OECD publishing

BUSINESS MODELS FOR SUSTAINABLE RESEARCH DATA REPOSITORIES

OECD SCIENCE, TECHNOLOGY AND INNOVATION
POLICY PAPERS
December 2017 **No. 47**

ICSU CODATA

OECD
BETTER POLICIES FOR BETTER LIVES

This paper was approved and declassified by the Committee for Scientific and Technological Policy (CSTP) on 24 October 2017 and prepared for publication by the OECD Secretariat.

Note to Delegations:

This document is also available on OLIS under the reference code:

DSTI/STP/GSF(2017)1/REV1/FINAL

This document, as well as any data and any map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

© OECD 2017

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for commercial use and translation rights should be submitted to *rights@oecd.org*.

FOREWORD

In 2007, the OECD Principles and Guidelines for Access to Research Data from Public Funding were published and in the intervening period there has been an increasing emphasis on open science. At the same time, the quantity and breadth of research data has massively expanded. So called "Big Data" is no longer limited to areas such as particle physics and astronomy, but is ubiquitous across almost all fields of research. This is generating exciting new opportunities, but also challenges. The promise of open research data is that they will not only accelerate scientific discovery and improve reproducibility, but they will also speed up innovation and improve citizen engagement with research. In short, they will benefit society as a whole.

However, for the benefits of open science and open research data to be realised, these data need to be carefully and sustainably managed so that they can be understood and used by both present and future generations of researchers. Data repositories - based in local and national research institutions and international bodies - are where the long-term stewardship of research data takes place and hence they are the foundation of open science. Yet good data stewardship is costly and research budgets are limited. So, the development of sustainable business models for research data repositories needs to be a high priority in all countries.

Surprisingly, perhaps, little systematic analysis has been done on income streams, costs, value propositions, and business models for data repositories, and that is the gap this report attempts to address, from a science policy perspective. The principal audiences for the report are science policy makers and those involved in making decisions on investment in data repositories in ministries, funding agencies, and institutions. It is also aimed at those involved in developing and implementing business models for repositories - data repository managers and all those involved with the stewardship of research data.

This work was requested by the OECD Global Science Forum (GSF) in October 2015. The project analysed a broad variety of data repositories and was conducted pursuant to the following task statement:

1. Identify and describe existing revenue sources as well as emerging and innovative business models, including methods of cost optimisation and reduction.

2. Test more rigorously possible business models with various stakeholders and against budgets and willingness and ability to pay.

3. On the basis of these findings, make policy recommendations to promote sustainable business models for research data repositories.

At the same time, GSF initiated another project on Coordination and Support of International Research Data Networks, which has produced a separate policy report. As the titles indicate these two reports have a different focus, i.e. individual repositories versus international networks, but they are complementary. After all, sustainable data repositories are the essential building block for international data networks.

This project has been carried out in partnership with the Committee on Data for Science and Technology of the International Council for Science (ICSU-CODATA). It also has benefitted enormously from strong links with the Research Data Alliance (RDA). A joint RDA - World Data System (WDS)

survey of data repositories has been incorporated into and expanded upon in this project. A two-day project workshop was held in Brussels in March, 2017 with generous support from the European Commission DG-RTD.

In addition to these organisations, many individuals from across the world have shared their knowledge and expertise to realise this project. An international Expert Group, co-chaired by Simon Hodson and Ingrid Dillo (see Appendix A for the full membership), has overseen the project and the writing of the report. Other experts from many different data repositories (see Appendices C and E) generously shared their experiences via telephone interviews and many of these and additional stakeholders participated in the two international workshops that were a critical part of the project. The report itself was drafted by Paul Uhlir and John Houghton, who also acted as economics consultant for the work and led the economic analysis. The work was supported by members of the OECD Secretariat - Giulia Ajmone-Marsan, Taro Matsubara, and Carthage Smith.

This publication is a contribution to the OECD Going Digital project, which aims to provide policymakers with the tools they need to help their economies and societies prosper in an increasingly digital and data-driven world.

For more information, visit www.oecd.org/going-digital

#GoingDigital

TABLE OF CONTENTS

FOREWORD....3
ABSTRACT....7
EXECUTIVE SUMMARY....8
Background and context....8
Research data repository business models....8
Policy recommendations....10
1. INTRODUCTION....14
Why is this important?....14
The value and benefits of research data repositories....15
Economic benefits....15
Other benefits....16
The potential vulnerability of research data repositories....17
Focus and methodology....18
Scope of the project....18
Organisation of this report....18
2. LANDSCAPE OF RESEARCH DATA REPOSITORIES....19
Scope and characteristics of the repositories....19
A typology of research data repository revenue sources....20
Revenue sources and expectations about their future adequacy....22
Alternative revenue sources....24
Costs and cost optimisation....24
3. REPOSITORY BUSINESS MODELS IN CONTEXT....27
Research data repository business models in context....27
An economic background to the issues....29
An economics primer....29
The problem (market failure)....30
Overcoming market failure....30
Mixed business models for data repositories....33
Costs, cost drivers, and scalability....34
Business models, cost constraint and optimisation....35
To what level should data repositories be funded?....35
What are the incentives for cost constraint?....36
4. SUSTAINABLE BUSINESS MODELS....37
Analysis of research data repository business models....37
Criteria for design and evaluation....37
Analysis of data repository business models....38
Structural Funding: Central funding or contract from a research or infrastructure funder that is longer-term, multi-year in nature....40
Hosting or institutional support: Direct or indirect support from the host institution....41
Data Deposit Fees: Annual contracts and per deposit....43
Access Charges: Charging for access to standard data or to value-added services....45
Membership model (privileged group, club, or consortium)....47
Contract for services and project funding: Services to other parties or research contracts....47
Business models combining various revenue sources....48

Business models and characteristic combinations of revenue sources....48
5. BUSINESS MODEL INNOVATION AND OPTIMISATION....50
Emerging business models....50
Deposit-side models....50
Supply-side and value-adding services models....51
Packaging models....51
Disruptive entry....51
User-side models....52
Opportunities for cost optimisation and reduction....52
Data repository costs and issues to consider in their analysis....53
Opportunities for cost optimisation....55
Technological approaches....55
Management and organisational approaches....56
Policy and legal approaches....56
6. CONCLUSIONS....58
ENDNOTES....60
APPENDIX A. OECD-GSF/ICSU-CODATA EXPERT GROUP....67
APPENDIX B. GLOSSARY OF KEY TERMS....68
APPENDIX C. LIST OF REPOSITORIES INTERVIEWED....70
APPENDIX D. A SWOT ANALYSIS OF FUNDING SOURCES....73
APPENDIX E. INVITED PARTICIPANTS AT PROJECT WORKSHOPS*....79

ABSTRACT

There is a large variety of repositories that are responsible for providing long term access to data that is used for research. As data volumes and the demands for more open access to this data increase, these repositories are coming under increasing financial pressures that can undermine their long-term sustainability. This report explores the income streams, costs, value propositions, and business models for 48 research data repositories. It includes a set of recommendations designed to provide a framework for developing sustainable business models and to assist policy makers and funders in supporting repositories with a balance of policy regulation and incentives.

Keywords:

Research, Data, Repositories, Sustainability, Business models, Open science, Open data.

EXECUTIVE SUMMARY

Background and context

Recognising the many scientific, economic, and social benefits of more open science, research policy makers and funders around the world are increasingly likely to prefer or mandate open data, and to require data management policies that call for the long-term stewardship of research data. At the same time, there are ever more data being created and used within research, and access to data is playing an increasingly central role in many research fields. Indeed, there are a number of fields of research that depend almost entirely upon the availability of global data sources provided through research data repositories.

As a result, repositories for the curation and sharing of research data have become a vital part of the research infrastructure. It is thus essential to ensure that these repositories are adequately and sustainably funded. However, relatively little work has been done to date on the revenue streams or business models that might provide ongoing support for research data repositories.

This project was designed to take up the challenge and to contribute to a better understanding of how research data repositories are funded, and what developments are occurring in their funding. Central questions included:

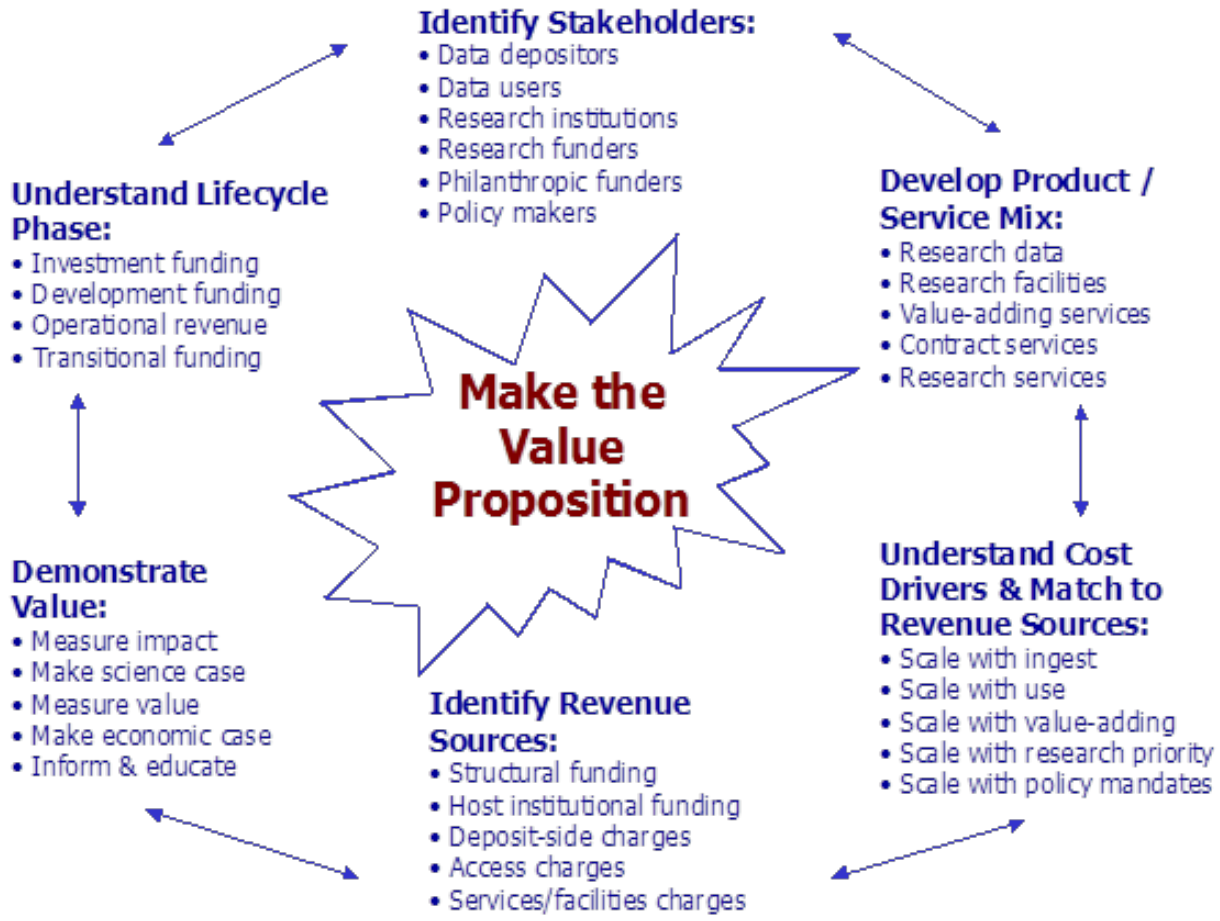
- How are data repositories currently funded, and what are the key revenue sources?
- What innovative revenue sources are available to data repositories?
- How do revenue sources fit together into sustainable business models?
- What incentives for, and means of, optimising costs are available?
- What revenue sources and business models are most acceptable to key stakeholders?

Forty-eight structured interviews were undertaken with repository managers from 18 countries and a broad range of research domains. They provided insights into key issues, which were further elaborated in two international workshops involving a variety of stakeholders - including repository managers, funders, and policy analysts.

Research data repository business models

A business model is “A plan for the successful operation of a business, identifying sources of revenue, the intended customer [stakeholder] base, products, and details of financing”[1]

Figure ES.1. Elements of a research data repository business model



Source: Authors' analysis.

The design and sustainability of research data repository business models depend on many factors, including: the role of the repository, national and domain contexts, the stage of the repository's development or lifecycle phase, the characteristics of the user community, and the type of data product this community requires (influencing the level of investment required in curating and enhancing the data).

All of these issues must be considered in choosing and developing appropriate business models for research data repositories, and revisited regularly throughout a repository's lifecycle. There is certainly no "one size fits all" solution.

The 47 data repositories analysed reported 95 revenue sources. Typically, repository business models combine structural or host funding with various forms of research and other contract-for-services funding, or funding from charges for access to related value-added services or facilities. A second popular combination is deposit-side funding combined with a mix of structural or host institutional funding, or with revenue from the provision of research, value-added, and other services.

Incentives, in the form of structural or institutional funding, or funding to support the payment of deposit-side fees, provide a foundation for sustainable business models for research data repositories. Regulation, in the form of policy mandates for open data, limits the potential for user-side funding models

and provides a foundation for deposit-side models. As data preservation and open data policies become increasingly widespread and influential, there will be more opportunities to develop deposit-side business models.

When repository activities grow in scale there will be more opportunities to optimise the "make versus buy" decision, through sourcing from specialist service providers, and we can expect to see more data repository business models based on supply-side services targeting data depositors and repositories, and user-side, value-adding services of various sorts.

Research data repositories themselves can take advantage of the underlying economic differences between research data, which exhibit public good characteristics, and value-adding services and facilities, which typically do not, to develop business models that support free and open data while charging some or all users for access to value-adding services or related facilities.

What someone is willing to pay for something depends on the perception of its value. For direct users of repository data and services, the value is clear to them and is revealed in their use. For stakeholders who are not direct users, but may be funders, it is more difficult for them to judge the value. Engaging and maintaining structural, institutional, philanthropic, or other funders depends on their understanding of the value proposition, and ensuring such engagement may involve repositories undertaking detailed benefit/cost, value, and impact analyses.

Currently, many research data repositories are largely dependent on public funding. A key policy question to be addressed is how this public funding is most effectively provided - by what mechanism and from what agency, ministry, or institution. In this context, it should be recognised that the value proposition is likely to be different for different public-sector stakeholders.

Policy recommendations

The following policy recommendations primarily target science policymakers and funders in OECD member states, as well as repository operators and managers.

Recommendation 1:
All stakeholders should recognise that research data repositories are an essential part of the infrastructure for open science.

Research data repositories provide for the long-term stewardship of research data, thus enabling verification of findings and the re-use of data. They bring considerable economic, scientific, and social benefits. Hence, it is important to ensure the sustainability of research data repositories.

Sustainability depends, inter alia, on a clearly articulated value proposition and the development of a "business model" (See Recommendation 2).

- Policy makers and research funders should take a strategic view of the data landscape and seek to ensure the appropriate provision of repositories. They can do this by ensuring that the researchers they fund have access to suitable and sustainable research data infrastructure, so that the research community can meet expectations for data preservation and sharing, and comply with open data mandates.

- Research data repository operators and managers need to study and understand the value proposition of their repositories, and clearly articulate it for all stakeholders in the research system.

- Research data repository operators and managers should continually review their business model as a repository evolves, and revise it accordingly.

Recommendation 2:
All research data repositories should have a clearly articulated business model.

Actions needed to develop and maintain a successful business model include (Figure ES.1):

- *Understanding the lifecycle phase of the repository's development* (e.g. the need for investment funding, development funding, ongoing operational funding, or transitional funding).

- *Developing the product/service mix* (e.g. basic data, value-added data, value-added services and related facilities, or contract and research services).

- *Understanding the cost drivers and matching revenue sources* (e.g. scaling with demand for data ingest, data use, the development and provision of value-adding services or related facilities, research priorities, and policy mandates).

- *Identifying revenue sources* (e.g. structural funding, host institutional funding, deposit-side charges, access charges, and value-added services or facilities charges; *Identifying who the stakeholders are* (e.g. data depositors, data users, research institutions, research funders, policy makers).

- *Making the value proposition* to stakeholders (e.g. measuring impacts and making the research case, measuring value and making the economic case, informing, and educating.

Because the context is dynamic, these actions should be revisited regularly throughout a data repository's lifecycle.

Recommendation 3:
Policy makers, research funders, and other stakeholders need to consider the ways in which data repositories are funded, and the advantages and disadvantages of various business models in different circumstances.

It is important to consider what system of allocation will best ensure that the optimal level of funding will be made available for research data repositories. For example:

- *Structural funding* typically involves a trade-off between funding for data repositories and funding for other research infrastructure or for research itself. That allocation will best be made by informed actors making choices, such as through a funding allocation process involving widespread research stakeholder participation, expert consultation, and “road-mapping”.

- Funding models depending on *deposit or access fees* bring the trade-off closer to the researchers, but their success in optimising allocation will depend on the extent to which the actors are

informed and on their freedom of choice. The latter may be constrained by open data mandates (regulation).

- *Host institutional funding* may divorce informed actors from the funding decisions or require additional processes to ensure greater stakeholder understanding of the value of the repository services.

Project funding often provides a mechanism to test the need for a data repository and the initial capacity to create one. However, as the repository matures and scales to provide an ongoing, reliable and quality service, a different funding model is likely to be needed.

- From an economic perspective, this is the distinction between *investment funding* to establish a business, and an *ongoing revenue source* during the operational phase.

- This distinction is not yet well made in the research data repository environment, but should form an important part in the design and evolution of repository business models.

Research data repository *costs* will change over time. As the global data repository infrastructure evolves there will be increasing learning and scale economies, which have the potential to reduce repository costs, although this needs to be balanced against increased data flows.

- Consequently, policy makers and funders should be wary of allocating a fixed percentage of research funding for research data repository infrastructure, as it would be very difficult to establish the appropriate level and very difficult to change it once established.

- The allocation of funds is likely to be better made when left up to those closest to their application (e.g. allocating funding to science and letting researchers and research managers meet open data requirements as best suits their needs).

Recommendation 4:
Research data repository business models are constrained by, and need to be aligned with, policy regulation (mandates) and incentives (including funding).

- Policy makers should be cautious of "un-funded mandates". They should combine regulation and incentives thoughtfully to achieve best results.

- Some business models depend on willingness to fund the repository in recognition of a strong value proposition (e.g. structural or host funding). Other business models are heavily dependent on strong policy incentives and regulation (e.g. deposit-side charges). Still other business models may limit data re-use and reduce the overall benefits that could be derived from research data curation and sharing (e.g. access-side charges).

- A key issue is matching funding and revenue sources to the main cost drivers, to ensure that revenue scales with demand and repository costs. These cost drivers can relate to the level of activities (e.g. deposits, access, and use), and/or to the level of curation (e.g. basic versus enhanced).

Recommendation 5:
In the context of financial sustainability, opportunities for cost optimisation should be explored in order to be able to effectively manage digital assets over time.

Therefore, policy makers, research funders, and repository managers should:

- Obtain greater clarity concerning costs, in order to fully understand and manage them.
- Consider cost optimisation system-wide (throughout the whole data lifecycle), rather than simply focus on cost savings at the repository level, as there is a risk that repository cost saving may only lead to cost shifting and/or a reduction in overall access to data.
- Consider the effect a funding model has on cost constraints, as the more a funding model depends on or creates low price elasticity of demand, the lower the incentive for cost constraints will be.
- Monitor the research landscape for emerging opportunities. As data repository activities grow and develop, there will be increasing opportunities to buy services from specialist providers, potentially enabling greater cost optimisation.
- Take advantage of economies of scale. For example:
  - By encouraging or funding the establishment of lead organisations for open research data at the national level, and encouraging those organisations to collaborate globally.

