Understanding and unlocking the value of public research data

OzNome social architecture report

Todd Sanderson, Andrew Reeson and Paul Box
# Contents

Acknowledgments.................................................................................................................. 7

Executive summary .................................................................................................................. 8

1 Introduction ........................................................................................................................... 11

2 The economics of public research data .............................................................................. 14
   2.1 What is the value of public research data? ................................................................. 14
   2.2 Realising the value of research data ............................................................................. 15
   2.3 Costs of disseminated research data products ............................................................ 21
   2.4 Reconciling value potential with the costs of disseminated research data products.......................................................................................................................... 26
   2.5 Institutional aspects of data sharing ............................................................................. 32
   2.6 The limits of open research data ................................................................................ 37

3 A case study of the CSIRO Data Access Portal .................................................................... 40
   3.1 The DAP ......................................................................................................................... 40
   3.2 Values and motivations: depositors ............................................................................. 43
   3.3 Values and motivations: users ..................................................................................... 48
   3.4 The costs of disseminated research products ............................................................... 56
   3.5 Reconciling the values and costs of the DAP ............................................................... 57

4 Unlocking the value of public research data: recommendations and approaches .......... 61
   4.1 Unlocking value by strengthening research data product supply chains ............... 61
   4.2 Unlocking value from enhanced dissemination processes and infrastructure ....... 62
   4.3 Value from cultivating research data sharing networks ............................................. 63

5 Conclusion ............................................................................................................................ 64

Appendix A: Survey of data depositors .................................................................................. 66

Appendix B: Survey of data users .......................................................................................... 68

References ............................................................................................................................... 71
Figures

Figure 1: Components of Total Economic Value.............................................................. 15
Figure 2: Research data generation ................................................................................. 17
Figure 3: Research data dissemination ............................................................................ 18
Figure 4: Components of use value of a “data + service” product ..................................... 19
Figure 5: Costs of data sharing ....................................................................................... 21
Figure 6: Cost models of data + research services + dissemination services ................. 22
Figure 7: Possible total cost schedule for DAP with sequential investments ....................... 23
Figure 8: Possible total cost schedule for an individual data collection, subject to incremental service addition .................................................................................................................. 24
Figure 9: Possible average cost (per user or data collection) schedule for DAP with sequential investments .................................................................................................................. 25
Figure 10: Reconciling the value potential of data with the cost of provision ...................... 27
Figure 11: Reconciling total value and total cost in the presence of service addition thresholds ................................................................................................................................. 28
Figure 12: Average value vs average cost for incremental investments in the DAP .......... 29
Figure 13: Alternative research data sharing models ........................................................ 31
Figure 14: socially optimal strength of exclusive use rights .................................................. 36
Figure 15: Innovation-strength curve (Tabarrock 2013) .................................................... 37
Figure 16: Alternative data access models, adapted from Broad (2015) ............................ 38
Figure 17: DAP datasets by download frequency (top 100) .............................................. 41
Figure 18: DAP data collections by type ........................................................................... 41
Figure 19: Average download frequency by type (all) ....................................................... 42
Figure 20: Average download frequency by type (all outside top 10 rank for each category) .... 42
Figure 21: Download count (log adjusted) by length of time published (public data collections) ................................................................................................................................. 43
Figure 22: Which of the below best describes your organisational affiliation? .................. 49
Figure 23: How did you first encounter CSIRO’s Data Access Portal? ............................ 50
Figure 24: Please identify one or more of the following which reflect the reason(s) you are accessing data from the CSIRO’s Data Access Portal ......................................................... 51
Figure 25: What do you intend to produce with the data you have accessed from the CSIRO’s Data Access Portal? .................................................................................................................. 51
Figure 26: On a scale from 1 (low) to 7 (high), how critical is the existence of this data to your work efforts? .......................................................................................................................... 52
Figure 27: On a scale from 1 (small) to 7 (large), what are the likely time-savings that the availability of this data creates in your work efforts? ................................................................. 53

Figure 28: On a scale from 1 (small) to 7 (large), given your current uses for this data (research, product development, etc.), what is the worth (value) of this data to your current work efforts? ............................................................................................................................... 54

Figure 29: On a scale from 1 (small) to 7 (large), considering possibilities for your future work, given your possible uses for this data (research, product development, etc.), what is the worth (value) of this data to your future work efforts? ............................................................................................................................... 55

Figure 30: On a scale from 1 (small) to 7 (large), to what extent, if any, has your use of CSIRO’s Data Access Portal changed your work efficiency (i.e., time saved compared to if no Data Access Portal existed)? .................................................................................................................................................. 56

Figure 31: DAP average cost of disseminated data products by download ............................................ 57

Figure 32: DAP average total economic value vs average cost .............................................................. 58

Figure 33: Australian Commonwealth Public Service value of time savings ........................................ 59

Figure 34: Average net value for Direct + Indirect use value, by download count ............................. 60
Acknowledgments

The research presented in this report was funded by CSIRO Land and Water as part of the OzNome Initiative. The authors would like to acknowledge this important strategic investment that contributes to an improved understanding of economics of research data.

In addition, the authors acknowledge the contribution of the following individuals and organisations to the development of this report: Cynthia Love (CSIRO/ANDS), Greg Laughlin (ANDS), David Lemon (CSIRO), Anne Stevenson (CSIRO), Tim Capon (CSIRO), Melanie Ayre (CSIRO), Gavin Walker (CSIRO) and Professor John Houghton (Victoria University). Their input and feedback has been of enormous value in the development of this work.
Executive summary

Data is a cornerstone of the digital economy. Successful economic management therefore increasingly requires optimising the value of data, from the organisational to societal level. In order to make the best use of data it is necessary to understand its value; however, this far from straightforward. This report focuses on the value of research data generated and held by the CSIRO and other public research institutions. The Australian Government has a commitment to making data open by default, which extends to publicly-funded research organisations.

There are many alternative arrangements through which data can be shared, ranging from simply depositing data files onto internet-accessible servers through to building models which collate and interpret data. Each of these incurs different costs (such as IT infrastructure, metadata documentation or modelling) and confers different benefits to potential users. Collectively these arrangements are defined by the prevailing socio-technical environment within which data is generated, held and used. Within these environments, the incentives faced by individuals and organisations strongly influence the value potential of data collections. While government, or institutional, policy may promote data sharing, in practice there are many ways in which this can be frustrated if data providers are not on board.

Values and costs of sharing data

Data has value because users value what they could do with it now and what they might do with it in the future. In current use, value could accrue directly from the time savings of users accessing data in highly interoperable form, it may also accrue indirectly through applications and services developed from the data. In an uncertain world data also has option value, in that it may find additional uses in the future.

Dissemination services and infrastructure are important for cultivating networks of users and depositors. Developing around particular data collections, networks play a significant role in stimulating innovation both in user applications of data and in the research products of data providers.

The optimality of enhancing the value of data collections through additional services must be evaluated by considering the costs of establishing and maintaining those services. For data depositors the lowest cost model is simply to deposit data online in whatever format they happen to be in. However, this imposes costs on users in terms of interpreting and transforming the data. Overall value can be enhanced if depositors invest some effort into making their data more accessible (e.g. annotating, formatting, documenting metadata).

Data depositors may seek to further increase the value of their data to potential users by adding further services such as modelling or applications. However there are trade-offs between making data easy to use (e.g. by wrapping it in an application) and maintaining flexibility in how it may be used (which is particularly important for maintaining option values). There is therefore value in maintaining alternative versions of data collections, including both the raw data and the highly-transformed data product applications.
High transaction costs are a key source of discouragement to experimentation and innovation, particularly in cases where the potential benefits of data access are uncertain. Minimising such costs is therefore an important goal for those responsible for disseminating data. Investment in dissemination infrastructure can significantly reduce the costs incurred by both depositors and users of data, and so facilitate sharing.

**Evolving institutions and incentives**

The extent to which the potential values of data sharing can be realised in practice is determined in large part by the prevailing institutional arrangements; “institutional” covers the organisations, rules and behavioural norms which apply to data sharing. The institutional arrangements in public research organisations straddle the gap between two seemingly incompatible objectives; there is a desire for world-leading research prestige but they are obliged to act in the public good by making the data products of that research available for public use. There will often be tensions between an organisational goal to share data widely in order to maximise its value, and the incentives for individual researchers to keep their data private in order to gain a competitive advantage in generating research papers or projects. This means that simply encouraging, or even mandating, individual researchers to share data may be insufficient to promote widespread pro-sharing behavioural change.

Identifying successful institutional arrangements for disseminating research data involves solving fundamental behavioural questions, as data sharing will only work if the researchers who generate and hold the data are willing and able to share it. There are a number of interventions available to an organisation to encourage individual researchers to more fully share their data ranging from recognition to financial rewards.

Cultural practices associated with positive data sharing behaviours can be cultivated by an organisation. These may include promotion of data sharing for reasons of promoting transparency in research practice, thereby leading to enhanced research quality and accountability. Among late career researchers there are substantial benefits to be obtained by targeting bequest motives and developing a culture of gifting data to future research generations.

**Testing this in practice: Case study of the CSIRO’s Data Access Portal**

CSIRO makes its research data products available through its Data Access Portal (DAP). As of October 2016 the DAP has 1,928 published data collections, with a total of just over 70,000 downloads since November 2012.

Over the six years that the DAP has been in operation we have a good understanding of its total cost at $13.9M; this equates to $198 per download. Given the strong economies of scale this cost is likely to continue to decline. Benefits are far harder to establish. A survey of a small number of users suggested they realised time savings of many hours through downloading data from the DAP, which equates to several thousand dollars in wage costs. Users reported further values from indirect use and option value. Based on the lowest estimated total value from our survey of $4,802 per download, the annualised total economic value generated by the DAP to its users is in the order of $67 million. While this estimate is highly uncertain, the fact that it exceeds the annual...
costs by nearly two orders of magnitude suggests that the DAP provides an excellent return on public investment.

A key motivating question of this work was to understand the value of data, and consider how this value may be enhanced by modifying the socio-technical environment. Based on the analysis of relevant theories and frameworks presented and the responses to DAP depositor and user surveys, there are a number of opportunities distributed throughout the data supply chain.

Maximising the value of research data

Unlocking value by strengthening research data product supply chains. Data dissemination objectives should be embedded in research data management procedures. Data produced within a research project should be made available at key stages during its transformation, not just at the end of the research project. Supporting this, there is a need to ensure researcher incentives are sufficiently aligned with their employer’s open data objectives. Incentives are an essential key to unlocking the latent value of research data collections which are currently closed.

Unlocking value from enhanced dissemination processes and infrastructure. Achieving scale in the data infrastructure is important, because the greater the number of collections stored, and downloaded, the lower the average cost. Achieving scale on the data supply-side may also be associated with positive network externalities as a critical mass of data collection is realised.

In the data sharing space trust is a significant source of competitive advantage and therefore of value. For example, it seems more than likely that the trust which users place in the quality of CSIRO’s data is a key driver of the value of its data collections. A position of trust can only be maintained by adopting the appropriate due diligence mechanisms to ensure, as best as possible, that data collections with errors and critical omissions are not released to the public.

Value from cultivating research data sharing networks. The value of research data can be enhanced by actively promoting the development of user-depositor networks around data collections such as the DAP. Value created in networks is determined by the nature of the links between groups: user to user, depositor to user, and depositor to depositor. Establishing links between groups or reinforcing existing links will have payoffs, although in early stages the formation of these networks may well require support. In order to support this, a more systematic approach to collecting information on users and depositors is required to facilitate more detailed analysis of the research data marketplace.
1 Introduction

It is increasingly clear that the successful economies of the 21st century will be defined by the strength of the competitive advantages which arise from the ability to generate value from information and data assets\(^1\). It is argued that a key precondition of this strength is a society which embraces the openness of data, in particular those data held by its public institutions (World Bank 2014). Public data is increasingly being made openly available in Australia, with the Australian Commonwealth Government recently reaffirming its commitment to make non-sensitive public data open by default\(^2\).

Australian research organisations, including the CSIRO, are also increasingly adopting an open data philosophy driven, in part, by extrinsic pressures to make publicly funded research data open (Houghton and Gruen 2014). These efforts are supported \textit{inter alia} by the Australian National Data Service (ANDS).

In this context, research data may be defined as:

“the factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated” (OECD 2007).

Discourse around the movement towards open data often expresses the need to unlock the value of data as an important objective. This is a complicated notion, because it raises questions about the potential value of data and the barriers to realising that value. Data has a value because users value what they could do with it now, and what they might do with it in the future. Visitation and download counts, which tend to be the current proxy measure of value, are only likely to capture a small part of the potential value of a data collection. A more rigorous approach to valuation is needed to expand our knowledge of the value of research data and associated dissemination and sharing services.

