# Open Research Data

Report to the Australian National Data Service (ANDS)

November 2014

**John Houghton**
Victoria Institute of Strategic Economic Studies

&

**Nicholas Gruen**
Lateral Economics

# About the authors

**John Houghton** is Professorial Fellow at Victoria University's Victoria Institute of Strategic Economic Studies (VISES). He has published and spoken widely on information technology, industry and science and technology policy issues, and he has been a regular consultant to national and international agencies, including the Organisation for Economic Cooperation and Development. John's research is at the interface of theory and practice with a strong focus on the policy application of economic and social theory. Consequently, his contribution tends to be in bringing knowledge and research methods to bear on policy issues in an effort to raise the level of policy debate and improve policy outcomes. In 1998, John was awarded a National Australia Day Council, Australia Day Medal for his contribution to industry policy development.

**Nicholas Gruen** is a policy economist, entrepreneur and commentator on our economy, society and innovation. He advised two cabinet ministers in the 1980s and 1990s, taught at the Australian National University, and sat on the Productivity Commission from 1993 to 1997. He directed the New Directions project at the Business Council from 1997 to 2000. He is CEO of Lateral Economics, and Chairs: the Australian Centre for Social Innovation, Peach Financial, Deakin University's Arts Participation Incubator, and the Open Knowledge Foundation (Australia). He chaired the Federal Government's Innovation Australia until 2014, and in 2009 chaired the Government 2.0 Taskforce.


### Acknowledgements

A number of the estimates presented in this report are based on previous work on the costs and benefits of UK research data centres, done in collaboration with Neil Beagrie of Charles Beagrie Ltd., as well as Peter Williams of the University College London and Anna Palaiologk of Ingenieria y Soluciones Informaticas, and we acknowledge and thank them for their contribution. We would also like to thank Dr. Greg Laughlin of the Australian National Data Service for his help and support.


### Disclaimer

# Main points

Research data are an asset we have been building for decades, through billions of dollars of public investment in research annually. The information and communication technology (ICT) revolution presents an unprecedented opportunity to 'leverage' that asset. Given this, there is increasing awareness around the world that there are benefits to be gained from curating and openly sharing research data (Kvalheim and Kvamme 2014).

Conservatively, we estimate that the value of data in Australia's public research to be at least $1.9 billion and possibly up to $6 billion a year at current levels of expenditure and activity. Research data curation and sharing might be worth at least $1.8 billion and possibly up to $5.5 billion a year, of which perhaps $1.4 billion to $4.9 billion annually is yet to be realized. Hence, any policy around publicly-funded research data should aim to realise as much of this unrealised value as practicable.

## Aims and scope

This study offers conservative estimates of the value and benefits to Australia of making publicly-funded research data freely available, and examines the role and contribution of data repositories and associated infrastructure. It also explores the policy settings required to optimise research data sharing, and thereby increase the return on public investment in research.

The study's focus is Australia's Commonwealth-funded research and agencies. It includes research commissioned or funded by Commonwealth bodies as well as in-house research within research-oriented agencies wholly or largely funded by the Commonwealth. Government data or public sector information is a separate category of publicly-funded data – although there is some overlap at the margins (e.g. Commonwealth Government funding for Geoscience Australia).[1]

## Main findings

For the purposes of estimation, we explore a range of research funding and expenditure from total Australian Government funding support for research to the sum of government and higher education expenditure on research by sector of execution. The lower bound estimates are based on the labour-cost share of research funding and expenditure ($4.3 billion to $6.4 billion per annum), and upper bound estimates on total research funding and expenditure ($8.9 billion to $13.3 billion per annum).

### The value of data in Australia's public research

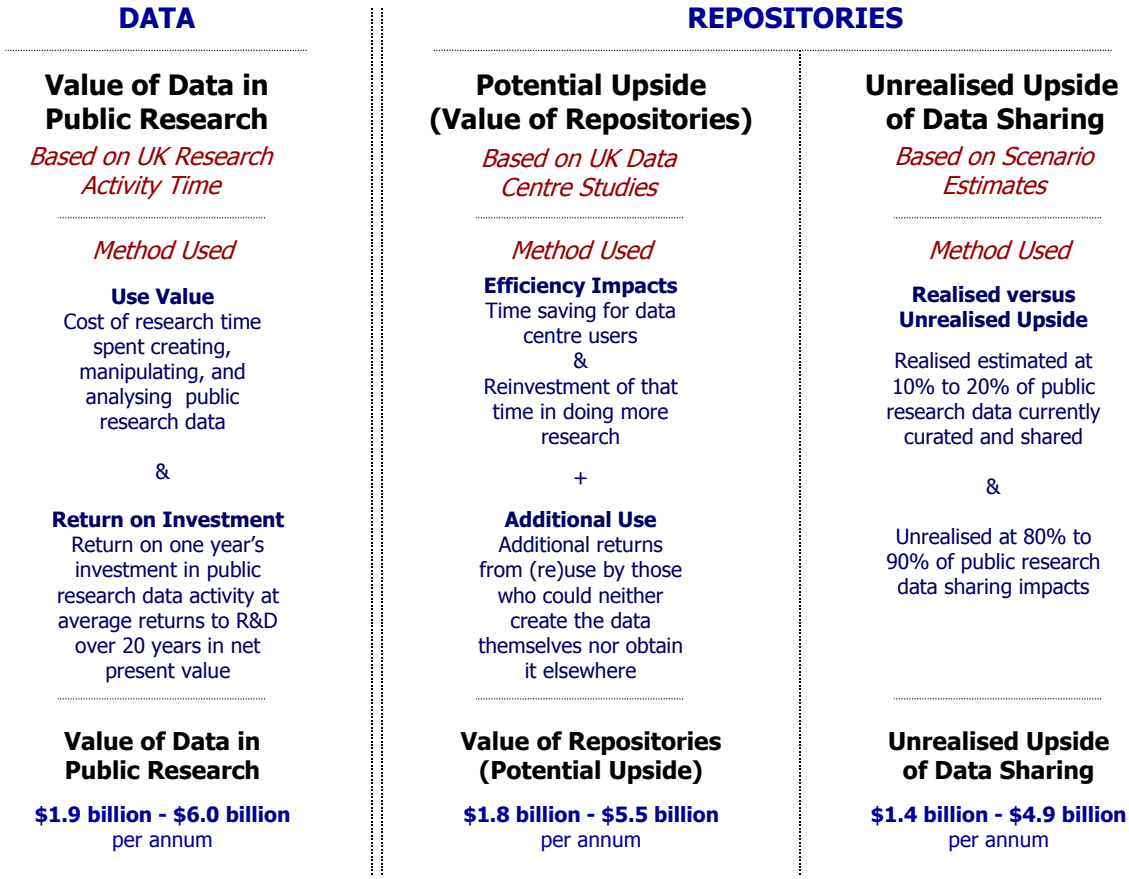
We present two alternative estimates of the value of data in Australia's public research. The first explores the 'use value' of research data (i.e. the cost of the research time spent creating, manipulating and analysing data), while the second estimates the return on investment in public research data activities (i.e. the return on one year's investment in research data activity at average returns to R&D over 20 years in net present value). Both are conservative (See Annex II).

---

1 Estimates by Gruen et al. (2014, p27) implied that the potential value of open government data might be around double that of open research data.

**Both approaches suggest that the value of data in Australia's public research is at least $1.9 billion per annum and possibly up to $6 billion per annum** – at 2012-13 levels of expenditure and activity (Figure 1).

Figure 1 Summary of estimates, methods and findings




Note: Range estimates are based on the labour-cost share (lower bound) and total research funding and expenditure (upper bound) using conservative methods.
Source: Authors' analysis.

## The value of repositories and related infrastructure

A series of studies of UK research data centres spanning the humanities, social and natural sciences identified two major impacts arising from research data curation and sharing: (*i*) significant efficiency impacts for the users of the data centres; and (*ii*) substantial additional (re)use of the data by users who could neither recreate the data themselves nor obtain it elsewhere.

Extrapolating from these UK studies and scaling to the Australian context, **our estimates suggest that the potential value of research data repositories for Australia might be at least $1.8 billion and possibly up to $5.5 billion per annum**.

### The potential benefits of having national collections

When exploring the benefits of curating and openly sharing research data, averages derived from other studies tend to disguise the importance of having national data collections that enable researchers to address national issues of importance – be they local issues (e.g. household expenditure patterns) or the local implications of global issues (e.g. climate change).

While national collections are a part of the UK-based estimates presented above, Australia's unique geography, climate, fauna and flora suggest that there may well be additional value associated with Australian national data collections.

### The potential upside value of data curation and sharing

Simply adding the estimated direct efficiency savings and reinvestment of those savings into further research to the returns from additional use facilitated by data curation and sharing suggests **a total potential annualised impact of $2.3 billion to $7.2 billion at 2012-13 levels of activity, of which $1.8 billion to $5.5 billion might accrue within Australia**.

Of course, it could be more if we in Australia do more or more effectively curate and share research data than have the UK data centres upon which these estimates are based.

### Currently realised and unrealised potential value

Without detailed study it is difficult to guess how much public research data is currently curated and shared. But, hypothetically, **if current data curation and sharing is in the range of 10% to 20% of the research data being produced, then some $1.4 billion to $4.9 billion in annualised benefits may remain as yet unrealised**.

### The potential cost of research data curation

Previous studies report *institutional data repository* operating costs equivalent to around $300,000 to $500,000 per annum, *national data centre* operating costs equivalent to $1.5 million to $6 million per annum, and UK *disciplinary data centre* costs equivalent to $420,000 to $3.3 million per annum.

A number of disciplinary research data centres are funded by UK Research Councils and a study by the Office of Science and Innovation found the running costs of these data centres to be remarkably consistent across the Research Councils – at between 1.4% and 1.5% of the total research expenditure of the research council. **Simply extrapolating this to the scale of Australia's public research funding would suggest national *disciplinary data repository* costs of some $130 million to $200 million per annum**.

Hence, while material, the cost of research data curation and sharing are small in comparison with the potential benefits, which can be measured in the billions rather than millions.

## Enabling policies

Optimising the policies and institutions to drive innovation around open data requires two things that are in some tension with each other. On the one hand, the architecture of the various

institutions needs to be articulated, and the relations between them. This is largely a 'top-down' policy exercise. On the other hand, top-down approaches can be inimical to innovation from the bottom up, which is at a premium where, as here, there is a high level of complexity and rapid change.

## Policy, institutions and culture

The starting point should be government, research funder, and institutional mandates stating the expectation or requirement for open research data as the default. Recognising that research is a global activity, with many cross-institutional and international collaborations, mandates should seek to maximise the national and international harmonisation of policies, in order to reduce compliance costs by ensuring that compliance involves the same actions across policy jurisdictions (Chan et al. 2013). Further, pride should be cultivated in Australian contributions to the global data commons, together with the use of such contributions, to encourage the development of a mutually reciprocating community of national practice (Cutler & Company 2008).[2]

At the same time, systems for resourcing and encouraging good work in the management or use of data should foster an environment which is open to experimentation and new approaches, and which rewards talent and intrinsic motivation. Open competitive bidding for projects is likely to have some role in such a system, but we should guard against over-reliance on those allocating funds being able to determine relative research merit before the research is conducted, or even their ability to assess it immediately upon completion. It is for this reason that a quite common management practice in some leading private sector innovators, such as Google and Atlassian, is '20% time', which deliberately creates scope for those who are more junior in the system to follow their own intrinsic motivations, back their own judgement and to collaborate with others of like mind, even amid the scepticism of those in higher positions.

## Enabling infrastructure

It is important to recognise that there can be costs associated with making research data openly available, and to provide appropriate data repository infrastructure funding through such schemes as the National Collaborative Research Infrastructure Strategy, the Super Science Initiative, and so on.

The National Collaborative Research Infrastructure Strategy (NCRIS) is currently under review, and we defer to the findings of that review as to whether the NCRIS model remains the most appropriate funding model for eResearch infrastructure, including research data repositories. However, from an economic point of view there are clear advantages to any scheme that creates incentives for institutions to collaborate, thus encouraging co-investment. This leverages greater overall funding by bringing forth investments that would not otherwise have happened, as well

---

2 *Venturous Australia: A Review of the National Innovation System*, Recommendation 7.14: To the maximum extent practicable, information, research and content funded by Australian governments – including national collections – should be made freely available over the internet as part of the global public commons. This should be done whilst the Australian Government encourages other countries to reciprocate by making their own contributions to the global digital pubic commons.

as ensuring that all parties have some 'skin in the game'. Moreover, given the characteristic scale and learning economies noted, encouraging collaboration will likely be more effective than competition for limited funding.

### Constraints on data openness (privacy and confidentiality)

All parties must realise that, while the default is open access, there may be privacy, ethical, security, commercial or other constraints on the open release of research data. Addressing these constraints at an early stage in the research process is crucial, and it is essential to ensure that there are clear data access and management guidelines (e.g. clear processes for meeting any privacy, confidentiality and security concerns, whilst imposing the minimum obstacles on worthwhile research being conducted).

In this regard, the focus should be on protecting data providers and those whose data is used from foreseeable injury, rather than obtaining consents from them for each and every research use of their data. Building research freedoms around consents will necessarily foreclose many opportunities for re-use and the discovery of serendipitous uses for data which, though they may generate huge benefits, were not contemplated at the time the data were collected.

### Intellectual property management

Universities and research organizations should establish and maintain enabling and harmonised intellectual property (IP) policies (perhaps incorporating AusGOAL), which explicitly include research data, as a foundation for IP management and licensing arrangements. Holding IP in the data keeps control and maintains the ability to make it open on one's own terms, but it is important to avoid locking-up IP too early (e.g. by overly encouraging patenting, noting the problems associated with, and critiques of, the US Bayh-Dole Act (Boettiger and Bennett 2006)).

IP management must be facilitative rather than blocking, and it may be worth doing further work to determine the principles by which IP can be kept maximally open, thereby enabling the maximisation of returns to public investment in research.

### Guidelines, standards and services

Policies must seek to maximise discoverability and usability by encouraging the use of open formats (i.e. to the extent practicable, platform neutral, machine readable, and open standards-based) and open source software for manipulating the data, and minimising technological barriers to access and use through supporting infrastructure-related open standards and services, and ensuring that data is supported by open standards-based, fit-for-purpose metadata and contextual information, which is published in a publicly-accessible repository.

These are the elements of a policy encompassing both the hard and soft infrastructure necessary to support research data curation and sharing, and provide the structure of incentives necessary to make it happen and make it sustainable.

# Contents

# 1. Introduction

Paul David (2013) tells of the birth of modern science in openness. Great patrons seeking to aggrandise their courts would seek to attract the stars of the firmament of natural philosophy. But without deep knowledge of the field themselves, they could only protect themselves against bad appointments (of scientific cranks) by opening up science for peer review. And the widest possible publicity was necessary for the reputation and careers of scientists, as well as the aggrandisement of courts (Box 1).

### Box 1 The historical origins of 'Open Science'

Economist and economic historian Paul David argues that the precondition for 'take-off' in modern science was the culture of peer review within a community of openness. But where on earth might such a culture have come from given the ancestry of science in the secrecy of military engineering and the cults of alchemy?