By encouraging or funding collaboration and federation. Not all research data repositories need to perform specialised curation and preservation tasks. Similarly, not all institutions or organisations need to create individual repositories. Collaboration and federation can help to manage and reduce costs.

1. INTRODUCTION

Why is this important?

There are many reasons for writing this report. The data deluge is enormous and growing both continuously and rapidly. Data have become a new currency of the global economy and of the research and innovation process. Researchers have developed different ways of managing and using factual information, leading to entirely unprecedented areas of data-driven inquiry and "data science" (Hey et al., 2009). These and many other reasons given throughout this report provide evidence of the importance of preserving research data and making them openly available.

Over the centuries, libraries, archives, and museums have shown the practical and policy advantages of preserving sources of knowledge for society. Research and other types of data constitute a relatively new subject that requires our serious attention. Although some research data repositories were founded in the 1960s and even earlier, the data that are now being generated have resulted in the establishment of many new repositories and related infrastructure. Societies need such repositories to ensure that the most useful or unique data are preserved over the long term.

At the same time, governments and others are struggling to keep up with the demand to help support the new repositories and those that already exist. Groups from all walks of life, but especially from the research community, have called for a better understanding of the stakeholders and for economic analyses of research data repositories and how they are funded (Ember and Hanisch, 2013; BRTF, 2010).

With the volume and variety of data increasing, and many of the budgets to manage these data unable to keep pace, investments in digital curation must be strategic and targeted to ensure the best value for money. Transparency of digital curation costs will help data repositories identify greater efficiencies and pinpoint potential optimisations. Insight into how and why peers target their investments can lead to the better use of resources, help identify weaknesses and drivers in current practices, and inspire innovations (Baker, 2012).

There is thus clearly a need to improve our understanding of research data repository costs, and where and how costs may be restrained. However, for sustainability, it is also important to explore alternative cost recovery options and a diversification of revenue sources (Berman and Cerf, 2013; Baker, 2012). This project contributes to strategic thinking in these areas, and aims to help develop an understanding of current and possible future revenue sources and business models for research data repositories.

Relatively little analysis has been done on this topic thus far. Notable, previous work has included the U.S. National Science Foundation's Blue Ribbon Taskforce (BRTF) report on *Sustainable Digital Preservation* (2010), which explored the need to understand the value proposition for research communities, and a report on *Sustaining Domain Repositories for Digital Data* (Ember and Hanisch, 2013), which identified the need to understand how research data repositories for scientific domains are funded. A similar study, *Funding Models for Open Access Repositories* (Kitchin et al., 2015), was recently published for the European context. This project builds on this previous work to contribute to a better

understanding of how research data repositories are funded, and what new developments are occurring in this regard.

The value and benefits of research data repositories

There are many benefits to preserving and making research data openly available. There has been a steady move towards openness in research over the past two decades that has accelerated in the last few years, with some significant changes in the global policy environment in which research is conducted (Bicarregui, 2016). National and international funders of research are increasingly likely to mandate open data and demand data management policies that call for the long-term stewardship of research data.

This trend is consistent with many recommendations and reports by both the members of the OECD and the organisation itself.[2] In fact, in many countries research data are increasingly viewed as an essential part of the research infrastructure. At the same time, adapting research practices to fully implement new open data policy requirements will require a reinvigorated or new infrastructure to support it (Bicarregui, 2016).

Even under the best of circumstances, however, in which governments have a default rule of open data, not all research data can or should be made broadly available. Data may be subject to various restrictions, such as the protection of personal privacy, national security, proprietary concerns, or other forms of confidentiality, complicating the decisions to save them and then to make them available. Many datasets are not of requisite quality, are not adequately documented or organised, or are of insufficient (or no) interest for use by others.

Nonetheless, there are various reasons, summarised below, why data that are generated in or for research should follow a default rule of openness. Open data and their organisation and curation in research data repositories can generate multiple benefits (Uhlir, 2006).

Economic benefits

There has been an increasing awareness across governments of the key role that government data or public-sector information (PSI) plays in supporting the goals of research and innovation, specifically, and in economic terms more broadly. Among the studies looking at the value of PSI and its current or potential wider economic impacts, there are a few canonical reports that have gained widespread attention (PIRA, 2000; Weiss, 2001; Dekkers et al., 2006; DotEcon, 2006; Pollock, 2009; Vickery, 2010). While these and other studies document the economic benefits derived from PSI, including open data produced or used in research, they are generally beyond the immediate scope of this report.

With regard to research data repositories, a recent series of UK-based studies combined qualitative and quantitative approaches to measure the value and impact of research data curation and sharing (Beagrie et al., 2012; Beagrie and Houghton, 2016, 2014, 2013a, 2013b). These studies have covered a wide range of research fields and practices, looking at the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS), the British Atmospheric Data Centre (BADC), and the European Bioinformatics Institute (EBI).

Two outcomes from these studies stand out. First, there are substantial and positive efficiency impacts, not only reducing the cost of conducting research, but also enabling more research to be done, to the benefit of researchers, research organisations, their funders, and society more widely. Second, there is

substantial additional reuse of the stored data, with between 44% and 58% of surveyed users across the studies saying they could neither have created the data for themselves nor obtained them elsewhere.

While these studies tend to provide a snapshot of the repository's value, which can be affected by the scale, age and prominence of the data repository concerned, it is important to note that in most cases, data archives are appreciating rather than depreciating assets. Most of the economic impact is cumulative and it grows in value over time, whereas most infrastructure (such as ships or buildings) has a declining value as it ages. Like libraries, data collections become more valuable as they grow and the longer one invests in them, provided that the data remain accessible, usable, and used.

The users of these data repositories come from all sectors and all fields – close to 20% of respondents to the ESDS and EBI user surveys 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 research sector alone.

Both the scale and extent of these impacts are reflected in citation analyses. Such analyses show widespread and increasing dataset and repository citation spanning a number of years in both academic publications and patent applications, also attesting to both research and industry use (Bousfield et al., 2016).

Other benefits

There are many other values, in addition to the economic benefits, that are promoted through the long-term stewardship and open availability of research data. These include better research, enhanced educational opportunities, and improved governance (CODATA, 2015).

Among the most important benefits are for research itself - both enabling new research and reproduction of completed research. A fundamental principle for research quality and integrity is the ability of others to verify the results by checking the data used to derive the research findings and to avoid fraud. The underlying data need to be broadly available for verification and reproducibility of results, sometimes even many years after their publication (Doorn, 2013; NRC, 2009b, 2004, 2003, 1999, 1997).

Interdisciplinary and international research, including participation of less-developed countries, can be enhanced. Much research now is data intensive and access to many different kinds of data is an essential part of the research process (Hey et al., 2009). Well-curated and open data allow unhindered data mining and automated knowledge discovery; that is, to have machines find, extract, combine, and disseminate the data with minimal or no human intervention (NRC, 2012a). The rapidly expanding area of artificial intelligence (AI) relies to a great extent on saved data. Open data also permit legal interoperability, which is necessary for the generation of new datasets (RDA-CODATA, 2016).

Downstream applications and commercial innovation are stimulated by open access to upstream data resources, leading to the creation of new wealth from research, some of which was documented in the preceding section. However, new opportunities are also emerging through collaborative innovation based on data access and "data driven innovation for growth and well-being".[3] The beneficial economic effects of open data extend to the research process itself by reducing inefficiencies, especially in the avoidance of research duplication (CODATA, 2015).

Furthermore, new types of research are promoted that are important in their own right and can lead also to serendipitous breakthroughs (Arzberger et al., 2004). For example, the collection and open sharing of all kinds of data are fundamental to the rise of citizen science and crowd sourcing approaches, with the data made available through public repositories (Benkler, 2006; Uhlir, 2006). Such approaches have been adopted in many domains, including the space sciences,[4] ornithology (Lauro, 2014; Robbins, 2013), environmental studies,[5] and even search and rescue missions (Barrington, 2014). Moreover, greater openness supports the reputational benefits of those who compiled the datasets by making such information more broadly available and generally democratising research (CODATA, 2015).

Education, across all ages and disciplines, can be enhanced by access to data from open repositories. At the secondary and even primary education levels students can use open data repositories to further their scientific understanding and skills.[6] University students need open data to experiment with or to learn the latest data management techniques (CODATA, 2015). And of particular interest to government policymakers, data management and curation skills, which require a good educational foundation, are a growth area for employment in an era of shrinking job opportunities (NRC, 2015).

Finally, the role of open repositories of research data in supporting good governance should not be overlooked. Openness of public information strengthens freedom and democratic institutions by empowering citizens, and supporting transparency of political decision-making and trust in governance. It is no coincidence that the most repressive regimes have the most secretive institutions and activities (Uhlir, 2004). Open factual datasets also enhance public decision-making from the national to the local levels (Nelson, 2011), and open data policies demonstrate confidence of leadership and generally can broaden the influence of governments (Uhlir and Schröder, 2007). Countries that may be lagging behind socio-economically frequently can benefit even more from access to public data resources (NRC, 2012b, 2002).

The potential vulnerability of research data repositories

Despite these many benefits that accrue from preserving datasets and making them openly available through research data repositories, there are significant forces that work to restrict investments in such repositories and in the broader data infrastructure. With the volume and variety of data increasing rapidly, budgets for data stewardship can struggle to keep pace despite falling storage costs.

Researchers themselves sometimes have reasons not to make their data openly available through data repositories. Some also eschew using “other peoples’ data”, preferring to generate their own data in the course of their research. Such attitudes, more prevalent or justified in some research domains than in others, can negatively affect the potential user base and make the establishment of research data repositories less worthwhile.

There also may be substantial financial or other pressures against saving and making available datasets through research data repositories. The budgets of most governments are constrained and ministries are looking for ways to save money. Moreover, the research community itself often resists calls to redirect funds from research to infrastructure services, when budgets for research itself are flat or declining. Researchers everywhere want to maximise the amount of the grant that is spent on actual research, and the infrastructure to support future research (often by others) is not yet seen as an essential part of that. There is thus a tension between short-term research considerations and longer-term and overarching research infrastructure, including data repositories.

The reality is that research institutions frequently have no idea how many datasets are held in ad hoc systems or how they are preserved. Many if not most of the digital data created or used in research over the last century have been lost because no long-term repository or other safeguards existed.[7] There have been

several recent instances of terminated repositories, or of data centres that have moved from an open to a closed or partially closed subscription business model in order to survive.

Focus and methodology

This project built on the recent work of a Research Data Alliance and World Data System (RDA-WDS) Working Group (the RDA-WDS Working Group), which was published in March 2016 (RDA-WDS, 2016). The RDA-WDS project and the current study of the OECD's Global Science Forum (GSF), which has been carried out in cooperation with the Committee on Data (CODATA) of the International Council for Science (ICSU, the GSF-CODATA study), each used in-depth interviews with managers of research data repositories in different research domains. The interviews focused on identifying existing approaches to cost recovery, the range of revenue sources available, and current and potential business models. The interviews were held in person or telephonically, and were guided by a structured questionnaire.[8]

This empirical work was then followed by two workshops of the Expert Group and other invited experts (Appendix E), as well as by extensive analysis and further research.

Scope of the project

Data repositories take many forms. For example, in 2005, the U.S. National Science Board (NSB) identified three types of collections of research data: (i) simple research data collections, which are the products of one or more research projects and usually having limited curation; (ii) resource or community data collections that serve one research community; and (iii) reference data collections, intended to serve large segments of the scientific and education community (NSB, 2005). We chose to focus our inquiry mainly on the more robust types of research data collections defined in second and third type of collection described by the NSB. Selection criteria included geographical spread, disciplinary coverage and funding model. The sample included generic institutional repositories and domain specific repositories, as well as a small number of private not-for-profit and for-profit entities (see Appendix C).

Organisation of this report

Chapter Two presents the main results from our in-depth interviews and a short description of the principal types of revenue sources used by research data repositories.

Chapter Three outlines the analytical context, providing an overview of the contextual and structural influences on the selection and development of data repository business models, and an economic background to subsequent analysis.

Chapter Four looks at the various business models used by research data repositories. The focus is an economic analysis of the business models identified.

Chapter Five explores the opportunities for, and dynamics of, innovation in repository business models, and at the opportunities for cost constraint and cost optimisation.

Finally, Chapter Six presents some of the key conclusions arising from this study. The recommendations arising from the study are presented in the Executive Summary, and are not repeated in the body of the report.

2. LANDSCAPE OF RESEARCH DATA REPOSITORIES

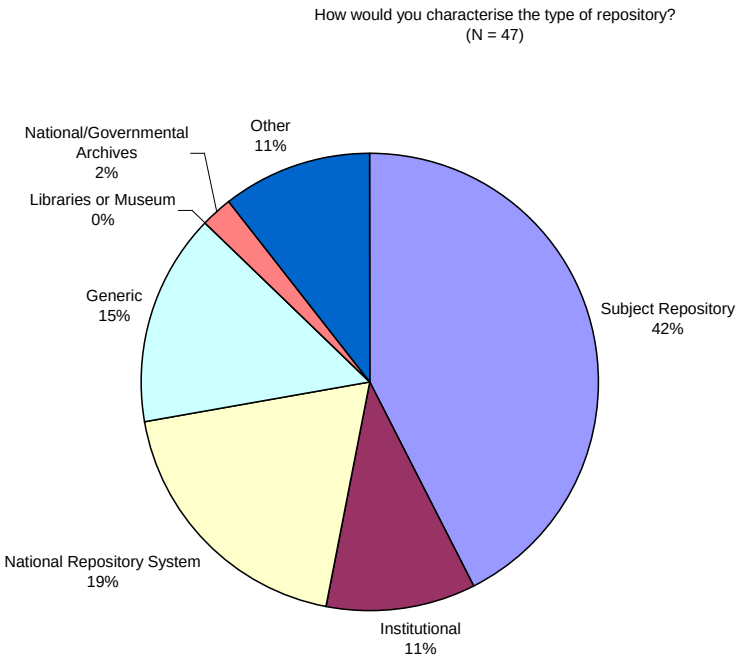
This chapter explores the research data repository landscape, based on the interviews undertaken for this study and the earlier RDA-WDS project. It looks at the scope and characteristics of the repositories interviewed, and examines research data repository revenue sources and expectations about their future adequacy, possible alternative revenue sources, and costs and cost optimisation. These themes are picked up and discussed more fully in Chapters Four and Five.

Scope and characteristics of the repositories

In selecting research data repositories for inclusion in this study, the aim was to achieve both geographical and disciplinary spread, and to include a balance of repository types. The sample selection was carried out by international experts, but should not be considered as statistically representative of the whole global landscape of data repositories.

Forty-eight interviews were undertaken with repository managers from 18 countries. Around half of the repositories interviewed focus primarily on the natural sciences, around one-third reported a mixed or multi-disciplinary focus, and close to 15% reported a focus on the social sciences and humanities. Around 40% are subject-matter repositories, 20% national repositories, 15% generic repositories, and 10% institutional repositories (Figure 1 and Appendix C).

Figure 1. Characteristics and focus of the repositories (N=47)

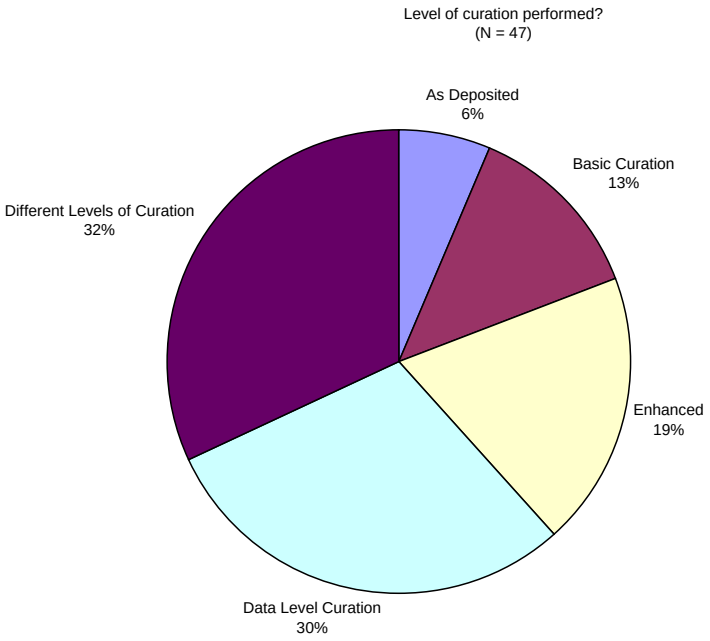


Source: Authors' analysis.

More than three-quarters of the repository managers interviewed reported having some latitude to determine the repository's mission and collection policy, with just 13% reporting that they had no such latitude.

The majority of repository managers reported undertaking relatively high levels of data curation. Half reported undertaking enhanced or data-level curation, around one-third reported undertaking different levels of curation, and just one-fifth reported basic level curation or making the data available as deposited (Figure 2).

Figure 2. Levels of curation performed by the repositories (N=47)



Source: Authors' analysis.

A typology of research data repository revenue sources

The survey of repositories undertaken for this and the previous RDA-WDS study classified the principal research data repository revenue sources as follows:

- *Structural funding* (i.e. central funding or contract from a research or infrastructure funder that is in the form of a longer-term, multi-year contract). We use the term “structural” to underline the difference between this and project funding. The research data repository is considered as a form of research infrastructure or as providing an ongoing service. Although the funding may be regularly reviewed, it is a form of funding that is substantively different to project funding.

- *Host institution funding and support* (i.e. direct or indirect support from a host institution). Some research data repositories are hosted by a research performing institution, e.g. a university, and receive direct funding or indirect (but costed) support from their host.

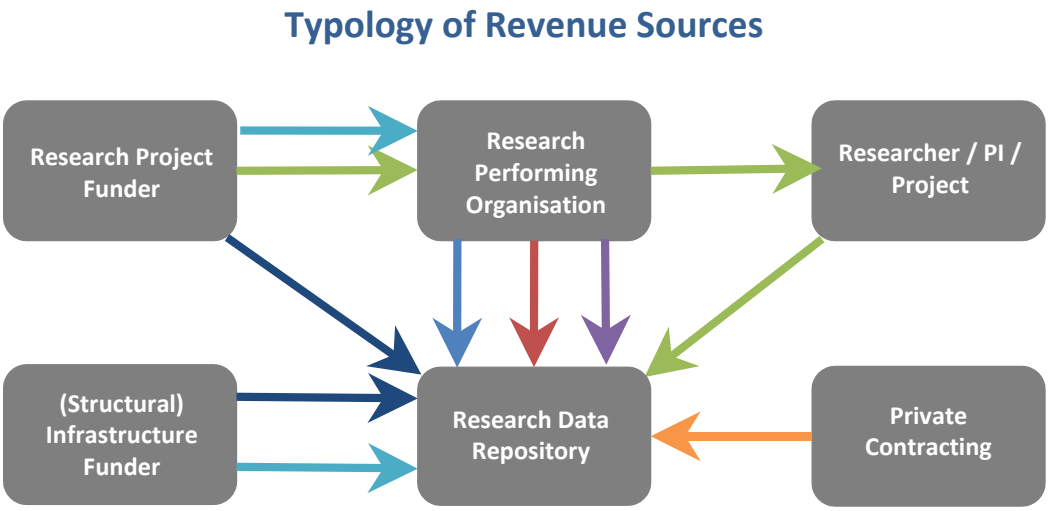
- *Data deposit fees* (i.e. in the form of annual contracts with depositing institutions or per-deposit fees). As indicated, this can take the form of a period contract or a charge per deposit. In either case, the cost is borne by the entity that wishes to ensure that the data are preserved and curated for the long term.

