Jisc

March 2014

The Value and Impact of Data Sharing and Curation

A synthesis of three recent studies of UK research data centres

The Value and Impact of Data Sharing and Curation

A synthesis of three recent studies of UK research data centres

Neil Beagrie
(Charles Beagrie Ltd)

and

John Houghton
(Centre for Strategic Economic Studies, Victoria University)

Contents

Executive summary

In the UK, substantial resources are being invested in the development and provision of services for the curation and long-term preservation of research data. It is a high priority area for a range of stakeholders, universities, researchers and research funders. There is strong interest in establishing the value and sustainability of these investments.

Although a number of studies have looked at methods for determining cost-benefit and broad indicators of value for research data sharing, there remain significant challenges. Only a relatively small number of socio-economic studies have focused specifically on the impact of research data sharing, or research data infrastructure. Moreover, their results have largely concentrated on qualitative indicators rather than quantification of value in economic terms.

This synthesis aims to summarise and reflect on the combined findings from a recent series of independent investigations, produced by the same authors, into the value and impact of three well established UK research data centres or services.[1] Its intended audiences are those interested in a brief overview of the key findings and lessons from the series as a whole. It provides a summary of the key findings and reflects on:

- The methods that can be used to collect data for such studies (Section 2.1);
- The analytical methods that can be used to explore value, impacts and benefits (Section 2.2);
- The measurable value, impacts and benefits of the research data centres and the research data curation and sharing that they support (Section 3); and
- The lessons learnt (Section 4) and recommendations arising (Section 5) from the series of studies as a whole.

The studies covered the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS), and the British Atmospheric Data Centre (BADC).[2] Each report was commissioned independently and at different times over a period of two years by the Economic and Social Research Council (ESRC), Jisc, and Jisc and the Natural Environment Research Council (NERC), respectively. There are, therefore, differences in the studies arising from varying requirements, timing, and levels of funding. Readers should refer to the individual study reports for greater detail on the context and findings from each data centre.

Nevertheless, all three studies combined quantitative and qualitative analytical approaches in order to quantify value and impacts in economic terms and explore other, non-economic benefits. Uniquely, the studies cover both users and depositors of data, and we believe the surveys of depositors that we have undertaken are the first of their kind.

Our key findings from the series of studies follow.

The economic analysis indicated that:

- Very significant increases in research, teaching and studying efficiency were realised by the users as a result of their use of the data centres;
- The value to users exceeds the investment made in data sharing and curation via the centres in all three cases; and

1 Elsewhere in the report, for simplicity, we use the generic term 'data centres' to refer to the group.
2 All three reports are available online: ESDS (**study page**) (**report PDF**), ADS (**study page**) (**report PDF**), BADC (**study page**) (**report PDF**).

- By facilitating additional use, the data centres significantly increase the measurable returns on investment in the creation/collection of the data hosted.

The qualitative analysis indicated that:

- Interviewees underlined the value seen by users and depositors in the data centres. Overall feedback was very positive. There were also some constructive suggestions about where improvements could be made;
- Surveyed academic users reported that use of the centres was very or extremely important for their academic research. A majority of respondents (between 53% and 61% across the three surveys) reported that it would have a major or severe impact on their work if they could not access the data and services; and
- For surveyed depositors, having the data preserved for the long-term and its dissemination being targeted to the academic community were seen as the most beneficial aspects of depositing data with the centres.

A unique feature of the ADS Impact Study was the inclusion of an analysis of the evolving, cumulative value of the centre. That analysis indicated that the value of ADS data and services has increased as the collection has grown and the service has developed. This is an important reminder that the maturity of collections and services at the date of evaluation may be an important factor to consider in future studies of less established centres.

Overall, the three studies show the benefits of integrating a range of quantitative economic approaches to measuring the value and impacts of research data archiving and sharing, with qualitative approaches exploring user perceptions and wider dimensions of value.

An important aim of each study has been to contribute to the further development of impact evaluation methods that can provide estimates of the value and benefits of research data sharing and curation infrastructure investments. This involved the use of a number of methods that we developed in the light of experience over the course of the series of studies. In this synthesis, we reflect on accumulated lessons learnt and provide a set of recommendations that could help develop future studies of this type.