He argues that science emerged from the unique conjunction of several factors. Firstly princes sought to aggrandise their court by attracting to it 'stars'. In a bid for self-aggrandisement they went in pursuit of 'merit goods' – and they found them in the arts and what was then called 'natural philosophy' or science.

The culture of openness then arose from emerging stars' need to advertise their achievements to distant princes in the hope of patronage. Galileo exploited his ability to prepare superior telescopes for the Grand Duke of Tuscany, Cosimo de' Medici the Second and urged his patron to present these to other European princes, whereby they too might observe the new-found moons of Jupiter that Galileo had proclaimed "the Medicean stars."

If publicity could fuel a scientific career in this world of merit goods, the patron had another problem. He could see for himself the difference between a painting by Georgio Vasari and one by Michelangelo. But Galileo's telescope notwithstanding, princes had a much harder time sorting the crank scientists from the Galileos. And so they asked other scientists… and peer review emerged.

And so, just as something as beautiful as the lyrebird's tail grew from the simpler stuff of survival of the fittest or competition for a mate, the glories of modern science grew out of the tenacious fight for prestige.

Source: Comments on David, P.A., (2013) The Historical Origins of 'Open Science': An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution (January 30, 2013). *Capitalism and Society*, 3(2) 2008. http://ssrn.com/abstract=2209188

There are many advantages to openness in science. Governments can reap significant economic benefits from the release of research data, through the productivity growth and job creation derived from innovation, and through better-informed policy and research (Science-Metrix 2013). The private sector can gain greater access to fundamental research information, which reduces the cost of and enhances opportunities for innovation.

Research funders can realise both economic and scientific benefits from open research data, and there are a growing number of funders mandating open data. Research institutions (e.g.

universities and research centres) benefit from the enhanced visibility of their contribution that open data brings and from a reduced risk of inadvertently playing host to scientific fraud.

Research users benefit from access to and use of research data, with many reporting significant efficiency savings (Beagrie and Houghton 2014), as well as the extension and enhancement of their work. Open research data can also be used in teaching and education, where anecdotal reports suggest that students are more engaged when working with real data than when using hypothetical textbook examples (Box 2).

### Box 2 Open Science, open source, open learning

Chris Raimondi was one of the early winners on Kaggle, the Australian start-up that hosts global data prediction competitions. The competition he won involved building a model from real world data that had been made openly available on Kaggle in order to optimise a predictive model. The aim of the model was to predict the rate at which HIV load would increase in patients from week to week given specific genetic markers in different patients.

Raimondi built his predictive model from data that Kaggle had made open on its site. He had no prior experience in bio-infomatics and no formal training in statistics, but became interested in data science running a small search engine optimisation firm he operated in Baltimore on the other side of the world to Kaggle's then headquarters in Melbourne. He taught himself data science using YouTube videos and open source data modelling tools.

Within a week and a half of the beginning of the competition, Raimondi's work was exceeding the predictive efficiency of the model that was the state of the academic art, and by the time the competition closed two months later he had taken the state of the art from 70% accuracy to 77%.

Second place went to a team of analysts at the Thomas J. Watson Research Centre at IBM.

Source: Authors' analysis.

Scientific bodies, academies, etc. can realise many benefits in terms of efficiency, transparency, reliability/trust, enhanced peer review, and the reduction of research misconduct, from over-fitting data to outright fraud. And all of these things bring widespread advantages to society in terms of improved healthcare, access to innovative products and services, greater efficiency in government spending, improved policy and better informed policy and business decisions.

From an economic perspective, research data are an asset that we have been building for many years, through public investment in research worth billions of dollars each year, and the information and communication technology (ICT) revolution presents us with an unprecedented opportunity to 'leverage' that asset. Indeed, the ICT revolution has far-reaching consequences and raises many new opportunities. Stiglitz et al. (2000) and Gruen (2014), among others, have suggested that the theoretical underpinnings of the private versus public trade-off shifts as the economy moves toward a digital one, with a larger public role in the digital economy.

There are a number of dimensions along which the public 'shared base' has been expanding. In research, there has been an expansion of those sectors undertaking 'pre-competitive' research, and the emergence of 'open innovation' (Chesbrough 2003; Chesbrough et al. 2006), open access publishing and open research data. In electronics, there has been a push towards greater

availability and use of shared IP cores, including OpenCores, and there is ever greater use of open source software. And, for a wide range of what are increasingly the most valuable economic activities, such as research, education, internet intermediary activities, etc., we are seeing an expansion of fair use/fair dealing, and exceptions to copyright (Houghton and Gruen 2012). Gruen (2012) suggested that the digital world has illustrated something analogous to a 'phase transition' in physics, as in a range of areas ICTs push transactions costs towards zero (Box 3).

## Box 3 Openness and phase transitions around collapsing transactions costs

In the monopolistically competitive world of most networks like our phone networks, as digital technology slashes costs, it takes years to pass it on. With long-distance calls and text now much cheaper in most 'telco' packages, the new frontier is outgoing international calls and the still astronomical cost of international roaming – calling on your Australian mobile in overseas markets.

The internet is a different world. Digital from the start, it works by routing 'data packets': Each is 'addressed' and makes its own opportunistic way through the net depending on network conditions. So although the internet is built from the same reciprocal service agreements between service providers as the phone network, if someone will not negotiate reasonably, other options are always available. And, since no one is indispensable, few are tempted to negotiate unreasonably.

And so, miraculously, all those transaction costs between service providers negotiating reciprocal access to each others' services collapse. Fully 99.5% of reciprocal access agreements on the internet occur informally, without written contracts. Paradoxically, as competition becomes more intense or 'perfect', it becomes indistinguishable from perfect co-operation – a neat trick demonstrated in economists' models a century ago.

What does this mean for efficiency and productivity? If internet transit prices were stated in an equivalent voice-per-minute rate, they would be less than a millionth of a cent per minute – one hundred thousandth of typical voice rates. And as transaction costs have collapsed in cyberspace, new possibilities, new social and economic formations have burgeoned. It is an extraordinary world in which anyone – including (crucially) any innovator – can access the network without requiring the permission of monopolistic gatekeepers – as one must with telephone or TV networks, for instance. So any one of the 2 billion plus people now connected to the net can collaborate with any other.

We already know of the power of Google, Wikipedia, Twitter and open-source software such as Linux. But that is just the beginning. Healthcare, education, even finance will be revolutionised, though the rate at which it happens will still depend on the extent to which the monopolistic gatekeepers impede the progress of the innovators, the visionaries and the barbarians at the gates.

Source: Gruen, N. (2012) 'Telcos reciprocate and market is a net winner', *Sydney Morning Herald*, November 14, 2012. http://www.smh.com.au/business/telcos-reciprocate-and-market-is-a-net-winner-20121113-29adq.html

Perhaps the most important, certainly most visible, open platform for public goods is the internet, which is characterised by its multiple levels of openly accessible platforms – from communications and servers, to Google, Twitter, Facebook, and so on. These platforms are

potentially excludable, but would not be anything like as valuable if excludability were enforced. Indeed, as Gruen (2010) notes:

> *Google and Facebook could close their platforms and charge you for access to them. But… they would not be anywhere near as socially valuable if you were charged – because participants add value on social networking platforms. In fact, they add so much value that private profit seeking builders of such platforms leave them open to all. They generate such vast social value that way that if the builders can monetise just a small fraction of that value they can become rich beyond their wildest dreams*.[3]

In this context, there is increasing awareness of the benefits of curating and openly sharing research data (Kvalheim and Kvamme 2014). The reasons for sharing and enabling re-use of data are many. Curating and sharing research data: encourages scientific enquiry and debate, promotes innovation and potential new data uses, leads to new collaborations between data users and data creators, maximizes transparency and accountability, enables scrutiny of research findings, encourages the improvement and validation of research methods, reduces the cost of duplicating data collection, increases the impact and visibility of research, provides credit to the researcher as a research output in its own right, and provides great resources for education and training.[4]

The importance of open research data policies was recognized internationally in 2004 by the Ministers of Science and Technology of the then 30 OECD countries, and of China, Israel, Russia, and South Africa – at a meeting chaired by Australia. The Ministers asked the OECD to develop a set of principles and guidelines to facilitate cost-effective access to digital research data from public funding. The OECD's guidelines, published in 2007, state that access to research data from public funding should be easy, timely, user-friendly and through the internet, where the marginal costs of transmitting data are close to zero (Science-Metrix 2013). The opportunity is significant and it is available to us now.

## 1.1 Scope of this study

This study seeks to provide estimates of the value and benefits to the Australian economy of freely-available research data, and to examine the role and contribution of data repositories and associated infrastructure. It also explores the policy settings that might encourage greater research data sharing, at national, sectoral, and institutional levels, and thereby increase the return on public investment in research.

The scope of the study is Australia's Commonwealth-funded research and agencies. It includes research commissioned or funded by Commonwealth bodies as well as in-house research within research-oriented agencies wholly or largely funded by the Commonwealth. Therefore, it applies to:

3 http://clubtroppo.com.au/2010/09/16/mr-gruen-goes-to-washington-again-2/

4 http://www.data-archive.ac.uk/create-manage/planning-for-sharing/why-share-data

- Research projects funded through grants, such as the Australian Research Council (ARC), National Health and Medical Research Council (NHMRC), Department of Innovation, and research and development corporations;
- Research directly undertaken within Commonwealth agencies, such as the Commonwealth Scientific and Industrial Research Organisation (CSIRO), the Australian Nuclear Science and Technology Organisation (ANSTO), and Geoscience Australia;
- Research commissioned through consultancies to Commonwealth agencies;
- Research infrastructure facilities, such as the National Collaborative Research Infrastructure Strategy (NCRIS); and
- Research conducted in Australia's higher education institutions.

### Box 4 'Open Science' and the advantages of openness

The crucial message is this: the research was accelerated by being open. Experts identified themselves, and spontaneously contributed based on what was being posted online. The research therefore inevitably proceeded faster than if we had attempted to contact people in our limited professional circle individually, in series. Perhaps this is not surprising, but if it is the case that 'none of us is as smart as all of us' and if we wish to reach scientific goals quickly, why is so much science not practised this way?

Besides speed, there are several other advantages of conducting science in the open:

- The process is transparent, meaning the public can be assured that funding for science, arising from their taxes, is being used responsibly and there is no suggestion of political interference in the scientific process.
- Secondly, in open projects everything is available on the web; the project need not cease with the graduation of students, the termination of a grant or the demise of a principal investigator. Funding for the kernel effort of such a project, crucial in generating activity to which others may respond, can leverage extra input that is unfunded, and this should be attractive for funding agencies keen to maximize the impact of the relevant science.
- Open science is subject to the most rigorous peer review because the review process never ends, essentially because there will always be a commenting function on results, and a mechanism for the community to police those comments.
- The results of open science, freely available on the web, can still be published in pre-publication peer-reviewed journals that accept work that has previously been made public, because this serves as an important mechanism to summarize the research for future participants, and to reward those who have contributed with authorship along a traditional model.

Source: Woelfle, M., Olliaro, P. and Todd, M.H. (2011) Open science is a research Accelerator, *Nature Chemistry* 3, pp745–748 (2011). doi:10.1038/nchem.1149

For the purpose of this study, research data is considered to be the factual digital information that is an input to, or output from, the research process, and which forms the basis of scholarly publications and their validation. It includes the contextual information and metadata to support

the usability of the data (e.g. instrument calibrations, concept definitions, and descriptive information), but does not include physical objects, laboratory notebooks, or plans for future research.

## 1.2 Outline of this report

This report presents estimates based on extrapolation from existing studies. Consequently, the estimates are no more than indicative. Rather than generating a precise number, the aim is to seek a broad appreciation of the potential value of freely-available research data and its associated repository infrastructure to the economy and to the community, where possible expressed as a numerical range and where this is impossible or more likely to mislead than clarify, to express any effects qualitatively.

The following section presents a brief summary of the evidence from previous studies of the value and benefits of shared research data and related data centres and services. It begins with a review of recent studies, drawing out the main approaches and findings. It then turns to an exploration of what these previous studies say about the value of research data and of related research data repositories and infrastructure, in both qualitative and quantitative terms.

The third section explores the implications for Australia, presenting estimates of the value of data in Australia's public research, the value of research data repositories and related infrastructure, and estimates of the potential upside for Australia of research data curation and sharing. The section also includes some commentary on the value and benefits of having data to address problems of national importance, and the costs of research data repository infrastructure and likely trends in those costs.

The fourth section examines the policy settings and infrastructure for public research necessary to enhance the return on public investment in research. It provides an outline of what might be required to make widespread research data curation and sharing happen, and how best to support its development.

# 2. The evidence to date

While there is an extensive body of work on the value and impacts of publicly-funded research, less attention has been paid to the value of research data and the contribution of related infrastructure and facilities. One crucial issue is to distinguish between the value of the data, on the one hand, and that of the related infrastructure and facilities, on the other. While articulating this distinction is possible, in practice it is extremely difficult. Hence, we explore studies in which the value and impacts of the data and related facilities are examined in combination or without distinction, before trying to tease out key findings in relation to the data and infrastructure elements.

## 2.1 The value of research data, facilities and services

There is a growing body of literature on the value and impact of science facilities, with an emphasis on 'Big Science' facilities rather than data repositories and infrastructure per se. 'Big Science' facilities are often single-site facilities (e.g. the Australian Synchrotron), but may also be distributed facilities (e.g. the Square Kilometre Array), networked facilities (e.g. the National Computational Infrastructure), or virtual collections (e.g. the Terrestrial Ecosystem Research Network).

### *2.1.1 Science facilities*

'Big Science' facilities are typically focused on the generation of research data, but they may also host and curate data. The majority of economic impact assessments of such facilities follow a broadly similar approach, wherein evaluators take expenditure and employment data and feed them into an input-output (IO) analysis to estimate the direct and indirect benefits of public expenditure. Such evaluations arrive at economic multipliers that typically range between 2 and 3, which is to say that every $1 million in public expenditure is generating an additional $2 million to $3 million in wider economic activity through onward purchases within supply chains and the personal consumption of employees (Technopolis 2013, p6).

An alternative and complementary approach involves case studies, which often follow innovation impacts on suppliers and users through surveys and/or through tracing the development of spin-off firms and the use of information derived from the science facilities. Examples include CERN studies, which have reported the value of supplier contracts and the ways in which these have facilitated the development of new products or processes. Similarly, NASA's spin-off database reports on the number and revenue of spin-off firms emerging from their research activities (Technopolis 2013, p47).

### *2.1.2 Keeping research data safe*

Among qualitative studies of research data collections are the series of projects, undertaken by UK-based Charles Beagrie Ltd., under the general heading Keeping Research Data Safe (KRDS). The initial KRDS study investigated the medium to long-term costs to UK higher

education institutions of the preservation of research data, and provided a brief overview of the potential benefits from such preservation. They developed a framework and guidance for determining costs, consisting of: a list of key cost variables and potential units of record; an activity model divided into pre-archive, archive, and support services and divided into the major phases from the activity model and by duration of activity; and a resources template including major cost categories (Beagrie et al. 2008).