- *Access charges* (i.e. charging for access to standard data or to value-added services and facilities). This covers charges of various sorts (e.g. contract or per-access charges) and can be levied either for standard data or value-added services. In all cases, the cost is borne by the entity that wishes to access and use the data.

- *Contract services or project funding* (i.e. charges for contract services to other parties or for research contracts). This covers short-term contracts and projects for various activities not covered above (i.e. these are not contracts to deposit or access data, but cover other services that may be provided). Similarly, this category of funding is distinct from structural funding because, although it may come from a research or infrastructure funder, it is for specific, time- and objective-limited projects, rather than for ongoing services or infrastructure.

For research data repositories, all of these revenue sources can, and often do, come from the public purse and, directly or indirectly, out of research and higher education budgets. Thus, at a national research system level, the meta-question is how best to allocate this funding, including which entities should have responsibility for it and what selection mechanisms should they use?

Figure 3. A typology of research data repository revenue sources



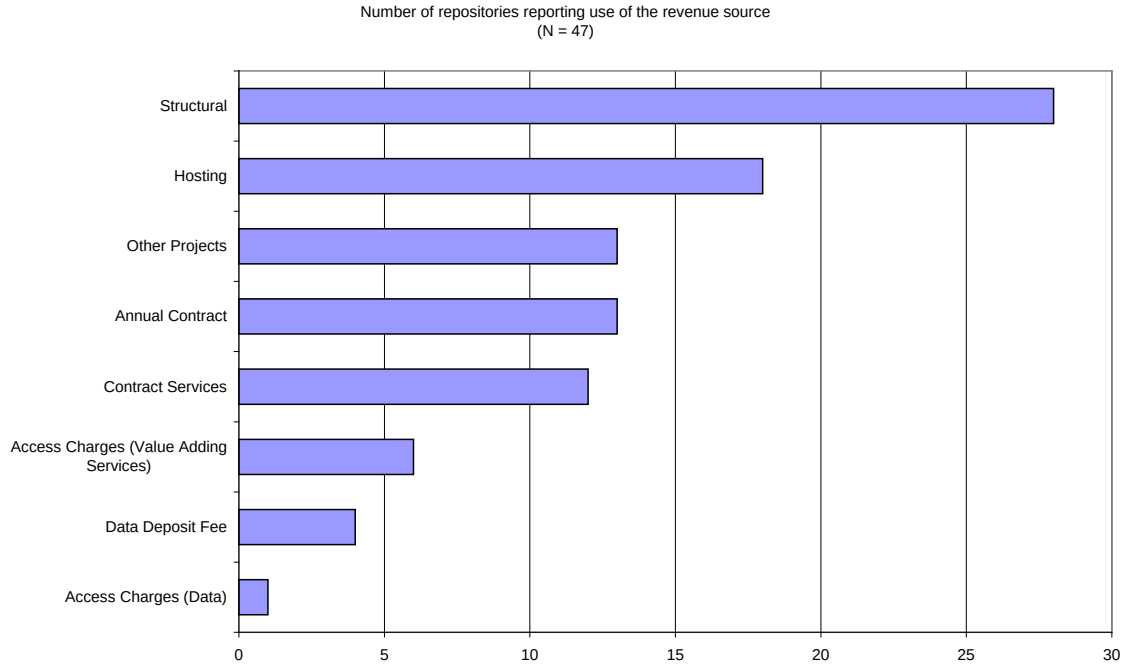
Source: http://bit.ly/revenue_source_diagram

Revenue sources and expectations about their future adequacy

The 47 data repositories analysed reported 95 revenue sources, an average of two per repository. Twenty-four repositories reported funding from more than one source, and seven reported more than three revenue sources. Combining revenue sources is an important element in developing a sustainable research data infrastructure.

Twenty-eight of the repositories interviewed reported some dependence on structural funding, 25 reported some dependence on funding from contract or other services, 18 reported some host or institutional support, 17 reported some dependence on deposit-side charges, and seven reported some dependence on access charges (i.e. for data or value-added services) (Figure 4).

Figure 4. The number of repositories using these revenue sources



Source: Authors' analysis.

Around 60% of the repository managers interviewed expected these revenue streams to remain stable, over the near future (e.g. five years). Asked if they expected these revenue sources to be sufficient for the tasks the repository will need to perform in the future, around half of the repository managers said they did.

Nevertheless, there was some sense of funding constraint. More than 80% of the repository managers interviewed said there were activities that the repository would like to be doing, but cannot because current revenue sources do not provide sufficient funding. A key factor was keeping pace with growing demand. Comments included:

> *Could be doing more in terms of both quality and volume with more funding, and we are experiencing increasing demand... because data will grow much larger and faster.*
>
> *As data become more complex and larger in scale, we will need additional resources to be able handle these. Also, if funders and journals... increasingly require deposit of data at repositories... we will need additional resources.*
>
> *There is already too little funding to do what is essential. There are insufficient funds to build the necessary capacity to curate all the available data properly. The infrastructure is aging and there appears to be little will to upgrade or even maintain the infrastructure.*
>
> *We need support for maintaining the repository and scaling-up activities.*

> *Increasing IT-costs are not automatically covered, which creates a significant uncertainty and risks.*

Alternative revenue sources

Three-quarters of the repository managers interviewed reported that they are exploring alternative revenue source.

A large majority (more than 80%), said they would *not* be considering any revenue sources that are incompatible with the open data principle. This is comparable with the number reporting that their repository does not currently have a revenue source that is incompatible with the open data principle.

Asked to explain what alternative revenue sources were being considered and why, responses suggested they are exploring a range of deposit-side charges, charges for value-added services, and philanthropic funding. Comments included:

> *For greater funding diversity and risk management, and to expand what we can do.*
>
> *Universities might suggest to funders to increase the overhead on research grants to cover additional infrastructure and service costs related to research data. As funders increasingly accept applications for funding of data management tasks, there might be more willingness on [the] researchers' side to cover (part of) the costs of data management from their grants. Third-party users could (also) provide additional income which would help to share certain costs.*
>
> *Looking at value-added services, such as offering Dropbox-style data management services that would be more oriented to desktop use. Also interested in perhaps offering computational services – possibilities around analysis of web archives – developing and providing tools around that.*
>
> *Primarily support from foundations beyond the NSF. Also considering fees for value-added services.*
>
> *Due to the... gap between science and data management demands we are about to test donations as [an] alternative income stream.*

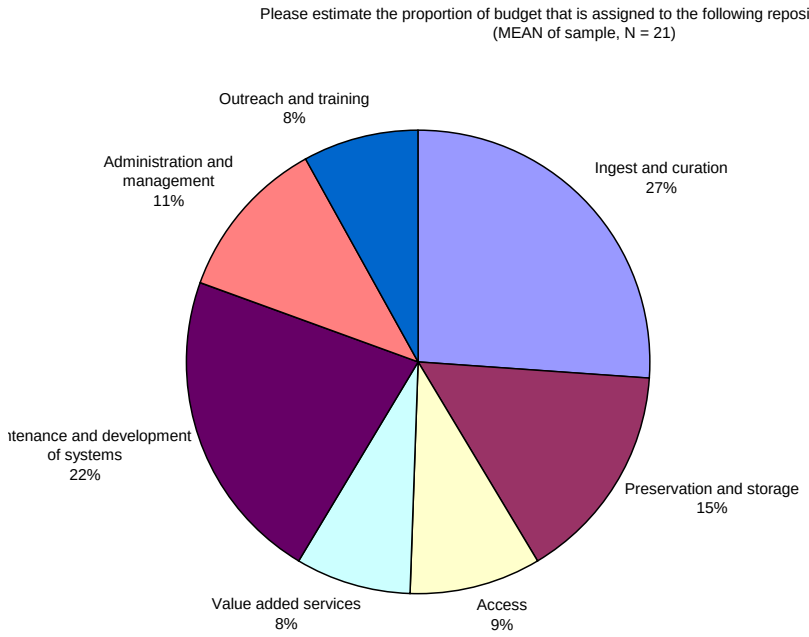
Costs and cost optimisation

Repository costs and expectations for costs savings and cost optimisation were not a part of the original RDA-WDS study. Consequently, there were only 25 repository interviews covering these topics.

Among the 16 repository managers reporting their overall operational budget, the mean expenditure was around USD 4 million per annum.

Across the 21 repositories reporting a breakdown of their operating costs, data ingest and curation accounted for the largest share (27%), followed by the maintenance and development of systems (22%), preservation and storage (15%), administration and management (11%), providing access (9%), and value-added services, and outreach and training (8% each) (Figure 5).

Figure 5. Proportion of repository budget allocations by activity (N=21)



Source: Authors' analysis.

Asked which of the repository activities listed above was most likely to be susceptible to cost optimisation in the near future, answers focused on technological developments and various forms of automation, as well as learning economies (i.e. efficiencies gained through experience), and collaboration and shared services. Comments included:

> *Ingest and curation by depositors. [And] Software development may get more voluntary contributions from the community (a Linux model).*
>
> *Automation on ingest, curation and preservation, and access automation.*
>
> *Labour is likely the next area for optimization (efficiency, effectiveness, and new value to be realised as work becomes more routine and staff and users move up the learning curve).*
>
> *The ingest part will probably benefit from the introduction of the new interface (intermediate repository), and the maintenance part benefits from the growing maturity of the solutions in place.*
>
> *Technology side – storage costs. Staffing side is very hard to optimise. Collaboration can build pools of expertise to draw on.*
>
> *Important aspect is shared services for benefits of projects, not reinventing the wheel, rebuilding from scratch. Shared services for administration and management: national bodies... who can run these sorts of services can take away significant cost, duplication, and inefficiency.*

Asked to what extent they thought cost optimisation can play a significant role in helping a repository achieve a sustainable business model, almost 90% said that it could play a role. When this was explored further to understand how cost optimisation may or may not play a role, repository managers offered a range of insights:

> *Tools and curation – looking to help depositors do more of the work for themselves as the tools we are developing make it easier for them to do this.*
>
> *Focusing on ingest and curation costs, looking for savings/efficiencies. May also be potential for data access mechanisms to be more streamlined to reduce staffing costs associated with access.*
>
> *We are still learning from new use cases and often manual interventions are necessary. We hope that we will be able to increase the level of automation of more processes... and in some cases even self-serviced workflows.*
>
> *Exploring partnership... to find more efficient ways of doing curation; some outsourcing of curation to post docs; have already migrated to Amazon web services, savings through using virtual machines, etc. Thinking about software development practices to make them more sustainable and efficient to maintain the code.*
>
> *One of the goals... is to automate key processes. This includes: identification of datasets, sharing them internally for risk assessments, tracking approvals and changes, and finally publishing.*
>
> *Reducing time of data cleaning, automation and self/institutional uploading of datasets from the depositors may reduce the costs.*
>
> *As storage becomes cheaper will be able to hold more data. Long-term preservation will become more viable. But this is balanced by staffing requirements that grow.*
>
> *Information technology is progressive enough to maintain or step up the services at relatively lower cost. Price of cloud computing is falling.*
>
> *We have been improving efficiency... to make our systems more and more automated, reducing human costs. Moving to cloud platforms offers another possible efficiency that might be realized. With an increasing amount of data that is not offset with increasing base support our current model is not sustainable for another 10 years.*

These themes are picked up and discussed more fully in Chapter 5.

3. REPOSITORY BUSINESS MODELS IN CONTEXT

This chapter outlines some of the important elements of the context in which research data repositories operate. It provides an overview of the contextual and structural influences on the selection and development of data repository business models, an economic background to subsequent analysis, and an exploration of the drivers of innovation in business models, as well as the drivers of repository costs and incentives for cost optimisation.

Research data repository business models in context

The sustainability of various research data repository business models depends, *inter alia*, on the role, context, type of data and repository, and stage of its development.

The role that data repositories play in scientific and scholarly communication and how that role is evolving differs between fields of research, regions, and countries. In some fields, research is inherently global and the core data and tools form the basis for research, requiring data repositories to be large-scale and internationally supported (e.g. bioinformatics). In other fields, research data collections may be more local and can be operated at a smaller scale (e.g. some fields of the humanities and social sciences work almost entirely with national data). These domain differences have profound implications for the design and sustainability of repository business models.

Closely related to the data repositories' role are data collection types. Research data collections span a wide spectrum of activities from focused collections for an individual research project at one end of the scale, to reference collections with global user populations and impact at the other. Along the continuum in between are intermediate collections, such as those derived from a specific facility or centre (NSB, 2005: 9). The latter may be well matched with an institutional repository with host institutional funding, as might individual project collections. In contrast, reference collections with global user populations are likely to require multinational structural funding (e.g. from research, infrastructure, and philanthropic funders).

Countries and regions vary greatly in their structure and organisation of research itself and research funding. In some there are long-standing subject-based research funding councils and specialist research infrastructure funding councils or mechanisms, while in others there is much more diversity in funding sources and there may be no special focus of funding support for research infrastructure. The split between capital and operational costs and responsibilities is important in many countries, but much more so in some than in others. These contexts and practices of research and research infrastructure funding will affect the funding options available to repositories, and what is a sustainable business model in one country may not be viable elsewhere.

Countries and regions also vary in their scholarly publishing practices, implying differences in data repositories if the data are to seamlessly support publication. For example, the SciELO Open Access publishing platform, which originated in Brazil but has now been taken up in a number of other countries,[9] is quite different to publishing practices in, for example, the United Kingdom (e.g. a "Gold Preferred" publishing model). Hence, while a repository charging deposit fees for the data supporting publications may seem a natural extension of charging article processing fees for publishing in some countries and disciplines, it may not be so elsewhere.

There is also a continuum of perspectives, with some seeing data repositories as a publishing function and others as a part of research infrastructure, with data repository functions often closely linked to core research infrastructure and equipment, such as at CERN and at various synchrotron facilities (e.g. ESRF in France or the Diamond Archive in the UK). Hence, there may be cases in which a repository business model that replicates publishing is seen as natural (e.g. deposit fees or contract subscriptions with depositing and/or using institutions), but there may be other cases where a business model that replicates research infrastructure funding may seem more natural (e.g. structural funding, such as that through the United Kingdom's Science and Technology Facilities Council (STFC), or Australia's National Collaborative Research Infrastructure Strategy (NCRIS)).[10]

What is a sustainable business model for certain types of repositories may not be so for others. For example, host institution funding is likely to be a sustainable basis for institutional repositories, but may not be a sustainable model for national and international repositories. The move of arXiv as an institutionally-funded global repository to Cornell, and subsequent efforts to expand the funding base, is an illustrative example.[11] Conversely, the Australian Data Archive (ADA) is a national repository supported by a university, but the Australian National University has a unique national role and mandate, and the host funding model may well be sustainable in this case. Indeed, the ADA has been operating for more than 30 years.

Domain repositories host data relating to a specific field of research, rather than an institution. In this case, structural funding, such as that from a disciplinary research funding body, may provide the basis for a more suitable business model. However, research domains vary in how they are funded, which may well affect data repository funding models. For example, a research domain that has a more fragmented funding landscape may feature data repositories that combine a wider range of funding sources and models. Other repositories are likely to be closely associated with an organisation (e.g. NASA or CERN). In such cases, a business model based on either structural or host funding may be suitable.

The research and social contexts are also important. For example, expectations for open science and mandates for open data, may make a business model that imposes excludability (e.g. via copyright and restricted licensing) to underpin the collection of access charges unsuitable. Such a model may not be sustainable either, because the widespread adoption of open data mandates will limit the extent of the potential market.

The stage of development of a repository, its institutional or disciplinary context, its scale, and level of federation are also important determinants of what might be a sustainable business model. Referring to the dynamic of the evolution of firms, some economists draw a human parallel, talking of the phases as births, deaths, and marriages (and sometimes divorces). All phases are needed and should be accommodated. Indeed, sometimes it may not be desirable, effective, or efficient for a repository to be sustainable - provided that the data can continue to be hosted elsewhere.

Project funding often provides the initial capacity to create a data repository and may be a good start, with institutional or other relatively short-term support coming in during the development phase. However, as the repository matures and scales to provide an ongoing, reliable, and quality service, a different model is likely to be needed. From an economic perspective, this is the distinction between investment funding to establish a business (e.g. fitting out a shop and stocking it prior to opening), and ongoing revenue sources during the operational phase (e.g. daily sales takings). This distinction is not yet well made in the research data repository environment, but should form an important part in the design and evolution of repository business models.

In a report to the Wellcome Trust , Bicarregui (2016) noted that the European Open Science Cloud (EOSC) communication identifies fragmentation of infrastructure provision as a barrier to maximising the use of data. In the United States, similar issues are being addressed by the Big Data to Knowledge (BD2K) initiative of the National Institutes of Health.

> *Technological fragmentation arises for two reasons: technically different infrastructures lead to low interoperability, and separate governance and funding arrangements lead to heterogeneity of provision. Currently, data infrastructure is provided by a mixture of horizontal and vertical services. The advantage of vertical service provision is that it can be dedicated to the needs of particular research fields. However, it can also lead to a lack of interoperability with other vertical infrastructures. Horizontal services, on the other hand, are more likely to lead to cross-disciplinary homogeneity, however it is more difficult for horizontal services to be tailored to particular researchers' needs. Domain specific vertical infrastructures are currently serving the needs of their particular communities well. However, unless there are strong incentives to provide interoperability between infrastructures, provision will remain fragmented and opportunities for cross-disciplinary research and innovation will be lost* (Bicarregui, 2016: 8).

The need is not only for technical interoperability, but also for semantic interoperability, to ensure that the data are really in an understandable and re-usable format, wherever they are stored and used. It is important to include a focus on re-use and re-usability, and what it actually costs to make data interoperable, which requires high levels of expertise and competence.

Data repository business models must also be responsive to the key issues and trends affecting research data generation and use. For example, in some fields, such as the social sciences, the challenge is growth in demand as growing expectations and increasing mandates are leading to much greater demands for data curation and sharing than has been typical to date, so to be sustainable a model must be able to scale. In other fields, such as bioinformatics, while ingest remains a challenge, most data are already curated in repositories and shared, so different models that are not so responsive to scaling may be suitable. However, in such cases, there sometimes remain issues regarding the quality of the data and their levels of curation, which may also imply the need for different business models as the mix and focus of activities (e.g. hosting, curating, and value-added services) may be different.

All of these, and other, issues must be considered in choosing and evolving research data repository business models. There is certainly no "one size fits all" solution to be had.

An economic background to the issues

The aim of an economic analysis is to explore what data repository business models might provide sufficient revenue to make the repositories sustainable in the longer term. As many data repositories begin with the support of short-term project funding or depend on funding that is subject to changing budgetary conditions, such as annual government funding, developing a sustainable business model can be a significant challenge.

An economics primer

The market is generally considered to be the best way to allocate "scarce" (finite) resources, because it ensures that the resources go to the highest value use - the highest bidder. However, information exhibits public good characteristics: most notably in that it is not exhausted in consumption (i.e. it can be consumed

many times without being diminished), and it may be inefficient to exclude potential users.[12] This non-rivalrous nature of research data has important implications.