1. Introduction

1.1 Scope and context

The three studies were undertaken in very different disciplinary fields and the research data centres we studied offer a wide spectrum of services across a range of data types and user communities. The studies covered the Economic and Social Data Service (ESDS), the Archaeology Data Service (ADS), and the British Atmospheric Data Centre (BADC). Each report was commissioned independently and at different times over a period of two years by the Economic and Social Research Council (ESRC), Jisc, and Jisc and the Natural Environment Research Council (NERC), respectively. Differences in services and data collections, and tailoring in application of methods for the study of specific data centres, mean that this synthesis cannot and should not be used as a comparison between the centres. Rather it is intended to draw out generic findings and lessons from the results and application of the methods.

The studies focus primarily on the value for, and impact on, the user and depositor communities of each data centre. These largely, but not exclusively, consist of researchers, with the data centres supporting academic research and, to a lesser extent, academic teaching and study. Although some of the secondary use and wider impacts on society arising from research based on data from the centres are reflected in our analysis, they are not directly measured. However, their possible measurable value over time is explored in one approach, in terms of the additional return on the investment originally made in the data creation/collection resulting from the additional use of the data facilitated by the centres.

1.2 The data centres

The three data centres studied span three different academic traditions; humanities (ADS), social sciences (ESDS), and environmental sciences (BADC), although each supports inter-disciplinary research across these and other boundaries. There is considerable variability between the data centres in areas such as: operating budgets; the target audiences and their size; the diversity and range of usage behaviours; the diversity of activities for which they are funded; the ways in which data are selected and ingested, and the diversity of data collections and access methods. Moreover, all three data centres have been established for a decade or more and have evolved and changed administrative arrangements over time.

Table 1: Key characteristics of the three data centres

	Economic and Social Data Service (ESDS)	Archaeology Data Service (ADS)	British Atmospheric Data Centre (BADC)
Operational history	2003 – 2012. Formed from a number of pre-existing collections including the UK Data Archive. Now part of the UK Data Service launched in 2012.	1996 – to present. Previously, ADS formed part of the Arts and Humanities Data Service.	1994 – to present. Previously the Geophysical Data Facility at Rutherford, Oxfordshire. Combined with NERC Earth Observation Data Centre in 2005 to form the Centre for Environmental Data Archival (CEDA).
Operational budget	£3.3m pa (£16.7m over 5 years to 2011).	£698,000 pa (circa 2012).	£2m pa (circa 2012).
Users	Registered users: 23,000 Active users (excl. school and undergraduate students): 18,098.	Registered users: 3,000. Active users (once a week or more): 11,020.	Registered users: 22,500. Active users (once a year or more): 5,959.
User survey profile	Academic: 80%.	Academic: 31%. Note the substantial non-academic user community: with private individuals 27% (i.e. general public), and other sectors 42% (local and central government, charities, and archaeological contractors).	Academic: 61%.

Source: *Authors' analysis based on data in the original studies.*

Each centre provides its data collections free at point of use, but at ESDS and BADC a range of resources can only be accessed by registered users, while others are freely available via the Web without registration. All resources at the ADS can be accessed without registration, but the ADS operates a relatively new voluntary registration system to assist with user profiling and customised services. All data centres have profile information for their registered users, but these registers are cumulative and information may be historic rather than current.

2. Approaches and methods

2.1 Data collection

In selecting conceptual approaches for the studies, we took account of the practical limitations of collecting the necessary data through desk research, survey and interview techniques, and sought to maximise economy in data collection through commonality (i.e. the same data can be used to inform more than one of the methods).

For all three studies, we combined:

- Desk-based analysis of existing evaluation literature and reports, looking at both methods and findings;
- Existing data from Keeping Research Data Safe (KRDS)and other studies[3] of the costs and benefits of research data infrastructure and services;
- Existing management and internal data collected by the data centres, such as user registration and access statistics, deposit records, internal operational and financial reports; and
- Original data collection in the form of online surveys of users and depositors, and semi-structured interviews.

For ESDS we used three case studies to explore the impact that research based on data sourced through ESDS has, in terms of debate and media coverage of major social issues and 'pathways to impact'. For ADS we held a focus group with a range of stakeholder representatives, to establish any change of perception of the ADS amongst participants as a result of our study. We also sought their views on how the ADS Impact Study results and benefits that we identified might best be presented, and we used this input in producing a KRDS Stakeholder Benefits Analysis and other dissemination materials. For BADC, we only used the standard combined set of approaches outlined earlier.