It is important to understand that the full value of data is determined by features of the socio-technical systems (Hanseth and Monteiro 1998). These systems enable the creation, sharing, use, and (re)combination of data. Socio-technical systems refer to the interaction of people and technology, with particular implications for the behaviour of people in alternative technical system designs.

OzNome: enabling information supply chains

In government, research and industry, there is a demand to extract greater value from data in the context of rapidly changing business models that place many organisations at risk of disruption. In order to meet this demand, society must address the complex challenges inherent in the information value chains, through which data is collected, stored, integrated, processed, and

---

\(^1\) For example, McKinsey (2013), \url{http://www.wired.com/insights/2014/07/data-new-oil-digital-economy/}

shared. Key blockages in these information value chains need to be identified to ensure that costs, benefits and incentives optimise information flow to where it can create the most value.

The costs of discovering, accessing, combining and using data, from sources requiring significant human intervention, is prohibitive. However, with rapidly advancing computational and network capacity and advances in the application of statistics, mathematics and autonomous systems, there is enormous opportunity to extract more value from data.

Recognising these opportunities and challenges, CSIRO initiated an ambitious research program called OzNome\(^3\) to address the set of complex interrelated challenges in order to assist in unlocking the value in data. This initiative will seek to address the significant inefficiencies and challenges that exist across many information supply chains and transform the infrastructure, institutions, policy, and behaviours around sharing and use of information.

The key issues to be addressed through OzNome are:

1. **Barriers to discovery, access and use:**
   
   Data intensive workers spend significant proportions of their time ‘wrangling’ data (i.e. attempting to find, access and use it). Data is often difficult to find. Access is regularly impeded by data either not being online or stored in non-standard ways. Access can also be impeded by inappropriate or overly restrictive licencing conditions imposed by data custodians.

   Once access is attained there are still many barriers to use including difficulty in understanding the data, poorly described structure or semantics, being unfit for purpose or incompatible with a tool of choice.

2. **Lack of incentives to share data:**

   In many cases, there are insufficient incentives to share data as the costs of making data easier to find, access and use are borne by the data provider while each user benefits. There exist many opportunities for individuals and organisations to collaboratively develop the mechanisms that ensure these contributions are recognised and valued. This may be reflected directly in agreed performance metrics within an organisation, or indirectly through enhanced recognition of the value of data sharing within the research community.

3. **Complex social and institutional arrangements:**

   The social and institutional context is complex, because there exist many interacting forces at levels from individual to the organisational and global. Among these, concerns may arise from ethical issues (privacy, confidentiality and security), the social license to operate, and the interaction of legislation, regulation and policy. The inability to rapidly develop and adopt agreements that enable sharing of data and models is a major impediment to efficient utilisation of data assets.

   Solving these challenges requires resolving the interwoven social, institutional, economic aspects of data sharing in order to inform the implementation of viable technical solutions. This is being achieved through the development of ‘social architecture’, a deliberative approach to designing environments conducive to the transformation of information supply chains. The research

\(^3\) https://research.csiro.au/oznome/
presented in this report was undertaken as part of the broader social architecture research agenda, providing an important economic perspective on the barriers to and incentives for sharing data based on the costs and value associated with data.

This report examines these issues from the perspective of economics. Section 2 considers the various sources of value for research data, and the ways in which research and dissemination services enhance or constrain the potential to realise value. Data is a highly heterogeneous good, which in the case of research data is bundled with a variety of information services. There exists a need to develop a framework to value data in the presence of the bundled information services created by research. We examine the institutional arrangements within which research data providers operate, and explore some implications for how the socio-technical environment that supports data supply chains can be positively influenced.

In Section 3 this framework is applied to understand the value of the data collections and dissemination services associated with CSIRO’s Data Access Portal (DAP), a data repository established for the purpose of sharing research data. We report the results of surveys with data depositors and users, and interpret these results in the context of the theoretical framing from Section 2. Finally Section 4 considers how these theoretical and empirical insights may be applied to unlock the value of research data.
2 The economics of public research data

There is a small, albeit rapidly expanding literature on the economics of data assets and the associated socio-technical (technical infrastructure and social, institutional and economic) systems within which data sharing is embedded. We reviewed this literature to help us answer two key economic questions: (1) what is the value of public research data to society and, (2) how is the realization of that value impacted by socio-technical systems and bundled information and infrastructure services?

2.1 What is the value of public research data?

In simple terms, data has value because users value what they could do with it now and what they might do with it in the future. If use is the key measure of value, then this may suggest that the rate of download of a particular collection provides a proxy of its value. However, while download statistics may reflect popularity they say little about value. There is a need for a more rigorous approach to valuation to supplement our knowledge around the value of research data and associated services.

Economists attribute the full range of values which establish the worth of something, total economic value, to the sum of use value (direct, indirect and option) and non-use value (existence and bequest) (Figure 1). Direct use values are those which arise from the direct primary consumption of the good. For example, consider the different kinds of value associated with an apple tree. An apple tree has a direct use value because it produces apples. Similarly, data has a direct use value, because people can directly use and reuse the data for alternative applications. Indirect use values arise from those who benefit from the use of the good by others, for example, third party consumers. An apple tree provides indirect use values when someone enjoys an apple pie. Data provides indirect use values when someone benefits from an app developed from the analysed and transformed data.

Option values attach to the ability to use the good for some purpose in the future, this may be for a known future use or an as yet unknown use (for example, Macauley 2005). We might value the apple tree in terms of the future option of converting it into timber. Data likely has a significant option value, because every future application of the data that we can envisage and many we cannot yet foresee add to its option value. As uncertainties about the future become more significant, the size of the option value will increase and possibly come to dominate other sources of value. In an environment of uncertain and rapidly changing technology and knowledge, it is likely that option values dominate these other use values.
Existence and bequest values are more abstract and may be less important in the current context. Existence values reflect the idea that people like that we have the particular good, but have no intention of using it. This may seem odd, but people do have values for national monuments which they never intend to visit. Our apple tree may well have existence value because people value its contribution to environmental amenity and biodiversity. Publicly funded research data collections have these values too, because the products of scientific endeavour are a kind of national asset (for example, Bhattachary and Dalziel 2012). Finally, bequest values reflect the idea that we can provide goods as a gift to future generations, which is like an inter-generational extension of the option values. There may be value in preserving the apple tree because of the fruits, timber and environmental amenity it confers to future generations. Curated data collections which are maintained indefinitely will have significant bequest values, especially where collections are contributed for purposes of scientific posterity.

In any case, measuring the size of these values is a challenge for various reasons, not least because users are unlikely to be revealing the totality of their valuations by paying for access to research data collections. Even if they were, a user’s willingness to pay for access primarily reflects their use values and says little about their non-use values. Disentangling particular sources of value remains a complicated affair with a variety of approaches in common use, such as contingent valuation (Beagrie et al. 2016). Nevertheless, it is possible to use these various sources of value to attribute some magnitude of worth to the data collections, even if we can’t be certain about the exactness of the resulting number.

### 2.2 Realising the value of research data

In terms of open data, it may be argued that value is created as soon as the data itself is made available. There exist a multitude of ways in which data can be made available, some more ideal than others. In recent years a set of universal open-data principles has emerged, which it is argued are necessary for organisations which seek to facilitate the effective sharing of their data which
maximises its value potential (for example, OECD 2007; Donker et al. 2016). A condensed summary of these include:

1. **Completeness** – data with associated well-defined metadata, and to the highest level of detail possible, subject to privacy constraints;
2. **Primacy** – data is available as it was when collected at the source, including relevant information to allow verification by users;
3. **Timeliness** – data should be released as soon as is practicable;
4. **Access** – data should be accessible easily in physical or digital format;
5. **Non-proprietary formats which are machine-readable** – data should be presented in open file formats that allow machine-processing;
6. **Non-discrimination** – data is available to users with no requirement for identification or justification;
7. **Licensing** – data may be used without requirements for attribution;
8. **Permanence** – data should remain accessible indefinitely, with facility to version-track;
9. **Minimised usage costs** – data should be available at no charge, where possible.

However, simply making data as open as possible isn’t necessarily going to maximise its value creation potential⁴. Potential users will face impediments to utilising data, such as the inability to interpret or navigate data collections. This suggests a role for some level of additional service which enhances the interoperability of those open data collections. It seems likely that in order to realise the full value potential of a data collection we need to consider the nature of the services provided by the socio-technical systems within which the data is managed. Services here are the features of the socio-technical systems (Hanseth and Monteiro 1998) which leverage or constrain the value inherent in the data.

In this sense, the value of a data collection could be described as arising from the combination of data and its bundled services. These include the myriad of technical services that facilitate digital access to the data and the institutional culture which supports the generation and sharing of data. We might consider informational services, which provide users with the tools to be able to analyse and develop products from the data. A significant contribution comes from the human-intellectual services which include publications, analysis and expert interpretation. In fact, raw data only come into being because some combination of these services has been applied to design a collection regime (e.g. experiments, observation, survey), subject to guidance from theory and experience.

In the context of research data, a significant portion of these might be described as research services which are applied to initially raw data to generate research data products (Figure 2). These research services can range from the basic, such as the processing of data, to the advanced, such as modelling and simulation. The research services are incrementally adding value to the raw data as those services become more advanced.

---

⁴ The Economist (http://www.economist.com/news/international/21678833-open-data-revolution-has-not-lived-up-expectations-it-only-getting)
Services are also wrapped around data through the process of dissemination and sharing, which we might describe as the data dissemination services which are applied to research data products as we make those products open to the public (Figure 3). In other words, these services create the disseminated research data products. These are the services added by the data curation, distribution and storage infrastructure, which in the case of the CSIRO are the service domain of the DAP. Data dissemination services can also range from the basic, such making the data web accessible, through to the advanced, for example presenting linked and highly transformable data collections (Berners-Lee 2016). The data dissemination services are distinct from research services and are likely to be common to any public agency who is engaged in the process of making their data open. This distinction is important because we need to recognise that research services nor data dissemination services alone won’t produce high value open data. If the potential value of data is to be realised then the complementarity of research services and dissemination services needs to be recognised.

*Figure 2: Research data generation*
The sources of value are closely linked to the manner and extent to which research and dissemination services are bundled with data to create data products. The relationship between the degree of service addition and the contribution of various components of use value to the total value could be represented by Figure 4. Raw data may be characterised by limited direct use value and indirect use value in its pure form, although its value is likely to be strongly dominated by option values. At this stage option values are likely to be at the maximum because the flexibility to utilise the data product is at its greatest. Research and dissemination services create use value by breaking down impediments to use. Raw data may be unusable for the majority of potential users of a particular category/type of data, but the addition of research services can make it more usable. It seems quite likely the number of potential users grows in direct proportion to the addition of research services. So too will the addition of dissemination services make data products more usable by a steadily increasing number of users. In this manner, the further investment in bundling services in this framework make the data product more valuable in terms of direct and indirect use values. However, these may also have the potential to diminish the option value by steadily locking up the ways in which the data product might be used in as yet unforeseen ways in the future.
Figure 4: Components of use value of a “data + service” product

As an example, consider the relationship between raw weather data and an associated weather app. The app is a complex bundle of research and dissemination services, which could be thought of as being on the more advanced end of each of the spectra presented in Figure 2 and Figure 3. The raw weather data has an associated option value which is greater than a data product which is presented as an app. The data is embodied in the app, but the flexibility to manipulate that data is constrained by the parameters of the app itself. As we constrain the flexibility in the use of the data in this way, we diminish the option value. On the other hand, because the app is accessible to a greater range of users with embedded analysis and data visualisation tools, it is associated with substantially greater direct and indirect use values when compared to the raw weather data.

This may present a problem in the dissemination of research data products. If the data product is only made available at one stage of the development (service addition) process, we are foregoing some potential sources of value. If research data products are made available at all stages of the development process, then we may well achieve a situation in which we are simultaneously maximising all of these possible sources of value. This relates directly to the principles of (1) completeness and (2) primacy outlined in Donker et al. (2016). That is, in order to maximise the value potential, primacy requires that we make available research data at all stages of the research process, and completeness requires that we fully document each of these stages. There are obvious costs associated with this approach and social-institutional and technical challenges in managing provenance of related data and information products.

The multitude of ways in which research data products can be generated and disseminated with different degrees of embodied service have direct implications for the formation of data user and depositor networks (Bansler and Havn 2004). Networks in this sense refers to the distributed relationships between users, depositors, or a combination of both, which form around particular types of data. This is important, because value is also created by the nature of the network effects
which develop around particular types of goods, by which the interactions among users and/or supplier of the good enhance the value of the good itself. In other words, the more people who use or supply goods of this type the more valuable the good becomes. These are known as positive network externalities, or demand-side economies of scale. Usually, this property is observed in goods which are largely non-rival in their consumption, such as data and associated data collection infrastructure. Telephone networks are a frequently cited example of this phenomena in which we observe positive network externalities; the value of having access to the network grows as more users also have access. Each additional user of the network enhances the value of the network at a greater rate than the corresponding cost of network connection. The rate of value contribution to a network by additional users is not well known, although Metcalfe’s law postulates that the total value of the network is proportional to the square of the number of users \( (N^2) \) (Shapiro and Varian 1999) and Reed’s law posits that this growth is exponential \( (2^N) \) (Reed 2001). In short, an additional user to join the network may contribute value to the network which is an order of magnitude greater than the immediate previous joining user.