The problem (market failure)

The market *may not be* the best mechanism for the allocation of a public good, as any price above the marginal cost (of copying and distribution) will reduce net welfare – by locking out users and uses that do not have the capacity to pay or are not willing to pay. For digital information made available online the marginal cost is very low – close to zero. However, a price set at zero or close to zero will not be sufficient to cover full costs. Consequently, market forces lead to the under-production of public goods and the producers and distributors of public goods will find it very difficult to generate sufficient revenue to sustain their activities.

This is the situation facing research data repositories. To be sustainable, data repositories need to generate sufficient revenue to cover their costs, but setting a price above the marginal cost of copying and distribution will reduce net welfare. So, the key question is: which revenue source(s) and business models will have no, or the least negative, impact on net welfare – which will come closest to optimising research data production, distribution, and use?

The value of research data arises from its use and the more it is used the greater the social benefits and the higher net welfare. A user-pays model, directed at the users ("withdrawers") of data, which leads to pricing above the marginal cost of copying and distribution, is likely to have a significant impact on demand, lowering demand and, thereby, reducing net welfare. A user-pays model, directed at the "depositors" of data, is likely to reduce the incentive to deposit data, thereby reducing the amount of data made available for (re)use, and reducing net welfare.

This problem is not unique: there are many things that exhibit public good characteristics. Like information, some of these things are not exhausted in consumption and generate positive externalities that arise in consumption. Examples include street lighting or lighthouses, which can provide benefit to many users. Economists are familiar with how these things can be, and are, efficiently produced and distributed, both in principle and in practice.

Overcoming market failure

Such "market failures" require some form of intervention to achieve efficient allocation. The two most common forms of intervention are *regulation* and *incentives*. In practice, regulation and incentives are often combined.

Regulation can be used to promote demand where an individual's consumption has wider social benefits (e.g. requiring that children be immunised), or limit production where a producer's activities have wider negative impacts (e.g. the regulation of industrial pollution). Conversely, regulation can be used to limit demand where consumption has negative impacts (e.g. smoking and alcohol bans), or promote production by offering producers protection or incentives (e.g. renewable energy targets and subsidies).

While, for research data repositories, mandates are the most directly relevant form of regulation, they are by no means the only relevant form. An example can be seen in the UK-based Archaeology Data Service (ADS), which recently introduced depositor charges, as a significant proportion of hosted data arises from finds made in the course of building and construction industry activities. Such activities are

subject to heritage regulation mandating what must be done, in terms of cataloguing and recording, when any archaeological material is found. A further large share of ADS's data comes from projects funded by the UK's Arts and Humanities Research Council, which requires data curation and sharing. Hence, in this case, regulation takes the form of both funder mandates and heritage regulations affecting the building and construction industry.

Incentives can be used to promote or limit production or consumption (e.g. through taxation concessions or increases). Hence, for example, many countries impose tobacco taxes and carbon taxes, and offer tax concessions to promote R&D and lower licensing fees for electric vehicles.

Regulation and incentives can be, and are, used in support of research data repositories, to support efficient resource allocation and achieve an optimal level of production and consumption. Regulation typically takes the form of government or funder mandates for open data that require the producers of data to make them openly accessible.

Incentives can be positive or negative. In the case of research data repositories, positive incentives include giving the producers of data permission to draw on special funding to make data openly accessible, or more centrally funding and making the repository freely available to depositors. Negative incentives may also be used to encourage compliance, such as linking career progression or future research funding to compliance with open access and open data mandates – as is done by the Wellcome Trust and the National Institutes of Health (NIH).

While regulation and incentives are the most common forms of intervention, there are other forms of economic intervention and other business models with relevance to research data repositories. Often these relate to stakeholders who are not direct producers and users of research data, and they often overlap.

Privileged Group: Where there is an individual or group who can cover the cost of making a public good freely available because it is tied to a private benefit. An example might be local shop-keepers installing video surveillance in their street for security, benefiting other residents and street users who did not pay for it. There are a number of variations around this theme. A club model is similar except that access is sometimes limited to members of the club or certain groups that they are willing to support, such as public-sector research. A consortium model is also similar. The common feature is that of collective action (i.e. consciously and deliberately acting as a group).

In the case of research data repositories, such a “privileged” situation may arise for research funders, research centres, universities, or a disciplinary group or society that may gain recognition and further funding from supporting or hosting an open data repository. Hence, there are many institutional repositories supported by a host institution or by a consortium of institutions that contribute to making the data openly available as a part of realising their mission.[13] Similarly, philanthropy may be part of the group. An example might be the European Bioinformatics Institute (EBI), which combines European and national funding with “philanthropic” funding from the Wellcome Trust.

There can be many different kinds of groups or clubs with different conditions on access to data and services for members and non-members. For example:

- Groups or clubs where the main user community want a data repository so much that they are all prepared to pay a little (which may be via in kind support) to make it work.

- Groups or clubs where a limited number of big players pay the main costs, but the data and services are open to everyone.

- Groups or clubs where there is privileged access for club members, but some more limited access for everyone.

- Groups or clubs where access is restricted to members for a limited period, and then subsequently open (i.e. delayed open access).

However, simply having multiple funders and supporters does not make a club or consortium unless they consciously and deliberately act collectively.

Packaging: Where the costs of making something freely available are covered by packaging the public good with a private one. The most obvious example is advertising, such as that on free-to-air television. Another, but perhaps more sinister, example in the internet age is tracking online activities and data analytics (e.g. to link use of certain data types to targeted advertising). Both depend on the private benefits associated with providing information to the users of the freely available public good.

In the research data repository context, the most common form of packaging is that of data made freely available in combination with charges for access to facilities (e.g. virtual laboratories and computing facilities), or charges for value-added services (e.g. training, consulting and advisory services). This is possible because, unlike data, the facilities and services do not exhibit public good characteristics - being both excludable and rivalrous in consumption.

Box 1. User-pays solutions?

There may be a user-pays solution, if users are able and willing to pay a sufficient amount for data repository access and/or services to cover some or all of the costs without limiting use. Of course, willingness to pay is constrained by capacity to pay. This can have important social and economic consequences. Certain groups may be excluded where capacity to pay is extremely limited (e.g. students or researchers in least developed countries), and perhaps more importantly, certain uses may be excluded, such as:

- Where the nature of use means that a very large number of types of data or datasets are required as inputs and willingness to pay for individual items is close to zero (e.g. text or data mining).

- Where use is non-commercial, and the revenue generated from it is small or non-existent, limiting capacity to pay for the optimal level of use (e.g. education and research).

- Where use is transformative and faces uncertain value, and users cannot judge how much to spend on it, or it has high positive externalities, with the users not capturing the value of their use as it spills over to others (e.g. education and research).

A key problem with any user-pays (for access) model is that charging for data implies the need to protect those data from simple copy and resale by others. Hence, access charges are typically combined with licensing restrictions, which limit the uses to which the data may be put. Free gratis (zero price) is typically a pre-condition for free libre (unrestricted reuse), and it is the latter that enables the maximum re-use and realises maximum value.

Transaction costs: In the 1960s, the economist Ronald Coase explored hypothetical situations in which there were low or no transaction costs (i.e. where people could combine and negotiate at low or no cost). In the internet age, this is no longer hypothetical: examples include crowd-sourcing and crowd-funding. However, such funding is typically provided for a one-off project, and it is unlikely to be suitable as a basis for a sustainable business model for a research data repository.

Another possible business model depending on low transaction costs is that of micro-payments. Examples include Data.World, which is currently exploring a freemium model and micro-payments. A key issue here is whether the payment is purely voluntary or depends upon access barriers erected to collect payment, as meeting funder expectations for open data is likely to be an important determinant of sustainability.

Moreover, there are obvious issues with the uncertainty of voluntary payments and the lack of a clear link between such revenue and repository costs. Nevertheless, there is little to stop any repository adding a "donate" button to its entry webpage, with payments being a bonus rather than an expectation.

Mixed business models for data repositories

One key issue in the development of research data repository business models is the mix of data and services within the repository, which will vary from case to case. For example, a repository may be:

- Primarily based on data curation and sharing, offering only limited services to users (e.g. basic user support).
- Primarily based on equipment and facilities that generate the data, such as large-scale scientific equipment or virtual laboratories, but also host the data generated as a part of the facility services.
- Primarily based on a range of value-adding services (e.g. front-end discovery, or computational services, analytical or visualisation services), but also host the data as a part of those services.

This is important, because while data exhibit public good characteristics (e.g. they are not exhausted in consumption) facilities and services typically do not; they are more or less rivalrous in consumption (e.g. one party's use of a facility, such as a synchrotron beamline, prevents others from using it at the same time). This provides an opportunity for the development of repository business models that mix freely available data and charge some or all users for related facilities and services.

The mix of data and services may also change over time as the scale of data curation and sharing grows and the "market" for data and related services matures. In economics, there is a major issue regarding the make versus buy decision, which is an important part of the theory of the firm (i.e. what defines the boundaries of firms and markets). This also relates to economies of scale and scope, and transaction costs. To date, there have been limited opportunities for research data repositories to buy services. As it was all new, they had to do it themselves. As time passes, there is an increasing possibility for repositories to purchase specialist external services, such as data storage, again providing an opportunity for new and innovative business models to develop, and for increasing private sector participation.

Costs, cost drivers, and scalability

To be sustainable, a business model should include a revenue source or sources that scale with the demand for repository services. This is likely to be scaling primarily with the volume of data or number of datasets to be ingested and hosted, and more particularly with the largely human curation and preparation of those data, rather than the number and frequency of uses. However, it is important to identify the main cost drivers in each situation and to ensure an adequate understanding of the dynamics of costs.

Where a key cost driver is the amount of data ingested and hosted, a deposit-side model might be sustainable. This might include depositor fees or annual charges to depositing institutions, or research funder, government or institutional support allocated as a percentage of research funding. However, costs will scale differently for different repositories depending on the volume of data, the amount of human stewardship required, and any economies of scale, automation, and other context-specific factors. Hence, pricing deposit-side charges will be challenging.

On the other hand, for repositories building services on top of the data (which are likely to be more valued by data users), the development and maintenance of those services will be a significant cost factor. In this case, a user-side model (e.g. charges for access to value-added data, value-adding services, or facilities) may be a more suitable model, as it would better scale to the cost drivers.

However, there can be many other cost drivers, including:

- The scope and variety of users (e.g. across disciplines or language groups), which may require more detailed and varied data descriptions, help and support services, and other types of assistance.

- User expectations and the sort of intervention required to make the data optimally (re)usable, which may shift with developing expectations and the development of a research field.

- The mix of data and services and the levels and nature of value-adding (as above).

- The state of the input data and the extent of data curation after deposit, such as the preparation of ontologies or linking to other relevant information resources.

- The disciplinary scope (e.g. the need to have data accessible and usable across fields or just within a limited research domain). Data may require more curation if the scope of users is broader, and they may require different metadata and greater support.

- The mix of closed versus open spaces, because some repositories provide project spaces wherein data are shared between project researchers, but are not open to the public until publication, project completion, or there is some other reason for restriction with a subsequent release trigger (e.g. some data policies allow the data producers to enjoy a period of exclusive use). This mix of open and closed spaces provides a foundation for a mix of charged and free access, and for a range of revenue sources and business models to be pursued simultaneously, either now or in the future (e.g. the closed project spaces are rivalrous in consumption, opening the door to user charging models). However, the operational complexity of such activities inevitably increases costs.

- National versus international focus and funding, as a repository focused on national data collections may have a clearer mission and a more direct link to a source or sources of revenue, but may also face a limited number of possible funding sources; while an international repository may have a wider range of possible revenue sources, but a greater coordination problem and costs in securing funding.

- The frequency of deposit, update, and use may also be an important cost driver in some cases, although it may also offer greater opportunity for automation (e.g. automated ingest).

- The degree of overlap of the data creators, data depositors, and data users: where there is overlap there may be more inclination to share the costs than might be the case if the data creators and depositors are not the major users.

- The number and variety of existing (and possible) funding sources. A wider range of possible revenue sources may imply greater funding opportunities, but will increase transaction and coordination costs.

Business models, cost constraint and optimisation

A key issue is that, in the absence of market or price signals, it is very difficult to know what level of service and associated funding is optimal.

To what level should data repositories be funded?

There are studies that have explored the costs and benefits of data repositories (Beagrie and Houghton, 2016, 2013a, 2013b, 2012) and many studies focusing specifically on the costs of data repository functions (see Box 3), but understanding current costs and that the benefits exceed those costs does not help to answer this crucial question.

Too little funding being made available for research data repositories will lower the level of curation and sharing, reducing the benefits. But more is not necessarily better, as too much funding will also be inefficient. So, what system of allocation will best ensure that the optimal level of funding will be available for research data repositories, and what incentives for cost constraint do the various business models imply?

Preference Theory would suggest that optimisation will best be served by informed actors making choices or trade-offs between valued things and activities.[14] Some of the business models discussed below are more likely to involve this than are others. For example, structural funding typically involves a trade-off between funding for data repositories and funding for other research infrastructure or for research itself. That allocation will best be made by informed actors making choices, such as through an infrastructure funding allocation process driven by widespread research stakeholder participation, expert consultation, and “road-mapping”.

Other business models, such as those depending on deposit or access fees, bring the trade-off closer to the researchers, but their success in optimising allocation will depend on the extent to which the actors are informed and on their freedom of choice. The latter may be constrained by open data mandates (regulation). Other business models may divorce informed actors from the funding decisions or require

additional processes to ensure greater stakeholder understanding of the value of the repository services (e.g. structural or host institutional funding).

What are the incentives for cost constraint?

A further issue is the internal effect that a business model has on cost constraint or cost optimisation. The more a business model depends on or creates low price elasticity of demand the lower is the incentive for cost constraint by the repository operators.

Price elasticity of demand refers to the relationship between changes in price and the level of demand. Demand is said to be inelastic if relatively large changes in price have a relatively small impact on demand. Examples include indispensable items, things to which consumers may be addicted, things for which there are no close substitutes, and things for which consumption is mandated.

Hence, mandated deposit is likely to be relatively price inelastic, implying that there may be limited cost constraint on repositories relying on depositor fees, especially where those fees also have direct and explicit funding support. The UK-based Archaeological Data Service may be an interesting example, as the building and construction industry faces a strong mandate and publicly funded research projects often have funding support available for archiving. Services supporting journal publishing in cases where there are mandates requiring data sharing, such as Dryad and Mendeley, may also be interesting to watch in this regard. However, one would expect that research data repositories that charge deposit fees will be constrained in their pricing because they will, to some extent, be competing with repositories that do not charge deposit fees.

4. SUSTAINABLE BUSINESS MODELS

This chapter explores the sustainability of research data repository business models, looking at the underlying economics of the business model, the financial sustainability of the identified revenue source(s), the pros and cons of various business models in terms of stakeholder economic and other incentives and benefits, and the possible effects of these incentives on cost constraint and cost optimisation.

Analysis of research data repository business models

This section presents an analysis of the research data repository business models identified during this study. It begins with a brief review of the criteria for business model design and evaluation.

Criteria for design and evaluation

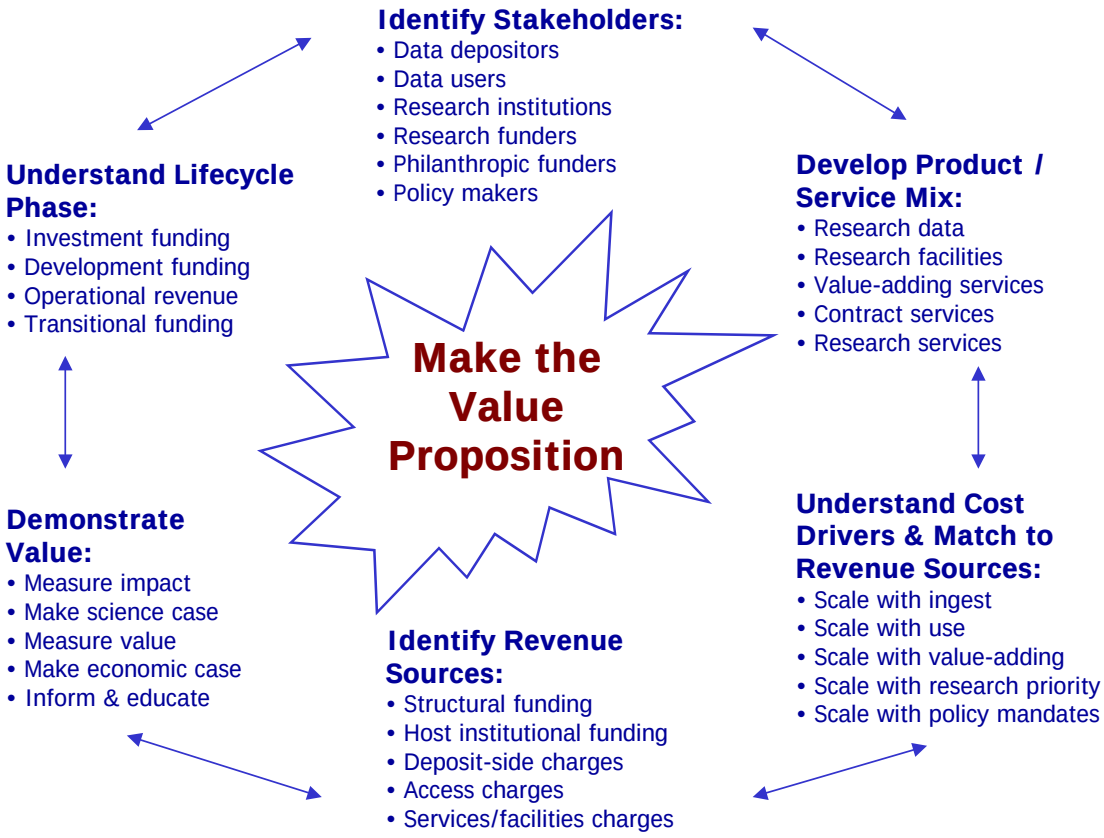
The development of a business model for research data repositories should be based on "a plan for the successful operation of a business, identifying sources of revenue, the intended customer [stakeholder] base, products, and details of financing."[15]

Actions needed to develop a successful research data repository business model include:

- *Understanding the lifecycle phase of the repository's development* (e.g. the need for investment funding, development funding, ongoing operational funding, or transitional funding)
- *Identifying who the stakeholders are* (e.g. data depositors, data users, research institutions, research funders, and policy makers)
- *Developing the product/service mix* (e.g. basic data, value-added data, value-added services and related facilities, and contract and research services)
- *Understanding the cost drivers and matching revenue sources* (e.g. scaling with demand for data ingest, data use, the development and provision of value-adding services or related facilities, research priorities, and policy mandates)
- *Identifying revenue sources* (e.g. structural funding, host institutional funding, deposit-side charges, access charges, and value-added services or facilities charges)
- *Making the value proposition* to stakeholders (e.g. measuring impacts and making the research case, measuring value and making the economic case, informing, and educating) (Figure 6).

As the context is dynamic, all these aspects of the business model should be revisited regularly during a repository's lifecycle.

Figure 6. Elements of a research data repository business model



Source: Authors' analysis.