### Box 5 Open Economics: reasons and background

*Reproducibility*: For economic research to be reliable and trusted, it should be possible to scrutinise and reproduce research findings. This is difficult, or impossible, if data and analysis is not made available. Making material openly available reduces to a minimum the barriers for doing reproducible research.

*Knowledge as a public good*: Data and code should be viewed as a public good, with the greatest benefit coming where it is available freely and openly. Publicly funded research is done in the public interest and should be openly available for the public to access.

*Stability and effectiveness of markets*: Transparent and available information can be central to well-functioning markets. The best way to ensure transparency and that information is available to all relevant parties, including regulators and researchers, is to make data open.

*Public engagement and trust*: Economics and specifically economic data and analysis, plays an important role in many areas of policy-making that directly affect all members of our societies. As such, public engagement and trust are important and openness is central to gaining and retaining trust and increasing engagement.

*Potential new uses of the data*: In many cases the best use of data may ultimately be found outside of its immediate use and making data available may generate new research and create new knowledge. By making material open we ensure that experimentation is easy and that it can be easily re-used and re-purposed.

*Equitable access*: Researchers and research institutions from around the world, including the Global South, can access economic research, data and analysis with no discrimination about their affiliation, research purpose or ability to pay for access.

*Higher impact of research*: Making economic research and data openly available delivers better dissemination of research outcomes and enhances the visibility and the impact of research.

*Democratisation of economics research*: Much of economic research is done with the purpose of improving the economy, policies and institutions. Open economic research will lead to higher citizen engagement leading to better policies and better lives.

*Better resources for education and training*: The opening up of economic research aids in the education of a new generation of economists and social scientists who will be able to produce high quality research.

*Better service delivery and new business models*: Open data can improve the quality and consistency of the public services by exposing inefficiencies and corruption and delivering new ideas on the effective use of public resources. It can also result in better integration of supply chains, harness innovation and revolutionise business models and stimulate entrepreneurship, generating knowledge externalities in the economy.

Source: Open Economics (2014) *Open Economics Principles: Statement on Openness of Economic Data and Code*. http://openeconomics.net/principles/

A second phase project (KRDS2), further developed the activity-based cost model, presented detailed cost information for four organizations, and developed a benefits framework illustrated with two benefit case studies from the National Crystallography Service at Southampton University and the UK Data Archive at the University of Essex. The study found that there can be significant benefits from research data curation and sharing to current researchers in the short-term, as well as long-term benefits to future research. They noted that the costs of a central data repository are an order of magnitude greater than that suggested for a typical institutional repository focused on e-publications alone – although likely less than the user and producer costs that would result from simply opening data, without appropriate curation (e.g. related metadata, sourcing information, and user guides) (Beagrie et al. 2010).

### 2.1.3 A benefit/cost framework

Fry et al. (2008) also sought to identify the benefits arising from the curation and sharing of research data. They suggested that potential benefits include: maximizing the return on investment in data collection; broader access where costs would be prohibitive for individual researchers/institutions; the potential for new discoveries from existing data, especially where data are aggregated and integrated; reduced duplication of data collection costs; increased transparency of the scientific record; increased research impact and reduced time-lag in realising those impacts; and new collaborations and new knowledge-based industries. They suggested that broader, indirect benefits might include: transparency in research and funding; use of data sets in education to enhance the data awareness of students; enhanced researcher skills through access to a broader range of data, tools and standards, having the potential to increase data quality; and increased visibility and the promotion of institutions and researchers making data available.

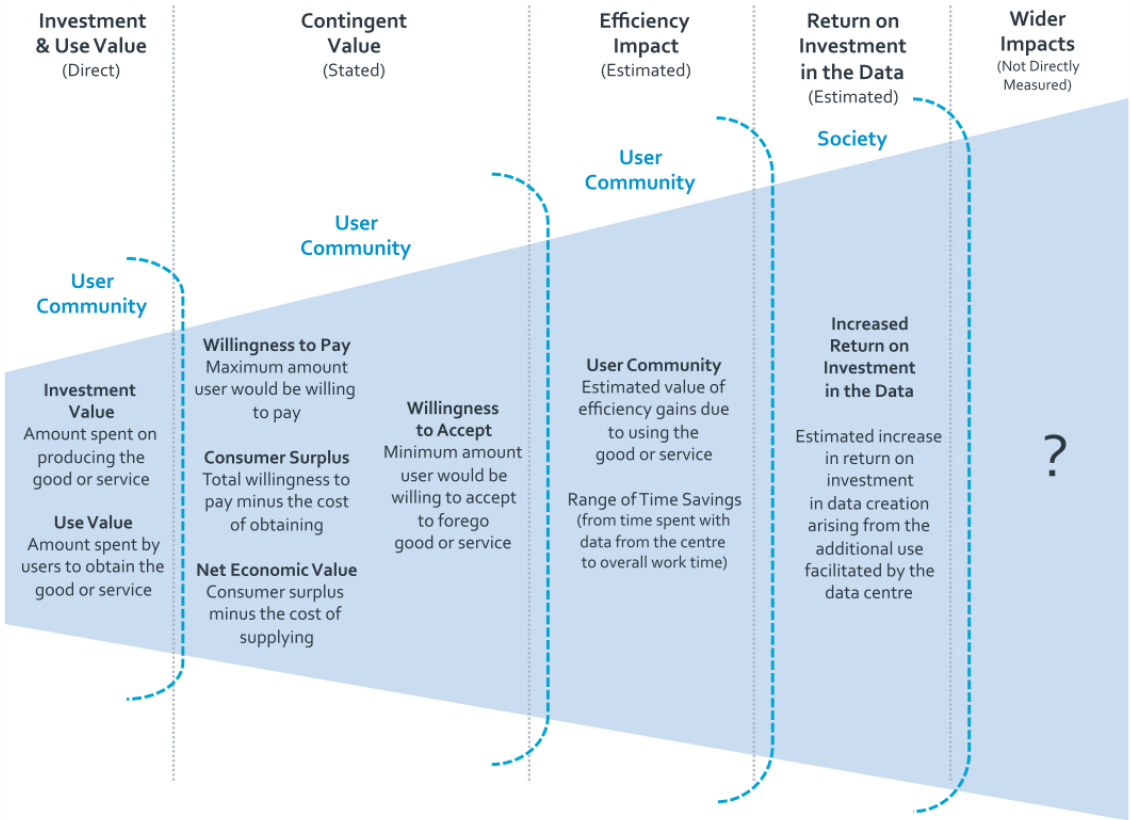
The study used a mixed-method approach, including a literature review and qualitative case studies to inform the development of a model on which to build a business case for data sharing in UK Higher Education. This was based on extensions of the research data preservation cost model proposed by Beagrie et al. (2008), to allow estimation of the benefit/cost to users depositing or accessing data. The case studies investigated included the European Bioinformatics Institute (EBI) and Qualidata, a part of the Economic and Social Data Service (ESDS). Based on the work of co-authors Houghton and Rasmussen, the report presented a simple example of cost-benefit analysis applicable to an individual dataset or repository, based on costs and potential cost savings. The approach was then extended to explore the more diffuse benefits of data curation and sharing at the institutional and disciplinary levels. Unfortunately, due to limited data availability, the study provided a framework for analysis without presenting any detailed analysis of the case studies.

### 2.1.4 UK data centre studies

A series of studies of UK-based research data centres has combined the largely qualitative KRDS framework with a number of quantitative approaches to measure the value and impact of research data curation and sharing. The studies covered the Economic and Social Data Service

(ESDS), the Archaeology Data Service (ADS), and the British Atmospheric Data Centre (BADC) (Beagrie et al. 2012; Beagrie and Houghton 2013a, 2013b; 2014).

## Figure 2 Methods for exploring the economic value and impacts of research data centres




Source: Beagrie, N. and Houghton, J.W. (2014) *The Value and Impact of Data Sharing and Curation: A Synthesis of Three Recent Studies of UK Research Data Centres*, Joint Information Systems Committee (Jisc), Bristol and London.

The quantitative methods (Figure 2) included:

- Estimates of investment value (i.e. the operational expenditure of the data centres plus the time and other costs for depositors submitting data), and use value (i.e. the cost of the time spent by users accessing the data and services);
- Contingent valuation (i.e. willingness to pay and willingness accept), using stated preference techniques (DTLR 2002), to explore the amount that users would be willing to pay to access the data and services, or would be willing to accept in return for giving up their access, in a hypothetical market situation;

- Welfare approaches to estimating consumer surplus (i.e. willingness to pay minus use value) and the net economic value (i.e. consumer surplus minus operational budget)[5] of the data and services provided by the data centres;
- An activity-cost approach to exploring the estimated work-time saving (i.e. efficiency) impacts of the research data centres among their user communities; and
- A macro-economic approach, using a modified Solow-Swan model (Houghton and Sheehan 2009), to explore the increase in social returns on investment in the original creation/collection of the data hosted, arising from the additional use of the data facilitated by the centres (i.e. the implied value of the data re-use by those who could neither have obtained the data elsewhere nor created/collected it themselves).[6]

In all cases, these impacts exceed the costs. The economic analysis from the three studies indicated that:

- A very significant increase in research, teaching and studying efficiency was reported by users of all three centres, with estimated efficiency gains ranging from 2 and up to more than 20 times the costs (i.e. including operational, depositor and user costs);
- The value to users exceeded the investment made in data sharing and curation via the centres in all three cases, with what users pay in terms of their access time and what they would be willing to pay for access, being 2.2 to 2.7 times greater than the value invested in the centres (i.e. in terms of operational costs plus depositor costs); and
- The estimated the value of the increase in return on the original investment in the creation/collection of the data hosted, resulting from the additional use facilitated by the centres, ranged from twice and potentially up to 12 times the investment in the data centres (Beagrie and Houghton 2014).

The users of these research data centres came from all sectors and all fields – close to 20% of respondents to the ESDS user survey were from the government, non-profit and commercial sectors (i.e. non-academic), as were around 40% of respondents to the BADC user survey, and close to 70% of respondents to the ADS users survey. Consequently, value is realised and impacts felt well beyond the publicly-funded research sector alone.

## 2.2 The value of research data

Our review of the literature on the value of research data, related facilities and services reveals that most studies combine the elements and/or focus on the data centre or service. Nevertheless, some of these studies provide a basis from which to tease out the value of the data from that of the related facilities and services. In this section we explore evidence relating to the value of the data, while the next section examines the value of the related infrastructure and services.

---

5 While value may be considered to include both consumer and producer surplus, the data and data centre services in these studies were free. Hence, there was no producer surplus to consider.

6 Social returns refer to the sum of private and public returns (i.e. both the returns that can be captured by the creator/user and those that spill over to others).

### 2.2.1 The investment value of research data

All three of the UK data centre studies included surveys of both data depositors and users, which explored the creation cost of the data hosted as well as the operational expenditures of the centres – thus separating the investment value of the data from that of the data centres. However, it should be noted that all three studies confronted difficulties in estimating data creation/collection costs from depositor surveys – due to the relatively low number of survey respondents, the treatment of initial versus subsequent deposits of data series, and limitations in the base data available to enable the responses to be weighted to reflect the overall pattern of data deposits and holdings. Consequently, the following estimates can be no more than indicative.[7]

Asked about the creation/collection cost of the last data deposited (i.e. a critical incident question):

- Respondents to the Economic and Social Data Service (ESDS) depositor survey reported a mean cost of around £770,000. Converted to an annual average cost and weighted to reflect the overall pattern of data deposits, the total cost of creation/collection of the data hosted by ESDS circa 2010 was estimated at £794 million per annum[8] (Beagrie et al. 2012).
- Respondents to the Archaeology Data Service (ADS) depositor survey reported a mean data creation/collection cost of around £60,000 per dataset deposited. Converted to an annual average cost and weighted to reflect the overall pattern of data deposits, the total cost of creation/collection of the data hosted by ADS circa 2010 was estimated at just over £13 million per annum (Beagrie and Houghton 2013a).
- Respondents to the British Atmospheric Data Centre (BADC) depositor survey reported a mean data creation/collection cost of around £180,000 per dataset deposited. Converted to an annual average cost and weighted to reflect total data deposits, the total cost of creation/collection of the data hosted by BADC circa 2010 was estimated at around £30 million per annum (Beagrie and Houghton 2013b).

These studies give a sense of the orders of magnitude of investments made in the creation/collection of the data deposited with these data centres, which while important in their fields are by no means representative of the overall research data investment.

### 2.2.2 The use value of research data

The UK research data centre users were asked to estimate the approximate share of their total working time spent with data during the last twelve months (e.g. creating, manipulating and analysing data), and for their impression of what might be typical for others in the same sector or field. In all cases, respondents estimated very similar levels of data activity among others in

7 Beagrie and Houghton (2014), and the individual studies, present detailed discussion of methodological limitations and difficulties encountered, as well as the methodological and other differences between the three studies.

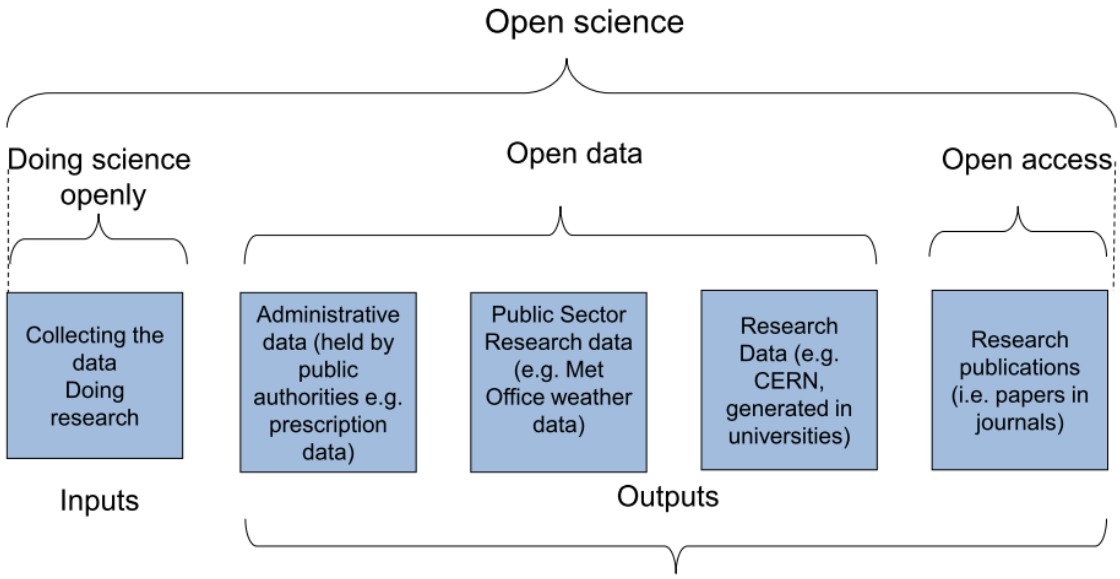
8 The ESDS hosts some very large national and international collections, such as UK census data, World Bank and International Monetary Fund data.

their sector and/or field. Converting the time to a cost gives a sense of how they value the data they use:

- Users of the Economic and Social Data Service (ESDS) reported spending an average of 44% of their time working with data, equivalent to a value of around £850 million per annum across the non-student user community;
- Research users of the Archaeology Data Service (ADS) reported spending an average of 38% of their time working with data, equivalent to around £140 million per annum across the user community; and
- Users of the British Atmospheric Data Centre (BADC) reported spending an average of 56% of their time working with data, equivalent to around £175 million per annum across the user community.