There are direct parallels to the data networks which are established around dissemination service facilities, which enable the interconnection between data collections, users and data providers. There are a number of ways in which positive network externalities may form. First, we may observe positive network externalities where the interactions of data users stimulate greater value, for example, through enhanced rates of innovation in product development delivery. In this case, positive network externalities are a key driver of innovation and their subsequent diffusion and adoption (Cabral 1990). Second, positive network externalities can occur on the data provider side, with the value of a contribution to the collection of data collections representing more than its individual value. Possibilities to combine data collections to develop new knowledge or new pathways for product innovation mean that the greater the connectedness of data collections, the greater the value of the data. This suggests that each additional data collection which is added to a data repository has the potential to contribute to the total value at a greater rate than the preceding collection.

Indeed, it is possible that the first and second types may interact with each other in which case we have a third option, that we may observe two-sided network effect. In these cases, an increase in usage by one type of user (type A) increases the value to another type of user (type B), which in turn causes user type B to behave in a way which enhances the value to type A users. These may also be called positive network externality feedback loops. For example, in a data sharing network the popularity of a particular data collection among users may stimulate enhanced value (for example, citation and prestige) among respective data providers, which in turn encourage them to deposit additional data and so enhance the value to the users of the data. Positive feedback loops can only be sustained when the flow of information between users and providers is well established. Providers may not know what features of a data collection are desirable from the perspective of a diverse and possibly anonymous group of users. So too, the ways in which users could create value for providers may not be adequately aligned to the institutional incentives of the providers, which means that the positive feedback loop is unlikely to be sustained.
2.3 Costs of disseminated research data products

The research and dissemination services involved in making data open are associated with costs, which vary widely depending on the data-holding organisation’s chosen sharing model. We can think of the costs of sharing data being divided among data providers and data users (*Figure 5*), among which the incidence of particular costs is dependent on the sharing model instituted by the organisation.

*Figure 5: Costs of data sharing*

We can further divide data providers among the research data depositors (data generators), who are engaged in research service provision, and the data dissemination service providers. Among these three groups involved in the data sharing process there are costs which are specific to each group, for example, the costs of dissemination infrastructure are borne entirely by the data dissemination service providers. It is difficult for dissemination service providers to shift these costs to other groups in the sharing process, although the possibility of cost recovery arrangements shouldn’t be precluded. In this way of thinking, research data product costs are specific to the research data depositors and data utilisation costs are specific to the data users. There are also costs which can be shifted from one group to another depending on the sharing model that is developed to disseminate the data. An example of this is the transaction costs of data acquisition, which include costs related to search and access. Better infrastructure translates into lower costs for data users.

In one sharing model, these costs may be borne almost entirely by users in a sharing model which involves very limited effort on the side of the data dissemination services. For example, the data is web accessible, but is hard to find, navigate and download. This involves high costs for data users, but relatively low costs for the dissemination services group. In an alternative sharing model, those costs may be borne...
largely by the data dissemination services group by investing heavily in the dissemination services which make the data highly accessible. In this case, the data user group bears smaller costs of data acquisition, because there is a lesser effort required to find, navigate and download.

**Research and dissemination services costs**

The question from an economics perspective is simply, which group is best placed to undertake particular tasks and bear the associated costs? The range of possibilities are captured in the alternative data sharing cost models, which are presented from the perspective of the data providers in *Figure 6*. In general, the higher the costs incurred by the data providers, the lower the likely costs to the data users, and vice-versa. Each of these incremental changes in the level of service generates greater value potential. Services make the bundled data product more accessible and potentially more tailored to the needs of particular users, thereby increasing the direct use and indirect use components of total economic value. However, each incremental addition to the level of service involves greater cost to the provider.

*Figure 6: Cost models of data + research services + dissemination services*

These costs may come in the form of one-off investments, or as a cost which is incurred for each additional user. For example, database infrastructure establishment costs, which economists refer to as sunk costs, tend to be invariant to the number of users, at least up to the point where congestion requires investment in infrastructure expansion. Those employed to curate the data
collection on a continuing basis may be thought of as a quasi-fixed cost, which may or may not be invariant to the number of users. Other costs, such as electricity to run the database infrastructure likely vary in direct proportion to the number of users. The majority of costs likely to be experienced in this instance are of the sunk and quasi-fixed cost form, because many of the costs for an institution such as CSIRO in making their data assets open are largely invariant to the number of users. The structure of costs of this type produce total cost schedules which tend to be discontinuous and stepped with incremental investment (A, B and C) commitments as user and/or depositor demand on the DAP requires service expansion (Figure 7). In fact, a typical cost schedule for an individual data collection, subject to the addition of services (Figure 2) would tend to follow a similarly stepped schedule with the complexity of service addition (Figure 8).

Figure 7: Possible total cost schedule for DAP with sequential investments
The dominance of sunk and quasi-fixed cost in this context raises the possibility that supply-side economies of scale may be realised. In simple terms, economies of scale of this form are present when the average cost of providing something decreases as more is provided. This may occur on both the depositor side and on the user side. First, as more data collections are added to the DAP the average cost of contributing additional collections decreases. This is reflected in Beagrie et al. (2010) who surveyed research data storage groups and found that the average cost per MB stored in collections tends to decrease as the total size of the collection grows. Second, as more users access and download data collections from the DAP, it is possible that the average cost of providing access to a collection would decline, albeit with a somewhat lumpy pathway as further investments are made.

Such economies of scale can help to drive a greater wedge between value (benefit) generation to users, and the cost of providing the data collections to those users. For example, for the DAP, subject to the incremental investment costs, we might expect to see a schedule of average costs per user, or per data collection, decline in the direction of increases in either of those Figures (Figure 9). We might expect a similar pattern of declining average costs to appear for individual data collections, for a given initial choice of service addition, but only with an increasing number of users. The presence of declining average costs helps us to answer the question about who is best placed to undertake particular tasks and bear the associated costs. In this case, the data dissemination services group can achieve increasingly lower costs as more data users access their services, which suggests they are best placed to perform these activities and thereby enhance the value of the data to society. If data users are required to perform these activities, we are unlikely to achieve the same social value outcomes.
Figure 9: Possible average cost (per user or data collection) schedule for DAP with sequential investments

### Transaction costs

Beyond the research data and dissemination service costs, sharing activities are associated with transaction costs, which may arise in any act of exchange between data providers and users. Donker et al. (2016) outline the sources of transaction costs through a possible sequence of a data acquisition exercise. From the user’s perspective, these costs can arise from (1) searching for the data supplier, (2) inquiring about the general conditions of exchange, (3) defining the exact characteristics of the data, (4) acquiring and verifying the quality of the data, (5) understanding relevant licensing and fee conditions, (6) obtaining the actual dataset, and (7) uploading acquired data into software, harmonising, and adapting the format. There are also transaction costs which may apply on the depositor’s organisational side, such as access negotiation and contract establishment, as well as enforcement of agreed conditions of use.

In some cases transaction costs (real or perceived) can be prohibitive, such as in the presence of substantial time costs, which may result in the costs overwhelming the potential benefits of having access to the data. In cases where the potential benefits of data access are uncertain, prohibitive transaction costs are likely to discourage use (Quiggin 2010). This is problematic in instances where data could be employed in innovative applications, which are inherently uncertain in the likely benefit generation. High transaction costs are a key source of discouragement to experimentation and innovation (Baca 2006).
Several factors may act to reduce transaction costs in the sharing of data, with the biggest factor being the provision of high quality database services such as the DAP. As the level of investment in the bundled dissemination services is increased, so too the transactions cost diminish from the perspective of the data users and the data depositors. It is relatively easy to imagine circumstances in which particular investments in such services make the process of data dissemination less burdened with transactions costs from the perspective of the depositor, for example, automation of the data sharing act. Increased effort by providers to invest in data standards has also been shown to play a significant role in reducing transaction costs. An analysis of the relative costs associated with provisioning and use of geospatial data supplied by multiple providers (Box et al. 2015) revealed that more effort by providers in standardising data reduced the total transaction cost of use. In data sharing models where providers use standardised exchange models and formats, the costs for providers are high and the costs for users are low. However, when these activities are undertaken at the provider level we are likely to observe economies of scale in their provision, which suggests the adoption of data standards is likely to enhance the social value of the data.

2.4 Reconciling value potential with the costs of disseminated research data products

When we have measures of the value potential of data (benefits) and some measure of the costs attending different options, we can perform a simple reconciliation to identify those options which are more or less preferred than others. Usually this is thought of as cost-benefit analysis (CBA), which is captured in the general idea that if the total costs exceed the total benefits then the proposed option should not be taken (Mishan and Quah 2007). CBA is frequently used to decide between several different options for investment with net positive benefits, in which case CBA prescribes choosing the option with the highest net value (net benefit), i.e. the greatest gap between total value and total cost. In the present case, we can use CBA to think about developing a strategy for expanding the value creation potential of the DAP, subject to the costs of investing the tools and services which enable that potential. For example, in Figure 10 we can think about a representative data collection in the DAP in terms of the total benefits and costs that attend the incremental addition of research services.
Figure 10: Reconciling the value potential of data with the cost of provision

In this case, total value increases with the incremental addition of research services, albeit at a diminishing rate. Functions of this form are consistent with data users who are largely satiated with service offerings at a basic or intermediate level and experience little additional benefit from more advanced service offerings. This need not be the case, and we may observe a piecemeal total value function which exists for values above zero only after a threshold level of service addition has been achieved, an example of such a case is illustrated in Figure 11. This substantially complicates the reconciliation of value creation and associated costs, especially if we aren’t entirely sure where threshold effects are likely to be found, if at all. It seems quite likely that threshold effects are present in the total value functions of open data collections, because the impediments to use are closely aligned to the absence of research and dissemination services (Zuiderwijk et al. 2012).
Applying the same thinking to the DAP as a whole may well be represented with similar illustrations (Figure 10 and Figure 11), and it seems possible that the value function is strongly dominated by threshold effects. These thresholds are likely driven by the nature of large-scale investments in the dissemination and sharing infrastructure of the DAP. Given the dominance of sunk and quasi-fixed costs in the DAP infrastructure, any reconciliation of values and costs is a challenge. We can make some useful inference about the nature of the returns on investments by considering the relationship between average value and average cost per user, for example, Figure 12. It is possible that average value per user is flat or it may increase with incremental investments in the DAP; in the very least, it seems likely that it is non-declining. In short, the form of the average value function is strongly contingent on the nature of the investments and we aren’t exactly sure what it looks like.
The question is then, how might we use this kind of information to develop an investment strategy for the DAP and its associated data collections? Following simple CBA logic, we might just choose options with the highest net value (total value minus total cost) or possibly the highest net average benefit. However as we have established above, the total value of the DAP is strongly contingent on the ways that users respond to the various data + service offerings which could be made. In the presence of strongly positive network externalities, the value function could increase at an increasing rate with additional users. This suggests that the net value increases strongly for any user or data capacity expanding investment. We don’t know what could trigger the initiation of positive network externalities and how strong these effects might become in terms of value creation. We are trying to make investment decisions in a state of uncertainty.

Economists have a lot to say about investments under uncertainty (for example, Dixit and Pindyck 1994, Chambers and Quiggin 2000), the main message of which is approach investments incrementally, where possible, while we wait for further information to emerge. In other words, we seek to manage our investment problems adaptively. A possible solution is captured in the concept of adaptive management, which has long been a feature of decision making over the management of natural resources (Holling 1978; Walters 1986). Management in these environments frequently involve circumstances in which we don’t know enough to develop a rigid strategy to ensure optimal outcomes. Frequently, these may be of the form where we do not know enough about the problem, solutions and possible outcomes, and we are especially unsure about what changes the future might throw at us. It sounds a lot like life in general. Approaches in the CBA family are notoriously poor at directing decision making in the absence of information, let alone situations in which we have evolving uncertainties. Whereas CBA might prescribe a single

![Figure 12: Average value vs average cost for incremental investments in the DAP](image)
solution that requires a commitment to that particular solution over time, AM prescribes a
solution that changes and evolves over time as new information becomes available (Christensen et al. 1996). This new information may be obtained by active learning, such as purposeful experimentation, or by passive learning which suggests a role for an ongoing process of monitoring the ways in which users interact with the DAP and experience value. CBA still has a place here, but only for the purposes of understanding the trade-offs as we make relatively small and incremental changes in the DAP (Randall et al. 2012).