Analysis of data repository business models

The interviews undertaken for this study reveal the following broad categories of business models being employed by research data repositories:

- Substantially structurally funded (i.e. central funding or contract from a research or infrastructure funder that is in the form of a longer-term, multi-year contract)
- Substantially supported by the host institution (i.e. direct or indirect support from the host institution)
- Substantially depending on data deposit fees (i.e. in the form of annual contracts with depositing institutions or per-deposit fees)
- Substantially funded from access charges (i.e. charging for access to standard data or to value-added services and facilities)

- Substantially supported through contract services or project funding (i.e. charges for contract services to other parties or for research contracts)
- Business models based on a combination of revenue sources

Box 2. Criteria to assess the sustainability of research data repository business models

There are a number of criteria by which one can assess whether a research data repository business model may be sustainable. These include:

- Does the business model make sense economically - is it a recognisable model from underlying economic principles?
- Does the model provide a source, or sources, of revenue that scale with demand and costs?
- What are the mechanisms or factors in each business model that help make it responsive, or not, to requirements?
- Which stakeholders benefit from the data being freely available, and how?
- What effects are different repository business models likely to have on data sharing and use?
- What are the mechanisms in each business model that ensure that any paying stakeholders have a clear understanding of the repository's value proposition?
- Are some business models more suitable for certain types of repositories than others?
- Are there optimal mixes of revenue sources for various types of data and types of repositories?
- Are there ways of grouping or dividing data and related services to better match single or combinations of revenue sources?
- What are the incentives within various business models for cost constraint?
- What are the mechanisms in the business model that might encourage cost optimisation?

These criteria form the basis for the following analysis.

Building on the previous RDA-WDS study, an analysis of the Strengths, Weaknesses, Opportunities, and Threats (SWOT analysis) associated with various revenue sources and combinations was conducted. An economic analysis has been incorporated to provide a more complete picture of the dynamics and incentives at play in the various data repository business models. This section presents a brief summary of this analysis (with a more complete summary of the SWOT exercise being presented in Appendix D).

Structural Funding: Central funding or contract from a research or infrastructure funder that is longer-term, multi-year in nature

Twenty-eight of the 47 repository managers interviewed reported some dependence on structural funding. No fewer than ten repositories reported being entirely supported by structural funding and a further six reported 90% to 97% structural funding.

Structural funding treats data repositories as research infrastructure, funding it in the same way as laboratories, equipment, and facilities. It is consistent with, and often conditional upon implementing, open data principles. Indeed, structural funding is frequently closely aligned with government or funder mandates for open data, and it avoids the difficulties that can arise when governments and funders issue "unfunded mandates".

A business model based on structural funding combines regulation (mandates) with incentives (including funding), and has no negative or limiting impacts on either data deposition or use. It can also align repositories with research fields or domains, with disciplinary research councils supporting a disciplinary or subject repository. In a similar way there can be synergies with philanthropic donors or private foundations, such as the Wellcome Trust, which supports the European Bioinformatics Institute (EBI). Hence it is a common model for large-scale subject or domain repositories. Such repositories can be national, regional, or global in scope. Their scale is likely to bring scale economies that help to keep a lid on repository costs, and improve interoperability, technical support, and related services.

A repository operating with the support of a number of international funders might benefit from fostering their collaboration, in effect transitioning to a club model, and it may be that such an approach is appropriate for large-scale international data repositories. However, the repository operators and managers should also consider the limitations of such models (e.g. limitations of scale, risks associated with a single funding source, possible loss of control, and opportunities to explore wider funding options).

The main strength of a business model based on structural funding is the long-term sustainability it can provide, which allows data centres to plan in advance and build an effective organisation. Fundraising does not consume too much time, with relatively few funders and low transaction costs involved. Some participants in the study workshops noted that there currently exists a window of opportunity, since data, their use and accessibility, are a "hot topic", and governments and funders may have increasing infrastructure budgets. Other participants were less confident that increasing infrastructure budgets were common. Clearly, there are significant national and regional differences.

At the same time, there is a danger that structural funding, often negotiated and fixed over multi-year contracts, will not keep pace with growing data volumes. Indeed, perhaps the primary weakness of this form of funding is its inflexibility, although that may provide an incentive for the repository to more carefully manage and seek to optimise operational costs. A further difficulty may be the global scale of some domain repositories, making the engagement and continued commitment of national funding agencies more difficult.

A threat posed by structural funding is that it can constitute a single point of failure, although of course, there can be more than one source of structural funding and it can be national and/or international in scope. Moreover, it is common to see data repositories combining the longer-term, but rather inflexible, structural funding with more flexible shorter-term funding (e.g. project funding or contract services revenue).

From the repository perspective, it is important to ensure that current and potential structural funders understand the value proposition. What someone is willing to pay for something depends on that person's perception of its value. For direct users of repository data and services the value is clear to them, and is revealed in their use. For stakeholders that are not direct users, it is more difficult for them to judge the value and understand what they may be willing to pay. Engaging and maintaining structural funders depends on their understanding of the value, and ensuring that engagement may involve repositories undertaking detailed benefit/cost, and value and impact analyses, perhaps synchronised with the funding renewal cycle(s) involved.

Hosting or institutional support: Direct or indirect support from the host institution

Eighteen of the 47 repositories analysed reported some dependence on host or institutional support. Examples of repositories reporting substantial host or institutional support include: the Center for Climate and Resilience (CR2), the Physiome Project Repositories (CellML at Auckland Bioengineering Institute), Purdue University Research Repository, the HSRC Research Data Service, The South Data Archive (SADA), the Australian Data Archive (ADA), UCT eResearch, and Harvard DataVerse.Not surprisingly, host or institutional support is a common and robust model for institutional repositories. In general, host or institutional support alone is unlikely to be a sustainable model for large-scale domain or international repositories. The example of Cornell University's hosting of arXiv is instructive, with increasing attempts to spread the load of funding beyond the host institution and onto user institutions.

At the opposite end of the continuum, it is common for there to be some element of host support (cash or in-kind) for repositories relying primarily on other funding sources. Host or institutional support may be suitable for data repositories, if there is sufficient commitment to the role. The scale and focus of the host is crucial (e.g. a single university versus an organisation such as CERN).

Table 1. A summary of the major pros and cons of the various funding sources

Funding Source	Pros	Cons
Structural funding	• Compatible with open data principles. • Longer-term stability. • Larger-scale and efficiencies. • Flexible as to allocation.	• Fixed, multi-year may not scale easily. • Competes with research funding. • Too many eggs in few baskets.
Host or institutional funding	• Compatible with open data principles. • Longer-term stability. • Efficiencies through sharing services. • Close to researchers (customers).	• Limited purview, with focus on local community. • May lead to fragmentation of domain data and lower interoperability. • Limited incentive to add value to data and develop related services.
Annual deposit-side contract	• Compatible with open data principles. • Demand oriented and scales with demand (data ingest).	• Unpredictable year-to-year and involves high transaction costs (managing contracts). • Limited engagement with, or focus on, data users. • May lead to fragmentation and loss of scale if many players, or monopolies if few players.
Data deposit fees	• Compatible with open data principles. • Demand oriented and scales with demand (data ingest). • Researchers price sensitivity ensures cost constraint. • Open data is part of research and its funding.	• Cost disincentive to depositing, so depends on strong mandates. • May lead to low level of curation to contain costs (price). • May be difficult for repository to compete for deposits with comparable repositories that do not charge.
Data access charges (subscriptions or use fees)	• Users pay for what they want, so funding reflects value. • More market-oriented approach may provide incentive for cost constraint.	• Not compatible with open data principles and many funder mandates, limiting the potential market size. • Charges limit use and will reduce the value of data. • Revenue scales with use and not ingest or curation costs. • Vulnerable to funding cuts.
Contract services and project funding	• Compatible with open data principles. • Can increase contact between staff and clients. • Potential for innovation.	• Short-term and not a sustainable sole revenue source. • High transaction costs, chasing money. • Revenue does not scale with data ingest or use (costs). • Inflexible as to allocation of funding.
Diversification of revenue sources	• No single source of failure. • Can maintain compatibility with open data principles. • Flexible and enables experimentation with new services.	• May lead to higher transaction costs (managing multiple funding sources). • May lead to Mission drift.

Source: Authors' analysis.

For institutional repositories, the major strengths of host or institutional support include longer term sustainability (e.g. universities and research organisations tend to have longevity), and convergence of interest between host and repository, such as where the data repository aligns with a strategic research priority of the research institution (e.g. ICPSR at University of Michigan, CIESIN at Columbia University, UKDA at Essex University, and ADS at York University).

Cost optimisation can be encouraged by sharing services within the institution, and repository support staff can be close to the researchers for on-hand support (e.g. the development of data management plans, or ingest support). Of course, the stability of the support depends upon the financial health and commitment of the host institution.

A weakness may be the focus on the local institutional community that needs to be served, which may lead to fragmentation of domain data, lower levels of curation, and lower interoperability. Moreover, researchers may find discipline-specific, specialised repositories more attractive, as they participate in the wider community of their research field rather than their institution, further exacerbating possible fragmentation. It may also be difficult for the repository to make a business case to the host, if the scale is relatively small and the remit multidisciplinary and internally focused.

As a business model, one based on host or institutional funding may also combine regulation (institutional mandates) with incentives (including repository funding and support), and has no negative or limiting impacts on data deposition. However, while there is no limit on use, there may be less focus on users as they will typically be outside the host institution. This may have an impact on user support and engagement, and lead to lower levels of curation and lower levels of data use. Hence, risking sub-optimal levels of data re-use.

Again, various mixes of revenue sources are available within business models, which can mitigate some of the potential weaknesses and threats relating to particular funding sources. Among our sample, only four repositories reported 100% host funding (e.g. the Purdue University Research Repository), while others mixed host with structural support, deposit-side charges, and a variety of project and contract services revenue. For example, Dryad currently mixes very limited host support with data deposit charges.

Data Deposit Fees: Annual contracts and per deposit

Repositories that substantially rely on deposit-side funding typically do so through annual contracts and support from depositing institutions, or through deposit fees levied on each deposition. These business models have somewhat different characteristics.

Annual contract (from depositing institution or organisation)

Contract funding can come from depositing institutions, such as those subcontracting for data hosting services from government agencies or universities. The business model is comparable to subscriptions to deposit data, paying the host an annual fee for data hosting.

Thirteen of the 47 repository managers interviewed reported some dependence on annual contract funding. Adding up the individual repository shares of funding and expressing the sources as a percentage shows that deposit-side contract funding accounted for around 15% of reported funding by source across the sample and an average of around 40% of the funding for those using the model. Examples of

repositories with substantial depositor-side contract funding include: 4TU.ResearchData, the Scholar's Portal, the Arctic Environment Research Center, Figshare, and Mendeley.

Such repositories have the potential to provide the scale and scope that institutionally supported and focused repositories may struggle to achieve, leading to economies of scale and greater opportunities for cost optimisation. That scale might also improve the level of curation, related services, and interoperability. They might also usefully fill gaps for "homeless data", support new journal publication policies, and possibly data journals.

However, contract funding can be unpredictable from year-to-year, and involve relatively high transaction costs in managing contracts. Funding difficulties could lead to limited curation and documentation, as there is an incentive to cut costs to lower subscription fees and attract more subscribers. In contrast to discipline, subject, or institutional repositories, there may be limited engagement with researchers, as users or depositors. Moreover, if the number of contracting organisations is small, such repositories may be too narrow: focused on creating solutions for one or a few clients that may not be useful for others.

From an economic perspective, the success of data-hosting contracts is likely to depend on regulation (mandates). There are a number of public repositories using this model to cover part of their funding. For example the Australian Data Archive provides a hosting service for government agencies that are incentivised to outsource non-core activities for which they have no experience or infrastructure. There are also emerging commercial and semi-commercial examples of repositories providing data-hosting to research organisations, universities, and journal publishers (e.g. Mendeley and Figshare).

The model is compatible with open data principles, but there may be some disincentives to deposit data (e.g. limited by the bounds of contracts) leading to lower levels of sharing and loss of re-use benefits – depending on the scope, strength, and enforcement of mandates. With some disassociation between the researchers who may be funded to deposit and the institutions paying the subscription there may be some transaction costs involved in handling the money and less clarity about who is willing to pay, for what, and how much. Hence, making the value proposition clear to all stakeholders is crucial.

Data deposit fee (charges for deposit)

Just four of the 47 repository managers interviewed reported some dependence on data deposit fees (i.e. Dryad, the Archaeology Data Service, 4TU.ResearchData, and the World Data Centre for Climate / Deutsches Klimarechenzentrum). Adding up the individual repository shares of funding and expressing the sources as a percentage shows that deposit fees accounted for one-tenth of reported funding by source across the sample, and an average of around one-third of the funding of those using the model.

In contrast to data access charges, data deposit fees are compliant with open data principles. The model puts a charge on the data depositor and, in principle, this works well with grant funding. The model is neutral as to the value of data to end-users and it is scalable with the demand for data deposition. Data depositors may be somewhat price sensitive, so there is incentive for cost constraint, although it is unlikely that a genuine market will emerge anytime soon.

At the same time, deposit fees are a barrier and disincentive to deposit, which may limit the level of data sharing and reduce the potential value of those data. On the repository cost side, there may be administrative overheads in handling many payment transactions. There is also a challenge in costing,

because such fees may be difficult to price as new data deposits need to cover the cost of ongoing support of an ever-growing body of legacy material, development, and other costs. There may also be an issue if the depositor-pays model encourages the repository to accept all paid data, leading to fragmentation of data collections, lower levels of curation and interoperability, and possible loss of focus.

High fees may put off depositors and encourage a rush to the cheapest option (i.e. a race to the bottom). That may lead to low levels of curation to contain costs, if researcher depositors see the money as a part of their research funding. Conversely, it may lead to high curation levels with high deposit fees for a niche market, if researcher depositors do *not* see the money as part of their funding. This will depend on the relative mix of regulation (mandates) and incentives (including funding).

A possible weakness of a business model based on deposit fees is that in some cases data repositories charging deposit fees will compete for data against broadly comparable repositories that do not charge fees to deposit. Managing the depositors' value perceptions and understanding their willingness to pay will be crucial for survival.

A depositor-pays model probably requires a very clear policy framework, including strong mandates (regulation) and support to cover reasonable deposit fees (incentives). To be sustainable, any depositor-pays model for research data repositories probably needs fully funded mandates, especially models that depend on direct per deposit fees.

However, strong regulation and a possible over-funding may lead to limited incentives to constrain repository costs. This may be a particular issue if a repository has a "captive market" of depositors who face very strong regulation and are funded to deposit. A possible example may be the UK-based Archaeology Data Service, which reports that deposit fees account for around half of its revenue, and for which research depositors face funder mandates and industry depositors face heritage regulations (e.g. governing what happens to archaeological finds during building and construction activities).

Access Charges: Charging for access to standard data or to value-added services

None of the repositories interviewed reported a dependence on *charging for access to standard or raw data*, while seven reported some dependence on *charging for access to value-added services*. However, the boundaries between raw data, value-added data, and value-added services are not always clear. For example:

- The Arabidopsis Information Resource (TAIR) charges for access to value-added data, and makes the raw data freely available after a one-year period.

- The Cambridge Crystallographic Data Centre (CCDC) also charges for access to aggregated value-added data *and* for related services.

- Inter-University Consortium for Political and Social Research (ICPSR) charges for access to some restricted use data, although these charges account for a very small share of overall revenue. Around 25% of ICPSR revenue comes from membership fees from data using institutions.

- The UK Data Archive (UKDA) is permitted to charge a cost recovery fee for access to certain secure data. The fee, which is not currently being enforced, is, strictly speaking, not for access to

data, but cost recovery for services provided by the UKDA for end users (e.g. software for data analysis).

Hence, there are a number of diverse repositories that do charge for access to data to some extent and in some form.

Data access charges are scalable and the charges can reflect the value of the data products to users. The major weakness of the model is that it is in conflict with open data principles, and may not meet the conditions set out by research funders and institutions.

Data access charges are a disincentive to use and will put some restraint on re-use, and reduce the value and impact of the data, thereby reducing the return on investment in the research activities from which the data arose, and reducing the efficiency and re-use advantages arising from openly sharing data. Nevertheless, there are attempts (e.g. by CCDC and TAIR) to mitigate the negative effects of charging, by offering reducing or waiving the cost for low income countries, data release after embargo periods, etc. Whether such mitigations are effective remains to be seen, but they are unlikely to overcome completely the negative impacts on use.

A problem with any user access charging model is that charging for data implies the need to protect it from simple copy and resale by others. Hence, access charges are typically combined with some licensing restrictions, which limit the uses to which the data may be put. Free *gratis* (zero price) is typically a pre-condition for free *libre* (zero use restrictions), and it is the latter that enables the maximum reuse.

Charging for access to value added services, while making the basic data freely available, does not carry the same disadvantages as data charges. However, revenue from such services scales with their use and is not closely linked to data ingest and curation costs. Moreover, value-added service charges could be vulnerable to economic downturns and funding cuts, as the basic data may be perceived to be sufficient in times of cut backs.

Nevertheless, there is an increasing number of research data repositories taking advantage of the underlying economic differences between research data, which exhibit public good characteristics, and value-adding services (e.g. training and advisory services), and facilities (e.g. virtual laboratories, computational and analytical facilities), which do not. This sees an increasing adoption of business models that combine freely accessible data with charged access to facilities or charges for value-adding services.

Such charges can be targeted to particular groups (e.g. differential charging for private sector and public-sector access). For example, CCDC's value-added services charges are significantly less for academic use than they are for industry use.

Introducing charges for newly available facilities or services may be feasible. A major limitation may be the difficulty of making a transition to charging (i.e. introducing charges where access had previously been free). An example may be the Swiss Institute of Bioinformatics (SIB), which recently introduced charges for a previously freely available facility-based service. The impact of this on user perceptions and levels of use will be interesting to observe, and may be instructive for repository operators and managers.

Crucially, in either form, access charges for data or for facilities or services are likely to lead to lower levels of access and use, and unequal access based on capacity and willingness to pay.

Membership model (privileged group, club, or consortium)

In the case of the membership model (privileged group, club, or consortium), where a group combine to cover the costs and the data and services remain freely available, an advantage is that there is a relatively stable and predictable income and that members have an influence on the repository that encourages it to be responsive to their needs.

However, membership revenue scales with the membership, and not with either data use or with ingest and curation costs. There can also be relatively high transaction costs involved in creating and maintaining the membership group. For example, it can be difficult to demonstrate value to people who are not directly data users or necessarily even close to users, and trying to do so can involve a lot of effort.

A common issue for membership models is that it is difficult to develop and maintain a group large enough to cover costs, affecting both scale and sustainability. From a different, but related, domain the examples of Knowledge Unlatched[16] and SCOAP3[17] open access scholarly publishing initiatives may be instructive. Both are niche solutions and are likely to remain so; the former because it is unlikely to scale to provide a major world-wide solution, and the latter because it is unlikely to be applicable across research fields.

Contract for services and project funding: Services to other parties or research contracts

Twelve of the 47 repository managers interviewed reported some dependence on funding from contract and additional services (i.e. services to parties other than the core funder), and 13 reported some dependence on other funding sources (e.g. training services), or on grants or contracts for research or development projects of various types.

Adding up the individual repository shares of funding and expressing the sources as a percentage shows that these sources accounted for around 14% of reported revenue by source across the sample. For those reporting contract services these represented an average of less than one-third of total revenue and for those reporting some dependence on other income sources these averaged less than one-fifth of total revenue. Hence, contract services revenue is generally used as a supplementary source of funding, rather than as a source of core funding.

Many repositories have some form of contract funding, such as for projects, training, and other services. For many, research funding supports innovation and development that cannot be funded from core structural or host funding, or charged directly to depositors or users. Hence, the principal advantage is support for innovation and development, and opportunities for further developments in the future.

The primary disadvantages include the short-term nature of such funding, lack of flexibility in its spending, and possible diversion of staff and effort from the core tasks and functions of repository operation. Contract services and project funding are not likely to be sustainable as the sole source of revenue, and involve relatively high transaction cost (i.e. spending time and money chasing money).