While no more than indicative, these estimates give some sense of the use value of the data hosted by the three UK research data centres surveyed, which, again, while important in their fields are by no means representative of the overall use value of research data.

Figure 3 Open science, open access and open data




Source: Boulton, G. (2013) Science as an Open Enterprise, The Royal Society, Presentation for Open Access Week, Edinburgh (October 2013).

## 2.3 The value of repositories and related infrastructure

The UK data centres studies combined both qualitative and quantitative analysis. Both elements shed light on the level and the nature of value of the data centre facilities and services.

### 2.3.1 Qualitative analysis

Asked if they thought they had saved time or money as a result of using the Economic and Social Data Service (ESDS), more than 80% of user respondents nominated the ability to find data from a single point of access as the biggest area of saving, followed by the quality of data (i.e. the level of preparation, validation and documentation associated with it) (66%), and the fact that it was beyond their ability to create or collect the data for themselves (58%). Asked to what extent they benefited from ESDS, it was clear that users saw the associated methods and documentation as a major benefit, followed by user support (e.g. user guides, helpdesk), and best practices (e.g. case studies and standards) (Beagrie et al. 2012, pp73-74).

For users of the Archaeology Data Service (ADS), the most widely cited factor contributing to savings was the ability to find data from single point of access (87%). That the data were beyond their scope to collect themselves (58%), long-term preservation of data (48%), and guidance on data quality through preparation, validation and documentation of data (41%) were also widely cited benefits. Fifty-eight per cent said that they derived a high or very high benefit from tools (e.g. search tools including ArchSearch, OASIS, web mapping), 36% said they derived high or very high benefit from guides to best practice and standards, and 21% said the derived high or very high benefit from methods and documentation (e.g. DataTrain) (Beagrie and Houghton 2013a, pp89-90).

Among users of the British Atmospheric Data Centre (BADC), 89% of survey respondents said they could not have created/collected the data themselves, with 71% saying they benefited from discovery of required data through single site, 59% from gaining access to multiple datasets/portals through one licence/account, and 52% from the long-term preservation of data offered by BADC. Moreover, almost 50% reported a high or very high benefit from dataset documentation, 34% from online help guides, and 32% from data discovery tools (Beagrie and Houghton 2013b, pp76-80).

### 2.3.2 Quantitative analysis

Quantitative analysis explored the efficiency impacts of accessing data and services from the centres and the additional use facilitated by the centres. In terms of efficiency impacts, Beagrie and Houghton (2014, pp14-16) noted that:

- The estimated efficiency impacts of the ESDS among its non-student user community might be worth around £100 million per annum;
- The estimated efficiency impacts of the ADS among its user community might be worth around £59 million per annum; and
- The estimated efficiency impacts of the BADC among its user community might be worth around £58 million per annum.

Exploring the potential impacts of the data centres on returns to investment in the data hosted and delivered, Beagrie and Houghton (2014, pp14-16) noted that:

- The ESDS facilitates additional use (i.e. by those who could neither obtain the data elsewhere nor create/collect it themselves), which realises additional returns that could

be worth £58 million or more over 30 years (net present value) from one year's investment expenditure;

- The ADS facilitates additional use which realises additional returns that could be worth £2.4 million to £9.7 million over 30 years (net present value) from one year's investment expenditure; and
- The BADC facilitates additional use which realises additional returns that could be worth some £11 million to £34 million (net present value) over 30 years from one year's investment expenditure.

Hence, both the efficiency impacts and the value of re-use facilitated by the data centres are significant.

## 2.4 Summary of the evidence

These studies reveal something of the dimensions and nature of impacts arising from increased access to research data and related facilities and services. They show:

- *Efficiency impacts*, primarily for data users, that can be quantified as time-cost savings. In some circumstances, these savings may be reinvested in the activity, thereby having further return on investment impacts (e.g. where research time saving results in more research being done, thereby increasing the return on investment in research);
- *Additional use* that would not otherwise have been possible, thereby increasing the return on investment in the data creation/collection (i.e. the additional uses of data by those who could neither create/collect the data themselves nor obtain it elsewhere);
- Potentially significant *wider impacts* that are more difficult, perhaps impossible, to fully measure (e.g. greater transparency of science leading to less risk of negative impacts arising from health or pharmaceutical research, making scientific misconduct easier to detect); and
- *New applications and combinations*, with unforeseen value and impacts emerging over time (e.g. new forms of analysis based on data mining, 'mashups' of datasets that had not previously been related to each other, including machine searching for hitherto undetected correlations).

Studies across research fields, disciplines and data types suggest that the benefits of more open access to data (i.e. free or at marginal cost, unrestrictive licensing, and standardised machine readable formats) substantially outweigh the costs.

Exploring the use, value and impact of five UK data centres (including the three noted above),[9] RIN (2011) highlighted the importance of repository infrastructure and services, saying:

> *The qualitative evidence showed very clearly that the benefits which flow from use of research data centres are closely related to the characteristics of these centres. It is not just about making research data available – although this is clearly an important and non-trivial precondition to further exploitation. Data centres provide many other services – and encourage many types of behaviour – which are crucial to achieving the research benefits and wider impacts… In some cases, these are the result of conscious policies; in others, they are side-effects of collecting, storing and managing vast quantities of data.*
>
> *Many researchers mentioned the value of having a large number – and wide range – of datasets in one place. This encouraged them to explore work that might be peripheral to their original interests, but which sometimes turned out to be important. Furthermore, the links that many data centres made between their holdings and a wider literature were highly valued by a number of researchers.*
>
> *Other benefits relate more to the quality of the data, and many respondents framed their comments on this in contrast to the likely outcomes if researchers were left to manage their own data. The quality of the data was felt to be higher because researchers knew it would be submitted to a data centre. Equally, the data centres were able to ensure good metadata, which enabled researchers to enjoy the efficiency benefits that they identified.*
>
> *The technical support provided by data centres was cited by many as an important research benefit… Some researchers also mentioned the value of training offered by the data centre, saying that this was important for the next generation of students* (RIN 2011, p55).

What all of these studies show is that, in the words of the RIN study, it is not just about making research data available, although this is clearly important and non-trivial. Data centres provide many other services, which are crucial to achieving the research benefits and wider impacts.

They also offer a down-payment on a certain cultural value – the value of openness and sharing in science – which evidence suggests is fundamental to science in some ultimate macro sense of ultimate validation, but is also likely to be of value at more micro levels, which is something that current macro-financial arrangements are inimical to – as they tend to encourage competition between scientists, rather than cooperation. Data centres create a structure for cooperation that may be a powerful cultural asset of much wider value.

9 The RIN survey covered the Archaeology Data Service (ADS), the British Atmospheric Data Centre (BADC), the Chemical Database Service (CDS), the Economic and Social Data Service (EDS), and the National Geoscience Data Centre (NGDC).

# 3. Implications for Australia

In this section we present estimates of the potential value of openly accessible Australian public research data, and the value and impacts of Australian research data repositories and related infrastructure. These estimates are based on activity costs and cost savings, and the use of a modified Solow-Swan model to generate estimates of the implied value of increased access to Australian public research data (See Annex I). They are based on extrapolation from international studies, primarily from the UK, scaled to the Australian context.

### Box 6 Research expenditure bases for the range estimates presented

Public research expenditure can be taken as relating to either the source of funds or the sector of execution. For the purposes of estimation, we explore a range of research expenditure from total Australian Government funding support for research ($8.9 billion), to the sum of government and higher education expenditure on research by sector of execution ($13.3 billion) (Australian Government 2014 and ABS 2014).

*Lower Bound*: As the focus is on research activity costs, we present lower bound estimates based on the labour-cost share of research funding and expenditure, which directly 'buy' researcher time. Labour costs account for around 48% of total public research spending ($4.3 billion to $6.4 billion per annum circa 2012-13).

*Upper Bound*: However, other current and capital expenditures also relate to research data activities and contribute to the extent and efficiency of those activities. Hence, we present upper bound estimates based of total research funding and expenditure ($8.9 billion to $13.3 billion per annum circa 2012-13).

Source: Authors' analysis.

## 3.1 The value of data in Australia's public research

In this section, we present two alternative estimates of the value of data in Australia's public research. The first explores the 'use value' of research data (i.e. the cost of the research time spent creating, manipulating and analysing data), while the second estimates the return on investment in public research data activities (i.e. the return on one year's investment in research data activity at average returns to R&D over 20 years in net present value). Both approaches are conservative (Annex II).

**Both approaches suggest that the value of data in public research in Australia is at least $2 billion per annum and possibly up to $6 billion per annum (at 2012-13 levels of expenditure and activity)**.

### *3.1.1 The use value of data in public research*

One simple approach to estimating the value of research data is to estimate the cost of the time spent producing and using it. Put simply, preference theory indicates that people express the value of something by how much time and/or money they spend on it, in preference to an

alternative activity. Hence, an approximate estimate of the value of data in Australia's public research can be derived from research activity times and expenditures. Such an approach is conservative, as the value of research data lies more in its effects than in its production and use (Annex II). Unfortunately, it is much more difficult, if not impossible, to measure those effects.

As noted, the UK research data centres studies (Beagrie et al. 2012; Beagrie and Houghton 2013a, 2013b; 2014), reported that research users of the Archaeology Data Service (ADS) reported spending an average of 38% of their time working with data, users of the Economic and Social Data Service (ESDS) reported an average of 44%, and users of the British Atmospheric Data Centre (BADC) an average of 56%. In all cases, respondents estimated very similar levels of data activity among others in their sector and/or field. Assuming a 37½ hour working week and taking a simple 'mean of the means' suggests that researchers across the disciplines may spend an average of around 17 hours and 18 minutes per week or 46% of their time working with data (i.e. creating, manipulating, and analysing data).

**If an average of 46% of research activity time is spent creating, manipulating and analysing data, then the activity-cost or 'use value' of data in Australian public research would be at least $2.0 billion per annum and possibly up to $6.2 billion per annum (at 2012-13 prices and levels of activity).[10]**

### *3.1.2 The return on investment in research data activities*

Another approach to estimating the value of data in Australia's public research is to explore the likely return on investment in the researcher activity time. There is an extensive literature on the returns to investment in R&D activities. Reported returns vary widely, but a characteristic finding is that returns are high – often in the region of 20% to 60% per annum over the useful life of the knowledge (Griliches 1995; Salter and Martin 2001; Scott et al. 2002; Shanks and Zheng 2006; Martin and Tang 2007; Sveikauskas 2007; Hall et al. 2009, 2010). Using a modified Solow-Swan model (Houghton and Sheehan 2009; Houghton et al. 2009), we can explore the likely return on investment in the research data activity time (Annex I).

As noted above, previous studies suggest that researchers spend an average of around 46% of their time working with data (i.e. creating, manipulating and analysing data), and average returns to R&D typically range from 20% to 60% per annum. A further issue is that data and research are global, not national, so we need to estimate the share of the impacts of data use that would accrue within Australia. There is an extensive literature on localization of returns to R&D, which suggests that something of the order of 66% are likely to accrue locally.[11]

---

10 An important simplifying assumption in this estimate is that the inward and outward inter-sectoral and international flows of data balance (i.e. that the value of data in public research is the same as the value of public research data). Consequently, these estimates are no more than indicative.

11 For example, a number of studies have looked at the issue of the relative impact of local research on local returns and/or the international spillover of R&D. Jaffe (1989) suggested that domestic knowledge is twice as important as foreign knowledge (i.e. 66% was local). Coe and Helpman (1993, 1995) adopted a trade weighting approach, and concluded that approximately a quarter of the benefits from R&D in G-7 countries accrued to their trading partners, and 75% locally. Verspagen (2004), citing Arundel and Guena (2004), suggested weights for domestic versus foreign sources of 73% for domestic and 27% for foreign sources.

Drawing on preliminary work on the UK R&D Satellite Account (Evans et al. 2008), and following the lead of the US R&D Satellite Account (Sveikauskas 2007), we depreciate publicly-funded research data at 10% per annum, and we set the useful life of the data created each year at an average of 10 years – although, of course, the useful life of data can be much shorter and/or much longer depending on data type and use. Following Mansfield (1991, 1998), we distribute the returns normally over five years from year 1 through year 5. Applying a 4% discount rate to estimate net present value, we then model the recurring increase in returns to expenditure on the data.

Table 1 Parameters used in modelling estimates

| *Item* | *Value* |
|---|---|
| Expenditure on public research (AUD millions per annum) | 13,336 |
| Australian government support for research (AUD millions per annum) | 8,935 |
| Labour-related expenditures on public research data (AUD millions per annum) | 6,370 |
| Labour-related expenditures on government supported research data (AUD millions per annum) | 4,268 |
| Returns to data creation/collection expenditure (per cent per annum) | 40% |
| Useful life of the data in years (averaged across all data types) | 10 |
| Rate of depreciation of the underlying stock of data (per cent per annum) | 10% |
| Discount rate to estimate Net Present Value (per cent) | 4% |
| Localisation of returns to R&D (per cent) | 66% |

Source: Authors' analysis.

**On that basis, we estimate the return on one year's research data activity time at between $2.9 billion and $9.1 billion over 20 years in net present value, of which some $1.9 billion to $6.0 billion might accrue within Australia.**

## 3.2 The value of repositories and related infrastructure

The effective curation and sharing of research data is a substantial task that requires data repositories and related infrastructure. While there have been a number of reports and studies of the benefits of research data curation and sharing, much of their focus has been on the research and 'scientific' benefits (e.g. transparency and accountability), rather than the economic benefits. An exception is the series of studies of UK research data centres undertaken by Beagrie and Houghton (Beagrie et al. 2012; Beagrie and Houghton 2013a, 2013b, 2014). Hence, we draw on the findings from these studies, extrapolating to the Australian context, to estimate the potential value and impacts of research data repositories and related infrastructure for Australia.

Across research fields, the UK data centre studies identified two major impacts: (*i*) significant research and work efficiency impacts for the users of the data centres; and (*ii*) substantial additional use of the data by users who could neither recreate the data themselves nor obtain it elsewhere. In this section, we focus on estimating the potential scale of these impacts for

Australia.[12] **Our analysis suggests that the potential value of public research data repositories for Australia might be at least $1.8 billion and possibly up to $5.5 billion per annum.**

### *3.2.1 The potential efficiency impacts*

Based on UK experience, we can estimate the direct value of the cost/time saved if all researchers in Australia had the access to research data and services enjoyed by the users of the three established UK data centres studied. However, that is not all of the benefits arising from the time saving, because the time saved would enable researchers to do more research, which would in turn create knowledge and value.[13] So the total impact is the sum of the direct time/cost saving and the additional returns to R&D resulting from the additional research that is done – as the time saved is re-invested in research.