There is certainly a compelling argument to be made for understanding the way in which incremental investments in research and dissemination services add value to the DAP. This should be approached by using a combination of adaptive management and cost-benefit analysis, where investments are made as incrementally as possible to actively learn about the market place for CSIRO’s data collections.

**Public data sharing models**

Choices among alternative investments in services here may be thought of as contributing towards a particular public data sharing model. In this way of thinking, a sharing model is a combination of focussed investments in particular bundles of services which determine the kind of research data products the organisation is disseminating. Consider, for example, our characterisation of the alternative cost models of research and dissemination services in *Figure 6*. We can use the same representation to map the range of alternative data sharing models as being bundled to a greater or lesser extent with combinations of alternative services (*Figure 13*). Public organisations may find themselves operating under a particular model as a result of external constraints/drivers, and resultant explicit or implicit decisions, which may entail a certain degree of irreversibility. Organisations may wish to, or be required to, capture further value from investments that have already been made, before embarking on a strategy towards a new data sharing model. An organisation could explicitly choose a model which reflects particular constraints on their operations, such as those involving budgets or data privacy. Certainly, there are a great many more factors which would locate an organisation within this field than we can consider here, nevertheless this representation presents an intriguing place to start such an analysis.
A possible characterisation of these models could reflect the degree to which investments are made in the particular services. For example, we might think about dissemination services as being broadly divided between models which choose to limit dissemination and those which choose to maximise dissemination. We might refer to models where limited data dissemination services have been chosen as operating like data monopolisers. The data monopolisers choose to limit dissemination to maintain control over their data products. This could reflect a desire to maintain some control over the value generated from their data, or some operational/institutional legal constraints on sharing. Alternatively, models which embrace the dissemination of data, as reflected in greater investment in dissemination services, could be described as positioning the organisation in the data ecosystem. The data ecosystem model involves the choice to disseminate their data products as broadly as possible. This may reflect the notion that the potential to generate public value from their data is maximised when the greatest number of users are able to access, transform and utilise. Within the data ecosystem diverse user-depositor networks drive innovation and value generation.

We can also identify particular distinctions among models arising from the degree of investment in research services, which result in the research data products (described in Figure 2). The research data products with limited research service investment could be described as data-led products, whereas those with greater research service investment could be called research-led products.
Investing in approaches which involve the internal development of products involving intensive modelling and/or analysis would appear towards the research-led end of the spectrum. This could be thought of as the products which involve a high degree of academic involvement in production, through significant research effort and innovation. Alternatively, approaches which involve limited research application to the data to develop products would fall towards the data-led end of the spectrum. Data-led products arise from situations in which the organisation cannot sufficiently justify the investment in the research to develop their products. This may be because there is limited internal research capability, or because the data doesn’t require research investment to be useful in the public domain.

The intersection among these alternative combinations of dissemination and research service investment suggest some intriguing possibilities within which to locate public data-holding organisations. This includes questions, such as, how could an organisation make strategic investments in developing its data assets to maximise the potential to generate public value?

2.5 Institutional aspects of data sharing

The extent to which the potential values of data sharing can be realised in practice is determined in large part by the prevailing institutional arrangements; “institutional” covers the organisations, rules and behavioural norms which apply to data sharing. This extends beyond the immediate governance of data archives to the broader socio-technical system, such as the incentives faced by individual researchers and managers of research organisations.

In the absence of formal rules we often observe complex arrangements of social norms filling the institutional gap, including for example, the notion of what constitutes acceptable behaviours for sharing or withholding data. Merton (1973) describes one such set of social norms in science as ‘communalism’, where the open and free sharing of scientific knowledge and data is a key behavioural expectation. The opposing position might be described as ‘individualism’. In general, some research disciplines are likely to be more inclined towards the communalism end of the spectrum, whereas others, perhaps including economists, are inclined towards the individualism end of the spectrum. Addressing undesirable and entrenched social norms is a significant challenge, although there are a number of examples of interventions which have been beneficial in promoting data-sharing behaviours among researchers. These include funding agency data-sharing mandates and journal data publishing requirements.

Individual motivations

An institutional perspective recognises that simply encouraging, or even mandating, individual researchers to share data may be insufficient to promote widespread behavioural change. In practice sanctions for non-compliance are often weak, and individuals could simply deposit incorrect or undocumented datasets which are of little or no value to other researchers. There will often be tensions between an organisational goal to share data widely in order to maximise its value, and the incentives for individual researchers to keep their data private in order to gain a competitive advantage in generating research papers or projects. This means that disseminating research data is in large part a behavioural question, as data sharing will only work if the
researchers who generate and hold the data are willing and able to share it (Kim and Stanton 2016).

The factors which motivate individual researcher behaviour within the institution are therefore key to understanding the willingness or inhibition of individuals to participate in data sharing and dissemination activities. Motivations may be intrinsic or extrinsic to the individual researcher. Intrinsic motivations relate to individual moral or ethical positions, for example if researchers believe it is their moral obligation to share data. Researchers may exhibit a desire to share data to make the research process more robust, thereby making better research which is good for the discipline and the nature of human knowledge more generally (Van den Eynden and Bishop 2014).

Extrinsic factors which influence the data sharing behaviour of researchers include things like the perception of career benefits and risks arising from data sharing, as well as the perceived effort costs involved with sharing (Kim and Adler 2015). Career benefits can arise directly through recognition, citation and authorship opportunities, but may also arise from the expectation of future reciprocal sharing of data. The extent to which any of these translate into career benefits are in part determined by the organisational environment within which researchers operate. We might assume therefore, that stronger organisational incentives which reward citation and authorship are positively related to positive sharing behaviours.

However, these motivations are likely to be moderated by perceptions of career risk, relating to the concern that sharing of data may result in lost opportunities for publication or research funding. Researchers may prefer limited sharing arrangements whereby they can make direct agreements with users to maximise the potential return of benefits. Strong institutional incentives which reward authorship and traditional publication will act to enhance the individual perception of career risk and so may overwhelm perceptions of career benefits. End-stage career researchers are more likely to actively share data with other researchers, especially those in the early career stages, perhaps because the intrinsic motivations for positive data sharing behaviours are less likely to be overwhelmed by the career progression risks.

Some intrinsic motivations might also inhibit data sharing, for example the frequently cited concerns among researchers that open sharing could result in the misinterpretation and misuse of data (Michener 2015), and that confusion may arise from multiple interpretations of the same data (Bhattachary and Dalziel 2012). This reflects the intrinsic concern that others may fail to appreciate the nature of the data and furthermore, fail to handle it according to the same scientific principles that the researcher would apply, or perhaps reach conclusions with which the researcher disagrees (either scientifically or ethically). A more significant (de)motivation likely relates to extrinsic concern that misinterpretation is a direct critique of their research ability and may jeopardise their potential career progression.

**Incentivising data sharing**

Mandated data-sharing requirements by funding agencies have been a significant driver of data-sharing behaviour among researchers (for example, Pham-Kanter et al. 2014, Michener 2015, Kim and Adler 2015). In simple terms, funding agencies tie the granting of research funding to the sharing of resulting data collections. In this instance there are direct financial, reputational and career progression benefits from agreeing to share research data, because doing so is necessary to
receive the grant. This kind of institution is strongly incentive-compatible, however it doesn’t guarantee that researchers will share data benevolently. In the absence of further requirements concerning issues like, for example, the production and standardised presentation of metadata, some researchers may attempt to game the system by sharing incomplete or obscured data collections. In order to achieve the full potential of this kind of intervention, there exists a complementary role for institutions which establish oversight in the way data is subsequently shared so that it conforms as much as possible to the principle of Completeness (Donker et al. 2016).

Publications are a key source of prestige, reputation and career progression potential for researchers, among which peer-reviewed journals are perhaps the pinnacle. Increasingly, journal publishers in various disciplinary fields of science are requiring authors to make their data available to the public through either data repositories or upon direct request (Michener 2015). This has become a dominant source of coercion for scientists to engage in data sharing behaviours (Kim and Stanton 2016). This approach raises similar complications to those identified above, because this approach doesn’t guarantee researchers will engage benevolently with the requirements. Relying on publishers to drive data sharing behaviours may not correspond to a socially optimal institution, because the data is unlikely to be truly open. There are likely to be issues around freedom of Access if the data is held by a publisher behind a subscription wall (Van den Eynden and Bishop 2014) and potentially issues around Non-discrimination if there exist requirements for identification and justification by the data holder (Donker et al. 2016).

There are a number of interventions available to an organisation to encourage individual researchers to more fully share their data. Rewards for positive data sharing behaviours could range from simple to more complex actions on the part of the organisation, such as providing recognition of a researcher’s efforts or providing financial grants. Punishment mechanisms, such as penalising those who don’t conform to data sharing mandates or targets, are generally not ideal because they may directly increase evasive behaviours (Dong et al. 2016).

Career progression potential is a key driver of individual behaviour in organisations, the elements of which may be described as visibility, reciprocity and reassurance (Van den Eynden and Bishop 2014). Visibility relates to the formal and informal ways in which data sharing behaviours could be recognised by others. This could include the recognition of data collection publishing as a form of research output on par with other forms of publishing, such as reports, journal articles, digital media, etc. There is some evidence that data sharing is positively related to visibility via increased citation of associated research publications (for example, Piwowar et al. 2011, Parsons et al. 2010). Reciprocity may be developed through institutions which require data re-users to appropriately cite data collections. Reassurance is particularly relevant for early-career researchers, whose inhibitions to share data may relate to perceptions of career risk.

Cultural practices associated with positive data sharing behaviours can be cultivated by an organisation. These may include promotion of data sharing for reasons of promoting transparency in research practice, thereby leading to enhanced research quality and accountability. This can also be extended to developing a research culture which limits the redundancy of reproducing data, instead emphasising the benefits of collaboration and data resource sharing. Among end-stage career researchers, there are substantial benefits to be obtained by targeting bequest motives and developing a culture of gifting data to future research generations. Finally, developing
cultural practice which promotes the feedback mechanisms between data depositors and users is crucial to ensuring data is developed in areas of demand.

Supporting technical infrastructure is an important feature of positive data sharing behaviours because of the way it impacts the transaction costs of data sharing. As identified above, these costs can create substantial impediments to data sharing (Quiggin 2010). Technical infrastructure which is designed and developed with data depositors’ needs in mind is important, especially if those developments make data sharing easier. There may also be a role for financial support from the organisation to support engagement with the technical infrastructure, for example, to facilitate research data management throughout the research process (Van den Eynden and Bishop 2014).

**Data exclusivity periods**

The risks of data sharing to individual researchers might be mitigated by allowing a short period of exclusive use of their data, analogous to a patent right (Bhattachary and Dalziel 2012). There are costs and benefits to such exclusivity which must be balanced to find the socially optimal solution. The Productivity Commission (2016) examines this problem in the context of optimal intellectual property (IP) rights setting, where benefits and costs of the arrangement vary with relatively more or less stringent rights. In this instance, we are specifically talking about the social costs and benefits arising from alternative arrangements of a data rights regime, where the arrangements relate directly to the length of time over which exclusive use rights would apply (*Figure 14*).

Identifying the socially optimal exclusive use rights regime is a long-running challenge, because we are trading-off the incentives to create (social benefits) against the monopoly distortions (social costs). Complicating this, is the likelihood that optimal arrangements may not be the same for all types of data.

![Diagram showing social value, social cost, and social benefit over time](image-url)
The benefits of moving to more stringent (longer periods of exclusive use) rights regimes are stronger incentives for the generation of research, particularly in cases where there are large upfront costs of producing the research products. Research in the pharmaceutical industry is frequently cited as an example of this, where there are substantial upfront costs of developing new pharmaceuticals and exclusive use rights are necessary to help generate a return on investment. This feature is less likely to be the case in publicly funded research, where the returns on investment are largely social rather than private, as in the pharmaceutical case. However, benefits of more stringent rights regimes are likely to be observed through researcher effort. The potential for a researcher to capture the prestige and reputation benefits of some period of exclusive use of data which they have generated is likely to be positively related to their research effort. In short, we might expect benefits to be an increasing, albeit at a diminishing rate, function of the length of time over which the exclusive use right is held.

The costs of more stringent exclusive use rights regimes are likely to be experienced directly through the suppression of innovation potential in the broader society. Tabarrok (2013) argues that weak exclusive use rights can be innovation enhancing, and that with increasing stringency they become innovation inhibiting (Figure 15). Assigning an exclusive use right over research data confers a monopoly to the data producer, which is associated with possible distortions such as the under-supply of the products of that data to society. These kinds of rights regimes are a direct impediment to innovation, because the re-users of data will have to either delay their work or reach some agreement with the data producer. Each of these are the direct social costs of more stringent rights regimes for data, which increase with the length of time of the exclusive use right. More costly inputs to innovation are naturally innovation reducing.
2.6 The limits of open research data

Clearly, there are reasons why all data cannot or should not be shared openly. Protection of privacy means that restrictions have to be placed on access to some data. In other cases organisations and/or individual researchers will choose to restrict access for their own purposes, for example to levy a fee for access. A possible classification which is framed in terms of degree of openness (Figure 16) provides a useful way of thinking about these arrangements. Data may be closed, shared or completely open under alternative business/governance arrangements, and various access forms under these specify the degree of control which is maintained by the data depositor or organisation.

