitness instructor for personal value. My genome, government-held medical record, personal health-sensing device, and food purchases might make a highly personalised health service possible.

Individuals can benefit directly and indirectly from the integration and sharing of their data. Direct value might come in the form of personalised financial, fitness, health, or Internet services. Indirect value comes from

the non-personal use of data, in the form of improved public services, a stronger civil society, and scientific and technological benefits for all society. Most current data integration exercises are for interests other than the individual. Google does data integration to sell it to third parties and extract financial value from it. In return they provide an email service or a search service. They don't let me use my personal data (the story they are building about me) nor allow me to share it with third parties of my own choosing. I'm lucky if I even know the story they have about me! This is the same for a range of other government and non-government organisations: they collect and integrate data for their own interests and I have no control over the process or ability to see the story they are developing and share it with whom I like.

The primary objective of a Person Data Commons is to create a platform for safe data integration and reuse within the power of individuals, enabling them to form data sharing relationships as they see fit for their personal value.

The main locus of decision-making about access for personal uses of data should be the main beneficiary or risk-holder – persons themselves.

Data about people is often co-produced (second parties)

Sometimes we capture data about our lives ourselves. I buy a camera and use it to take snaps, so nobody else is involved. At other times, data is “co-produced”. My financial transactions going through my bank are a co-production between my spending activity and my bank's ability to capture and store that.

In these cases, whereas an individual has a special first-party relationship with data referring to them, there is also a second party with a special interest: the co-producer, who is often custodian, user, steward, or enabler of data about me.

This means that, for example, although I have a right to forget, my bank may be required by law to hold financial data for the tax agency for seven years. How well the commons works will be judged in part by how well these co-production relationships and interests can be managed.

Co-producers are often the siloed first user of the data to provide the initial personalised service. The bank, government, app provider who is providing a service and needs to know something about you will capture your servicing data to help administer the service.

Third parties want to integrate and reuse data about me to provide me with a service Third parties are interested in integrating and reusing (and potentially adding new data to) your existing data for a new purpose unrelated to its original capture.

Sometimes they may wish to “add value” to existing data – to have a better way of doing photo analytics to identify and categorise objects in your

How well the commons works will be judged by how well these co-production relationships and interests can be managed.

photos, for example, or a new way of analysing your heart rate to detect medical conditions. So a personal health service provider might wish to use your medical record and your fitness sensing device to build a new kind of exercise monitoring service. Or a budget support manager might wish to integrate your financial data with your budget goals to create a new service for you.

They have a newly formed second-order relationship with you based on the new data they have produced, or the value they have added to your existing data. But they are a third party in respect to the existing data about you that was co-produced between you and other providers.

Interests in non-personal reuses of data about me (fourth parties)

Individual people's stories and their engagement with the world can also be used in non-personal ways: data about me may be used in ways not special to me, but for other purposes. This is a fourth-order relationship, in the sense that is unrelated to particular applications for you.