As noted, the users of the UK data centres reported spending an average of around 17 hours and 18 minutes per week or 46% of their research/work time working with data (i.e. creating, manipulating and analysing data). While the exact wording of the questions varied a little between surveys, all were asked how much their use of data and services from the data centre had changed their research efficiency.[14] Average responses ranged from time savings of 28% to 44% across the surveys, with a simple 'mean of the means' of around 37%. Hence, taking the labour-related share of public research funding as the lower bound, the direct efficiency time saving of similar facilities would be worth between $720 million and $1.1 billion per annum (at 2012-13 costs and levels of activity). Taking total research spending as the upper bound, the direct efficiency time saving would be worth between $1.5 billion and $2.3 billion per annum (at 2012-13 costs and levels of activity).

Assuming that the time saved would be used to do more research, we can add the returns to the additional research – effectively, treating the efficiency saving as additional research expenditure. At an average return of 40%, the additional returns from one year's research spending would be worth at least $1.1 billion to $1.6 billion over 20 years in net present value, of which $710 million to $1.1 billion would be likely to accrue within Australia, and possibly as much as $2.2 billion to $3.3 billion over 20 years in net present value, of which $1.5 billion to $2.2 billion would be likely to accrue within Australia.

---

12 While these are simply extrapolations, there is sufficient similarity between the UK and Australian research, funding and disciplinary contexts to support such extrapolation.

13 For example, exploring the use, value and impact of five UK data centres, RIN (2011) noted that the most widely agreed benefit of data centres was research efficiency – data centres make research quicker, easier and cheaper, and ensure that work is not repeated unnecessarily. Time was the most important efficiency benefit reported by the survey respondents, with cost savings and reduction in duplication of effort significant but somewhat less important. Notably, 50% of respondents or more (up to 77% in one case) agreed that the data centre had enabled them to undertake a greater quantity of research.

14 For example, can you estimate what percentage of your total research time you save as a result of using the particular data centre's data and services?

**Adding these elements suggests a total annualised potential efficiency impact from the widespread curation and sharing of research data of at least $1.8 billion and possibly as much as $5.5 billion, of which perhaps $1.4 billion to $4.5 billion might accrue within Australia.**

### 3.2.2 The potential increase in return on investment

The second major finding from the UK data centre studies was the extent of additional use (re-use) of the data facilitated by the data centres. To the extent that there is additional use of the data facilitated by the repository infrastructure, there will be additional returns to investment in the data creation/collection.

The UK surveys asked users: (i) If the data centre '*x*' had not existed, would you have been able to obtain the last data you used from another source?, and (ii) If you could not have got the data elsewhere, would you have been able to collect/recreate the last data you used yourself?[15] For those saying 'YES' they could re-create and/or get elsewhere, the benefit is the saved time/cost of doing so plus the additional returns generated by the additional research done during the time saved (as above). For those saying 'NO' they could not recreate or obtain the data elsewhere, we assume that the research could not have been done and that the entire additional return to research is an attributable benefit of the data centres' curation and sharing. Across the three surveys, between 44% and 58% of respondents said they could neither recreate the data themselves nor obtain it elsewhere, with a simple 'mean of the means' of 52%.

**Based on estimates of current returns, the value of the returns to the additional use would be at least $560 million and possibly up to $1.6 billion from one year's expenditure in net present value, of which perhaps $370 million to $1 billion might accrue within Australia.**

### 3.2.3 The potential benefits of having national collections

Averages derived from other studies tend to disguise the importance of having national data collections that enable researchers to address national issues of importance, be they local issues (e.g. household expenditure patterns and trends) or the local implications of global issues (e.g. climate change).

The UK data centre studies from which the estimates presented above are derived, explored the value and impacts of UK data centres that host UK-oriented data collections. For example, the Archaeological Data Service (ADS) provides a repository for all UK construction industry-related archaeological finds and focuses on the UK archaeological evidence base. Similarly, the British Atmospheric Data Centre (BADC) host UK weather observations, as well as those from surrounding waters. Hence, the value of national collections are included in the estimates to some extent. Nevertheless, Australia's unique geography, climate, fauna and flora suggest that there may be additional value associated with Australian national data collections.

15 Note that these are critical incident questions to randomise the responses.

### Box 7 Case studies of Australian research data re-use

*Where in the world?* – Predicting where particular plants and animals are likely to be found is one of the most important problems in ecology. Re-use of publicly-funded research data enables species distribution modelling packages to determine the range of climatically suitable conditions for a particular species and analyse how particular species will react to changes in climatic variables.

*Maximising the benefits of high-resolution climate modelling* – Instead of simulating the climate of the whole globe at a coarse resolution, Regional Climate Models simulate the climate of a continent or smaller region at a much finer resolution. Regional Climate Models are therefore able to better simulate the local climatic effects and provide more realistic climate data to climate impacts assessments.

*Sensitive data sharing benefiting women's health* – The Australian Longitudinal Study on Women's Health (ALSWH) has been Government-funded and gathering data on the mental, physical, and social health of over 50,000 women since 1995. ALSWH adds considerable value to other data sources by supporting sub-studies and data linkage projects.

*Health policy needs data sharing* – Researchers, practitioners and communities must have access to public, health-relevant information through data-sharing mechanisms to continue improving the health of all Australians. Access to such data, which can be confidentialised, enables these groups to contribute to the evidence base that guides policy and program development, and monitors its progress.

Source: http://www.ands.org.au/discovery/reuse.html

One example is hydrological data and the Bureau of Meteorology's Australian Water Resources Information System (AWRIS), which provides free online access to water data, leading to improvements in the timeliness, quality and efficiency of water management and policy decision-making.[16] Inter alia, this provides the foundation for water allocation and trading, and is essential in the management of water allocations for agricultural irrigation and the efficient allocation of what is one of Australia's most scarce resources (i.e. water).

### *3.2.4 The potential upside value of data curation and sharing*

Adding the estimated efficiency and re-use impacts gives a sense of the potential upside of research data curation and sharing for Australia. However, there is an important caveat. These estimates are simply extrapolating from what is already being achieved in the UK as a result of data centre infrastructures that have been established for a decade or more. As such the estimates may be conservative, as there may be further benefits to be gained from the existing data infrastructure and/or we in Australia may do better than the UK has to date in curating and sharing research data and providing related infrastructure and services.

Nevertheless, simply adding the estimated direct efficiency savings and reinvestment of those savings into further research to the returns from additional use facilitated by data curation and sharing suggests a **total potential annualised impact of $2.3 billion to $7.2 billion at**

16 http://www.bom.gov.au/water/about/publications/document/InfoSheet_3.pdf

**2012-13 levels of activity, of which some $1.8 billion to $5.5 billion might accrue within Australia.**

In addition, there are wider benefits for science and society that cannot be captured fully in economic estimates, such as the transparency and accountability of science and speeding up the discovery and innovation processes, with all the potential economic, health and other benefits that that might have.

### *3.2.5 Currently realised and unrealised potential value*

This section explores the issue of what share of Australia's potential public research data curation and sharing benefits are already being realised in order to provide some estimate of the potential unrealised benefits. While based solely on scenarios derived from previous studies, these estimates are suggestive of the overall quantum of benefits that may result from a concerted effort to facilitate the curation and sharing of Australia's public research data.

Without undertaking a detailed study, we have little information about the current level of public research data curation and sharing. Clearly, there is active and widespread data sharing in some fields. But anecdotal evidence suggests that such sharing is characteristic of islands of activity, rather than being typical across the board.

Table 2 The annual value accruing within Australia from curating and openly sharing public research data (summary of estimates)

| *Estimate* | *Labour Costs (Lower Bound)* | *Total Expenditure (Upper Bound)* |
|---|---|---|
| **Current value of data in public research** | **$1.9 billion to $2.9 billion** | **$4.0 billion to $6.2 billion** |
| - Use value (cost of data activity time) | $2.0 billion to $2.9 billion | $4.1 billion to $6.2 billion |
| - Estimated return on investment (at 40%) | $1.9 billion to $2.9 billion | $4.0 billion to $6.0 billion |
| **Potential value of data repositories** | **$1.8 billion to $2.6 billion** | **$3.7 billion to $5.5 billion** |
| - Efficiency impacts | $1.4 billion to $2.1 billion | $3.0 billion to $4.5 billion |
| - Additional (re)use return on investment | $370 million to $495 million | $690 million to $1.0 billion |
| **Unrealised upside of curation & sharing** | **$1.4 billion to $2.4 billion** | **$2.9 billion to $4.9 billion** |

Note: Lower bound estimates are based on the labour-cost share and upper bounds estimates on total research funding and expenditure.
Source: Authors' analysis.

**Hypothetically, if current data curation and sharing is in the range of 10% to 20% of the research data being produced, then on the estimates presented above some $1.4 billion to $4.9 billion in annualised benefits may remain unrealised**.[17] Of course, it may be more, if we do more and/or better than the existing UK research data centres studied.

17 Implicit in these estimates is an assumption of linearity of returns to investment in research. However, there may be *diminishing returns* because people are likely to host and share the best and most valuable data first (i.e. prioritise their curation and sharing). Conversely, there may be *increasing returns* because the culture of re-use develops as the available collections develop, and such benefits as transparency, peer review and the identification of mistakes, the reduction of fraud, the reduction of duplicative research, etc. grow as the share of data available grows. Hence, the assumption of linearity seems an acceptable simplification.

Estimated returns to one year's investment in Australia's public research data creation activities are currently worth at least $960 million and possibly up to $3 billion net present value. At average returns to research expenditure, these estimates suggest that each multiple of use of Australia's public research data might be worth an annualised $960 million to $3 billion, of which $635 million to $2 billion might accrue with Australia. Hence, the potential upside impact of research data curation and sharing noted in the previous sections is around 2.8 times estimated current impacts. Effectively, these estimates suggest what might be possible if public research data were re-used an across-the-board average of 1.8 times.

## 3.3 The potential costs of research data curation

Of course, research data curation and sharing is not without costs. But without conducting a survey, it is difficult to estimate the current costs of research data curation in Australia.[18] Consequently, this section briefly explores existing studies of research data repository costs, extrapolating the results for Australia.

### *3.3.1 Institutional, disciplinary and national data repository costs*

The initial Keeping Research Data Safe study (Beagrie et al. 2008, p70) explored the costs of a small number of *institutional data repositories*, noting that:

- A University of Cambridge repository reported employing 4 staff (FTE) and spending £58,764 per annum on equipment and other costs – equivalent to a total cost of around $510,000 per annum; and
- A Kings College London repository reported employing 2.5 staff (FTE) and spending £27,564 per annum on equipment and other costs – equivalent to a total cost of around $303,000 per annum.

Looking at *national data centres and collections* the second Keeping Research Data Safe study (Beagrie et al. 2010) noted that:

- The UK Data Archive had approximately 50 staff (FTE) and held more than 5,000 datasets – equivalent to a cost of around $6 million per annum; and
- The National Digital Archive of Datasets, operated by the University of London on behalf of the National Archives, had 13.34 staff (FTE) – equivalent to around $1.5 million per annum.

Summarising the findings of the RIN survey of UK *research data centres*, Collins (2011) noted the following annual operating costs:

- The Archaeology Data Service, circa £640,000 – equivalent to around $1.1 million per annum;

18 In part, because it is not clearly capital or current spending, as data repository infrastructure can be established, even supported on an ongoing basis, by project expenditure.

- The British Atmospheric Data Service, circa £1 million – equivalent to around $1.7 million per annum;
- The Chemical Database Service, circa £250,000 – equivalent to around $417,000 per annum;
- The Economic and Social Data Service, circa £2 million – equivalent to around $3.3 million per annum; and
- The National Geoscience Data Centre, circa £350,000 – equivalent to around $583,000 per annum.

More fully investigating the operational costs of UK research data centres and the specifically data sharing-related costs faced by their data contributors, Beagrie and Houghton found that:

- The annual operating cost of the Economic and Social Data Service was around £3.3 million, and an estimated £7 million per annum was invested by data depositors in the preparation for deposit and deposit of data (Beagrie et al. 2012) – equivalent to a total of around $17.2 million per annum;
- The annual operating cost of the Archaeology Data Service was around £700,000, and an estimated £465,000 per annum was invested by data depositors in the preparation and deposit of data (Beagrie and Houghton 2013a) – equivalent to a total of around $1.9 million per annum; and
- The annual operating cost of the British Atmospheric Data Centre was around £2 million, and an estimated £2 million per annum was invested by data depositors in the preparation and deposit of data (Beagrie and Houghton 2013b) – equivalent to a total of around $6.7 million per annum.

### 3.3.2 National disciplinary research data centres

Beagrie et al. (2008) noted that a number of *national disciplinary research data centres* were funded by UK's research councils. The Natural Environment Research Council (NERC) was then funding eight disciplinary data centres, the Economic and Social Research Council (ESRC) was funding the UK Data Archive, and the Arts and Humanities Research Council (AHRC) was funding five service providers within the Arts and Humanities Data Service until 2008 – thereafter it was funding archaeology alone. A study by the Office of Science and Innovation working group for preservation and curation found that the running costs of these data centres across the three research councils to be remarkably consistent – at between 1.4% and 1.5% of the total research expenditure of the research council (Beagrie et al. 2008, p71).

**Simply extrapolating this UK result to the scale of Australia's public research funding would suggest national disciplinary data repository costs of some $130 million to $200 million per annum.**

While material, the costs of research data curation and sharing are small in comparison with the potential benefits, which can be measured in the billions rather than millions. Given similarities between the Australian and UK public research sectors, there seems no reason to suppose that

the direct and measurable benefits of research data curation and sharing through repositories and related infrastructure would not be as great or greater than those identified above.

### 3.3.3 Trends in research data sharing costs

Looking at cost trends over time, in general terms, it is clear that staff costs account for the largest share of overall costs – typically more than 50% and up to 90% of the costs in some cases. Hence, to a significant extend, costs will move inline with research administration or academic pay rates. ICT costs are an important, but smaller component of costs. Moreover, the trend is towards ever-cheaper storage and equipment capacity. To some extent, ICT may also be considered a sunk cost for many of the potential research data hosting agencies. Thus adding relatively little in terms of data curation and sharing specific costs.

Research data repository-related activity costs reflect phases of activity, from the pre-archive phase (e.g. outreach and initiation) to the archive phase (e.g. acquisition, ingest, storage, preservation, and access), and include a range of shared services (e.g. administration and common services). Examining the UK Archaeological Data Service, Beagrie et al. (2010, p33) note that:

> *Looking at the distribution of staff costs over five major cost categories… (pre-archive, acquisition, ingest, archive, and access), the largest proportion is accounted for by the access category (31%). However, the activities leading up to and including ingest of the materials into the archive collectively account for 55% of total staff costs. Somewhat surprisingly (compared to some public perceptions), the process of actually preserving the materials (archive category) accounts for only 15% of total staff costs.*

Moreover, there appear to be significant scale economies, with per unit archived costs falling as collections grow. Once archival capacity has been set up, the marginal cost of adding another mega-byte of content seems to be quite low. The cost of setting up and maintaining an apparatus for getting material into the archive seems to be much greater than the cost of setting up and maintaining an apparatus for preserving these materials over the longer term (Beagrie et al. 2010).