*Figure 15: Innovation-strength curve (Tabarrock 2013)*
Broad (2015) provides definitions for each of these access forms, including: (1) Internal access, data that is closed is data that can only be accessed by its subject, owner or holder, (2) Named access, is data that is shared only with named users or organisations, (3) Group-based access, is data that is available to specific groups who meet particular criteria, (4) Public access, is data that is available to anyone who signs-up under specific terms and conditions, and (5) Open data, is data that anyone can access, use and share which is consistent with the principles of Donker et al. (2016). Among these alternatives, the business models for sharing become less about asserting data depositors’ proprietary rights as we move towards open access. Strategic value is created by researchers and organisations choosing to share data in proprietary ways (copyright, patent) and non-proprietary ways (open source, non-patented) (Rossi 2010).

In the absence of commercial-in-confidence restrictions, the decision to designate a data collection as restricted named access likely reflects an expectation that there is value to be obtained by controlling access. At the individual researcher level, maintaining control over data may help to maintain a research advantage over competitors and enhance directly the prospects for career progression. This control comes with a direct cost to society, because data has public good properties and will maximise its value by open sharing. Public research organisations straddle the gap between these two extremes; they desire world-leading research prestige but are obliged to act in the public good by positioning their data assets towards the open end of the spectrum (Figure 16). This implies a careful balancing act between maximising the incentives for high quality research with the need to make the products of that research available for public use. So the question is, what kind of control arrangements does the research organisation need to establish to (1) maintain sufficient incentive at the individual researcher level to produce world-leading research, and (2) meet the public service obligations.

There is no prescribed method for achieving a precise balance in these cases, although the Productivity Commission (2016) identifies four criteria (effective, efficient, adaptable, and accountable) which could be applied to guide the development of data sharing arrangements.
given the social economic and institutional contexts for data sharing. Each of these criteria are seeking to balance some aspect of a socially optimal arrangement for data sharing:

1. Effectiveness: providing sufficient incentive to generate new research and data, while maintaining the socially desirable dissemination of data and other research products.
2. Efficiency: ensuring data is being generated by those who can do so at lowest cost and is finding its way to those who can use it to generate the highest social returns.
3. Adaptability: maintain sufficient certainty in the arrangements for data providers and users, whilst maintaining capacity for those arrangements to evolve in response to social and technical changes.
4. Accountability: establish transparent and open dealings among data providers and users without discouraging the generation or use of data.
3 A case study of the CSIRO Data Access Portal

This research has been motivated by an attempt to understand the value of public research data and the ways in which the associated socio-technical environment influences that value. It is therefore useful to examine these features in the context of an example from a public research organisation, such as the CSIRO’s Data Access Portal (DAP). Using a case-study allows us to investigate in more detail the motivations and values of both data depositors and users who engage with data dissemination institutions. The DAP has detailed information on the nature of the costs involved for the various services which are bundled with data collections. Data dissemination service costs are more visible and easier to obtain than research service costs, which are largely hidden within other research costs. Taking these values and costs together, we develop an analysis mechanism to arrive at an estimate of the value of the DAP. This exercise is of great value in supporting some inference of the scope to which the DAP, and public research data collections more generally, can unlock the value of research data subject to adjustments in the associated socio-technical environment.

3.1 The DAP

The predominant use of the DAP elsewhere in CSIRO is the broad dissemination and sharing of their research with the public. An overview of the DAP user visitation information, is provided by tools in Google Analytics, which have been anonymously tracking user activities since November 2012. For example, among the 1,928 published collections, we can identify 565 data collections which have recorded download events greater than zero since November 2012. Among these, 501 are public (open access) and 64 have some degree of public access restriction. The total downloads for these collections across this period is 70,282, of which 70,046 instances relate to collections open to the public. The top one hundred collections ranked by downloads for the period November 2012 to October 2016 is presented in Figure 17. This reveals a handful of very popular collections with the highest recorded at 9,874 and the top 14 recording more than 1,000 downloads. However, this popularity diminishes rapidly with the 50th ranked collection recording 205 downloads and the 100th just 57.

---

1 CSIRO Astronomy and Space Science deal with large data collections and predominantly use the DAP as a data storage and transfer conduit with their research partners. Given this, we have excluded Astronomy and Space Science data from our analysis.
Data collections in the DAP can be categorised by type, for example, as being observational, modelling outputs, or images. Observational data collections are those which report measurements of some phenomena, such as the result of experimentation or survey activities. Modelling outputs are data collections which are the product of computational processes, such as the application of algorithms to observational data. These collections may include modelling data presented as maps, simulations or projections. Images are data collections which are presented as various forms of visual media, such as photographs, images scanned materials, or film. Among the 565 collections, 281 could be classified as modelling data, 231 as observational data, and 33 as images (Figure 18).
Consider, for example, the relationship between the type of data collection and the average frequency of download across the analysis period. The average download frequency for collections identified as images is 484, which is substantially greater than collections identified as modelling or observational data with 104 and 109 respectively (Figure 19). It is worth considering the extent to which the top ranked collections in each of the three data types is driving these results, for example, Figure 20 presents average downloads for collections in each type outside of the top 10. The resulting pattern is virtually the reverse from the preceding Figure, with images below modelling and observational data. This suggests that very popular image data collections are driving the high average download Figures when all collections are considered.

Figure 19: Average download frequency by type (all)

![Figure 19](image1.png)

Figure 20: Average download frequency by type (all outside top 10 rank for each category)

The length of time over which the data collection has been published may have some positive bearing on the observed download count (Figure 21). This Figure is presented as a logged data series, so a strong positive relationship between the duration of publication (number of days) and
the download count would produce a cloud of markers distributed roughly from the bottom left to the top right of Figure 21. The observation that the cloud is largely flat and indifferent on the left and right hand sides of the Figure suggests that there is likely to be a weak relationship between downloads and duration of publication. When we remove extreme outliers, we can identify a statistically significant, although weak relationship, with an increase on average of 14 downloads per year. This appears to be the case irrespective of the particular type of data collection (image, modelling, or observational).

![Figure 21: Download count (log adjusted) by length of time published (public data collections)](image)

### 3.2 Values and motivations: depositors

The initial phase of the analysis involved semi-structured interviews with depositors of research data collections on the DAP, which is entirely composed of researchers within CSIRO⁶. Depositors were interviewed to elicit their assessment of the value potential of their data collections, and the implications of social, institutional or economic constraints on the realisation of the value potential of their data.

Using the Google Analytics tool, depositors were selected on the basis of their collection’s access and download counts over the period March 2015 to April 2016, which were ranked in order of highest to lowest. Collections which had access or download counts of zero during this time were excluded, resulting in a sample size of 452. Depositors associated with the 10 highest ranked collections were invited to participate in an interview, thereafter a random sample of 20 additional depositors were identified from the remaining 442 collections. A total of 15 data depositors from a wide variety of research disciplines accepted invitations and were interviewed. These interviews were conducted exclusively over the phone for a period of 30 minutes. In order

---

⁶ CSIRO Social Science Human Research Ethics Committee Approval: project 055/16 “Data Access Portal – costs and benefits for depositors and users”
to maintain their anonymity, responses have been clustered together in the below reporting of the results.

Guiding questions for the semi-structured interviews are presented in Appendix A, among which included questions on: *Use Values* “Are you aware of the benefits that others have received from using the data you have submitted? If so, what are they and how might they be quantified (e.g. in terms of money and/or time saved, etc.)”, *Indirect Values* “Have there been any benefits to you the depositor based on the sharing/re-use of your data (e.g. funding, citations, collaborations, recognition or reward, etc.)”, *Option Values* “Beyond the possibilities for current use of the data, do you think the data will have benefits to future researchers with changes in future research technology/knowledge/capability? If so, how might this occur?”, *Existence and Bequest Values* “Have you submitted data to the DAP for the purposes of use by future generations or scientific posterity?”.

1. Background of depositor:
   i. **What is your research field?** Surveyed researchers were drawn from research fields and domains including: environmental science, climate science, social science, biological science, astronomy, geography, economic geography, energy and agriculture.
   
   ii. **What are the main outputs of your research?** See iii below.
   
   iii. **What type of data does your research generate?** Output types included modelling/simulation outputs, images, raw and semi-processed observational data, and maps.
   
   iv. **To what extent is your research data-centric, and how important are databases and other data storage services to your research?** All respondents confirmed some degree of data-centricity of their research, with the vast majority indicating that their research is highly data centric. This may well be a somewhat skewed population as a consequence of contacting depositors as a result of their data contribution to the DAP. Care should be taken when extrapolating this to the broader population of researchers within CSIRO. The importance of database and data storage services, such as those available through the DAP was confirmed by all respondents. Some indicated a preference for using local data storage options (for example, portable hard-drive, or network drive) over the DAP.
   
   v. **Is it possible for you estimate what proportion of your total research output is submitted/shared?** A range of estimates was provided here, for some respondents only a small proportion of their overall research output was submitted/shared through the DAP, around 5-10%. The majority of respondents indicated something in the region of 20-40% of their research output was shared in various forms, including through the DAP. None of the respondents indicated that more than 50% of their research output in any given year was submitted or shared.

2. Depositor engagement with the DAP:
i. How do you use the DAP for your own data? For example, for data storage and backup, for sharing and disseminating your data, or for posterity purposes, etc. Among the respondents, more than half indicated that they used the DAP for data storage and backup. The overwhelming majority used the DAP as a mechanism for the sharing of their data and promotion of their associated research activities/program.

ii. Is the submission/storage/sharing of data a requirement of the data service, funding body, disciplinary norms or your own preference? Several respondents indicated that their sharing of research data was a requirement of a relevant funding body, and these funding bodies were predominantly Commonwealth Government agencies. At least one respondent suggested that data sharing was part of an emerging disciplinary norm, and data sharing was necessary to be seen to be a ‘good disciplinary citizen’. The majority of respondents indicated that their own preference was the key driver in sharing their research outputs, especially those through the DAP.

iii. Do you use the CSIRO Data Access Portal (DAP) to access the data of other researchers? The vast majority of respondents indicated that they don’t use the DAP to access the data of other researchers. This is likely a consequence of the manner in which survey subjects were selected – as depositors, rather than some other category of DAP user.

iv. If you weren’t able to use the DAP to store/share data, would you be using another service? If so, what would that service be and would you be using it in a different way to your current use of the DAP service? Most respondents indicated that they regularly use alternatives to the DAP. The use of alternatives to the DAP reflects perceived or actual deficiencies in the capabilities of the DAP. For example, one respondent indicated that the DAP was unable to accommodate spatial data in the way that they would need to use it. So they employ a third-party alternative. Others indicated that they still use non-DAP storage and sharing options out of habit, or because they didn’t want to engage in the approvals process to deposit the data in the DAP.

3. Value (benefits) of data:
   
a. Use values:
   
i. Own data: Are you actively using the data you have deposited, or has the usefulness of the data been exhausted for your current research purposes? Several respondents indicated that they were actively using the data they had deposited, several others indicated that the data they had deposited on the DAP was out-of-date and they had their own copies of up-to-date data. The remainder indicated that their submission to the DAP was the end product of a research activity, and they hadn’t visited the data since their submission, nor did they have any intention of revisiting the data in the future.

   ii. Own data: Are you aware of the extent to which the data you submit has been re-used? If so, how much and by whom? The majority of respondents didn’t know the full range of who was using their data, nor were they aware of the extent to which their collections had been accessed and downloaded. Many were aware of several instances of a user
contacting them to discuss their data, but were generally unable even to estimate the extent to which their data had been re-used by these users known to them. Some respondents were aware of users due to collaborative research arrangements, and were using the DAP to exchange data. It was difficult for many respondents to estimate the extent to which their data has been re-used, because they hadn’t seen it explicitly cited in peer-literature.

iii. **Own data**: Are you aware of the benefits that others have received from using the data you have submitted? If so, what are they and how might they be quantified (e.g. in terms of money and/or time saved, etc.)? Where specific users were known to the respondent, many could provide an example of how a user has, or could have, generated benefits from the data. The majority of these benefits were reflected in time-savings, for example, where data was directly useful in a procedure the user was engaged with and avoided the outlay of time to generate the same or similar data. Time-savings were also experienced through the enhanced ease of data acquisition, for example, exchanging high quality imaging of an insect collection rather than the collection itself. In this case, the advent of digital exchange rather than physical collection exchange has meant a large saving in avoided transportation cost.

iv. **Own data**: Is it possible to estimate the extent to which the DAP has facilitated, or hindered, the use of your data by others? For example, has the DAP made the use of your data easier, and thereby enhanced the rate of access and utilisation by others? Overwhelmingly respondents indicated that the DAP has been a significant facilitator of data exchange. This was especially true of the ability to share very large file sizes with ease, something that would have previously been bothersome.

v. **Others data**: If you use the data of others deposited on the DAP, can you describe the benefits from use that have accrued to your research program? For example, in terms of money and/or time saved, etc. We didn’t identify anyone among the respondents who use the data of others deposited in the DAP.