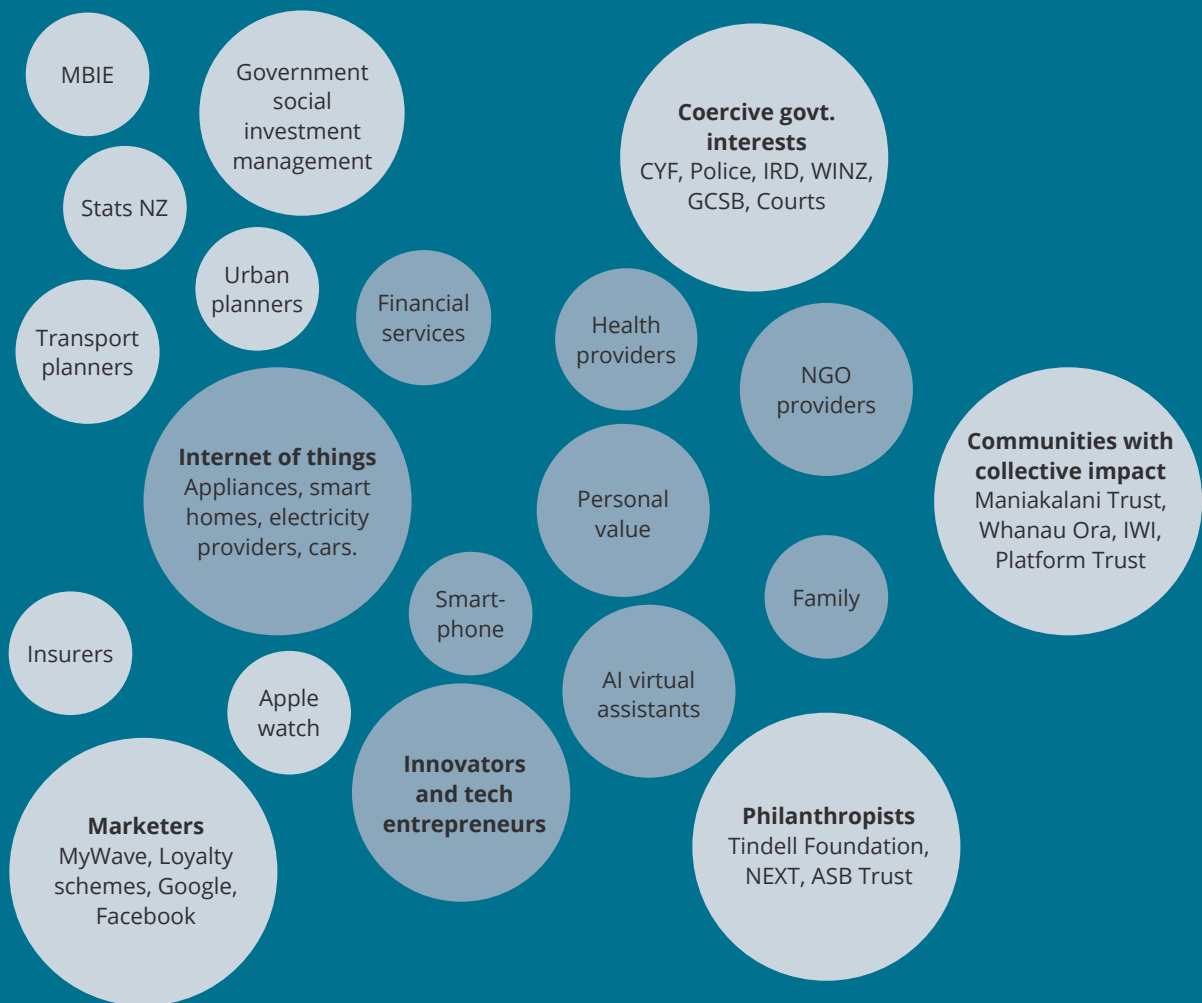
Often data about people is lumped together and used anonymously. A group of people tells a collective story with their collective data. "Our story" might be about a medical condition that is studied by scientists, or a study of what is working or not working for our community. Appropriately anonymised data that is aggregated into groups of people can be used to answer research questions, measure the effectiveness of social services, monitor social service providers, identify improvement opportunities for government, etc. In this way, the common goal that we share is to understand the stories of groups of people and their life paths and engagement with the world to better understand what can be done to improve lives in general. The individual person is not directly targeted by non-personal uses of their individual data, but can be an indirect beneficiary or can help others by adding their story to the collective narrative.

What this illustrates is that the difference between personal and non-personal data is the use to which it is put. Data about people can be used for personal purposes, such as curating a personalised service portfolio for an individual person. But it can also be used in a non-personal way by NGOs, scientists, social investors, entrepreneurs, and others to make decisions based on quality, multidimensional data. The personal data sharing platform needs to be capable of enabling both these purposes in a way that is safe and trusted.

The aim is to develop an information exchange that empowers people whilst also delivering more value for business, scientific, and common interests, so creating a stronger and more prosperous civil society. Here the commons thinking applies. My data can be used in non-personal ways (ways that do not impact me directly) if this is for wider common benefits. Bearing in mind the levels of distance from first-order through to fourth-order relationships to the subject of the data – you! – we turn now to particular kinds of interest in your data. What *kinds* of interest are there in personal data?

Who wants to use my data?

Who has an interest in (re)using integrated data about people?



○ Non-identifiable access? ● Limited identifiable access?