These observations suggest a number of things. First, scale economies and the high share of staff costs suggest that cooperation, collaboration and coordination of activities might help to reduce overall costs, through the dynamics of scale and learning. Although this needs to be balanced with consideration of disciplinary differences and specifics, as well as individual and institutional incentives and responsibilities.

Second, the high labour intensity of research data curation suggests that some activities may become automated over time, especially those aspects of the data curation process with very high staff costs, such as pre-ingest and ingest (Beagrie et al. 2010, p45). As research data curation and sharing becomes an integral part of the research process, therefore, one can envisage much greater automation from capture to ingest, further reducing unit costs.

# 4. Making up lost ground

This section outlines the policy settings, infrastructure and services necessary to more fully realise the benefits of open research data. It suggests that funding is just one of the problems that must be solved to optimise the total value of data public goods. There are also difficult trade-offs and design questions involving, for instance, ensuring the quality of that public good and its accountability to its most intensive users and ultimately to optimising social value.

### Box 8 Permissions, information innovation and serendipity

Free access to information and serendipity are closely related. A central fact about the human condition, ignored in many economic models, is that even at our most sophisticated we are only boundedly rational. A person or group cannot consider all possible propositions and information states they could encounter. Thus, the possible outcomes of any research project, large or small, can never be fully anticipated. Serendipity is central to our relationship to information.

Many serendipitous discoveries arise when a prepared mind makes a previously unnoticed connection between seemingly disparate pieces of information. The number of such discoveries that are possible in a given information network depend on the number of people with access to the network and on the number of connections they can potentially make. This is of the order the square of the number of pieces of information accessible to each member of the network.

Even seemingly moderate restrictions on the freedom of information may drastically reduce the potential for serendipitous discovery. This is true whether we are talking about freedom as in availability without payment or in another sense of the freedom to copy and tinker with others' work and ideas.

Suppose that requirements for paid access, or practices that put off participation reduce the number of network participants by 80% (this seems likely given the general pattern in which most value accrues to the top 20% of participants in any activity) and that each participant only accesses 20% of the information that would be available in the absence of those restrictions. Then the number of observed connections potentially available is only 0.8% (0.2*0.2*0.2) of those that would be available without restrictions. While this is a purely illustrative example, there is no reason to suppose that it overstates the loss of potential discovery associated with restricting the size of networks.

In policy terms, the ubiquity of serendipity and the inherent impossibility of predicting serendipitous discovery or connection implies that there must always be a presumption in favour of free inquiry, free discussion and therefore of free access to information. This presumption may be rebuttable in particular cases, but the burden of proof should always be firmly on those arguing to restrict freedom.

Source: Professor John Quiggin, Federation Fellow, University of Queensland. Cited in the Report of the Government 2.0 Taskforce available at http://www.finance.gov.au/publications/gov20taskforcereport/.

Key ingredients for success include:

- Recognising the importance and benefits of open access to research data;
- Recognising that there can be costs associated with making research data openly available, and supporting appropriate infrastructure funding;

- Establishing a structure of incentives to encourage and/or enforce open data and ensure that it is sustainable;
- Ensuring that institutional use of the intellectual property system facilitates open access to research data;
- Recognising that there will be privacy and confidentiality issues that must be addressed, and ensuring that there are clear guidelines to follow that provide strong protections to possible mischiefs, such as privacy breaches, whilst minimising obstructions to data use;
- Ensuring that the necessary 'top-down' management of the data infrastructure is complemented with measures to encourage bottom-up innovation and experimentation; and
- Encouraging the use of open formats and minimising technological barriers to access and use through supporting infrastructure-related standards and services.

## 4.1 Enabling policy

Optimising the policies and institutions to drive innovation around open data requires two things that are in some tension with each other. On the one hand, the architecture of the various institutions needs to be articulated, and the relations between them. This is largely a 'top-down' policy exercise. On the other hand, top-down approaches can be inimical to innovation from the bottom up, which is at a premium where, as here, there is a high level of complexity and rapid change.

**Policy institutions and culture**: The starting point should be government, research funder, and institutional mandates stating the expectation or requirement for open research data as the default. Recognising that research is a global activity, with many cross-institutional and international collaborations, all such mandates should seek to maximise the national and international harmonisation of policies, in order to reduce compliance costs by ensuring that compliance involves the same actions across policy jurisdictions (Chan et al. 2013). Further, pride should be cultivated in Australian contributions to the global data commons, together with the use of such contributions, to encourage the development of a mutually reciprocating community of national practice (Cutler & Company 2008).[19]

At the same time, systems for resourcing and encouraging good work in the management or use of data should foster an environment that is open to experimentation and new approaches, and which rewards talent and fosters intrinsic motivation. Open competitive bidding for projects is likely to have some role in such a system, but we should guard against over-reliance on those allocating funds being able to determine relative research merit before the research is conducted,

19 *Venturous Australia: A Review of the National Innovation System*, Recommendation 7.14: To the maximum extent practicable, information, research and content funded by Australian governments – including national collections – should be made freely available over the internet as part of the global public commons. This should be done whilst the Australian Government encourages other countries to reciprocate by making their own contributions to the global digital pubic commons.

or even their ability to assess it immediately on completion. It is for this reason that a quite common management practice in some leading private sector innovators, such as Google and Atlassian, is '20% time', which deliberately creates scope for those who are more junior in the system to follow their own intrinsic motivations, back their own judgement and to collaborate with others of like mind, even amid the scepticism of those in higher positions. There have always been problems in getting compliance with government policies of open data, though progress has steadily been made in improving compliance. One expedient worth considering would be to allow institutions and/or individuals to enforce access rights to others' data in the system in appropriate circumstances.

Further, we are not far begun on the task of open data in research and, in view of this, it seems appropriate to devote some resources to cultivating discussion on the potential and opportunities for open data. The UK have moved in this direction more vigorously than Australia – something attested to by this study's reliance on UK data rather than Australian data. The UK has also led the way with the Open Data Institute and the Connected Digital Economy Catapult.

Another issue of importance is the way in which the falling costs of storage are enabling some data curation to occur in retrospect at the time the data is interrogated. This does not vitiate all the advantages of appropriate curation, but in some circumstances it does raise the spectre of a penumbral class of data, that could be stored at minimal cost, with standards of curation that are lower than currently practiced or are indeed minimal. Thus raw data might be stored from all manner of sources at low cost on the assumption that the benefits of preserving the option value of searching such data are likely to exceed the very modest costs of storage.

**Enabling infrastructure**: It is important to recognise that there can be costs associated with making research data openly available, and to provide appropriate data repository infrastructure funding through such schemes as the National Collaborative Research Infrastructure Strategy, the Super Science Initiative, and so on. In some fields of research (e.g. climate modelling, particle physics, astronomy) data sets may too large to download and use locally in any meaningful way, and in such cases it may be preferable to provide access to the data in situ via such means as applications and/or application interfaces (APIs) to enable users to re-sample, subset, or downscale the data.

The National Collaborative Research Infrastructure Strategy (NCRIS) is currently under review, and we defer to the findings of that review as to whether the NCRIS model remains the most appropriate funding model for eResearch infrastructure, including research data repositories. However, from an economic point of view there are clear advantages to any scheme that creates incentives to collaborate, thus encouraging co-investment and both leveraging greater overall funding by bringing forth investments that would not otherwise have happened and ensuring that all parties have 'some skin in the game'. Given the characteristic scale and learning economies noted above, encouraging collaboration will likely be more effective than simple competition for limited funding.

## Box 9 The National Collaborative Research Infrastructure Strategy

The National Collaborative Research Infrastructure Strategy (NCRIS) model has been defined by how it has emphasised collaboration, strategically identified priorities through consultative road-mapping, facilitated the development of capability plans, and funded skilled staff and operating costs.

In 2010, an evaluation of NCRIS found it to be an efficient, effective and appropriate model for developing important research infrastructure in Australia. Key findings included:

- *Appropriateness:* The NCRIS model is appropriate for funding medium to large scale research infrastructure, and is superior to previous models. It greatly improved the allocation of resources.
- *Effectiveness:* NCRIS was cost-effective and met research infrastructure needs (as defined for funding purposes).
- *Efficiency:* NCRIS was managed efficiently with appropriate administrative costs. However, it needed greater transparency around how access fees for infrastructure are charged.
- *Integration:* NCRIS engaged successfully with the commonwealth, state and territory governments and government agencies.
- *Performance assessment:* NCRIS adequately assessed performance for NCRIS capabilities.
- *Strategic policy alignment:* NCRIS aligned with the Australian Government's broader policy objectives.

Source: Department of Industry (2010) National Collaborative Research Infrastructure Strategy (NCRIS): Evaluation 2010. https://education.gov.au/2010-evaluation-national-collaborative-research-infrastructure-strategy-ncris

**Privacy and confidentiality**: All parties must realise that, while the default is open access, there may be privacy, ethical, security, commercial or other constraints on releasing research data. Addressing these constraints at an early stage in the research process is crucial, and it is essential to ensure that there are clear data access and management guidelines (e.g. clear processes for meeting any privacy, confidentiality and security concerns whilst imposing the minimum obstacles on worthwhile research being conducted).

In this regard, the focus should be on protecting data providers and those whose data is used from foreseeable injury, rather than obtaining consents from them for each and every research use of their data. Building research freedoms around consents will necessarily foreclose many opportunities for re-use and the discovery of serendipitous uses for data which, though they may generate huge benefits, were not contemplated at the time the data were collected. Thus for instance if someone's primary school records can be used to uncover some regularity – for instance between primary school performance and health in later life – the presumption should be that this should be possible without further permission providing privacy and other appropriate matters are handled satisfactorily.

Consent based approaches seem compelling in public debate – appealing as they do to notions of personal property in one's own data – with the alternative of maximising the public good from all data being a more abstract and technocratic proposition. It may be possible to make progress by seeking to introduce deliberatively democratic forms of governance (citizens'

juries) rather than top down codes of ethics – which can obstruct reasonable trade-offs between ethical values.

**Intellectual property management**: Universities and research organizations should establish and maintain enabling and harmonised intellectual property (IP) policies (perhaps incorporating AusGOAL), which explicitly include research data, as a foundation for IP management and licensing arrangements. Holding IP in the data keeps control and maintains the ability to make it open on one's own terms, but it is important to avoid locking-up IP too early (e.g. by overly encouraging patenting, noting the problems associated with, and critiques of, the US Bayh–Dole Act (Boettiger and Bennett 2006)).

IP management must be facilitative rather than blocking. It may be worth doing some further work to determine the principles by which IP can be kept maximally open whilst earning sufficient revenue to maintain it and/or curate the data.

**Guidelines, standards and services**: Policies must seek to maximise discoverability and usability by encouraging the use of open formats (i.e. to the extent practicable platform neutral, machine readable, and standards-based) and open source software for manipulating the data, and minimising technological barriers to access and use through supporting infrastructure-related standards and services, and ensuring that data is supported by standards-based, fit-for-purpose metadata and contextual information, which is published in a publicly-accessible repository.

These are the elements of a policy encompassing both the hard and soft infrastructure necessary to support research data curation and sharing, and provide the structure of incentives necessary to make it happen and make it sustainable.

# References


ABS (2014) *Research and Experimental Development, Government and Private Non-Profit Organisations*, Australia, 2012-13, Cat No 8109.0. http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/8109.0Main%20Features32012-13?opendocument&tabname=Summary&prodno=8109.0&issue=2012-13&num=&view=

ABS (2014) *Research and Experimental Development, Higher Education Organisations*, Australia, 2012, Cat No 8111.0. http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/8111.0Main%20Features22012?opendocument&tabname=Summary&prodno=8111.0&issue=2012&num=&view=

ACIL Tasman (2008) *The Value of Spatial Information*, Spatial Information Systems Limited. www.crcsi.com.au/uploads/publications/PUBLICATION_324.pdf

Arundel, A. and Geuna, A. (2004) 'Proximity and the use of public science by innovative European firms,' *Economics of Innovation and New Technology* 3(6), pp559-580.

Australian Government (2014) *The Australian Government's 2012-13 Science, Research and Innovation Budget Tables*, Canberra. http://www.industry.gov.au/AboutUs/Budget/Documents/SRIBudgetTables2012-13.pdf

Barabási, A-L. and Albert, R. (1999) 'Emergence of scaling in random networks,' *Science*, 286 (5439): 509–512. arXiv:cond-mat/9910332

Beagrie, N., Chruszcz, J. and Lavoie, B. (2008) *Keeping Research Data Safe*, JISC, London and Bristol. http://www.beagrie.com/krds.php

Beagrie, N. (2009) *Draft Guide to Cost/Benefit Analysis for Research Data Services*, Charles Beagrie, Salisbury. http://www.beagrie.com/DMIcost&benefit_programmeguidev1.pdf

Beagrie, N., Lavoie, B, and Wollard, M. (2010) *Keeping Research Data Safe 2*, JISC, London and Bristol. http://www.beagrie.com/jisc.php

Beagrie, N., Houghton, J.W., Palaiologk, A. and Williams, P. (2012) *Economic Evaluation of Research Data Infrastructure* (ESDS), Economic and Social Research Council, London. http://www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf

Beagrie, N. and Houghton, J.W. (2013a) *The Value and Impact of the Archaeology Data Services: A Study and Methods for Enhancing Sustainability*, Joint Information Systems Committee, Bristol and London. http://www.jisc.ac.uk/whatwedo/programmes/preservation/ADSImpact.aspx

Beagrie, N. and Houghton, J.W. (2013b) *The Value and Impact of the British Atmospheric Data Centre*, Joint Information Systems Committee and the Natural Environment Research Council UK, Bristol and London. http://www.jisc.ac.uk/whatwedo/programmes/di_directions/strategicdirections/badc.aspx

Beagrie, N. and Houghton, J.W. (2014) *The Value and Impact of Data Sharing and Curation: A Synthesis of Three Recent Studies of UK Research Data Centres*, Joint Information Systems Committee (Jisc), Bristol and London. http://repository.jisc.ac.uk/5568/

Boettiger, S. and Bennett, A.B. (2006) 'Bayh-Dole: if we knew then what we know now,' *Nature Biotechnology* 24(3), pp320-323 (doi:10.1038/nbt0306-320). http://www.nature.com/nbt/journal/v24/n3/full/nbt0306-320.html

Boulton, G. (2013) Science as an Open Enterprise, The Royal Society, Presentation for Open Access Week, Edinburgh (October 2013).

Buchanan, M. (2001) *Ubiquity: Why Catastrophes Happen*, Crown Publishers, New York.

Chan, L., Houghton, J.W. and Swan, A. (2013) *Financial Implications of a Tri-Agency Harmonized Open Access Policy*, Report to the Canadian Funding Agencies, by the University of Toronto (October 2013).

Chesbrough, H. (2003) *Open Innovation: The new imperative for creating and profiting from technology*, Harvard Business Scholl Press, Boston MA.