**b. Indirect use values:**

i. **Own data**: Have there been any direct benefits to you the depositor based on the sharing/re-use of your data (e.g. funding, citations, collaborations, etc.)? Many respondents identified various combinations of funding and collaborative opportunities that have arisen from sharing their data through the DAP. The generation of citations has been less significant, although many respondents believed that the promotion of their research through the DAP was conducive to generating citations for their research in the longer term, albeit not necessarily for the data itself.

ii. **Own data**: What recognition and reward, if any, do you receive from your funders, employing institution, peers and disciplinary community for these activities (teasing out what and from who)? The vast majority of respondents could not identify any instances of recognition and reward they had received from their institution (exclusively CSIRO) from
having shared their data. Some reported recognition from their research peers, especially in discipline areas where data sharing is more substantially embedded in the culture, such as climate science and astronomy.

iii. **Others data**: *Do you experience any benefits from others depositing and accessing data in the DAP, which don’t involve you using the data directly? For example, has the use of DAP by others enhanced the international standing of CSIRO, and thereby translated into enhanced opportunity for research funding and collaborations, etc.?* Several respondents believed that the sharing of data through the DAP has been helpful in developing the international standing of the CSIRO, although it was hard for them to identify particular instances where they believed this was indicated.

**c. Option values/Bequest values:**

i. **Own data**: *Beyond the possibilities for current use of the data, do you think the data will have benefits to future researchers with changes in future research technology/knowledge/capability? If so, how might this occur?* Respondents associated with more raw types of data, such as observational data and scanned images, identified several ways in which their data could be used in the future in ways in which it isn’t used at the moment – particularly with changes in computational power. Overall, respondents with these forms of data identified value in holding the data to respond to/manage future uncertainties.

ii. **Own data**: *Have you submitted data to the DAP simply for the purposes of scientific posterity?* Most of the respondents identified significant value in the permanence of the DAP for the purposes of ensuring the scientific posterity of their contributions. This was enhanced by the assignment of a DOI, which gives the data collection itself a publication identity. Some depositors were actively going back through old research to extract data to be placed on the DAP in order to maintain a copy for future generations of researchers.

iii. **Others data**: *Do you value the collection of data deposited by others in the DAP for the purpose of your future research activities? For example, intend to use data at some stage in the future, but not yet ready to do so.* Some respondents indicated they would like to review some of the data collections on the DAP for their own research activities, but haven’t yet done so.

4. **Costs of providing/preparing the data:**

i. **How ‘raw’ is the data that you submit?** None of the surveyed respondents had deposited was raw in the pure sense, with most having done some basic cleaning/processing to tidy-up the data before submission. Of those close to the raw end of the spectrum, most indicated that they did this because they wanted to ensure that the data was trustworthy, so they needed to check and clean errors from the data prior to submission. The remainder of the data collections submitted had been substantially removed from its raw state, either by modelling or by production into some kind of software etc.
ii. **What effort goes into generating these data (e.g. time, personnel, and learning new skills)?** Can you give specific numbers (e.g. number of hours or percentage of total time or of time doing research)? In general the data was a direct and integral product of the research efforts, it wasn’t clearly obvious how to attribute time to these activities. In some cases, where old research collections are being digitised and presented in a ‘modern’ computational file format to be placed on the DAP, it was estimated that it took two FTEs several months to prepare the data. In other cases, it is as easy as changing the production file format on the software.

iii. **What additional work is required to prepare the data for submission to third parties for sharing (i.e. over and above necessary work on the data for use within your own immediate projects and/or over and above good research practice, etc.)?** In general this question referred to the creation of metadata, transformation of data to usable forms, and the conversion of outputs into a more universal or modern file format. At most the preparation of metadata was reported as 1 FTE for a week, at least it was 1 hour. The transformation of data and file format arrangement time cost varied from 5 minutes, to several months. This appeared to depend on the age of the data, and whether some conversion from some antiquated storage type (i.e. tape) to more modern storage methods was required.

iv. **How much effort is involved, how much time does it take in hours or as a percentage of total project time, total research time, etc.?** Preparation of metadata was universally considered to be a very small proportion of total project time, because it is something that should have been happening anyway. Many respondents mentioned the need to ensure good data management processes were firmly embedded in their research activities, which necessarily included the keeping of high quality metadata.

### 3.3 Values and motivations: users

The second phase involved a structured online survey of users with a greater focus on eliciting their assessment of the value of the data collections they were using. The survey was administered using a Survey Monkey application, to which a link was presented in a banner on the DAP webpage inviting users to participate. At February 2017, the response rate was 23.

**Question 1: Which of the below best describes your organisational affiliation?**

We are faced with a particular problem in understanding the motivations and values of users, in that we don’t know who they are or what sectoral background they come from. This question is concerned with helping us to develop a sectoral profile of potential users. Among the users surveyed, approximately 41 percent come from the research sector (for example, universities and
public research agencies) and 32 percent come from the public sector (for example, Commonwealth, state and local government agencies). The residual, approximately 27 percent were unaffiliated individuals or associated with the business sector (Figure 22).

![Figure 22: Which of the below best describes your organisational affiliation?](image)

**Question 2: Are you or your organisation based in Australia?**

In analyses of social value it is often important to establish whose values are to be included, where a typical boundary on these values is often defined at a national border. In this instance, 95 percent of respondents were based in Australia, and the residual 5 percent were based outside of Australia.

**Question 3: Are you visiting CSIRO’s Data Access Portal to access a specific data collection?**

The intent of this question was to get a sense of what proportion of users were visiting the DAP with a specific intention to access a data collection, as opposed to those users engaged in informal browsing of the collections. Those with the specific intention to access a particular data collection are highly likely to have had prior knowledge of the existence of a particular collection, possibly indicating a close connection to CSIRO research activities. Approximately 83 percent of respondents answered ‘yes’, with the remaining 17 percent answering ‘no’.

**Question 4: How did you first encounter CSIRO’s Data Access Portal?**

The manner in which potential users find their way to the collections in the DAP is of great interest from the perspective of connecting with the widest possible range of users. Among the respondents approximately 83 percent were from some kind of referral, either from a CSIRO employee/collaborator or from a publication (Figure 23). This high number suggests few potential users are alerted to the existence of the DAP unless they are engaging with other CSIRO research products or people. Only a small fraction of user encounters was associated with web search.
discovery, which accounted for approximately 13 percent. The limited extent to which web search is helping potential users encounter the DAP needs to be examined in more detail.

**Figure 23: How did you first encounter CSIRO’s Data Access Portal?**

**Question 5: Please identify one or more of the following which reflect the reason(s) you are accessing data from the CSIRO’s Data Access Portal:**

The ways in which potential users intend to employ the data that they access from the DAP provides some idea of the kinds of research data products they may be seeking. Data which is intended for use in a research activity usually has different desirable qualities than that sought for the development of a commercial product or service. For example, approximately 13 percent of respondents accessed the data for the ‘current development of a commercial product or service’, whereas 57 percent access the data for ‘current research activities’ (Figure 24).

In this question we were also seeking to identify what proportion of users were accessing data to store and use at future time. This reflects the distinction between direct/indirect use values and the option values as components of total economic value (Figure 1). For example, ‘use in a current research activity’ reflects the direct and indirect use values, whereas the ‘possibility to use in a future research activity’ reflects the option values of users. Approximately 57 percent of respondents selected current research use as opposed to 39 percent for future research use. These categories don’t distinguish between future planned research use and future as-yet-unknown research use, so we can’t clearly determine current use vs future use values in this context.
Question 6: What do you intend to produce with the data you have accessed from the CSIRO’s Data Access Portal?

This question sought to identify the kind of products potential users may be producing from the DAP’s data collections. The kinds of products that users produce is closely related to the values they place on the availability of those collections. Approximately 17 percent of respondents are utilising materials they had accessed to develop software products, with 61 percent intending to produce some kind of written materials (Figure 25).

Figure 24: Please identify one or more of the following which reflect the reason(s) you are accessing data from the CSIRO’s Data Access Portal

Figure 25: What do you intend to produce with the data you have accessed from the CSIRO’s Data Access Portal?
Question 7: On a scale from 1 (low) to 7 (high), how critical is the existence of this data to your work efforts?

This question sought to identify the kind of value users may have for data collections in the DAP, without referring directly to value. Presumably if the existence of the data is highly critical to the work efforts of user, then the associated value estimate should also be reasonably high. This allows us to verify subsequent responses in the survey where respondents are asked to provide estimates of values for data collections they have accessed.

Respondents were provided with grounding values to help them locate the degree of criticality of data on the scale (Figure 26). For example, a value of 1 was described as ‘low – I could get similar data elsewhere’, a value of 4 as ‘medium – could obtain similar data, but it would be costly and/or time consuming’, and 7 as ‘high – absolutely critical, I couldn’t get this data elsewhere’.

Approximately 78 percent of respondents identified a value of 6 and above, with 44 percent identifying the highest value of 7.

Figure 26: On a scale from 1 (low) to 7 (high), how critical is the existence of this data to your work efforts?

Question 8: On a scale from 1 (small) to 7 (large), what are the likely time-savings that the availability of this data creates in your work efforts?

The intent of this question was to elicit some measure of the direct use values of users because time savings are directly reflected in this component of total economic value (Figure 1). Grounding values were provided to respondents to help them locate their time savings on the scale. For example, a value of 1 was described as ‘small – the availability of this data doesn’t save me much time in my work’, a value of 4 as ‘medium – the availability of this data saves me several hours of work’, and a value of 7 as ‘large – the availability of this data saves me months/years of work’. A total of 65 percent of respondents nominated a value of 6 or greater, with 52 percent of respondents nominating the highest value of 7 (Figure 27).
Question 9: Can you provide an estimate of the likely time-savings (in hours) that the availability of this data creates in your work efforts?

This question sought to get respondents to provide some more specific scale to the size of their estimated time savings. Responses ranged from a low of 38 hours, to several estimates of 100 hours, through to a range of years and decades. This closely reflects the scale of responses in question 8.

Question 10: On a scale from 1 (small) to 7 (large), given your current uses for this data (research, product development, etc.), what is the worth (value) of this data to your current work efforts?

In common language, asking a respondent to identify the current value of something usually means the direct use and indirect use values are confounded. Therefore, the intent of this question was to elicit some measure of the combined direct and indirect use values of users as components of total economic value (Figure 1). In principle, with some value scale on responses to questions 8 and 9, we may be able to net out direct use value and identify the indirect component.

Respondents were provided with grounding values to help locate themselves on the value scale, for example, a value of 1 was described as ‘small – we use the data but it isn’t important for our output creation’, a value of 4 as ‘medium – data is very useful for our output creation and is of substantial value’, and a value of 7 as ‘large – data is critical for our output creation and is of enormous value to us’. A total of 76 percent of respondents nominated a value of 6 or greater, with 48 percent of respondents nominating the highest value of 7 (Figure 28).

Figure 27: On a scale from 1 (small) to 7 (large), what are the likely time-savings that the availability of this data creates in your work efforts?
Question 11: Can you provide an estimate of the worth/value (in dollars) of this data to your current work efforts?

This question sought to get respondents to provide some scale to the size of their estimated value of data in their current work efforts. Responses ranged from a low of $5,000, to $10,000, $40,000, several estimates of $100,000, through to values of millions of dollars. This closely reflects the scale of responses in question 10.

Question 12: On a scale from 1 (small) to 7 (large), considering possibilities for your future work, given your possible uses for this data (research, product development, etc.), what is the worth (value) of this data to your future work efforts?

The intent of this question was to elicit some measure of the option values of users (Figure 1), which capture the values arising from intention for future use. Grounding values were provided to respondents to help them locate their option (future use) values on the scale. For example, a value of 1 was described as ‘small – we might use the data but it isn’t likely to be important’, a value of 4 as ‘medium – data might be very useful and could be of substantial value’, and a value of 7 as ‘large – data might be critical for us and could be of enormous value’. A total of 65 percent of respondents nominated a value of 6 or greater, with 48 percent of respondents nominating the highest value of 7 (Figure 29).
Question 13: Can you provide an estimate of the worth/value (in dollars) of this data to your future work efforts?