Marketing interests

Much of the technology around big data has grown out of the opportunity to monetise Internet and social network data by selling it to marketing companies. Google, Facebook, and many other high-tech companies both in New Zealand and overseas integrate personal data to develop new marketing methods; Flybuys is a familiar example of this. Marketers profit from the integrated personal information that allows them to target services that inform consumers about products. The people whose data is being used receive a free service such as email, a social media platform, or rewards points, and some better-targeted advertising, but the value of the integration of data itself (the insight itself about the person) is sold off to others to use for their benefit, not for the person whose profile is being built through integrated data. People are not receiving the true value of their personal data.

So none of this is high-trust or high-value to the individual; the people whose data is being used and sold have no access to their own data, and they have no control over what it is used for. They are excluded from the value chain; in most cases there is little to no transparency around the use and re-use of their data.

Embryonic attempts to hand back some control to people is emerging. Privacy settings controls are rudimentary attempts to hand back limited control. Ability to delete search histories on Google provides another layer of control reflecting “right to forget”. But none of this comes close to what is possible in terms of control.

A better approach is the development of Personal Information Management Systems (PIMS) such as MyWave here in New Zealand. These kinds of system allow people to capture and derive value from their own data. Users can upload high-quality data about themselves to a PIMS and create a consent-based relationship with marketing companies: a direct relationship that returns value back to the individual, often in the form of loyalty schemes, discounts, and bonuses.

Making this relationship direct so that transactions take place between the individual and the marketing company in a transparent setting gives the person generating the data more control, and includes them in the transaction. These relationships are an improvement on traditional loyalty schemes for both parties. Marketing companies can simply ask for the data they want, and use it with its subject’s consent, which means they can derive

more value from it. Individuals have control over who uses their data, have a transparent relationship with the company using it, and receive benefits in exchange.

Not-for-profits, iwi, community groups, and service provider networks

Any group with an interest in understanding and integrating data about its individual members, in order to mobilise for and orient community-based activities, can benefit from access to a peer-to-peer market for data exchange and sharing. The exchange of data facilitated by a commons platform would support activities such as advocating for policy, or monitoring the health of the community. There are opportunities for both non-personal uses of shared data in this space, such as research into housing needs for informing an advocacy programme, and personal uses, if individuals within a community consent to their data being used to solve particular problems.

Examples of organisations who would benefit from using data in this way include the Platform Trust and Te Pou (who are mobilising around mental health and addiction) education entrepreneurs such as the NEXT Foundation, the Manaiakalani Trust, Tuhoe, and the Canterbury health alliance network. Low-cost data sharing and integration, and interoperability across communities of interest on a high-trust Data Commons that is in the control of the participants, will improve the ability of communities to mobilise the value of their own data and community.

At present, the information landscape of the NGO sector is a patchwork quilt of information systems that would yield enormous benefits for their owners if they could be joined up for easy sharing and access. There is a great willingness to share information amongst NGOs, as it's widely acknowledged that service delivery could be improved for clients with more and better access to data. There is scope for operational effectiveness gains, the development of new and better commissioning and funding models, innovation, and collective action.

There are several key obstacles to data sharing in the NGO sector, including disparity in information-gathering systems developed in isolation from one another; and control over who can see the data, for protection both of clients' privacy, and of NGOs' business interests in a competitive contracting environment. Narrow interests trump the collective and local value of a more general solution.

A Data Commons solution for the sharing and integration of data about people would address both these issues.

A Data Commons solution for the sharing and integration of data about people would address both these issues.

Disparate data –in terms of both content and format – is an obstacle to sharing and integration, as NGOs have a wide range of data entry and storage processes. Many reporting metrics are of little to no value in assessing service performance or individuals' outcomes, particularly in isolation. By sharing and integrating data about people, a Data Commons would provide NGOs with opportunities to improve their services, become more responsive, learn about their clients, take a more holistic approach to clients' wellbeing, and collaborate with one another using shared measurement systems to work towards common goals.

The control principle that would be applied to the Data Commons is of particular relevance for the NGO sector, as NGOs' sharing of data with one another would need to be a controlled transaction to protect the privacy of the individuals whose data is being shared. Additionally, NGOs are unlikely to want to share their data without direct control over who can see it, and what they can see. With a decentralised network model that enables direct peer-to-peer sharing, NGOs could be confident in their control over who can access their data, and secure in the safety of a sharing system that does not depend on a vulnerable central hub. With a guarantee of control over the sharing of NGO data, the Person Data Commons can enable the creation of enormous value for NGOs, and for the public who use their services.