Chesbrough, H., Vanhaverbeke, W. and West, J. (eds.) (2006) *Open Innovation: Researching a New Paradigm*, Oxford University Press, Oxford.

Coe, D.T. and Helpman, E. (1993) *International R&D Spillovers*, NBER Working Paper 4444, National Bureau of Economic Research, Cambridge MA.

Coe, D. T. and Helpman, E. (1995) 'International R&D Spillovers', *European Economic Review* 39, pp859-887.

Collins, E. (2011) 'Use and Impact of UK Research Data Centres,' The International Journal of Digital Curation 6(1) 2011, pp20-31. http://www.ijdc.net/index.php/ijdc/issue/view/12

Cutler & Company (2008) *Venturous Australia: Building strength in innovation*, Review of the National Innovation System, January 2008. http://www.industry.gov.au/science/policy/Documents/NISReport.pdf

Denison, E.F. (1985) *Trends in American Economic Growth, 1929-1982*, Brookings Institution, Washington DC.

DTLR (2002) *Economic Valuation with Stated Preference Techniques*, London: Department of Transport, Local Government and the Regions. http://www.communities.gov.uk/documents/corporate/pdf/146871.pdf

Evans, P., Hatcher, M. and Whittard, D. (2008) 'The preliminary satellite account for the UK: a sensitivity analysis', *Economic & Labour Market Review* 2(9), September 2008, pp37-43.

Franklin, S. (2000) Review of Stuart Kauffman's "At Home in the Universe, The search for laws of self-organization and complexity", *Times Literary Supplement.* http://www.msci.memphis.edu/~franklin/kauffman.html

Fry, J., Lockyer, S., Oppenheim, C., Houghton, J.W. and Rasmussen, B. (2008) *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes*, JISC, London and Bristol. http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/jiscdataproposal-public.pdf

Griliches, Z. (1995) 'R&D and productivity: Econometric Results and Measurement Issues,' In Stoneman, P. (Ed.) (1995) *Handbook of The Economics of Innovation and Technological Change*, Blackwell, Oxford, pp52–89.

Gruen, N. (2014) *Government as Impresario: Emergent Public Goods and Public Private Partnerships 2.0*, NESTA, London. http://www.nesta.org.uk

Gruen, N., Houghton, J.W. and Tooth, R. (2014) *Open for Business: How Open Data Can Help Achieve the G20 Growth Target*, A Lateral Economics Report commissioned by the Omidyar Network (June 2014). http://www.omidyar.com/sites/default/files/file_archive/insights/ON%20Report_061114_FNL.pdf

Hall, B.H., Mairesse, J. and Mohnen, P. (2009) *Measuring the returns to R&D*, NBER Working Paper 15622, NBER, Cambridge MA.

Hall, B.H., Mairesse, J. and Mohnen, P. (2010) Measuring the returns to R&D, in eds. Hall, B.H. and Rosenburg, N. (2010) *Handbook of the Economics of Innovation*, North Holland.

Houghton, J.W. and Sheehan, P. (2009) 'Estimating the potential impacts of open access to research findings,' *Economic Analysis and Policy* 39(1). http://www.eap-journal.com/vol_39_iss_1.php

Houghton, J.W., Rasmussen, B., Sheehan, P.J., Oppenheim, C., Morris, A., Creaser, C., Greenwood, H., Summers, M. and Gourlay, A. (2009) *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits*, Report to The Joint Information Systems Committee (JISC). http://www.jisc.ac.uk/publications/reports/2009/economicpublishingmodelsfinalreport.aspx

Houghton, J.W. (2011) *Costs and benefits of data provision*, Report to The Australian National Data Service, Canberra. http://ands.org.au/resource/cost-benefit.html

Houghton, J.W. and Gruen, N. (2012) *Exceptional Industries: The economic contribution to Australia of industries relying on limitations and exceptions to copyright*, A Lateral Economics Report for The Australian Digital Alliance (ADA). http://www.digital.org.au/content/LateralEconomicsReports

Houle, D., Diddahally, R., Govindaraju, and Omholt, S. (2010) Phenomics: The Next Challenge, *Nature Reviews Genetics* 11, pp855-866 (December 2010). doi:10.1038/nrg2897.

Jaffe, A. (1989) 'Real effects of academic research,' *American Economic Review* 79, pp957-970.

Kaufmann, S.A. (2000) *Investigations*, Oxford University Press.

Kvalheim, V. and Kvamme, T. (2014) *Policies for Sharing Research Data in Social Sciences and Humanities: A survey about research funders' data policies*, International Federation of Data Organizations (IFDO). http://www.cessda.net/news/ifdo_survey_report.pdf

Lateral Economics (2014) *Open for Business: How Open Data Can Help Achieve the G20 Growth Target*, A Lateral Economics report commissioned by Omidyar Network. http://www.omidyar.com/sites/default/files/file_archive/insights/ON%20Report_061114_FNL.pdf

Mansfield, E. (1991) 'Academic research and industrial innovation,' *Research Policy* 20(1), 1991, pp.1-12; and Mansfield, E. (1998) 'Academic research and industrial innovation: an update of empirical findings,' *Research Policy* 26(7/8), 1998, pp.773-776.

Martin, B.R. and Tang, P. (2007) *The benefits of publicly funded research*, SWEPS Paper No. 161, Science Policy Research Unit, University of Sussex, Brighton.

OECD (2007) *OECD Principles and Guidelines for Access to Research Data from Public Funding*, Paris. http://www.oecd.org/science/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm

RIN (2011) *Data centres: their use, value and impact*, A Research Information Network Report to JISC, London and Bristol. http://www.jisc.ac.uk/publications/generalpublications/2011/09/datacentres.aspx

Royal Society, The (2012) *Science as an Open Enterprise*, The Royal Society, London. https://royalsociety.org/~/media/policy/projects/sape/2012-06-20-saoe.pdf#page=1&zoom=auto,0,842

Salter, A.J. and Martin, B.R. (2001) 'The economic benefits of publicly funded basic research: a critical review,' *Research Policy* 30(3), pp509-532.

Science-Metrix (2013) *Open Data Access Policies and Strategies in the European Research Area and Beyond*, EC DG Research and Innovation. http://www.science-metrix.com/pdf/SM_EC_OA_Data.pdf

Scott, A., Steyn, G., Geuna, A., Brusoni, S. and Steinmueller, E. (2002) *The Economic Returns to Basic Research and the Benefits of University-Industry Relationships*, Report to the Office of Science and Technology, London.

Shanks, S. and Zheng, S. (2006) *Econometric modeling of R&D and Australia's productivity*, Staff Working Paper, Productivity Commission, Canberra.

Sveikauskas, L. (2007) *R&D and Productivity Growth: A Review of the Literature*, US Bureau of Labor Statistics, Washington DC., Working Paper 408.

Solow, R.M. (1957) 'Technical Change and the Aggregate Production Function,' *Review of Economics and Statistics* 39, pp312-320.

Solow, R.M. (1987) Growth Theory and After. *R.M. Solow – Prize Lecture*, Nobel e-Museum Laureates.

Stiglitz, J.E., Orszag, R.R and Orszag, J.M. (2000) *The Role of Government in a Digital Age*, Computer and Communications Industry Association, Washington, DC.

Technopolis (2013) *Big Science and Innovation*, Report to the Department of Business, Innovation and Skills, London. https://www.gov.uk/government/publications/big-science-and-innovation--2

Tung, J.Y., Do, C.B., Hinds, D.A., Kiefer, A.K., Macpherson, J.M., Chowdry, A.B., Francke, U., Naughton, B.T., Mountain, J.L. Wojcicki, A. and Eriksson, N. (2011) 'Efficient Replication of over 180 Genetic Associations with Self-Reported Medical Data,' *PLOS One*, August 17, 2011. DOI: 10.1371/journal.pone.0023473

US Department of Commerce (2014) *Fostering Innovation, Creating Jobs, Driving Better Decisions: The Value of Government Data*, Department of Commerce, Washington DC.

Verspagen, B. (2004) *The Impacts of Academic Knowledge on Macroeconomic Productivity Growth: An Exploratory Study*, Eindhoven Centre for Innovation Studies, Eindhoven.

# Annex I A modified Solow-Swan model

It is possible to gain some sense of the scale of potential benefits arising from open research data by using a modified Solow-Swan model. This annex describes the modification developed by Houghton and Sheehan (2009).[20]

## Returns to R&D in a simple Solow-Swan model

In the basic Solow-Swan model, the key elements are a production function:

(1) $Y = A^{\eta} K^{\beta} L^{\alpha}$

where A is an index of technology, K is the capital stock and L is the supply of labour, with both K and L are taken to be fully employed by virtue of the competitive markets assumption, and an accumulation equation:

(2) $\dot{K} = sY - \delta K,$

where $\dot{K}$ is the net investment or the change in the net capital stock, equal to gross investment less depreciation, and $\delta$ is a constant depreciation rate. Substituting (1) into (2) gives

(3) $= sA^{\eta} K^{\beta} L^{\alpha} - \delta K.$

From (3) it is possible to determine the conditions for steady state growth in the capital stock.

Re-arranging, taking logarithms, differentiating with respect to time and imposing the condition that for steady state growth:

$$d/dt(\ln \dot{K}/K) = 0$$

gives:

(4) $$\dot{K}/K = \frac{\eta}{1-\beta} \dot{A}/A + \frac{\alpha}{1-\beta} \dot{L}/L$$

where $\dot{K}/K = \dot{C}/C = \dot{Y}/Y$, is the single constant steady state rate of growth of capital stock, consumption and output, respectively.

The main features of the Solow-Swan model are apparent from equation (4). Firstly, if technology and labour supply are fixed, the steady state growth rate is zero. That is, there is no endogenous growth in the model, growth being driven in the steady state by change in the exogenous variables. Secondly, if one of technology and population show positive growth then the steady state growth rate of the economy is proportional to the growth rate in that variable; if both rates are positive the economy's growth rate is a weighted average of the two. Thirdly, the steady state growth rate does not depend on either the level of savings or of investment in the economy. An economy that continuously saves and invests 20% of national income will have a higher level of output than one investing 5%, but it will not have a higher steady state growth

20 Houghton, J.W. and Sheehan, P. (2009) 'Estimating the potential impacts of open access to research findings,' *Economic Analysis and Policy* Vol. 39, Issue 1. http://www.eap-journal.com/vol_39_iss_1.php

rate. Thus the broad economic message of the Solow-Swan model is that steady growth is possible in a purely competitive world, provided that there is growth in either population or technology, or both.

## Contributions to growth and total factor productivity

Solow (1957) further developed this model in a way that provided the foundations for subsequent 'growth accounting'. Starting with total differentiation of the production function (1), and substituting for the partial derivatives of Y from (1) with respect to each of its arguments, yields:

(5) $\dot{Y}/Y = \eta\dot{A}/A + \beta\dot{K}/K + \alpha\dot{L}/L.$

Equation (5) can then be used in two main ways in the empirical study of growth.

Given that in the competitive model capital and labour are paid their marginal products and assuming constant returns to scale, $\beta$ and $\alpha$ can be estimated from the relative shares of capital and labour. A variant of (5) with those weights can then be used to estimate the relative contribution of capital, labour, technology and other factors to growth. Solow made pioneering estimates in 1957, the results of which he later described as "startling" (Solow 1987), and these have been much refined and amplified by Denison (1985) and others. Solow found that $7/8^{th}$ of the growth in real output per worker in the US economy between 1909 and 1949 was due to "technical change in the broadest sense" and only $1/8^{th}$ to capital formation. Denison's 1985 estimates covered the US economy for the period 1929 to 1982. Of the growth in real business output of 3.1% per annum over that period, he found that the increase in labour input with constant educational qualifications accounted for about 25% and capital input for 12%. Most of the remainder is accounted for by technological progress and by the increased human capital of the workforce. What was "startling" about these results was the relatively minor contribution to output growth arising from the increase in the traditional factors of production, capital and labour.

The other related use of equation (5) is to estimate the "Solow residual", or total factor productivity. This is defined as the difference between output growth and the weighted sum of the growth rates of factor inputs (K and L), using constant return to scale weights. That is, total factor productivity growth (TFP) is given by:

(6) $TFP = \dot{Y}/Y - \beta\dot{K}/K - \alpha\dot{L}/L,$

where $\beta = 1 - \alpha$, and $\beta$ and $\alpha$ are derived from the shares of capital and labour in total income.

Total factor productivity is thus the growth in output not accounted for, on these assumptions, by the growth in capital and labour inputs. This method is now used very widely around the world in measuring productivity. This recent use has confirmed the broad Solow-Denison findings, in that for most modern economies total factor productivity growth is significantly more important than expansion of inputs in explaining total output growth. However, it must be remembered that the method rests on the assumptions embedded in the Solow model and that, as a consequence, the finding that the larger proportion of growth is to be explained by an exogenous "technical change in the broadest sense" constitutes something of an admission of defeat for economic analysis.

## Estimating the rate of return to R&D

While there are recognized limitations to the traditional growth model approach, this basic framework has been widely used in estimating the rate of return to R&D. The standard approach to estimating returns to R&D is to divide the technology variable A in (1) into two components, a stock of R&D knowledge variable R and a variable Z that represents a matrix of other factors affecting productivity growth. The production function then becomes:

(7) $Y = K^{\alpha} L^{\beta} R^{\gamma} Z^{\eta}$,

and the counterpart of equation (5) becomes:

(8) $\dot{Y}/Y = \alpha \dot{K}/K + \beta \dot{L}/L + \gamma \dot{R}/R + \eta \dot{Z}/Z$.

That is, the rate of growth of the R&D knowledge stock (*i.e.* accumulated R&D expenditure or R&D capital) contributes to output growth as a factor of production, with elasticity $\gamma$. The rate of return to knowledge ($\partial y/\partial R$) is that continuing average per cent increment in output resulting from a one per cent increase in the knowledge stock. This can be readily derived from the elasticity $\gamma$ by

(9) $\partial y/\partial R = \gamma . (Y/R)$.

The normal approach to creating a measure of the stock of R&D knowledge, for a given industry or for the economy as a whole, is to use the perpetual inventory method to create the knowledge stock from the flows of R&D, using the relationship:

(10) $R_t = (1 - \delta) R_{t-1} + R\&D_{t-1}$,

where $\delta$ is the rate of obsolescence of the knowledge stock. This method also requires some starting estimates ($R_0$) of the knowledge stock, and estimates can be sensitive to that assumption.

Then the capital stock at time *t* is given by:

(11) $$R_t = (1 - \delta)^t R_0 + \sum_{i=0}^{t-1} (1 - \delta)^i R\&D_{t-1}$$

Given a series for R and for the variables Z, it is then possible to estimate $\gamma$ by either of the two methods noted above: estimate equation (8) with the parameters $\alpha$ .. $\eta$ unconstrained, or obtain estimates of the parameters $\alpha$ and $\beta$ (constrained to be equal to one) from the factor shares of capital and labour, calculate TFP by a variant of (7) and regress R and Z on TFP to obtain $\gamma$.