That being said, many repositories have a high dependency on project funding in their early development phase, because they start out as a project and aim to transition into a service (e.g. DRI, CCDC, Dryad, and TAIR). Project funding is often the mechanism for identifying the need for new and innovative data resources, but a number of repositories remain reliant on project funding for a large proportion of their revenue for some time. One of the key challenges for the sustainability of data

repositories is to be able to manage the transition from being a project to being a service, when such initiatives have proved their worth, and to have processes that apply appropriate criteria to that judgement.

In economic terms, this transition is the equivalent of the distinction between one-off investment funding to establish a business, and ongoing revenue once established. This is the time at which the repository needs to develop a clear business plan, which identifies the users (including research data depositors) and other stakeholders (e.g. research or infrastructure funders, or philanthropic funders) who may be willing to pay, and to establish how much they might pay. Again, being able to make the value proposition clear to target groups will be crucial.

Business models combining various revenue sources

The 47 data repositories analysed reported 95 revenue sources, an average of two per repository. Twenty-four repositories reported funding from more than one source, and seven reported more than three revenue sources. Combining revenue sources into a sustainable business model is an important element of developing a sustainable research data repository.

Adding up the individual repository shares of funding and expressing the sources as a percentage, shows that structural funding is the largest single source at around one-quarter of total, and combining the forms of deposit-side sources (i.e. data deposit fees and annual contracts with depositors) shows that they account for just less than one-quarter. Access-side charges, primarily for access to value-added services, account for around one-fifth of total, while host or institutional support accounts for just less than one-fifth of total.

Diversification of revenue sources is attractive, as it has no single point of failure and the financial risks are spread. It offers flexibility to experiment with new services and markets and can stimulate innovation, which might not be supported from core funding sources.

The flipside is likely to be a relatively higher administrative overhead. With different funders, with possibly diverse and shifting interests, attention could be drawn away from the core mission of the data repository or lead to mission creep or confusion. Moreover, making the business case to a range of funders with different stakeholder perspectives may prove difficult.

Business models and characteristic combinations of revenue sources

Typically, repository business models combine structural or host funding with various forms of research and other contract for services funding, or funding from charges for access to related value-added services or facilities. Indeed, there were 46 reports of structural and or host funding, 25 reports of contract and other services funding, and six reports of funding from charges for access to value-added services, accounting for around 80% of reported funding sources.

When one explores the strengths and weaknesses of business models built on these combinations the attraction to them is clear, from the repositories' perspective. The security, but relative inflexibility, of structural and host institutional funding combines well with the flexibility, but short-term orientation, of research and other project and services funding. This allows core funding to cover core activities, and funding for research and development activities and value-adding and other services to support the further development of value-adding efforts.

As a business model, there is considerable scope to develop packaged offerings based on combining free access to data with charges for access to value-added services (e.g. training, consulting, and advisory services), and charges for access to and use of facilities (e.g. virtual laboratories, and analytical and computational facilities).

Indeed, as noted above, six of the 47 repository managers interviewed reported some dependence on funding from charges for access to value-added services, 12 reported some dependence on funding from contract and additional services (i.e. services to parties other than the core funder), and 13 reported some dependence on other funding sources (e.g. training services), or on grants or contracts for research or development projects of various types. Examples include: the Digital Repository of Ireland (DRI), the Neuroscience Information Framework (NIF), ICPSR, the SAEON shared data platform, and the World Data Centre for Climate (Deutsches Klimarechenzentrum).

The second, but smaller, combination is deposit-side funding combined with a mix of structural or host institutional funding (seven cases), or with research, value-added and other services funding (six cases).

The first combination suggests steady, but perhaps somewhat inflexible funding that, because the deposit-side element will scale to demand for deposition, to some extent. The latter combination may be less certain, but may allow for the development of more commercially oriented repository services featuring value-adding services overlaying the basic data.

As business models, these are further examples of packaging freely accessible data with charging for value-added services, or combinations of deposit-side and value-added services charges.

There are many possible combinations, building on the underlying economic characteristics of research data, which exhibit public good characteristics, and facilities and services, which do not. The choice of combination of revenue sources used to build such business models is likely to depend on the research domain and stage of repository development, among many other factors.

5. BUSINESS MODEL INNOVATION AND OPTIMISATION

This chapter explores the issues of innovative and emerging business models, cost constraint and cost optimisation. It builds on the interview findings presented in Chapter Two and the economic discussion of these issues presented in Chapter Three.

Emerging business models

One can see innovation in repository business models in five main areas:

- As data preservation and open data policies become increasingly widespread and influential, there will be more opportunities to develop deposit-side business models.
- As repository activities grow in scale and "the market" matures we can expect to see more business models based on charging for supply-side services to repositories (e.g. hosting) and user-side, value-adding services of various sorts.
- Similarly, research data repositories themselves can take advantage of the underlying economic differences between research data, which exhibit public good characteristics, and value-adding services and facilities, which typically do not (i.e. by charging for the latter).
- There may also be opportunities for the more disruptive entry of big players, possibly using some form of a packaging business model.
- As commercial and semi-commercial players enter the space, one can expect to see increasingly innovative attempts to develop user-side charging models.

Hence, repository business models that make data freely available, but charge for access to facilities or for the supply of value-adding services, may become increasingly common. Models focusing on deposit-side fees, based on strong policy mandates and funding support (i.e. regulation and incentives), might also become more prevalent. At the same time, there are likely to be increasing opportunities for private-sector participation and the development of packaged and user-side charging models.

Indeed, it is clear from the interview data and workshop discussions during the course of this project that there are a number of innovative business models emerging already.

Deposit-side models

As data preservation and open data policies become increasingly widespread and influential, there will be ever more opportunities to develop deposit-side models, either directly in the form of depositor fees or deposit-side subscriptions, or in a top-sliced form as structural funding, based on strong policies that include regulation and incentives (i.e. open data mandates and funding support).

Indeed, as noted in the previous chapters, there already is increasing use of deposit-side business models, such as those based on depositor fees (e.g. the Archaeological Data Service), deposit-side

contracts directed at research performing institutions and government agencies who may produce data but may lack the facilities or expertise to curate data (e.g. the Australian Data Archive), or something closer to structural funding from research or research infrastructure funding agencies (e.g. the Swiss Institute for Bioinformatics (SIB) or the European Institute for Bioinformatics (EBI)).[18]

There are also more commercially oriented deposit-side business models emerging (e.g. FigShare and Mendeley), relying on the emergence of a market for research data hosting and curation services.

Supply-side and value-adding services models

As repository activities grow in scale and "the market" matures there will be more chances to optimise the "make versus buy" decision, and we can expect to see more supply-side services (e.g. cloud hosting, ingest, and related services), and more user-side, value-adding services of various sorts.

Indeed, in the last few years, many data repositories have started to use commodity services for basic storage of their data. In terms of specialist services, a lot of data repositories rely on the Open Archive Information System (OAIS) reference model for recommended standards and practices for archival processes and adherence to community defined metadata standards.[19] These things, particularly the standards, are often themselves looked after by a not-for-profit organisation into which stakeholders, including the data repositories themselves, pay (e.g. the Data Documentation Initiative Alliance).[20] Such activities spread the costs involved and enable greater efficiency of research data infrastructure and management.

Supply-side infrastructure inputs, such as cloud hosting services at institutional, domain, national, and international levels are likely to emerge more strongly in the coming years, and may grow into something closer to full service providers over time. This type of development would provide further opportunities for cost optimisation, through economies of scale, and increasing opportunities for private-sector participation. These will need to be balanced against potential risks in terms of long-term sustainability, maintenance of service provision and access.

Packaging models

Research data repositories themselves can also take advantage of the underlying economic differences between research data, which exhibit public good characteristics, and value-adding services (e.g. training and advisory services), and facilities (e.g. virtual laboratories, computational and analytical facilities), which typically do not. An increasing number of repositories are already doing so, and this sees the adoption of business models that combine freely accessible data with charged access to facilities or with charges for value-adding services. Such charges may be for specific user groups.

Disruptive entry

As also noted, there may be opportunities for the more disruptive entry of big players, possibly using some form of a packaging business model. This could be in the form of taking over significant elements of repository activities and making them available through revenue from other related activities.

Examples may include such players as:

- Google, perhaps offering extensions of Google Scholar and related academic metrics more explicitly and extensively to data, or providing large-scale cloud hosting and data searching, linking, and analytics capabilities
- Microsoft, perhaps offering large-scale cloud hosting and preservation services
- Major scholarly publishers, extending their current offerings for research data hosting to much more fully developed data hosting and curation, search, linking, and analytical services
- Major research equipment suppliers, perhaps adding research data hosting and curation services to their existing equipment and services offerings

Any such developments would bring new opportunities for economies of scale and scope, and thereby cost optimisation. They may also introduce new risks, particularly in the longer term, as the policies, aims and ownership of commercial companies are liable to fluctuate. These risks need to be assessed and taken into account.

User-side models

There are also a number of emerging initiatives on the user-side, with various business models looking to user-side payments, while attempting to mitigate the impacts of access barriers and be compliant with policy demands for open data.

For example, The Arabidopsis Information Resource (TAIR) seeks to promote the use of its data by less well-funded researchers. TAIR uses affordable pricing with a sliding scale according to usage level, with metered access (some free page views each month), which provides open access for infrequent users. The data resources are free for the lowest income countries and for students using the database in a course (Huala, 2016).

On the important issue of licensing restrictions, where researchers need to publish subsets of data in the course of their work, TAIR has adopted academic license terms that allow researchers to publish and re-use limited excerpts. The repository also freely releases all datasets after one year - essentially a delayed open access model.

User-side micro-payments are also being explored. For example, Data.World is exploring a freemium model and micro-payments, but is currently reliant on investor funding and it is too early to say if this is a viable model.

Inevitably, of course, any access and use restrictions will have an impact on the level of use, and will tend to reduce the value and impact of the data. The problem with any user-side charging model is that charging for data implies the need to protect it from simple copy and resale by others. Hence, access charges are typically combined with licensing restrictions, which limit the uses to which the data may be put.

Opportunities for cost optimisation and reduction

Financial sustainability is an essential attribute of any trusted and reliable data repository. While research data repository managers may see requests by funders for cost optimisation as ways to “cut budgets”, they should be conscious of ways to constrain and optimise costs. Organisations that understand their cost structure will be more able to effectively control and manage their digital assets over time. They

may also be able to create new cost-effective solutions and services. For these reasons, it is important to identify areas where efficiency improvements are possible.

In the preceding chapters, we have discussed the incentives for cost constraint implied within various repository business models. In this chapter, we describe some strategies for optimising and reducing the core and discretionary costs that research data repositories face, building on information obtained during the interviews and workshops undertaken for this project (see Chapter Two).

Data repository costs and issues to consider in their analysis

Most data repository managers are conscious of their operational costs and try to limit them in various ways. Each repository has its own cost structure, but there are certain categories of costs that are common to all. Within cost categories, there are fixed costs and variable costs, as well as establishment costs, which are incurred at the beginning of the lifecycle, operational costs, which are incurred throughout the operational lifecycle, and some additional costs incurred later in more mature systems (e.g. converting legacy data to new formats and standards).

There have been many studies and cost models produced for research data repositories in recent years, each with a somewhat different way of presenting categories of costs and the underlying issues. Some of the most informative of these are listed in Box 3, along with the links for access to them.

A recent study by DANS/NCDD (the Data Archiving and Networked Services in collaboration with the Netherlands Coalition for Digital Data) examined the proportions of the budget assigned to various repository activities. The activities with highest proportions were data ingest and curation, as well as the maintenance and development of technical systems. Activities with the smallest budgetary proportions (or more variation) included value added services, administration and management, and outreach and training (Dillo, 2016, and also Figure 5).

The European 4C project compared ten existing cost models and concluded that:

- The cost analysis processes put in place should not only support accounting, but more importantly should enable effective budgeting.
- They should be able to handle various types and amounts of assets and use cases.
- The cost models should accurately define the costs and enable modelling of cost variables with easy to use tools, preferably with default values or settings.
- They should support the assessment of the benefits and value of digital curation as well as the costs.

While a lot of work has been done on repository costs and cost models, few studies have focused on cost optimisation.

Box 3. Digital Data Preservation Costing Initiatives

- LIFE 1, 2 and 3. Projects to explore digital preservation costing, and develop costing models. http://www.life.ac.uk/
- Cost Model for Digital Preservation (CMDP): Project at the Royal Danish Library and the Danish National Archives to develop a new cost model. Currently covers Planning, Migrations and Ingest http://www.costmodelfordigitalpreservation.dk
- Keeping Research Data Safe 1 and 2 (KRDS): Cost model and benefits analysis for preserving research data http://www.beagrie.com/krds.php
- Presto Prime cost model for digital storage http://prestoprime.it-innovation.soton.ac.uk/
- Cost Estimation Toolkit (CET): Data centre costing model and toolkit, from NASA Goddard http://www.pv2007.dlr.de/Papers/Fontaine_CostModelObservations.pdf
- Cost Model for Small Scale Automated Digital Preservation Archives (Strodl and Rauber) http://www.ifs.tuwien.ac.at/~strodl/paper/strodl_ipres2011_costmodel.pdf
- APARSEN Project activity focused on digital preservation costing http://www.alliancepermanentaccess.org/index.php/knowledge-base/digital-preservation-business-models/costbenefit-data-collection-and-modelling
- EPRSC and JISC study on Cost analysis of cloud computing for research http://www.jisc.ac.uk/media/documents/programmes/research_infrastructure/costcloudresearch.pdf
- Cost forecasting model for new digitisation projects (Excel and web tool under development) (Karim Boughida, Martha Whittaker, Linda Colet, Dan Chudnov) http://www.cni.org/wp-content/uploads/2011/12/cni_cost_boughida.pdf
- DP4lib is developing a business and cost model for a digital preservation service http://dp4lib.langzeitarchivierung.de/downloads/DP4lib-One-Pager-08-eng.pdf
- DANS Costs of Digital Archiving Volume 2 Project, focusing on preservation and dissemination of research datasets http://www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access http://brtf.sdsc.edu/
- Economic Sustainability Reference Model (draft) http://4cproject.eu/community-resources/outputs-and-deliverables/ms9-draft-economic-sustainability-reference-model
- ENSURE Project - Enabling Knowledge Sustainability Usability and Recovery for Economic value http://ensure-fp7-plone.fe.up.pt/site
- 4C. EU funded project on costing, led by JISC. http://4cproject.eu/
- Cost Model for Electronic Health Records (Bote, Fernandez-Feijoo, and Ruizb) http://www.sciencedirect.com/science/article/pii/S2212017312004434
- TCP: Total Cost of Preservation (California Digital Library) https://wiki.ucop.edu/display/Curation/Cost+Modeling
- Cost model for digital preservation (National Archives of the Netherlands) http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf

Source: http://wiki.opf-labs.org/display/CDP/Home.

One important question is for whom the costs are being optimised? From an economic perspective, *cost optimisation must be at the system level*. If a repository cuts services to save money, how will it affect the users? If the funders reduce their financial support, can the repositories still perform their key functions? If the data depositors do not have the resources to adequately curate their data in advance, will the repository be able to operate effectively?

For example, cost reduction at the repository level (e.g. by shifting more data preparation and ingest work to the researchers) may well increase overall system costs (e.g. through fragmentation of effort and loss of expertise). Hence, we must be clear about the differences between cost reduction, cost shifting, and cost optimisation.

It is also important to note that research data repository infrastructure costs will change over time. During the early years of research data infrastructure development there will be significant establishment costs. As the global data repository infrastructure evolves there will be less need to establish new repositories and there will be learning economies (e.g. greater familiarity and experience with repository operations) and scale economies (e.g. through individual repository growth and activity growth), which will reduce repository costs per unit. Consequently, we should be wary of allocating a fixed percentage of research funding to research data repository infrastructure, as it would be both very difficult to establish the appropriate level and very difficult to change it once established.

Opportunities for cost optimisation

The following comments build on the interview findings presented in Chapter Two, and further develop the ideas for cost optimisation that emerged during interviews and workshops.

Technological approaches

An obvious area for cost optimisation or saving is technological. Storage capacity has increased rapidly, while costs have fallen. So far, this has enabled efficiencies of scale and cost savings for existing repositories as the volume of data has exploded. For example, it has been reported that without the storage technology cost reductions, the existing European Bioinformatics Institute's operations and services would be impossible (Apweiler, 2016).

But what will happen if, or rather when, such technological progress slows down? Many analysts have observed that storage capacity growth has fallen well short of Kryder's law predictions.[21] Although there are technologies on the horizon, such as atomic or holographic storage that may continue or even increase storage efficiency, there is no certainty about their reliability or longevity, or when they may become operational (Kalff et al., 2016). At the same time, there is little doubt that the pace of data production in almost all areas will continue to expand rapidly (Cisco, 2016).

Data repositories may be able to automate some data management and curatorial tasks. For example, the DANS/NCDD study cited above found that the repository was able to optimise costs mostly in the technology areas. These were in automation of ingest (up to 50%), information technology techniques, use of open source tools, and, of course, cheaper storage (Dillo, 2016). In the future, using improved artificial intelligence techniques in such areas as metadata development, could offset a slowdown in some of the storage cost efficiencies (NRC, 2015: 57-60).

Management and organisational approaches

One approach for improving management and organisational efficiencies can include taking a closer look at the types of data that are retained by repositories, and for how long. Better rationalised and more realistic retention and purging, or dark archiving criteria, can be developed in consultation with the research communities who generate and use the data. Tiered storage is a technique that has been used by some larger, well-established repositories to reduce costs. Data that have not been used for some time can be saved in deep archives, where costs can be minimised.

Research data repositories could distribute or decentralise certain tasks to other players in the research community, including those in the private sector. For example, repositories can require depositors to submit only well organised, fully documented, quality assured, and standardised data sets. Other costs, such as various value-added products or services that may not be very economical to undertake in-house, can be pushed to users downstream. However, these may be examples of cost shifting, rather than cost saving, and the potential contribution to cost optimisation must be examined closely.

Federated organisations for data may have a number of attributes that are particularly well-suited for large data volumes from a distributed set of sources and can lead to economies of scale (Diepenbroek, 2016). At the same time, the use of private-sector services or public-sector infrastructure, such as cloud storage, can be used to help reduce the cost of long-term data stewardship.

Cost reductions have also been found in infrastructure collaboration, shared services (e.g. minting of Digital Object Identifiers), and international collaboration on preservation. In the DANS/NCDD study, the cost-saving measures from management approaches accounted for 5% to 15% of savings to the budget (Dillo, 2016).

The physical, geographical location of data centres may also be important for costs, in terms of variation in electricity prices, for example. Data centres can consume a lot of energy, so energy-efficiency and cost-efficiency go hand in hand, and might also be an area to look into when thinking about how to reduce data repository costs. This is also a matter of environmental sustainability in a world where the data volumes are growing exponentially, and may be an important consideration in the development of a sustainable research data repository business model in the future.

Research data repositories are a relatively new institutional development. While some repositories have been in place for many decades, most repositories have existed only a few years and many disciplines and countries have yet to establish long-term data preservation institutions. There is clearly an opportunity for certain countries or communities to take the lead, learning from successes and failures to date, in developing new institutions for the long-term stewardship of research data.

Policy and legal approaches

Policy and legal strategies can also have significant cost implications. Chapter One summarised the benefits of open research data. One benefit that was not mentioned, however, is the reduction in costs when one does not need to administer and enforce proprietary or other restrictive data access and use regimes (Houghton, 2011). For example, the enforcement of regulations for segregating and managing restricted data was cited as being the main cost for the European Bioinformatics Institute (EBI) repository (Apweiler, 2016).