2.2 Assessing value, impacts and benefits

We undertook literature reviews of previous studies of the value and impact of information services, research publications and data for each of the studies. They suggested that no single analytical approach has dominated across the different but related fields. Consequently, we combined quantitative and qualitative approaches in the studies. These included a range of economic methods, such as contingent valuation, welfare economics and growth accounting, and qualitative methods such as the Keeping Research Data Safe (KRDS) Benefits Framework, in order to quantify the value and impacts of the three research data centres in economic terms, and also explore other non-economic benefits.

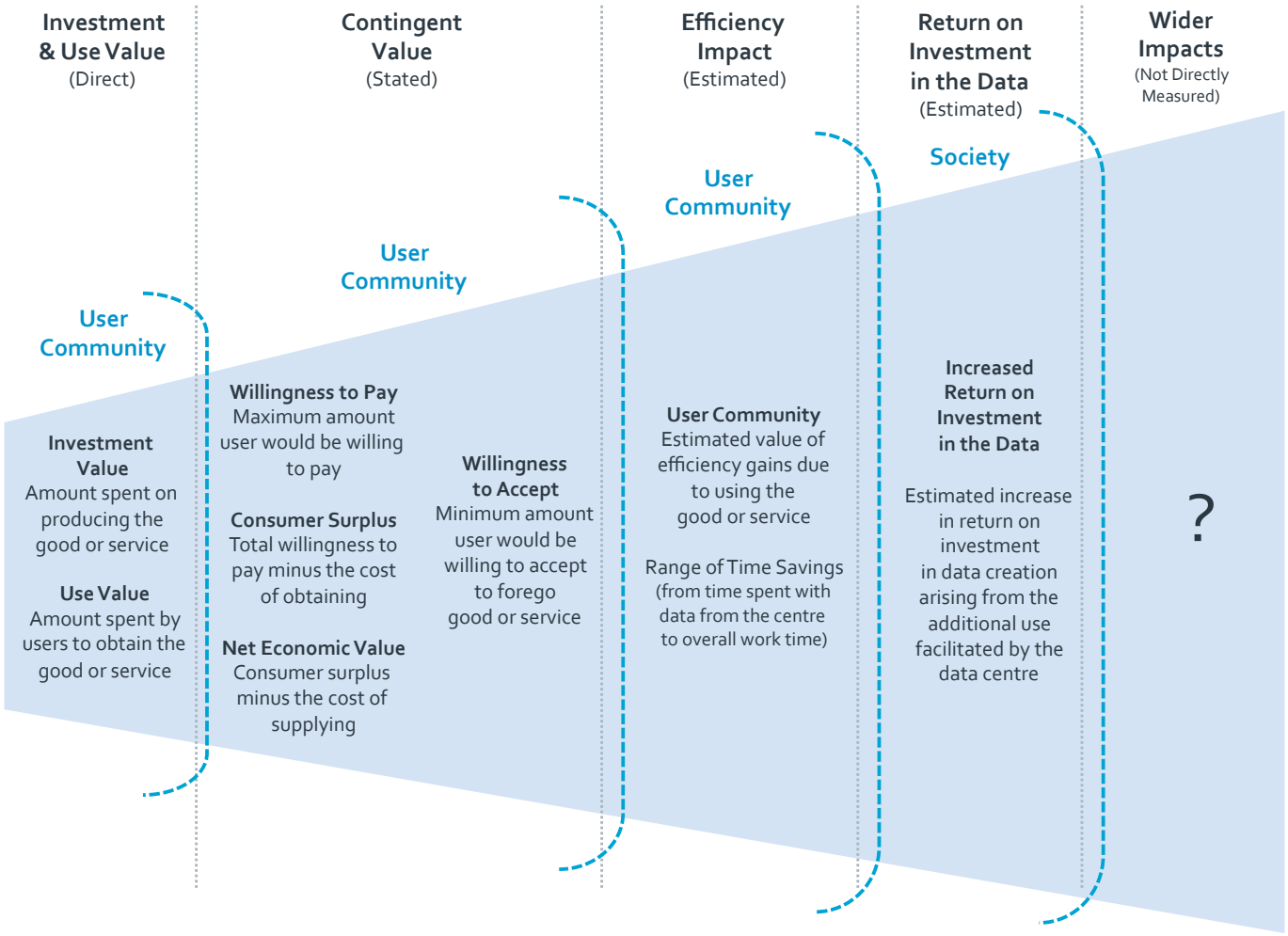
Our aim was to use methods that could assess the value and impact at the level of the data centre (i.e. all of the data centre's data collections and services, rather than individual collections). However, many of the methods could be used at a more disaggregated level, such as an individual data collection, if that were the primary focus for analysis.