This question sought to get respondents to provide some scale to the size of their estimated option value of data in their future work efforts. Responses ranged from a low of $60,000, to $100,000, through to values of millions of dollars. This closely reflects the scale of responses in question 12.

Question 14: On a scale from 1 (small) to 7 (large), to what extent, if any, has your use of CSIRO’s Data Access Portal changed your work efficiency (i.e., time saved compared to if no Data Access Portal existed)?

This question was an attempt to identify the extent to which the DAP has facilitated reductions in the transaction costs involved with users accessing CSIRO’s data products (Figure 5). Grounding values were provided to respondents to help them locate the degree of change in work efficiency on the scale. For example, a value of 1 was described as ‘small – our work efficiency hasn’t been impacted by the existence of the data access portal’, a value of 4 as ‘medium – our work efficiency has been positively impacted by the existence of the data access portal’, and a value of 7 as ‘large – our work efficiency has improved significantly due to the existence of the data access portal’. A total of 55 percent of respondents nominated a value of 6 or greater, with 30 percent of respondents nominating a value of 7 (Figure 30).
3.4 The costs of disseminated research products

Following our framing of the production process of a disseminated research data product, the costs of that production process can arise from research services (Figure 2) and dissemination services (Figure 3). Some information about research service costs has been obtained through the depositor survey and details of investments in dissemination services have been obtained from the DAP.

Research service costs are hard to identify, because we are attempting to determine the degree to which additional research services have been bundled with a particular data collection before release through the DAP. We know from the depositor survey that many data collections had been upgraded from their raw state to various degrees, although the vast majority of these services had been applied through the original research use of the data. In instances where some additional research service was applied to enhance the useability of the data, estimates ranged from a few hours to a few months. Preparation of metadata may be considered a part of the research service addition to facilitate sharing, and ranged from estimates of 1 hour to 1 week, although many depositors already had well developed metadata from their original research use of the data.

Dissemination service costs are easier to identify, because they are directly reflect in the operations of the DAP. It has been estimated that over a 6 year period, a total of $13.9 million was invested in infrastructure and associated administration and support services. As these costs may be considered sunk we can represent this information as an expression of average costs per download (Figure 31). This representation emphasises the idea that the DAP possesses significant economies of scale, with the average cost per download decreasing rapidly with increasing downloads. For example, given the 70,282 recorded downloads as observed for our 565 data collections, the averaged provision cost per download is $198.

Figure 30: On a scale from 1 (small) to 7 (large), to what extent, if any, has your use of CSIRO’s Data Access Portal changed your work efficiency (i.e., time saved compared to if no Data Access Portal existed)?
3.5 Reconciling the values and costs of the DAP

How can we reconcile the values and costs for the services provided by the DAP? When we describe the DAP in this context, it is important to remember that the DAP is both the provision services and the data collections which it contains. This is important, because as we have established above in Figure 3, the value of a disseminated research data product is the result of a complex bundle of the data itself with the DAP’s dissemination services. It isn’t possible to disentangle the particular contributions of one or other of these features to the total value of a disseminated research data product. In this analysis we will consider the values and costs of the DAP in their totality.

There is insufficient information to inform our analysis including research service costs. The survey of depositors indicated that in most cases there were only small additional research service costs to publication. A handful of cases indicated more substantial costs of cleaning and preparing data for publication, as such care should be taken when interpreting the totality of data sharing costs. Focussing on dissemination costs, we know the function for averaged provision cost per download, as given by Figure 31. In order to examine the net value, we need to also understand something about the nature of the total economic value, which may be expressed as a value averaged against downloads. Consider the illustration in Figure 32, where Average Total Economic Value is given as the sum of average direct use value, indirect use value and option value. The reconciliation of values and costs of the DAP comes about by assessing the size of the gap between our average total economic value and average cost – the greater the gap, the greater the value the DAP contributes to CSIRO, Australian society and the World. Given the average provision cost per download is $198, average total economic value estimates which greater than that cost will result in positive net value.
How do we put some scale on the components of average total economic value? We have a range of estimates for these various components from the DAP user survey, with often large ranges of values. One approach is to examine this problem by looking for the most conservative estimate of the respective values, in this case the minimum identified value, and establishing a lower bound estimate of average total economic value. The implication of this approach is that uncertainties in the value estimate will be concentrated on the upside, meaning that the actual average total economic value will likely be higher than our estimate.

First consider the average value of direct use, which closely reflects the value of time saved. We have a range of estimates from Question 9 of the DAP user survey, for example, a minimum of 38 hours and several estimates of 100 hours, through to estimates of years. The question is then, what is the value of time? The answer to this question varies considerably between professions within the Australian economy and even more so among different countries and societies. An example in the Australian context, comes from a publicly available estimate of the value of time in 2014 given by the Commonwealth Public Service for employees at various employment levels\(^9\) and associated hours\(^10\). For the time savings in hours reported from the DAP user survey, the value of time savings are reported in (Figure 33) for various public service employment levels. Consider the minimum time savings estimate of 38 hours at the lowest paid APS3 position, then the direct use value could be estimated to be $1,558 associated with the download of an average data collection. This value of saving 38 hours may be as high as $3,496 in the instance of the highest


paid EL2 level. Together these values establish a plausible lower bound range for direct use values arising from time savings.

The sum of direct and indirect use values are captured in Question 11 of the DAP user survey, which sought to identify the current use values of respondents. Responses ranged from a low of $5,000, to $10,000, $40,000, several estimates of $100,000, through to values of millions of dollars. Consider the lowest of these value estimates at $5,000, which is marginally higher than the value of time savings estimates from above. This suggests a reasonable scale on the estimate of the minimum sum of average direct and indirect use values is $5,000.

Finally, option values estimates are provided by Question 13 of the DAP user survey, where responses ranged from a low of $60,000 to $100,000, through to values of millions of dollars. Taking the minimum here would give us an average option value of $60,000, and an average total economic value of $65,000. There remain greater uncertainties with the option values, as respondents may be confounding future use values for collections as a whole with the value of particular download events. As such, we should interpret the future use value with some caution.

If we considered only the value of current use in our analysis, then we would be comparing the average total cost of the DAP to the sum of components 1 + 2 (direct + indirect value) in Figure 32. Assuming these average values apply across all downloads from the DAP, then given our estimate of the average provision cost per download is $198, we are looking at minimum average net value of $4,848 (Table 1). If we consider the sum of all components 1 +2 + 3 (direct + indirect + option values), then we are looking at a minimum average net value of $64,848.
Table 1: Reconciling average values and costs of the DAP

<table>
<thead>
<tr>
<th>Source of value</th>
<th>Average minimum value estimate (per download)</th>
<th>Average net value (per download)</th>
<th>Break-even threshold (downloads)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct use</td>
<td>$1,558 - $3,496</td>
<td>$1,360 - $3,298</td>
<td>3,976 – 8,920</td>
</tr>
<tr>
<td>Direct + Indirect use</td>
<td>$5,000</td>
<td>$4,802</td>
<td>2,780</td>
</tr>
<tr>
<td>Option</td>
<td>$60,000</td>
<td>$59,802</td>
<td>232</td>
</tr>
<tr>
<td>Direct + Indirect +</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Option</td>
<td>$65,000</td>
<td>$64,802</td>
<td>214</td>
</tr>
</tbody>
</table>

An equivalent way in which we can approach this reconciliation is to ask the question, what is the minimum number of data collection downloads required to achieve positive average net value? The average provision cost per download decreases with more downloads, and by assumption the average value remains the same, which means that the average net value increases with downloads. For example, consider the case for average minimum direct and indirect use value of $5,000, which is illustrated in Figure 34. The average net value is zero once 2,780 downloads have been recorded for the DAP as a whole, which means it takes negative values below this threshold and positive values above. This suggests that given the nature of the DAP’s costs, at our estimate of $5,000 per download, that the DAP makes a positive net value above 2,780 downloads (Table 1). This is especially relevant where we might be concerned about our ability to define precisely a download event within the DAP. Given that there have been some 70,282 recorded downloads across our 6 year period analysis period, the value of the DAP and its data collections could be in the order of $337 million\(^{11}\) which is an annualised value of $67 million.

\(^{11}\) All values are presented in Australian Dollars.

---

Figure 34: Average net value for Direct + Indirect use value, by download count
4 Unlocking the value of public research data: recommendations and approaches

Motivated by the question of how best to understand the value of public research data, we have concluded that the value of data held by public research organisations can be increased through investment in data services and infrastructure. This is because of the central importance of these aspects of the socio-technical environment to the realisation of data’s potential value. There is certainly scope to increase data value, although the challenge is to identify where and how this can best be done. Based on the analysis of relevant theories and frameworks presented earlier, and the case study of the CSIRO’s DAP, there are a number of opportunities which should be investigated in more detail. There are opportunities throughout the data supply chain, however, some of these may be challenging to implement.

Our main recommendations below relate to the economic and related institutional levers to both realise and increase the value of an organisation’s public research data. Other technical approaches to maximising the value of data, such as, improvement of data supply chain technologies are beyond the scope of this report.

4.1 Unlocking value by strengthening research data product supply chains

Consider first the opportunities to improve the socio technical systems at the early stages of the research data supply chain, which primarily concern researchers. First among these is the adoption of research governance policies which explicitly embed data dissemination objectives in the data management procedures of the research organisation. Data produced within a research project should be made available at key stages during its transformation, not just at the end of the research project. This should occur for two reasons. First, as we have established above (for example, Figure 4), data is more valuable when users can access earlier versions of research project’s data, rather than just the end data product. This is because some users will want to manipulate the data in different ways to the researchers, others will want to use the raw data. Providing only the end data product will lock out these potential users, and so diminish the value of the data generation processes. Second, publishing various versions of data which are being produced and altered through a research process enhances the transparency of the research process itself. Fostering a process which enhances the transparency of the research within a research organisation is likely to have reputational and quality benefits for the research which is produced. Reputational benefits, for example, may be thought of as a form of indirect use value which accrues to the organization and to the associated researchers.

Managing researcher attitudes and expectations is a critically important factor for achieving the data dissemination objectives of an organisation and realising the value potential of that data. A key aspect of this management involves ensuring researcher incentives are sufficiently aligned with the organisation’s open data objectives. In many cases, it seems that an organisation’s desire to disseminate research data is inconsistent with the institutional incentives presented to
researchers. Mandates which require the open dissemination of researcher’s data collections will not work, not least of which because of the possibilities for inducing perverse behavioural results. Incentives are an essential key to unlocking the latent value of research data collections which are currently closed. This may be addressed in many ways, including treating data publication as a form of publication for the purpose of measuring key performance indicators of researchers. Financial incentives, such as awards for outstanding data publishers, could be applied to induce and/or support positive data dissemination behaviour among researchers. More advanced incentive focused data sharing arrangements concern managing periods of data use exclusivity, before which research data is openly disseminated. Identifying the socially optimal exclusive use rights regime remains a challenge, because of the complex trade-off among the social benefits (for example, arising from the researcher incentives to generate new and novel data) and the social costs (for example, foregone innovation among potential user groups).

4.2 Unlocking value from enhanced dissemination processes and infrastructure

The scale of data dissemination infrastructure matters, because costs are dominated by sunk and fixed costs. This means the greater the number of collections housed and downloaded from an infrastructure cluster, the lower the average cost. It seems that the scope for scale economies of this kind are substantial. The implication of this is the gap between value (benefits) generation and cost grows with more user and depositor activity. Possibilities to increase the scale of collections held within public research organisations should be investigated, among which the incorporation of like-minded collections across the Australian public research sector should be considered. Achieving scale on the data ‘supply-side’ may also be associated with enhanced scale economies on the ‘demand-side’. In other words, if we can develop a critical mass of research data collections then network effects may exert positive influence on the structure of the average value schedule.

A good market for the exchange of data among providers and users requires a two-way flow of information. Many data providing organisations don’t have a clear understanding of who their user groups are, perhaps only recognising instances of use when they appear in subsequent research or commercial products. Collecting information on users and depositors at their point of contact with the data dissemination infrastructure is crucially important in understanding the value of data. The case study presented in this report has sought to shed some light on the values and motivations of users and depositors as they interact in various ways with CSIRO’s DAP. We have approached this applying various survey and interview methods, which have provided some crucial insights to our analysis. However, there is a need to establish mechanisms which gather more detailed information on depositors and users on an ongoing basis. For example, on a systematic basis it would be desirable to ask depositors questions about the research service costs which have been incurred up to the point of data upload to the data dissemination infrastructure. So too, it would be of great value to ask users their intended use of data collections and associated value estimates at the point of download from that infrastructure.

The ability to connect these specific values to particular collections would facilitate substantially more detailed analysis of the research data marketplace. So too, detailed information along these lines is crucial in unlocking the potential of the two-sided research data network effects, which could strongly drive the value of an organisation’s data collections. The mechanisms to achieve
this are technically within reach, although the time-frame over which this project has run was insufficient to realize their potential. Small investments on the CSIRO side may pay large dividends in the public domain when it comes to re-use of the data collections.