There are, of course, legitimate and unavoidable restrictions on data, based on the protection of personal privacy, national security, and other confidentiality concerns. The management of such restrictions leads to different cost optimisation strategies for various disciplines and are part of the reason why one approach does not fit all.

6. CONCLUSIONS

The design and sustainability of research data repository business models depend on many factors. These include the role of the repository, national and domain contexts, the stage of the repository's development and lifecycle phase, the characteristics of the user community and the type of data product this community requires (influencing the level of investment required in curating and enhancing the data). All of these issues need to be considered in choosing and developing appropriate business models for research data repositories. There is no "one size fits all" solution.

Research data repositories are an essential part of the infrastructure for open science. Research data repositories provide for the long-term stewardship of research data, thus enabling verification of findings and the re-use of data. They bring considerable economic, scientific, and social benefits. Hence, it is important to ensure the sustainability of research data repositories.

Many research data repositories are largely dependent on public funding. The key policy question to be addressed is how this funding is most effectively provided - by what mechanism and from what source?

Fundamental economic principles can be seen underpinning the sustainability of research data repository business models, as well as their operational and funding policies. As data preservation and open data policies become increasingly widespread and influential, there will be more opportunities to develop deposit-side business models:

- *Regulation* in the form of policy mandates for open data provides a foundation for deposit-side models, and limit the potential for user-side funding models.

- *Incentives* in the form of structural or institutional funding, or funding to support the payment of deposit-side fees, provide a foundation for sustainable revenue streams for research data infrastructure.

As repository activities grow in scale and “the market” matures, there will be more opportunities to optimise the make versus buy decision, through sourcing from specialist service providers, and we can expect to see more data repository business models based on supply-side services targeting data depositors and repositories, and user-side, value-adding services of various sorts.

Research data repositories themselves can take advantage of the underlying economic differences between research data, which exhibit public good characteristics, and value-adding services and facilities, which typically do not, to develop business models that support free and open data while charging some or all users for access to value-adding services or related facilities.

There are advantages and disadvantages of various business models in different circumstances that can greatly affect data repository operations. It is important to consider what system of allocation will best ensure that the optimal level of funding will be made available for research data repositories. Project funding often provides a mechanism to test the need for a data repository and the initial capacity to create one. However, as the repository matures and scales to provide an ongoing, reliable and quality service, a different funding model is likely to be needed.

Research data repository business models are constrained by, and need to be aligned with, policy regulation (mandates) and incentives (including funding). “Un-funded mandates" (i.e. requesting open data without ensuring that researchers have the necessary facilities or means to comply) can be a problem. It is important to combine regulation and incentives thoughtfully to achieve best results.

A clearly articulated business model is indispensable for all research data repositories. Developing a successful business model involves understanding the phase of the repository's development, developing the product/service mix, understanding the cost drivers and matching revenue sources, identifying revenue sources and stakeholders, and making the value proposition to those stakeholders). These elements change throughout the repository's lifecycle. A successful business model has to align with a repository's mission(s) and be sensitive to the context in which it operates and these may also be subject to change over time.

The value proposition needs to be clear for different repository stakeholders, including funders. What someone is willing to pay for something depends on their perception of its value. For direct users of repository data and services the value is clear to them, and is revealed in their use. For stakeholders who are not direct users, but may be funders, it is more difficult for them to judge the value. Engaging and maintaining structural, institutional, philanthropic, or other funders depends on their understanding of the value proposition, and ensuring that engagement may involve repositories undertaking detailed benefit/cost, value and impact analyses.

The collection of good quality information to demonstrate the economic value and impact of data repositories to potential funders can be challenging. There is also a parallel in making the research or scientific case, which may require the collection of impact-related information (e.g. publication and citation counts), and the conduct of impact case studies.

Cost optimisation efforts can help ensure the effective and sustainable management of digital assets over time. Greater clarity concerning cost optimisation system-wide (throughout the whole data lifecycle) is important, rather than simply to focus on cost savings at the repository level. The effect a funding model has on cost constraint and monitoring the landscape for emerging opportunities are key in this regard.

Finally, taking advantage of economies of scale is also important. This may involve encouraging or funding the establishment of lead organisations at the national level and encouraging those organisations to collaborate globally, and encouraging or funding collaboration and federation. Not all research data repositories need to perform specialised curation and preservation tasks. Similarly, not all institutions or organisations need to create individual repositories. Collaboration and federation can help to manage, share, and reduce costs.

ENDNOTES

1. See https://en.oxforddictionaries.com/definition/business_model.

2. Within the OECD, relevant foundational documents include the OECD Principles and Guidelines for Access to Research Data from Public Funding (2007) and Recommendation on Public Sector Information (2008). Also directly relevant to this report was Making Open Science a Reality (2015).

3. See http://www.oecd.org/sti/ieconomy/data-driven-innovation.htm.

4. See http://www.nasa.gov/open/plan/peo.html.

5. See also http://www.ngdc.noaa.gov/geomag/crowdmag.shtml.

6. See, for example, http://www.goes-r.gov/education/students.html#NOAA or https://www.globe.gov/.

7. See, for example, (Vines et al, 2013), which surveyed 516 studies in ecology and found that data supporting 80 per cent of the articles were inaccessible. Another measure is the difficulty that all research institutions have in estimating the volumes of data that are stored by research groups because of the prevalent use of ad hoc systems and external storage.

8. The questionnaire is available at: http://bit.ly/business_models_questionnaire, which redirects the reader to https://docs.google.com/a/codata.org/forms/d/1OZZ5npEdgTd8JHNMiZYMrKB7usuWwqlpDUNfAvllfk8/viewform.

9. About SciElo: http://www.scielo.org/php/level.php?lang=en&component=42&item=8.

10. For further discussion of these issues see (Parsons and Fox, 2013).

11. About ArXiv: https://arxiv.org/help/general.

12. The context for this study is the increasing demand for research data curation and sharing driven, in part, by the widespread adoption of open data policies. Hence, we forego any discussion of the other characteristics of public goods (e.g. excludability).

13. For example, the Open Library of Humanities is funded by a non-pay wall membership model. See https://about.openlibhums.org/libraries/supporting-institutions/.

14. See https://en.wikipedia.org/wiki/Preference_(economics).

15. See https://en.oxforddictionaries.com/definition/business_model.

16. See http://www.knowledgeunlatched.org/.

17. See https://scoap3.org/.

18. This latter can also take the form of what some workshop participants described as research infrastructure negotiated as a line-item in research budgets.

19. See https://en.wikipedia.org/wiki/Open_Archival_Information_System.

20. There are many metadata standards: the most mature and widely used (e.g. Dublin Core http://dublincore.org/, Darwin Core http://rs.tdwg.org/dwc/ and DDI http://www.ddialliance.org/) are looked after by foundations of one sort or another. Each of these foundations provides an essential component in the open data ecosystem.

21. See Kryder's Law: https://en.wikipedia.org/wiki/Mark_Kryder.

REFERENCES

Apweiler, R. (2016), *EMBL-EBI: Providing Bioinformatics Research Infrastructure for the Life Sciences*, Presentation at the Workshop on Revenue Sources and Cost Optimisation, 3-4 November 2016.

Arzberger, P. et al. (2004), "An International Framework to Promote Access to Data", *Science.* Vol. 303, pp. 1777-1778. March 19, 2004, AAAS, Washington D.C.

Baker, M. (2012), *Databases fight funding cuts, Online tools are becoming ever more important to biology, but financial support is unstable*, http://www.nature.com/news/databases-fightfunding-cuts-1.11347.

Barrington, L. (2014), *Crowdsourcing Satellite Imagery*, Earth Imaging Journal, http://eijournal.com/print/articles/crowdsourcing.

Beagrie, N. et al. (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. et al. (2008), *Keeping Research Data Safe*, Joint Information Systems Committee (JISC), London and Bristol. http://www.beagrie.com/krds.php.

Beagrie, N. and Houghton, J.W. (2016), *The Value and Impact of the European Bioinformatics Institute*, Full Report to EMBL-EBI by Charles Beagrie Limited, January 2016, http://www.beagrie.com/EBI-impact-report.pdf.

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*, JISC, Bristol and London. http://repository.jisc.ac.uk/5568/.

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.

Benkler, Y. (2006), *The Wealth of Networks—How Social Production Transforms Markets and Freedom*, Yale University Press, New Heaven and London, http://www.benkler.org/Benkler_Wealth_Of_Networks.pdf.

Berman, F. and Cerf, V. (2013), "Who will pay for public access to research data?", *Science, Vol 341*, pp. 616-617, 9 August 2013,

http://www.sciencemag.org/content/341/6146/616.full.pdf?keytype=ref&siteid=sci&ijkey=.e2Ezowko%2FxF2.

Bicarregui, J. (2016), *Building and sustaining data infrastructures: putting policy into practice*, in Report to the Wellcome Trust, https://dx.doi.org/10.6084/m9.figshare.4055538.

BRTF (2010), *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Bousfield, D. et al. (2016), *Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources*, F1000Research 2016, 5 (ELIXIR):160, https://f1000research.com/articles/5-160/v1.

Cisco (2016), *The Zettabyte Era: Trends and Analysis*, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf.

CODATA (2015), *The Value of Open Data Sharing*, Geneva: Group on Earth Observations, Geneva. http://www.earthobservations.org/documents/dsp/20151130_the_value_of_open_data_sharing.pdf.

Dekkers, M. et al. (2006), *MEPSIR: Measuring European Public Sector Information Resources*, European Commission, Brussels, http://www.epsiplus.net/psi_library/reports/mepsir_measuring_european_public_sector_resources_report.

Diepenbroek, M. (2016), *Federated Data Infrastructures*, Presentation at the Workshop on Revenue Sources and Cost Optimisation, 3-4 November 2016.

Dillo, I. (2016), Overview of Current Work on Repository Cost Analysis, Presentation at the Workshop on Revenue Sources and Cost Optimisation, 3-4 November 2016.

Doorn, P. et al. (2013), "Lies, Damned Lies and Research Data: Can Data Sharing Prevent Data Fraud?" *International Journal of Digital Curation* Vol. 8, No. 1, 2013, http://www.ijdc.net/index.php/ijdc/article/view/8.1.229.

DotEcon (2006), *The commercial use of public information (CUPI)*, http://www.oft.gov.uk/OFTwork/publications/publication-categories/reports/consumer-protection/oft861.

Ember, C. and Hanisch, R. (2013), *Sustaining Domain Repositories for Digital Data: A White Paper*. Output of the workshop, "Sustaining Domain Repositories for Digital Data," Ann Arbor, MI, June 24-25, 2013. Ann Arbor, MI: Inter-University Consortium for Political and Social Research, http://dx.doi.org/ 10.3886/Sustaining Domain Repositories Digital Data.

Hey, T. et al. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

Houghton J.W. (2011), *Costs and Benefits of Data Provision*, Australian National Data Service, Canberra. http://ands.org.au/resource/cost-benefit.html.

Kalff, F.E. et al. (2016), "A kilobyte rewritable atomic memory", *Nature Nanotechnology* 11:926-929, 18 July 2016, Macmillan Publishers Limited, http://dx.doi.org/10.1038/nnano.2016.131.

Kitchin, R. et al. (2015), *Funding models for Open Access Repositories*, Maynooth: Maynooth University. Dublin: the Royal Irish Academy and Trinity College Dublin. http://dx.doi.org/10.3318/DRI.2015.4.

Lauro, F. M., et al. (2014), *The Common Oceanographer: Crowdsourcing the Collection of Oceanographic Data*, PLoS Biol 12(9), 9 September 2014. http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001947.

National Research Council (NRC) (2015) *Preparing the Workforce for Digital Curation.* National Academies Press, Washington, D.C., http://www.nap.edu/catalog/18590/preparing-the-workforce-for-digital-curation.

NRC (2012a) *The Future of Scientific Knowledge Discovery in Open Networked Environments*, Paul F. Uhlir, rapporteur, National Academies Press, Washington, D.C., https://www.nap.edu/catalog/18258/the-future-of-scientific-knowledge-discovery-in-open-networked-environments.

NRC (2012b) *The Case for International Sharing of Scientific Data: A Focus on Developing Countries.* Kathie Bailey Mathae and Paul F. Uhlir, eds., National Academies Press, Washington, D.C., https://www.nap.edu/catalog/17019/the-case-for-international-sharing-of-scientific-data-a-focus.

NRC (2009a) *The Socioeconomic Effects of Public Sector Information on Digital Networks: Towards a Better* Understanding *of Different Access and Reuse Policies*, National Academies Press, Washington, D.C., http://www.nap.edu/openbook.php?record_id=12687&page=40.

NRC (2009b), *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. National Academies Press, Washington, D.C., http://www.nap.edu/catalog/12615/ensuring-the-integrity-accessibility-and-stewardship-of-research-data-in-the-digital-age.

NRC (2004) *Open Access and the Public Domain in Digital Data and Information for Science.* National Academies Press, Washington, D.C., http://www.nap.edu/catalog/11030/open-access-and-the-public-domain-in-digital-data-and-information-for-science.

NRC (2003) *The Role of Scientific and Technical Data and Information in the Public Domain.* National Academies Press, Washington, D.C.,. http://www.nap.edu/catalog/10785/the-role-of-scientific-and-technical-data-and-information-in-the-public-domain.

NRC (2002) *Down to Earth: Geographic Information for Sustainable Development in Africa*, National Academies Press, Washington, D.C., https://www.nap.edu/catalog/10455/down-to-earth-geographical-information-for-sustainable-development-in-africa.

NRC (1999) *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. National Academy Press, Washington, D.C., http://www.nap.edu/catalog/9692/a-question-of-balance-private-rights-and-the-public-interest.

NRC (1997), *Bits of Power: Issues in Global Access to Scientific Data*. National Academy Press, Washington, D.C. http://www.nap.edu/catalog/5504/bits-of-power-issues-in-global-access-to-scientific-data.

Nelson, J. S. (2011), *U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center—fiscal year 2010 annual report*. U.S. Geological Survey Open-File Report 2011–1057. http://pubs.usgs.gov/of/2011/1057/pdf/of2011-1057.pdf.

NSB (2005), *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*, National Science Foundation. https://www.nsf.gov/pubs/2005/nsb0540/.

OECD (2015), "Making Open Science a Reality", *Science, Technology and Industry Policy Papers*, No. 25, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jrs2f963zs1-en.

OECD (2008), *OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*, http://www.oecd.org/sti/44384673.pdf.

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

PIRA (2000), *Commercial exploitation of Europe's public sector information*, European Commission, Brussels, http://www.epsiplus.net/psi_library/reports/commercial_exploitation_of_europe_s_public_sector_information_pira_study.

Parsons, M.A. and Fox, P.A. (2013), "Is Data Publication the Right Metaphor?", *Data Science Journal*, Vol. 12, pp.WDS32–WDS46, http://doi.org/10.2481/dsj.WDS-042.

Pollock, R. (2009), *The Economics of Public Sector Information*, CWPE 0920, University of Cambridge, Cambridge, http://www.econ.cam.ac.uk/dae/repec/cam/pdf/cwpe0920.pdf.

RDA-CODATA Interest Group on the Legal Interoperability of Research Data (2016), Principles and Implementation Guidelines for the Legal Interoperability of Research Data, http://dx.doi.org/10.5281/zenodo.162241.

RDA-WDS Working Group on Cost Recovery for Data Centres, (2016), *Income Streams for Data Repositories*. Zenodo, https://doi.org/10.5281/zenodo.46693.

Robbins, J. (2013), "Crowdsourcing for the birds" in the *New York Times*. August 19, 2013. http://www.nytimes.com/2013/08/20/science/earth/crowdsourcing-for-the-birds.html?pagewanted=all&_r=0.

Uhlir, P.F. (2006), "The Emerging Role of Open Repositories as a Fundamental Component of the Public Research Infrastructure", in *Open Access: Open Problems*, Sica, G. ed., Polimetrica, http://eprints.rclis.org/9656/1/OpenAccess.pdf.

Uhlir, P.F. (2004), *Policy Guidelines for the Development and Promotion of Governmental Public Domain Information.* UNESCO. http://portal.unesco.org/ci/en/ev.php-URL_ID=15863&URL_DO=DO_TOPIC&URL_SECTION=201.html.

Uhlir, P.F. and Schröder, P. (2007), "Open Data for Global Science", *Data Science Journal*, Vol. 6, 17 June 2007, pp. 36-53, Ubiquity Press, London, http://datascience.codata.org/articles/abstract/10.2481/dsj.6.OD36/.

Vickery, G. (2010), *Review of recent studies on PSI re-use and related market developments*, European Commission, Brussels. http://www.paikkatietoikkuna.fi/c/document_library/get_file?uuid=b1ad5545-266e-4e1b-8970-b855dbcbf997&groupId=108478.

Vines T. H. et al. (2013), *The availability of research data declines rapidly with article age*, https://www.ncbi.nlm.nih.gov/pubmed/24361065.

Weiss, P. (2002), *Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts*, National Oceanic and Atmospheric Administration.

APPENDIX A.
OECD-GSF/ICSU-CODATA EXPERT GROUP

Country/ nominating organisation	Name	Affiliation
Australia	Andrew Treloar	Australia National Data Service and co-chair of RDA Technical Advisory Board
European Commission	Wainer Lusoli	EC, DG Research and Innovation
Finland and RDA/CODATA	Irina Kupiainen	CSC - IT Centre for Science, Espoo, Finland
France and CODATA/WDS	Mark Asch	Professor, Université de Picardie Jules Verne *He initially participated in the Expert Group from ANR.*
Germany and CODATA/ WDS	Michael Diepenbroek	Centre for Marine Environmental Sciences, Alfred Wegener Institute for Polar and Marine Sciences, Bremen, Germany
Japan	Kazuhiro Hayashi	Senior Research Fellow National Institute of Science and Technology Policy (NISTEP), Japan
Korea	Seo-Young Noh	GSDC (Global Science Data Center), KISTI (Korea Institute of S&T Information)
Korea	Sun Kun Oh	Professor of Physics, Konkuk University
Netherlands and CODATA/ WDS	Ingrid Dillo (co-chair)	Director, Policy Development, Data Archiving and Networked Services, Netherlands.
Norway	Roar Skålin	Norwegian Meteorological Institute (MET Norway), NORWAY
South Africa	Happy Sithole	Director, Center for High Performance Computing, Council for Scientific and Industrial Research (CSIR)
South Africa and CODATA/ WDS	Martie van Deventer	Manager of Information Services, Council for Scientific and Industrial Research,
Switzerland	André Golliez	Open Data Consulting, Impact Hub Zurich
Switzerland	Peter Grolimund	Tetradata
Switzerland	Philippe Cudre-Mauroux	Dept. of Informatics, University of Fribourg
USA	Allen Dearry	Director Office of Scientific Information Management National Institute of Environmental Health Sciences National Institutes of Health
CODATA	Simon Hodson (co-chair)	Executive Director, ICSU-CODATA, Paris, France
CODATA	Cameron Neylon	Based at Bath, UK (Centre for Culture and Technology, Curtin University, Australia)
RDA (India)	Devika P. Madalli	Professor, Documentation Research and Training Center (DRTC), Indian Statistical Institute (ISI)
RDA (USA)	Francine Berman	Computer Science Rensselaer Polytechnic Institute
RDA Europe (Ireland)	Natalie Harrower	Director Digital Repository Ireland
Consultant	*John Houghton*	*Australia*
Consultant	*Paul Uhlir*	*Washington DC*

APPENDIX B.
GLOSSARY OF KEY TERMS

Business model. "A plan for the successful operation of a business, identifying sources of revenue, the intended customer [stakeholder] base, products, and details of financing." (https://en.oxforddictionaries.com/definition/business_model)

Cost Optimisation. "An act, process, or methodology of making the cost of something as fully perfect, functional, or effective as possible." (Adapted from https://www.merriam-webster.com/dictionary/optimization)

Cost Restraint (or Reduction). A way of limiting, controlling, avoiding, or stopping an expense, now or in the future.