3 For further information on the KRDS projects, Benefits Framework and Toolkit and related projects and implementations see **beagrie.com/krds.php**

Quantitative methods

The economic methods we used can be seen as estimating a range of values, moving from those focusing on minimum values toward methods that measure some wider impacts (Figure 1). They include two ways of expressing return on investment in the data centres: the ratio of users' value to investment in the centres; and the ratio of value of the additional (re-)use of the data hosted to investment in the centres.

Figure 1: Methods for exploring the economic value and impacts of research data centres



Source: *Authors' analysis based on original studies.*

The methods (ordered from those focusing on minimum values through to wider values) include:

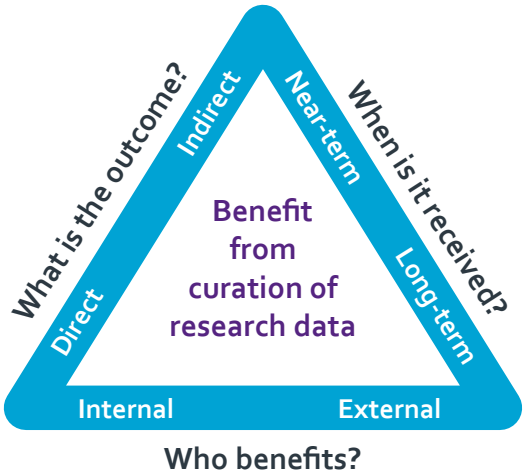
» Estimates of 'investment value' (i.e. the operational expenditure of the data centres plus the time and other costs for depositors submitting data), and use value (i.e. the cost of the time spent by users accessing the data and services);

- Contingent valuation using stated preference techniques[4] (a survey-based technique) to explore the amount that users might be willing to pay to access the data and services from the data centres in a hypothetical market situation, or be willing to accept in return for giving up their access;
- Welfare approaches[5] (a micro-economic approach to measuring net social welfare) to estimating consumer surplus (i.e. willingness to pay minus use value) and net economic value (i.e. consumer surplus minus operational budget) of the data and services provided by the data centres;
- An activity costing approach to exploring the estimated work-time saving efficiency impacts of the research data centres among their user communities; and
- A macro-economic growth accounting approach (using a modified Solow-Swan model)[6] to exploring the increase in social returns on investment in the original creation/collection of the data hosted, arising from the additional use of the data facilitated by the centres (i.e. the implied value of the data re-use by those who could not have obtained the data elsewhere or collected/created it themselves).[7]

Qualitative methods

In parallel with this, we used a number of qualitative approaches including: analysis of desk research, interview summaries and user and depositor survey responses; case studies (for ESDS) to explore the impact that research based on ESDS has had in terms of debate and media coverage of major social issues; the Keeping Research Data Safe Benefits Framework as a method for identifying, collating and presenting the broad spectrum of benefits analysed in the studies arising from a data centre's collections and associated services – and a focus group (for ADS) to explore the perceptions of benefits, value, and impact among various stakeholders.

Figure 2: The KRDS Benefits Framework



Source: *Beagrie et al 2010, and KRDS 2011.*

4 See DTLR (2002) *Economic Valuation with Stated Preference Techniques*, London: Department of Transport, Local Government and the Regions. Available **http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communities.gov.uk/documents/corporate/pdf/146871.pdf**

5 See **http://en.wikipedia.org/wiki/Welfare_economics** for a short overview.

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

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

The KRDS Framework was chosen as a method as it had been specifically developed for appraising qualitatively the benefits of digital preservation and curation of research data. The framework uses three dimensions to illuminate the benefits investments can potentially generate: each dimension has two main sub-divisions (Figure 2). These dimensions serve as a high-level framework within which thinking about benefits can be organised, and then sharpened into more focused presentations of value for specific contexts and communities. We have used the KRDS Framework to capture the wider value of the data centres in terms of different types of outcome, timescale, and beneficiaries. However, KRDS tools can also be used at a narrower level to select and quantify individual benefits[8], if that were the primary focus for analysis.