Business, innovators, and technology entrepreneurs

The market for personalised services is growing, with products such as Apple Health beginning to respond to demand for a high-trust, personalised health service. There are two components to this business model, representing both personal and non-personal use of users' data; the first is that Apple's hardware products provide an information and networking opportunity for the integration of medical records, test results, and personalised health information, integrating heart rate data and third-party data for a bespoke service that caters to the individual. The second, non-personal use of the data is the opportunity for users to donate their personal health data for scientific research, with their consent. This non-personal use of personal data yields indirect benefits to users and their communities, whilst protecting their privacy.

This is a great example of a fledgling Data Commons that offers different levels of access to different parties based on users' consent, delivers value for individuals, and yields a profit for business.

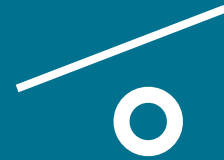
A Person Data Commons would deliver many of the same opportunities in terms of product and service development and monetisation for profit, but without the need for a single centralised repository placing control of the data in the hands of a single corporate entity.

A peer-to-peer data sharing platform that facilitates the integration of personal data from multiple sources would offer New Zealand businesses most of the same opportunities that Apple is capitalising on with Apple Health, but in a more democratised, low-risk fashion. By decentralising the network and keeping data ownership in the hands of those who generate it, the commons-based structure will keep individual peoples' data safer and provide more opportunities for multiple businesses to innovate off the back of the insights the integration of such data might generate. A Data Commons that is owned by its participants, rather than by a single provider such as Apple, is more inclusive and therefore offers more value to all participants.

An effective Data Commons will enable small to mid-sized technology and analytics entrepreneurs to grow on a level playing field – in the same way that the Internet protocol enabled new startup retail businesses to enter a market traditionally held by monopoly providers. The costs of scaling are diminished if the Data Commons is inclusive.

Scientists and researchers

Both traditional and big-data-based science, such as the precision health work being done in the United States and by Orion Health here in New Zealand, receive a huge amount of value from high-quality data on individual life pathways. A Person Data Commons would give scientists and researchers the ability to integrate their own research data with existing data on the commons.



Government

The government uses data about people to make policy and spending decisions. At the present, government is committed to the adoption of an Investment Approach at a high level: this is a methodology for the treatment of public service spending as an investment that yields ongoing value throughout a person's life. It requires rich, integrated, longitudinal data about people to fuel the analytic tools that are being developed to assess the effectiveness of government investment.

There are several core objectives and ideas behind the Data Commons approach as these relate to the government's Collective Impact work. The Data Commons working group accepts the findings of the New Zealand Data Futures Forum that New Zealand's data sharing ecosystem needs to be high-trust, inclusive, controlled by its participants, and of high value to those participants. Our work aims to take those principles and turn them into a practical roadmap for how to build an ecosystem.

Citizens using the Data Commons as their preferred platform for data sharing have ultimate control over the use of that data, including right to forget. In this high-trust environment more people are likely to contribute more (and more accurate) data, and more marginalised people are more likely to engage with this kind of platform than with a government data-sharing platform.

If a citizen is in control of their data relationship with potential service providers and innovators, those providers and innovators are more likely to have access to citizen data which will drive more innovation, collaboration, and use. This avoids ownership and control by centralising interests such as big business or government, and allows the citizen to form a direct relationship with potential providers of new services. The innovators don't need to ask permission of monopolising interests (such as the Ministry of Education), but rather have to form a high-trust relationship with the person whose data it is.

The value of this to government is that more marginalised people are more likely to put more accurate, more complete, and more interesting data onto the platform where the government can use it under limited licence to achieve its objectives of monitoring investments and developing policy and planning solutions. Because citizens are in control of their data, they are more willing to share more of it with government for their own value, because they know that they are still in control of it.

The government is likely to have only de-identified access to integrated data unless this suits the citizen. This means that coercive uses, such as detecting fraud, policing, and child protection, are unlikely to be possible using the Data Commons, but will have to be undertaken using traditional sources of data.