The value of data originating from an organisation is closely related to the trust users place in its quality. For example, the CSIRO is often referred to as a ‘trusted’ provider of research for the Australian public, a status which is maintained by the rigorous quality assurance processes of the organization. In the data sharing space trust is a significant source of competitive advantage and therefore of value. It seems more than likely that the trust which users place in the quality of CSIRO’s data is a key driver of the value of the DAP’s data collections. For any public research organisation a position of trust can only be maintained by adopting the appropriate due diligence mechanisms to ensure, as best as possible, that data collections with errors are not released to the public. This may also include review processes which a focus on the commercial and/or legal dimensions of data dissemination.

4.3 Value from cultivating research data sharing networks.

Public research organisations may directly enhance the value of their data collections by actively promoting the development of user-depositor networks. There is potential to enjoy positive network externalities from the data collections, although in early stages the formation of these networks may well require support. Networks are all about the nature of the links between groups (user to user, depositor to user, and depositor to depositor), any reasonable action which either establishes a link between groups or reinforces weak links between groups will have some network payoff. For instance, simple network strengthening improvements could include mechanisms which link data collections to the published products of that data. Among these are the publications produced through the original research application and where possible through subsequent applications of the data by users. Similarly, developing mechanisms which make links between complementary data collections could positively stimulate network formation.

Innovative network formation requires the possibility to develop random connections among user and depositor groups. In most cases the internet has facilitated the formation of random connections quite effectively, albeit that many public research data collections are hidden away in its dark corners. In the DAP case study, a somewhat surprising finding from the user survey was the absence of users who are encountering the DAP from unsolicited/random web-searching. We know that the vast majority of users are visiting the DAP because they have been directed there by a publication or a CSIRO employee. This may well be a problem common to public research organisations, which then raises the question about how best to raise the general visibility of their data collections.
5 Conclusion

Data is a cornerstone of the digital economy. As ever more powerful analytical techniques become available the value of data is likely to continue to increase. Value stems from using direct and indirect uses of data. In some cases users need raw data on which to build their own models whereas others use products and services such as models and apps into which the data are incorporated. Data can also be a source of substantial option value – even if it is not currently used its existence creates options for future use. Ensuring that raw data is available, rather than just the outputs of models or apps, will maximise option value as it offers future users the greatest flexibility in how they use the data. Yet to realise these potential values of data, a number of challenges must be overcome. These include barriers to data discovery, access and use, a lack of incentives for data sharing, and complex social and institutional arrangements.

Governments around the world are increasingly seeking to ensure open access to data, which will impact publicly funded research and research institutions. Scientific research values transparency; the ability to follow and build on the work of others is integral to the incremental improvement of knowledge. However, while most disciplines share knowledge openly through publication, the sharing of data is typically patchy and ad hoc. Creating the best possible value from data requires optimal sharing. Providing data with unclear documentation and formatting imposes additional costs on users. From an economic perspective it is likely to be more efficient if providers bear the cost of formatting and documenting data as it only has to be done once, rather than multiple times, and they can probably do it more easily. In some cases it will also be efficient for providers to include additional services. However, it is possible to over-invest; developing high-cost services for boutique user groups is unlikely to realise sufficient public benefit to justify the investment.

In practice, there are many ways in which data can be shared. To be useful to others data should be documented, with metadata describing how each variable was measured and collected. Documenting metadata can be costly to the data provider (as it takes time) but enhances the value to users. Similarly providing data in machine-readable form, as opposed to simply attaching documents to websites, makes data more readily accessible and useful, though again at some cost to providers. Efforts to disseminate data can also enhance its value through making it more discoverable and accessible. This requires investment in IT infrastructure and associated systems. This is costly, but there are significant economies of scale. The more datasets held in a repository the lower the infrastructure costs per dataset. The value of the data is also likely to be enhanced through greater discoverability and linkages to other datasets.

Determining how much should be invested in sharing and disseminating data requires an understanding of its likely value. To better understand the value of data held by public organisations, we investigated the time saved by users of CSIRO’s Data Access Portal, finding benefits orders of magnitude greater than then costs of the Portal. Our survey indicates that direct use values can be very high. Our estimate of $67m a year, while highly uncertain, is nearly two orders of magnitude greater than the costs of disseminating the data. However, for this (and many similar) data repositories very little is known about who is using the data and for what purposes.
Collecting such information would add to the costs of dissemination but is essential for evaluation and improvement.

The value of data is ultimately determined by the prevailing institutional arrangements. Open data is a public good, which makes it difficult for those who incur costs through providing data to be recompensed. In the competitive world of research there are incentives at both the individual and institutional level to keep data private in order to monopolise its use. Even if data sharing is mandatory there are many opportunities to comply in ways which stymie potential users. Effective sharing requires a combination of rules, norms (e.g. within disciplines or organisations) and incentives (e.g. recognition for sharing data). Tracking how often data is accessed and where it is re-used (e.g. through download statistics and data citations) would help target incentives of efficient sharing. It might also support intrinsically-motivated sharing, by demonstrating how much a researcher is contributing to the scientific community.

The value of data held by public research organisations can be increased by enhancing the function of the socio-technical environment through investment in data services and infrastructure. This value can be unlocked by (1) strengthening research data product supply chains, (2) enhancing dissemination processes and infrastructure, and (3) cultivating research data sharing networks. Digital data has remarkable economies of scale. Once it has been documented and deposited it can be downloaded many times at close to zero marginal cost. Data sharing infrastructure also has significant economies of scale; as more data is added the marginal costs decline, though there are ongoing maintenance costs and upgrades will be required from time to time as the volume of data expands. Data depositories should therefore aim to be large. They should also cultivate the networks among depositors and users which can both facilitate and motivate data sharing and re-use. In this way research data collections can become as integral to scientists in the 21st century as the library was in the preceding millennia. This is important because the success of the Australian economy in the 21st century will be increasingly determined by the role of the digital economy and the competitive advantages that follow from the ability to generate value from data.
Appendix A: Survey of data depositors

1. Background of depositor:
   vi. What is your research field
   vii. What are the main outputs of your research?
   viii. What type of data does your research generate? (For example: numerical data, images, sounds, maps)
   ix. To what extent is your research data-centric, and how important are databases and other data storage services to your research?
   x. Is it possible for you estimate what proportion of your total research output is submitted/shared?

2. Depositor engagement with the DAP:
   v. How do you use the DAP for your own data? For example, for data storage and backup, for sharing and disseminating your data, or for posterity purposes, etc.
   vi. Is the submission/storage/sharing of data a requirement of the data service, funding body, disciplinary norms or your own preference?
   vii. Do you use the CSIRO Data Access Portal (DAP) to access the data of other researchers?
   viii. If you weren’t able to use the DAP to store/share data, would you be using another service? If so, what would that service be and would you be using it in a different way to your current use of the DAP service?

3. Value (benefits) of data:
   a. Use values:
      vi. Own data: Are you actively using the data you have deposited, or has the usefulness of the data been exhausted for your current research purposes?
      vii. Own data: Are you aware of the extent to which the data you submit has been re-used? If so, how much and by whom?
      viii. Own data: Are you aware of the benefits that others have received from using the data you have submitted? If so, what are they and how might they be quantified (e.g. in terms of money and/or time saved, etc.)?
      ix. Own data: Is it possible to estimate the extent to which the DAP has facilitated, or hindered, the use of your data by others? For example, has the DAP made the use of your data easier, and thereby enhanced the rate of access and utilisation by others?
      x. Others data: If you use the data of others deposited on the DAP, can you describe the benefits from use that have accrued to your research program? For example, in terms of money and/or time saved, etc.
b. Indirect use values:

iv. Own data: Have there been any direct benefits to you the depositor based on the sharing/re-use of your data (e.g. funding, citations, collaborations, etc.)?

v. Own data: What recognition and reward, if any, do you receive from your funders, employing institution, peers and disciplinary community for these activities (teasing out what and from who)?

vi. Others data: Do you experience any benefits from others depositing and accessing data in the DAP, which don’t involve you using the data directly? For example, has the use of DAP by others enhanced the international standing of CSIRO, and thereby translated into enhanced opportunity for research funding and collaborations, etc.?

c. Option values/Bequest values:

iv. Own data: Beyond the possibilities for current use of the data, do you think the data will have benefits to future researchers with changes in future research technology/knowledge/capability? If so, how might this occur?

v. Own data: Have you submitted data to the DAP simply for the purposes of scientific posterity?

vi. Others data: Do you value the collection of data deposited by others in the DAP for the purpose of your future research activities? For example, intend to use data at some stage in the future, but not yet ready to do so.

4. Costs of providing/preparing the data:

i. How ‘raw’ is the data that you submit?

ii. What effort goes into generating these data (e.g. time, personnel, and learning new skills)? Can you give specific numbers (e.g. number of hours or percentage of total time or of time doing research)?

iii. What additional work is required to prepare the data for submission to third parties for sharing (i.e. over and above necessary work on the data for use within your own immediate projects and/or over and above good research practice, etc.)?

iv. How much effort is involved, how much time does it take in hours or as a percentage of total project time, total research time, etc.?
Appendix B: Survey of data users

1. Which of the below best describes your organisational affiliation? (choose one option)
   a. Individual (no affiliation)
   b. Public sector
   c. Research sector
   d. Business sector (small or medium enterprise <200 employees)
   e. Business sector (large company or organisation 200+ employees)
   f. Voluntary sector (e.g. charitable) organisation

2. Are you or your organisation based in Australia? (choose one option)
   a. Yes
   b. No

3. Are you visiting CSIRO’s Data Access Portal to access a specific data collection? (choose one option)
   a. Yes
   b. No

4. How did you first encounter CSIRO’s Data Access Portal? (choose one option)
   a. Web search
   b. Referral from a publication or presentation (journal article, report, etc.)
   c. Referral from a CSIRO employee or collaborator
   d. Referral from a non-CSIRO colleague or collaborator
   e. Other

5. Please identify one or more of the following which reflect the reason(s) you are accessing data from the CSIRO’s Data Access Portal: (tick boxes that apply)
   a. Use in a current research activity
   b. Use in the current development of a commercial product or service
   c. Use to verify or replicate research previously produced from the data
   d. Possibly to use in a future research activity
   e. Possibly to use in the future development of a commercial product or service
   f. Not sure yet
   g. Other
6. What do you intend to produce with the data you have accessed from the CSIRO’s Data Access Portal? (tick boxes that apply)
   a. Written materials (reports, journal articles, etc.)
   b. Digital media (e.g. websites etc.)
   c. Modified dataset
   d. Images
   e. Software
   f. Not sure yet
   g. Other

7. On a scale from 1 (low) to 7 (high), how critical is the existence of this data to your work efforts? (choose one option)
   a. 1 (low – I could get similar data elsewhere)
   b. 2
   c. 3
   d. 4 (med – could obtain similar data, but it would be costly and/or time consuming)
   e. 5
   f. 6
   g. (high – absolutely critical, I couldn’t get this data elsewhere)

8. On a scale from 1 (small) to 7 (large), what are the likely time-savings that the availability of this data creates in your work efforts? (choose one option)
   a. 1 (small – the availability of this data doesn’t save me much time in my work)
   b. 2
   c. 3
   d. 4 (med – the availability of this data saves me several hours of work)
   e. 5
   f. 6
   g. (large – the availability of this data saves me months/years of work)

9. Can you provide an estimate of the likely time-savings (in hours) that the availability of this data creates in your work efforts? (enter a value)
   a. Specify an estimated number of hours

10. On a scale from 1 (small) to 7 (large), given your current uses for this data (research, product development, etc.), what is the worth (value) of this data to your current work efforts? (choose one option)
   a. 1 (small – we use the data but it isn’t important for our output creation)
   b. 2
   c. 3
11. Can you provide an estimate of the worth/value (in dollars) of this data to your current work efforts? (enter a value)
   a. Specify an estimated value in dollars

12. On a scale from 1 (small) to 7 (large), considering possibilities for your future work, given your possible uses for this data (research, product development, etc.), what is the worth (value) of this data to your future work efforts? (choose one option)
   a. 1 (small – we might use the data but it isn’t likely to be important)
   b. 2
   c. 3
   d. 4 (med – data might be very useful and could be of substantial value to us)
   e. 5
   f. 6
   g. (large – data might be critical for us and could be of enormous value to us)

13. Can you provide an estimate of the worth/value (in dollars) of this data to your future work efforts? (enter a value)
   a. Specify an estimated value in dollars

14. Do you have any general comments about your experience with CSIRO’s Data Access Portal that you would like to share with us? (choose one option, enter comment if Yes selected)
   a. Yes (please specify)
   b. No
References