Curation. "The activity of managing and promoting the use of data from their point of creation to ensure that they are fit for contemporary [and future] purpose and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep them fit for purpose. Higher levels of curation will also involve links with annotation and with other published materials." (http://dictionary.casrai.org/Curation)

Data. Data are "[f]acts, measurements, recordings, records, or observations about the world collected [or generated] by scientists and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records." (http://dictionary.casrai.org/Data)

Financial sustainability. Financial sustainability is the ability of an organisation—whether public or private, or for-profit or not-for profit—to support a defined level of activity indefinitely.

Open Data. Open data and content can be freely used, modified, and shared by anyone for any purpose. (http://opendefinition.org).

Open Science. Open science commonly refers to efforts to make the output of publicly funded research more widely accessible in digital format to the scientific community, the business sector, or society more generally. (http://www.oecd-ilibrary.org/science-and-technology/making-open-science-a-reality_5jrs2f963zs1-en).

Operational Data. Data (see the definition of "data" in this Glossary) that are used primarily for supporting an official mission of an organisation, but also have research (and other) uses. For example, meteorological data are used by government ministries that generate them to issue weather forecasts and help protect the public safety, but the data may be repurposed for many other applications, including atmospheric studies and research on climate change.

Private good. A private good can be both depleted (consumed or used) and excluded (kept separately from others). Most tangible objects have private good characteristics.

Public good. A public good has two attributes: it cannot be exhausted in its use (non-depletable) and cannot be kept from others (non-excludable) (Wikipedia 2015 at https://en.wikipedia.org/wiki/Public_good). The warnings from a lighthouse are often cited as being a pure public good, but there are few goods or services that have both aspects in their entirety.

Quasi-public good. Data and information have "quasi-public good" characteristics. That is, they can be used without diminution, since an exact copy can be made and shared while the original remains intact, but they can be excluded, since it is possible to keep them restricted or even secret.

Repository. "Repositories preserve, manage, and provide access to many types of digital materials in a variety of formats. Materials in online repositories are curated to enable search, discovery, and reuse. There must be sufficient control [of the ongoing activity] for the digital material to be authentic, reliable, accessible and usable on a continuing basis." (http://dictionary.casrai.org/Repository)

Research Data. Data (see the definition of "data" in this Glossary) that are either produced in the course of research, or that are produced primarily for other applications, but also have significant uses in research.

Revenue (or Income) Source. The money that flows into an organisation from one or more sources, usually accounted for on an annual basis. An aggregate revenue source is the financial support from all sources, including in-kind resource contributions.

SWOT. This is an acronym that stands for "Strengths, Weaknesses, Opportunities, and Threats", an analytical method that is often used for structured planning.

APPENDIX C.
LIST OF REPOSITORIES INTERVIEWED

	Acronym	Data Repository	Country	Discipline/domain	URL
1		3TU.Datacentrum	Netherlands	Technical Science	http://researchdata.4tu.nl/en/home/
2	ADS	Archaeology Data Service	UK	Archaeology	http://archaeologydataservice.ac.uk/
3	CCDC	Cambridge Crystallographic Data Centre	UK	Molecular crystallography	http://www.ccdc.cam.ac.uk/
4	CSC	Centre for Scientific Computing	Finland	Generic	https://www.csc.fi/
5	DANS	Data Archiving and Networked Services	Netherlands	All / Social Science	http://www.dans.knaw.nl/nl
6		DataFirst	South Africa	Social Science	https://www.datafirst.uct.ac.za/
7		Diamond Archive	UK	Synchrotron	
8	DRI	Digital Repository of Ireland	Ireland	Humanities	http://www.dri.ie/
9		Dryad Data Repository	USA	General	http://datadryad.org/
10	ECDS	Environment Climate Data Sweden	Sweden	Climate Data	https://ecds.se/about
11	CERN	European Organization for Nuclear Research	Switzerland	High-energy Physics	http://home.cern/
12		figshare	USA	General	https://figshare.com/
13		Harvard Dataverse	USA	General	https://dataverse.harvard.edu/
14	ICPSR	Inter-university Consortium for Political and Social Research	USA	Social Science	http://www.icpsr.umich.edu/icpsrweb/landing.jsp
15	IEDA	Integrated Earth Data Applications	USA	Earth Science	http://www.iedadata.org/
16	NITRC	Neuroinformatics Tools and Resources Clearinghouse	USA	Neuroimaging	https://www.nitrc.org/include/about_us.php
17	NIF	Neuroscience Information Framework	USA	Neuroscience	https://www.neuinfo.org/

	Acronym	Data Repository	Country	Discipline/domain	URL
18	SAEON shared data platform	South African Environmental Observation Network shared data platform	South Africa	Environmental Observation	http://data.saeon.ac.za/about
19	SEDAC	Socioeconomic Data and Applications Center	USA	Earth and Social Science	http://sedac.ciesin.columbia.edu/
20	UKDA	UK Data Archive	UK	Social Science	http://www.data-archive.ac.uk/
21	WDCC/DKRZ	World Data Center for Climate / German Climate Computing Center	Germany	Climate Data	https://www.dkrz.de/daten-en/wdcc
22	WDC-RRE	World Data Center for Renewable Resources and Environment	China (People's Republic of)	Geoscience	http://eng.geodata.cn:8080/geonetwork/srv/en/main.home
23	AERC - NIPR	Arctic Environment Research Center - National Institute of Polar Research	Japan	Polar research (Earth Science)	http://www.nipr.ac.jp/aerc/e/index.html
24	ADA	Australian Dat Archive	Australia	Social Science	https://www.ada.edu.au/
25	CR2	Center for Climate and Resilience Research	Chile	Climate	http://explorador.cr2.cl/
26	DARIS - FORS	Data and research information services - Foundation for Research in Social Sciences	Switzerland	Social Science	www.forscenter.ch
27		ETH Data Archive	Switzerland	General	http://www.library.ethz.ch/Digital-Curation
28	GSJ - AIST	Geological Survey of Japan - National Institute of Advanced Industrial Science and Technology	Japan	Geological survey (Earth Science)	https://www.gsj.jp/en/index.html
29	GTN-P	Global Terrestrial Network for Permafrost	International	Earth Science	http://gtnp.arcticportal.org/about-the-gtnp
30	HSRC Research Data	Human Sciences Research Council Research Data	South Africa	Social science and Humanities	http://datacuration.hsrc.ac.za/search/browse
31	IRIS	Incorporated Research Institutions for Seismology	US	Seismology	https://www.iris.edu/hq/
32	LSDB Archive	Life Science Database Archive	Japan	Life Science	http://dbarchive.biosciencedbc.jp/index-e.html
33		Mendeley Data	UK	General	https://data.mendeley.com

	Acronym	Data Repository	Country	Discipline/domain	URL
34	NIO	National Institute of Oceanography	India	Oceanography	http://www.nio.org/index/title/HOME
35		Ontario Data Catalogue	Canada	General	https://www.ontario.ca/search/data-catalogue
36		Open data Kenya	Kenya	General	https://www.opendata.go.ke/browse
37		Physiome Project Repositories		Physiology	http://physiomeproject.org/data-repositories/about
38	PURR	Purdue University Research Resource	US	General Institutional	https://purr.purdue.edu/
39	RKMP	Rice Knowledge Management Portal	India	Agriculture	http://ring.ciard.net/rice-knowledge-management-portal
40		Scholar's Portal / University of Toronto	Canada	General Institutional	http://www.scholarsportal.info/
41	SADA	South African Data Archive	South Africa	General	http://sada.nrf.ac.za/
42	CDS	Strasbourg astronomical Data Center	France	Astronomy	http://cdsweb.u-strasbg.fr/
43	SFA	Swiss Federal Archives	Switzerland	General	https://www.bar.admin.ch/bar/en/home.html
44	SIB	Swiss Institute of Bioinformatics	Switzerland	Bioinformatics	http://www.sib.swiss/
45	TAIR	The Arabidopsis Information Resource	US	Plant biology	https://www.arabidopsis.org/
46	UCT eResearch	University of Cape Town eResearch	Australia	General	www.eresearch.uct.ac.za
47		World Data Center for Geomagnetism, Kyoto*	Japan	Geomagnetic Data (Earth Science)	http://wdc.kugi.kyoto-u.ac.jp/index.html
48		Zenodo	CERN, Switzerland	Multidisciplinary	https://zenodo.org/

(*) Interview with the repository was conducted after the others, and it is not included in the aggregated statistical analysis, although the information is taken into account in the qualitative analysis.

APPENDIX D.
A SWOT ANALYSIS OF FUNDING SOURCES

The following boxes present a summary of the Strengths, Weaknesses, Opportunities and Threats (SWOT) of the major funding sources identified in this study. This analysis was carried out in collective exercises with participants including mainly repository managers.

<table>
<tr><th colspan="2">Structural Funding</th></tr>
<tr><td>Strengths
Longer-term stability enables easier planning.
Tend to be large scale, which can encourage efficiency (economies of scale).
Is compatible with open data principles.
Stronger commitments and communication with stakeholders.
Larger chunk of investments can cover operational costs (core functions can be funded).
Up front funding can help plan budget and build effective organisation and can be more flexible as to its internal allocation to activities.
Largely immune to market and collateral effects, and less affected by vagaries of scientific fashion.
No need to spend too much time fundraising, as single or few points of contact.</td><td>Weaknesses
Fixed multi-year funding does not scale easily and is a weakness in the context of rapidly growing demand and volumes of data.
Could reduce efficiency, as there may be little incentive to improve (long evaluation cycles make you lazy!), and market forces are weak.
Inflexibility of funding, making it difficult for a repository to adapt easily to changing needs.
Competing with research for funds, with possibly different priorities, and may put researchers off-side, if perceived as reducing their research funding.</td></tr>
<tr><td>Opportunities
Data is a hot topic and funders are more amenable to providing structural funding.
Data can be recognised as infrastructure, and many funders appear to have an increasing budget for infrastructure.
Stay in touch with principal funder with links to research developments and priorities.
Flexibility as to allocation and spending.</td><td>Threats
Increase demand for data curation cannot be handled easily.
Not receiving structural funding because of big national initiatives with which you are not aligned (“Today it’s hot, tomorrow it’s not!”).
May be terminated with external decisions, and funder itself may be cut.
Not in control of your funding, and dependent on single or small number of sources.
The more public funding is guaranteed (e.g. by law) the less incentive there is to innovate.</td></tr>
</table>

Host or Institutional Support	
Strengths • Sustainability (universities tend to stay around for a long time). • Convergence of interest between host and repository (could also be a weakness, depending on the function of the repository). • Is compatible with open data principles. • The host lends the repository a strong identity, position and visibility. • Efficiency through sharing services in the university and an incentive to reduce costs. • Encourages collaboration with other universities to reduce costs and share expertise. • Close to the researchers for on hand curation and interventions.	**Weaknesses** • Challenge to get institutional commitment. • Dependent on the level of funding of the host institution, with possible resourcing uncertainties. • Host funding may be dependent on individuals within the host institution. • Limited purview, with focus on the local communities that need to be served. • May lead to fragmentation of domain data and lower interoperability.
Opportunities • Data increasingly seen as an asset by universities. • Brand opportunities for institutions (e.g. the Inter-university Consortium for Political and Social Research at the University of Michigan). • Data repository can be part of the institution strategy for research excellence, and may support collaboration with industry. • Possibilities to optimise cost by making use of university facilities. • When backed by the host there is more chances of success when looking for funding elsewhere. • Alignment and grant possibilities based on the high profile. • Repository can train local researchers in data science (i.e. training in data skills).	**Threats** • Cuts in funding of host institution may lead to funding insecurity, if depending on single source. • Divergence of interest between host and repository. • Universities are increasingly outsourcing all kinds of infrastructure and services, and seeking to divest themselves of non-core activities. • Universities concerned about services as revenue sources, and may try to use repository to raise revenue. • Potential lack of interoperability if multiple institutions set up their own repositories. • Researchers may find discipline specific, specialised repositories more attractive, and it is important not to duplicate work of specialist repositories. • Justification – who benefits, who pays? How to demonstrate the Return on Investment.

Annual Contract	
Strengths • Is compatible with open data principles. • Demand oriented, not supply oriented. • Funding can scale to data volumes. • Can be flexible and adaptive. • Can fill gaps for homeless data, support new journal publication policies, and possibly data journals.	**Weaknesses** • Unpredictability year to year. • Relatively high transaction costs in managing contracts. • May lead to limited curation and documentation as incentive to cut costs to lower subscription costs. • Limited engagement with researchers, as users or depositors. • Can be too narrow (e.g. creating narrow solutions for one client that may not be useful for others).
Opportunities • Similar to structural funding opportunities. • Can create new services, be adaptive and responsive to needs. • May identify new funding sources and models.	**Threats** • Larger threat of termination that is not controllable by the repository. • Data lose context and are separated from community of researchers. • If a number of players try to enter the field it may lead to fragmentation and loss of scale economies. Conversely it could also lead to monopolies.

Data Deposit Fees	
Strengths • Fits the funding models of research, costs are incurred by the project. • Part of the funding for research, and needs to be incorporated into the research proposals (e.g. data management plans). • Is compatible with open data principles. • Scalable as ties cost directly to activities. • Closely linked to the research community, responsive to research needs. • Price sensitivity so ensures cost constraint (but unlikely that there will be a genuine market).	**Weaknesses** • Barriers to data deposit and disincentive to deposit. • Challenge of costing - very difficult to price. • Administrative overheads in payment transactions.
Opportunities • Establishes a market where there are incentives to do the work upstream. • Additional services can be covered by the fee. • Included in the research project, and encourages funders to look at the full cost of research.	**Threats** • Researcher pushback (e.g. top-slicing research grant). • Rush to cheapest option (race to the bottom), unless supported by legislation. Needs very clear policy framework. • May lead to low levels of curation to contain costs, or high curation levels with high deposit fees for a niche market if researchers do not see the money as part of their funding. • May be difficult to compete for deposits against comparable repositories that do not charge deposit fees.

Access Charges	
Strengths • Consumer pays for what consumer wants, assuming their capacity to pay. • Income reflects value of data product to users who can pay. • If based on membership/subscription fees: it can provide a stable and predictable income and generate loyalty from the community of members, who have influence over priorities. • Model can accommodate licensing flexibility.	**Weaknesses** • Not compatible with open data principles. • Not affordable for people at poorer organisations, so unequal access. • Charges limit use and are welfare reducing. • Revenue scales with use, not with data ingest and curation costs. • Causes competition amongst repositories for high impact data, data poaching, and may limit possibilities to curate lower impact data. • Must think carefully about licensing terms and conditions, which can be expensive to manage, limit use and be welfare reducing.
Opportunities • Can monitor demand, change/improve services (responsive). • Possible to create free access for basic products, build consumer base and create demand for value-added services. • Most market-oriented approach, so may provide cost constraint.	**Threats** • Vulnerable to economic downturns and funding cuts. • Expectation is for free access, so other providers might undermine business.

Contract Services and project funding	
Strengths • Can support innovation and development. • Can increase contact between staff and clients.	**Weaknesses** • Not sustainable as the sole source of revenue. • Time limited, relatively short term and less flexible in allocation. • Less predictable year on year. • Less stable revenue that other sources, so difficult to plan. • Relatively high transaction cost, chasing money. • Can take attention away from core repository tasks.
Opportunities • For further, future innovation and development.	**Threats** • If services are to private sector, results may not be usable by others (e.g. one-of-a kind, proprietary solutions). • May lose expertise and divert best people from core functions.

Diversification of Revenue Sources	
Strengths • No single point of failure. • Flexibility to experiment with new services and markets. • Stimulates innovation which may not be covered by core funding.	**Weaknesses** • May lead to relatively high transaction costs (e.g. chasing contract money). • Research and contracting requires highly skilled staff, who may be diverted from core functions.
Opportunities • Opportunities to explore new revenue sources.	**Threats** • Mission Creep

APPENDIX E.
INVITED PARTICIPANTS AT PROJECT WORKSHOPS*

Workshop on Revenue Sources and Cost Optimisation
Paris, 3-4 November 2016

Name	Affiliation
Rolf Apweiler	European Bioinformatics Institute (EBI), UK
Kevin Ashley	Digital Curation Centre (DCC), Scotland
Magchiel Bijsterbosch	SURFsara, Netherlands
Ian Bruno	Cambridge Crystallographic Data Centre (CCDC), UK
Len Fishman	data.world, USA
Eva Huala	TAIR, USA
Mustapha Mokrane	ICSU-WDS International Programme Office
Mark Parsons	Research Data Alliance (RDA)
Jamie Shiers	European Organization for Nuclear Research (CERN)
Andrew Smith	Elixir, Switzerland
Dan Valen	FigShare
Matthew Viljoen	EGI, Netherlands

*Expert Group members, consultants and OECD staff (see appendix A) participated also in both of the workshops.

Workshop on Sustainable Business Models for Data Repositories
Brussels, 28-29 March 2017

Name	Affiliation
Kevin Ashley	Digital Curation Centre (DCC), Scotland
Grace Baynes	Springer Nature, UK
Neil Beagrie	Charles Beagrie Limited, UK
Niklas Blomberg	Elixir, UK
Ian Bruno	Cambridge Crystallographic Data Centre (CCDC), UK
Ron Dekker	Consortium of European Social Science Data Archives (CESSDA), Norway
Koenraad De Smedt	Common Language Resources and Technology Infrastructure (CLARIN), Norway
María Guillermina D'Onofrio	Ministry of Science, Technology and Productive Innovation, Argentina
Kylie Emery	Australian Health Research Council
Chiara Gabella	European life-science infrastructure for biological information (Elixir)
Josh Greenberg	Sloan Foundation, USA
Mark Hahnel	FigShare, USA
Robert Hanisch	National Institute of Standards and Technology (NIST), USA
Bjorn Henrichsen	Norwegian Centre for Research Data (NSD)
Wolfram Horstmann	Göttingen State and University Library, Germany
Tibor Kalman	Digital Research Infrastructure for the Arts and Humanities (DARIAH)
Robert Kiley	Wellcome Trust, UK
Damien Lecarpentier	EUDAT Collaborative Data Infrastructure
Hyungjin Lee	Korea Institute of Science and Technology Information (KISTI)
Mark Leggott	Research Data Canada
Natalia Manola	OpenAIRE
Mihaela Meresi	European Commission
William Michener	University of New Mexico, DataONE and Dryad
William Miller	Office of Advanced Cyberinfrastructure, US National Science Foundation
Mustapha Mokrane	ICSU-WDS International Programme Office
Martin Moyle	University College London
Dale Peters	ICT eResearch, University of Cape Town
Benjamin Pfeil	Ocean carbon data centre, Norway
Eva Podgorsek	European Commission
Steven Ramage	Global Earth Observation System of Systems (GEOSS)
Robert Samors	Belmont Forum
Paul Stokes	UK Joint Information Systems Committee (Jisc)
Jorge Tezon	CONICET, Argentina
Matthew Woollard	UK Data Archive