Overall, more lives will be saved, more economic value delivered, more children helped. Health and social outcomes are likely to improve by an order of magnitude where citizens have control over their data and can trust that it will not be used coercively, since this will allow them to form relationships with trusted parties (such as the Salvation Army, their GP, their teacher, their bank, their budget advisor, their fitness instructor). The personal value created for them by allowing their integrated data to be used by trusted parties will help them meet many of their needs. The few lives that would be saved by coercively trolling through the Data Commons to find child abuse or fraud will be vastly outweighed by the benefits to other children who will be able to engage trustingly with their GPs, mental health providers, etc., to get what they need. There is a high cost to the whole community when coercive uses of what would otherwise be freely shared data erode trust and the ability of citizens to form good relationships within and across their communities and with their providers.

Digital information about citizens will increase by several orders of magnitude in the coming years. Whilst up until the Internet, probably only the government and some businesses such as banks were collecting digital information about people, the situation has changed: we have personal sensing devices, the Internet of things, electricity providers, heart rate monitors, car GPS and driving performance monitors, and a host of other forms of data (see diagram below).

This data will be more accurate than government data. Some of it will be nearly real-time. It will be more content-rich. It will be more highly personal. There will be more data about people who do not typically engage much with government – precisely what is required to understand deprivation and what works to address it.

Government data is minimal, largely focuses on servicing, and is likely to be less accurate since people and government officials both collude in collecting it badly. It is probably data about only 2 to 10 per cent of a person's life depending on whether they are a high or low user of government services.

The government is only one sector or interest group that has an interest in personal data (see diagram below). People need to have data-sharing relationships with a wide range of interest groups such as non-government providers, financial institutions, friends and family networks, high-tech entrepreneurs, digital virtual assistants, etc.

Overall, more
lives will be
saved, more
economic value
delivered, more
children helped.

In summary, there are many groups of people with an interest in integrating and reusing data about people to target things at them (personal uses of data) and/or to generate insights about groups of people (non-personal uses).

What will a Person Data Commons look like?

It is hard to predict in advance exactly what the solution will look like but, if the six Data Commons principles are applied, it is almost certain to have several key elements (see diagram below).

A person is likely to have their own “personal data wallet”, a bit like a Bitcoin wallet, to facilitate their control. This will have several key functions:

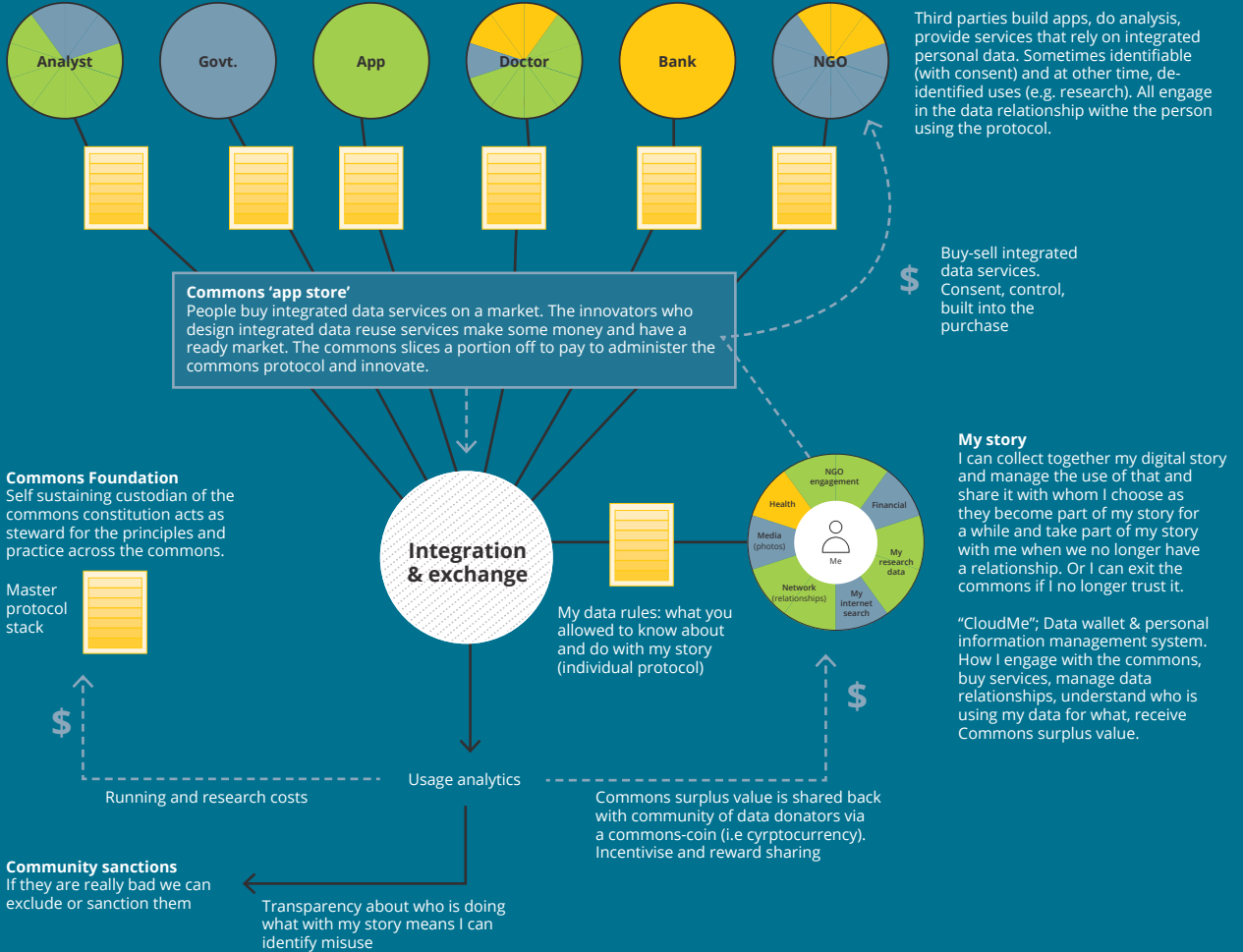
- It is a Personal Information Management System. I get to upload and manage my own data: I can set permissions for non-personalised uses of my story (who can use my data in a non-personal way); form and terminate data reuse relationships for individual personal uses of my story; and extract a copy of the data we have co-produced together (like number portability, I can take a copy of my data with me).
- It provides access to the market for service providers – rather like an app store where entrepreneurs can market various personalised reuses and services reliant on access to part of a person’s story. So a person could buy an app that uses the protocols shared by the commons.
- It may be the method by which any surplus value generated by the commons gets shared back with the individual who has a share in the commons (by sharing their story for non-personal uses).
- If the person terminates their relationship with the commons (for example if they lose trust because the government seeks to nationalise it), their data wallet may enable them to take a copy out and keep a record of their story, in case they want to hook back in or join a rival (more trustworthy) commons. It is possible that a person’s personal data wallet will be located on their own digital device (computer, cell phone) and can be added or removed from the commons as per their needs.
- Any use of their data on the commons might be tracked and monitored – showing who is using data for what – to provide transparency across the community.

This commons PIMS probably needs in the first instance to be commissioned by a commons foundation. However, it is just one portal onto the commons, and rival commons PIMS should be possible – if they meet the community protocol standards.

Commons advisory companies
Supports data entrepreneurs to develop data specific protocol rules and/or apply existing protocols to develop apps that meet community standards.

Our stories
We can share our collective story and remain anonymous to safely enable scientists, investors, researchers, service evaluators analyse our collective story for our collective good.

My service relationships
I can reversibly consent to have a relationship with somebody and they can integrate specific parts of my story for my benefit because I have a high degree of control over my data use and trust the commons protocol.



Sharing parts of my story for personal value

Second-party and third-party providers are likely to develop specific solutions via an "App Store"-style interface, where they can form relationships and obtain consent from customers to use specific elements of their data wallet to provide a particular service to them. So, for example, the Salvation Army might have a budgeting app which requests access to financial data and links this to a budget plan and a forecasting model. This would be an app that the person could choose to use and that obtained their consent. But the service provider/entrepreneur must adhere to the community rules and build into their app the right for that person to withdraw their data and to manage various levels of consent on an ongoing basis. They must provide consent mechanisms and an easily

understandable form, not a long laundry list of legalese. They would have no rights to on-share the data or copy it. The data must be stored in such a way that it is available for the person to withdraw it, and to hold it for themselves electronically should they do so.

Our Story

For fourth-party non-personal and non-targeted uses of personal data, scientists, researchers, public policy planners, or urban planners, would have access to integrated de-identified data. The commons community needs to determine the protocols around consent to reuse and whether it should be highly granular or a community-level decision.

We think, however, in accordance with the underlying principles guiding the Data Commons, licence to undertake this form of analysis should be withdrawable in specific cases of data misuse, or when people don't trust the user. And codes of conduct and ethical frameworks would likely be required for analysts using a Data Commons.