## Incorporating the efficiency of research and accessibility of knowledge

This standard approach makes some key simplifying assumptions. Here we note three in particular. It is assumed that:

- All R&D generates knowledge that is useful in economic or social terms (*efficiency of R&D*);

- All knowledge is equally accessible to all entities that could make productive use of it (*accessibility of knowledge*); and
- All types of knowledge are equally substitutable across firms and uses (*substitutability*).

A good deal of work has been done to address the fact that the substitutability assumption is not realistic, as particular types of knowledge are often specialized to particular industries and applications. Much less has been done on the other two assumptions, which are our focus.

We define an '*accessibility*' parameter $\varepsilon$ as the proportion of the R&D knowledge stock that is accessible to those who could use it productively, and an '*efficiency*' of R&D parameter $\phi$ as the proportion of R&D spending that generates useful knowledge. Then starting with a given stock of useful knowledge $R^*_0$ at the start of period zero, useful knowledge at the start of period 1 will be given by:

(12) $R^*_1 = (1 - \delta) R^*_0 + \phi R\&D_0,$

where the contribution of R&D in period zero to the knowledge stock is reduced by the parameter $\phi$ to allow for unproductive R&D. This means that the stock of useful knowledge at period *t* is given by:

(13) $$R^*_t = (1 - \delta)^t R^*_0 + \phi \sum_{i=0}^{t-1} (1 - \delta)^i R\&D_{t-1}$$

If the period over which knowledge is accumulated is long, so that $(1 - \delta)^t R^*_0$ is small relative to $R^*_t$, then $R^*_t$ can be approximated by $\phi R$. However, only a proportion of useful knowledge may be accessible, so that accessible useful knowledge at period *t* is $\varepsilon R^*_t$, and hence approximately $\phi\varepsilon R_t$, where $R_t$ is the stock of knowledge as calculated under the standard methods.

Using this approximation and noting that it is accessible useful knowledge that is the correct factor in the production function, (6) becomes:

(14) $Y = K^\alpha L^\beta (\phi\varepsilon R)^\gamma Z^\eta$

If $\phi$ and $\varepsilon$ are independent functions of time, then the results of estimating a linearized version of (14) that excludes them will be misleading. However, if we assume that these parameters reflect institutional structures for research and research commercialisation in a given country, and can hence be taken as fixed (and as less than or equal to one), then the standard results stand, but need to be reinterpreted. Again using R as the stock of knowledge calculated by the standard method (which assumes $\phi = \varepsilon = 1$) and $R^*$ as the corresponding accessible stock of useful knowledge, then $R = R^*/\phi\varepsilon$, and the rate of return to useful and accessible knowledge becomes:

(15) $\partial y/\partial R^* = \gamma . (Y/R^*) = \gamma/\phi\varepsilon . (Y/R) = \gamma . (Y/R).1/\phi\varepsilon.$

Thus, if $\phi$ and/or $\varepsilon$ are less than one, the rate of return to $R^*$ is greater than that to R by the factor $1/\phi\varepsilon$. This does not imply that the measured rate of return to R is biased, because $R^* = \phi\varepsilon R$.

Assume now that there is a one-off increase in the value of $\phi$ and $\varepsilon$, from the constant values of $\phi_0$ and $\varepsilon_0$ to new values of $(1 + \delta_\phi)\phi_0$ and $(1 + \delta_\varepsilon)\varepsilon_0$, respectively. Then the rate of return to $R^*$, that is:

(16) $\partial y/\partial R^* = \gamma . (Y/R) . (1/\phi_0\varepsilon_0)$

is fixed, but the return to R will increase:

(17) $\partial y/\partial R = \gamma . (Y/R) = \phi_1\varepsilon_1 \, \partial y/\partial R^* = \gamma . (Y/R) . (\phi_1\varepsilon_1 /\phi_0\varepsilon_0)$

$= \gamma . (Y/R) . (1 + \delta_\phi).(1 + \delta_\varepsilon)\varepsilon_0.$

It follows from (17) that, because the increase in efficiency and accessibility leads to a higher value of $R^*$ for a given level of R, the rate of return to R will increase by the compound rate of increase of the percentage changes in $\phi$ and $\varepsilon$.

The basic result of the foregoing is that, if *accessibility* and *efficiency* are constant over the estimation period, but then show a one-off increase, then, to a close approximation, the return to R&D will increase by the same percentage increase as that in the *accessibility* and *efficiency* parameters.

# Annex II A digression on non-linearity

The following section explains why we might expect the benefits of open data to be increasing at a possibly accelerating rate with the openness of data and the power of the systems we have to handle data.[21] We do not rely on this possible phenomenon in our report, but outline the issues here to make the conservatism of our own method clear.

Systems and networks can be thought of as collections of individual items and the linkages between them. The best way to think of individual bits of data is as points in such a system with the linkages being its relationship to other bits of data, either directly or as uncovered by the demands of the users. It is clear that each data point gets its value from its contribution to whatever understanding or insight these connections give us.

It is always difficult to value outcomes that are, to a significant extent unknown. In such cases, there can be a tendency to talk in terms of black swans or other metaphors and attribute subjective assessments. In the case of open data, however, there is reason to believe that the potential value may go well beyond the direct benefits that more and better information provide or what can be ascertained by extrapolating existing usage and returns. It may, in fact, be orders of magnitude greater.

This is because a collection of data may have properties that exist *at the level of the entire system* rather than at the level of the individual bits. In particular, it may share, with many other phenomena ranging from the web, earthquakes, brain functions, traffic jams and the like, the characteristics of a complex network, or what is known as *self organized criticality*. What this means is that once the collection becomes sufficiently large, the entire network is at a critical point. At such a point, any small change, such as an additional input of data, may cause it to rapidly shift to another state (Buchanan 2001).

One of the most significant indicators of self-organized criticality in a system is that linkages are scale free and exhibit a power law distribution (Buchanan 2001). A power law distribution means that the fraction of data points with $x$ connections, or inputs into other collections of data, are expected to follow the function

$$p(x) = bx^{-a}$$

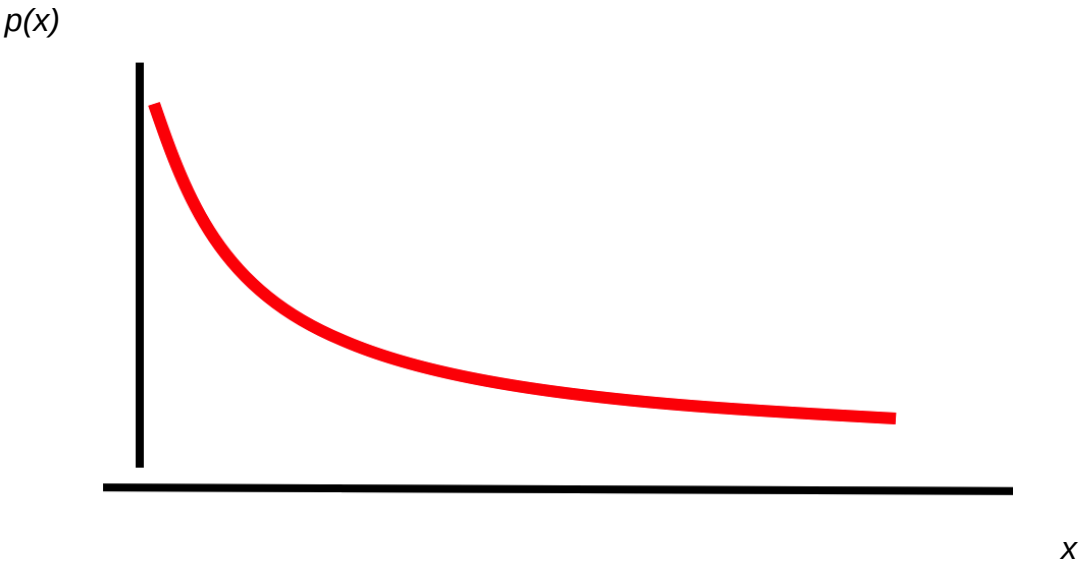
where $a$ and $b$ are positive numbers. For example, let $a = 2$ and $b = 1$. Then every bit of data will have one connection, half of the data will have two connections, one hundredth will have ten connections, and so on. This function is represented below (Figure A1). The horizontal axis is the number of connections and the vertical is the fraction of the data.

Although a full study is beyond our scope, there are strong reasons for suspecting that open data will exhibit a power law distribution. This distribution has been observed across the web, for example, in social media and networking websites, as well as in search engines. Similarly with

21 This annex is derived from a passage in Gruen, N., Houghton, J.W. and Tooth, R. (2014) *Open for Business: How Open Data Can Help Achieve the G20 Growth Target*, A Lateral Economics Report commissioned by the Omidyar Network (June 2014).

other data in all their varieties, some bits will rarely be used, say the amount spent on recycling lawn clippings in public parks, whereas other bits of data, such as economic growth rates and GDP, will be more widely used.

Figure A1 Power law distribution for data




Source: Authors' analysis.

Most data linkages will follow a process known as preferential attachment. In other words, links between bits of data will not be established at random but will depend on other links. The more useful a bit of data, the more likely it is to be linked. This attachment turns up in wealth distribution, the popularity of web sites, both for visiting and for contributing to, downloads of music and video clips and data searches on the web. For similar reasons, this would be expected in patterns of open data usage. Any distribution with preferential attachment also has the characteristics of a power law distribution (Barabási and Albert 1999).

### Implications for the value of open data

New 'big data' means of tackling knowledge problems are sometimes producing productivity improvements of several orders of magnitude. The fact that open data may exhibit self-organized criticality means that most of the standard valuation methods used in this report may (will probably?) return a figure that is lower than the expected future value, possibly orders of magnitude lower. If open data turns out to be critical, a small increase in the amount of data available should not be treated as simply an increment to the existing network. The important mathematical property of such a system is that any addition always has the potential to cause the entire network to jump to a new state in which the connections and the payoffs change dramatically, perhaps by several orders of magnitude.

Underestimations of this type seem to have been relatively common in relation to information and communication associated activities. Among the examples here are the demand for computers, the effect of mobile phone technology, the time required to sequence DNA and so

on. The idea of the runaway process built on power laws is behind the popular idea of 'the singularity'. The idea was put into play by mathematician John von Neumann in 1958, and since paraphrased as the process unleashed at the point that artificial intelligence exceeds human intelligence unleashing a process of "ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue".[22]

An experiment that gives some idea of how criticality works is the button and thread experiment (Kaufmann 2000). In the words of Franklin (2000):

> *Imagine strewing a multitude of buttons randomly about a bare floor. Now pick two buttons at random and join them by a thread. Put them back. Choose another two, connect and return them. Continue this process, keeping track of the number of buttons in the largest connected cluster, that is, the largest group of buttons that could be lifted together by picking up one of its members. Kauffman's computer models of this experiment show that this largest connected cluster grows slowly until the number of threads is a little more than half the number of buttons. Then, suddenly, it grows large very quickly. In models containing 400 buttons, this maximal cluster size goes from under 50 to over 350 as the number of threads rises from just below 200 to just above 250. Plotting a graph of maximal cluster size against number of threads yields a steep S-curve.*

In this experiment, the additional connection has a value, in terms of the number of buttons lifted, well in excess of what we might estimate using more standard extrapolation techniques.

In a similar way, open data always carries with it the probability of such a jump. Moreover, in the case of open data, the jump must always be to a more valued from a less valued state. As with the buttons, more threads can only cause the number lifted to increase. All this is speculative, but it is backgrounded in a well-established body of empirical and mathematical research, which allows estimation of probable returns. In trying to ascertain the value of open data it is always these returns that matter and the probability of large gains from criticality must be properly considered. To err on the side of conservatism, however, they have been omitted from this study, so that it is clear that we are not making over-optimistic claims for the benefit of open data.

### A worked example: Productivity disruption in genomic research

Using genetic information to identify potential health conditions (e.g. disease risks, drug resistance or susceptibility to drug side effects) involves two expenses. The most familiar is the cost tied to sequencing an individual's genetic code. In recent years, this has radically decreased in price from about USD 100 million in 2001 to about USD 5,000 in 2014 for full genome sequencing.[23] Using a genotyping approach, where single spot-differences in DNA are used to build a snapshot of the genome instead of sequencing every nucleotide, reduces the price again, down to below USD 100 per individual.

---

22 The Singularity, Wikipedia. http://en.wikipedia.org/wiki/The_Singularity.

23 National Human Genome Research Institute. https://www.genome.gov/sequencingcosts/

However, the second major cost has not reduced in line with declining individual sequencing costs; this is the expense associated with gathering the phenotypes (the physical traits, such as diseases), which are identified and linked to specific genetic variations (Houle et al. 2010). The traditional approach would be to identify a cohort of people with a condition of interest, usually via medical records. A cohort of about 50 individuals is a minimum starting point, and even this can present difficult and expensive problems for researchers to obtain medical records and contact individuals and meet the inevitable red tape involved in consents and ethics approvals. The cohort is then compared to a control group without the condition, also requiring 50 participants at a minimum. The whole genome of each individual would be sequenced, so that sequencing costs alone would begin at about USD 500,000 without taking into account the work needed to locate potential participants or the subsequent analysis costs.

23andMe have recently demonstrated a different approach where self-reported medical information provided by individuals was used to screen for conditions and build genotype-phenotype maps (Tung et al. 2011). These maps were then used to test whether the researchers could replicate the results of other studies that had used more traditional genome-wide phenotyping approaches. The researchers successfully replicated about 70% of the study results they tested. More work is clearly needed, but this is an impressive feat when considering the cost and efficiency savings. The approach adopted by 23andMe used about 20,000 individuals. They were then able to use the self-reported medical data to phenotype 50 conditions at once instead of just one condition.

Using a traditional approach, it would take a small team about twelve months to contact participants, sequence genomes and analyse results. The 23andMe team also took about twelve months, but they were able to examine a much larger set of phenotypes. Put simply, in terms of the conditions that can be phenotyped within one year, the new methodology represents a productivity increase of 5,000%. However, this compares a technology at start-up with all its fixed and learning costs with a mature technology – the incumbent, traditional one. It would be possible to set up the 23andMe database to run such scans on an ongoing basis, and effectively identify patterns as they appear in the data. This takes us somewhere near an infinite increase in productivity!

Similarly, the new methodology offers to almost eliminate the costs of new knowledge of this kind. The cohort of 20,000 people cost USD 2 million to genotype, but now 600,000 customers of 23andMe have paid USD 99 to be genotyped and have donated their own phenotype by completing 23andMe surveys for the range of *private benefits* this brings to them, which includes tapping into existing knowledge about genotype-phenotype associations of interest to them. The public good of more scientific knowledge was thus a by-product of a private investment decision. Given this, we have assumed that the eleven authors of the published study of the work each worked part time on the project doing USD 20,000 worth of work on it at a total cost of USD 220,000. From this, each of the 50 genetic associations was generated at a cost of USD 4,000 or at 0.89% of the cost of the traditional method, a productivity improvement of 10,000%. Even if one were to assume that the cost of this work also involved the retail cost of USD 99 for performing all the genomic sequencing of the full 20,000 people involved in the study, the productivity improvement is still more than 1,000%.