So, for example, in the specific case of the state sector's new interest in Collective Impact social investing, the Data Commons community may agree that the government can have a restricted licence to use personal data via the Statistics NZ integrated data infrastructure. Since this is a crown entity, the data is de-identified, and there are strict controls and close supervision over use. This could then be used to develop investment strategies and monitor provider performance for Collective Impact investing.

The commons foundation is also likely to have some kind of market regulation role: looking at use and misuse and administering sanctions; developing community protocols and standards; redistributing surplus back to the commons; and growing the commons.

Where to start? Kick-starting the Person Data Commons

Although individuals about whom data is collected are the primary Community of Interest for a Person Data Commons, a first customer is needed to kick-start the project and attract participants, beginning the virtuous cycle of an expanding network effect and attracting further participants so that data donators receive value in exchange for their data. The ideal first customer for a Person Data Commons is the NGO community, for several reasons.

NGOs want to share data to improve their services, learn about their performance, and coordinate mutually reinforcing activities in pursuit of shared goals. The current data landscape of the NGO sector is fragmented, and this is recognised as a problem. Significant benefits could be delivered very quickly by making more data accessible, with a standardised process for gaining access and a watertight mandate for use based on the consent of data subjects. NGOs report that their frontline staff and clients must collect the same data over and over again, a process that is time-consuming and expensive for the organisation, and tiring and humiliating for the client.

The ability to access a client's data through a Data Commons network would save money, time, and dignity, and make NGOs' service delivery more consistent and responsive.

The market is already producing various manifestations of a Person Data Commons. Apple Health and PIMS are just two examples of private enterprise capitalising on the potential to derive value from integrated data about people.

The main challenge for founding a Person Data Commons that adheres to the NZDFF principles of value, inclusion, trust, and control isn't the technology; it's stimulating the involvement of the first tranche of participants. The best way to create an incentive for communities to invest in demonstrating the value is to set about solving one specific challenge that is of high value to those first participants, but ensure that the methodology is accessible, adaptable, and scalable so that it can be expanded to accommodate other solutions for other communities. The challenge is in ensuring that the initial point solution can be scaled into a generalised solution that offers a diverse spectrum of value to multiple parties. There are already communities of interest in the social sector who are building the trust and shared vision required to coordinate their efforts around a common objective. Some current examples are the Manaiakalani Trust mobilising around housing and education needs, and Platform Trust, Te Pou, and others mobilising around mental health and addiction services with further interests in education and other social needs.

These alliances would benefit enormously from the ability to exchange data safely and efficiently, whilst maintaining the trust of the people they are working for. The key advantage of beginning with these groups as a first customer is that they have high trust with their clients, clear objectives, and an articulated need for efficient data sharing and interoperability. However, the challenge is that whatever we do here has to be of direct value to the people who are donating their data. They can be motivated by a safe harbour, but to get real traction will require demonstrable personal value.

Leveraging the value of the social investment approach

The government is interested in place-based funding models which focus on outcomes. To do this they need access to high-quality individual data. But therein lies the government's problem. People won't trust the government with this kind of micro-level data. So the Data Commons solution provides an effective way to bridge the gap between government interests and individual and community interests in micro-level personal data. Rather than seeing this as a place to start, we should provide the service to the Community of Interest, then use this as a value-added opportunity to engage the government about data sharing from a position of strength. If we go to government too early, the interests at the centre are likely to erode trust as well as the bargaining power of the community of people who have mobilised around their own data interests.

The market is already producing various manifestations of a Person Data Commons.

Governments of both persuasions have signed up to open data and have sought reuse of government assets to drive value and innovation. Medical records, education records, and the like are national assets and assets to individual citizens. If citizens could reuse their own co-produced government data to form other relationships, then you would see innovation in education and health, for example. On several occasions one of our members has been asked by social entrepreneurs, philanthropists, scientists, and technology entrepreneurs for a way in to citizen data – with citizen consent.

The government has a lot to gain by enabling and supporting a Person Data Commons. Doing so will improve trust, obtain better access to data to inform policy, open up and drive social and commercial innovation, and improve New Zealand science. It will also significantly lower IT budgets for large government agencies. There are no substantive downsides for a country in not adopting a commons-based approach to personal data coproduced through engagement with government – except in losses to individual institutions' level of power and control.