8 For worked examples see **beagrie.com/krds-i2s2/**

3. Synthesis of analysis and findings

3.1 Economic analysis and findings

Table 2 presents a brief description of the basic data used and approaches taken in the three data centre studies. Major differences relate to the extent of geographic coverage of the user surveys, inclusion or exclusion of undergraduate and school students, the basis for costing applied depending on the composition of the user communities plus date of study, the definition of active users, and the reliability of access/download data and consequent ability to weight the survey respondents' cost and value responses to reflect overall data centre use. There were also differences in the survey questionnaires applied across the studies as the series evolved, leading to minor differences in the data collected and treatment of it in efficiency impact and return on investment estimates.

Readers should refer to the individual study reports for ESDS (Beagrie et al 2012), BADC (Beagrie and Houghton 2013a), and ADS (Beagrie and Houghton 2013b), respectively, for more detailed description of the data used and approaches taken in each data centre study.

Table 2: Data and approaches used in the three studies

Economic and Social Data Service (ESDS)	Archaeology Data Service (ADS)	British Atmospheric Data Service (BADC)
Surveys and interviews		
User survey responses N=952 Depositor survey responses N=193 Interviews conducted N=25	User survey responses N=299 Depositor survey responses N=86 Interviews conducted N=15	User survey responses N=1,141 Depositor survey responses N=42 Interviews conducted N=13
Geographic coverage of surveys		
User survey geographic coverage was UK, Anglophone and Eurozone countries.	User survey geographic coverage was UK, Anglophone and Eurozone countries.	User survey geographic coverage was worldwide, including developing countries.
Student coverage of surveys		
School and undergraduate students were excluded.	Included undergraduates and a few school students.	Included undergraduates, but no school students.
Active users and how these were defined by the centres		
Active non-student users = 18,098 (Active is those who have either registered or renewed their registration in the preceding three years).	Active users = 11,020 (Active is using once a week or more during last year).	Active users = 5,959 (Active is using once or more during last year).
Metric for use		
Data Collections delivered = 56,777	Annualised visits = 170,757	Annualised access sessions = 22,608

Economic and Social Data Service (ESDS)	Archaeology Data Service (ADS)	British Atmospheric Data Service (BADC)
Basis for costings		
User community predominantly higher education. User and depositor costs based on UK TRAC fEC of 2.3 and average UK academic salary. (Mean salary and on-cost £64/hr).	User community diverse. User and depositor costs based on UK Treasury Green Book mark-up of 1.3 and average salaries for the UK archaeology profession. (Mean salary and on-cost £21/hr).	User community predominantly higher education, but geographically diverse. User and depositor costs based on UK Treasury Green Book mark-up of 1.3 and average UK academic salaries for the job levels. UK equivalent costings were used for developing country users. (Mean salary and on-cost £33/hr).
Basis for weighting user survey economic analysis		
User survey-based cost and value variables weighted on total datasets delivered in last year.	User survey-based cost and value variables weighted on total use visits in last year.	User survey-based cost and value variables remained un-weighted, due to limitations in the download data available.
Depositor community		
Depositors from last 3 years were selected, with results weighted.	Depositors from last 3 years were selected, with results weighted.	Depositors from last 3 years were selected, with the survey respondents treated as the population.
Metric for deposit		
Data collections acquired = 750 of which 453 new.	Datasets acquired = 444 of which 399 were estimated to be new. For the grey literature collection deposits are deposition events.	Survey respondents reported 42,748 deposits, of which 42,705 were updates (43 new).
Basis for weighting depositor survey economic analysis		
Depositor survey-based cost and value variables weighted on total datasets acquired over last 3 years.	Depositor survey-based cost and value variables weighted on total dataset deposit events over last 3 years.	Depositor survey-based cost and value variables remained un-weighted, due to limitations in the deposit data available.

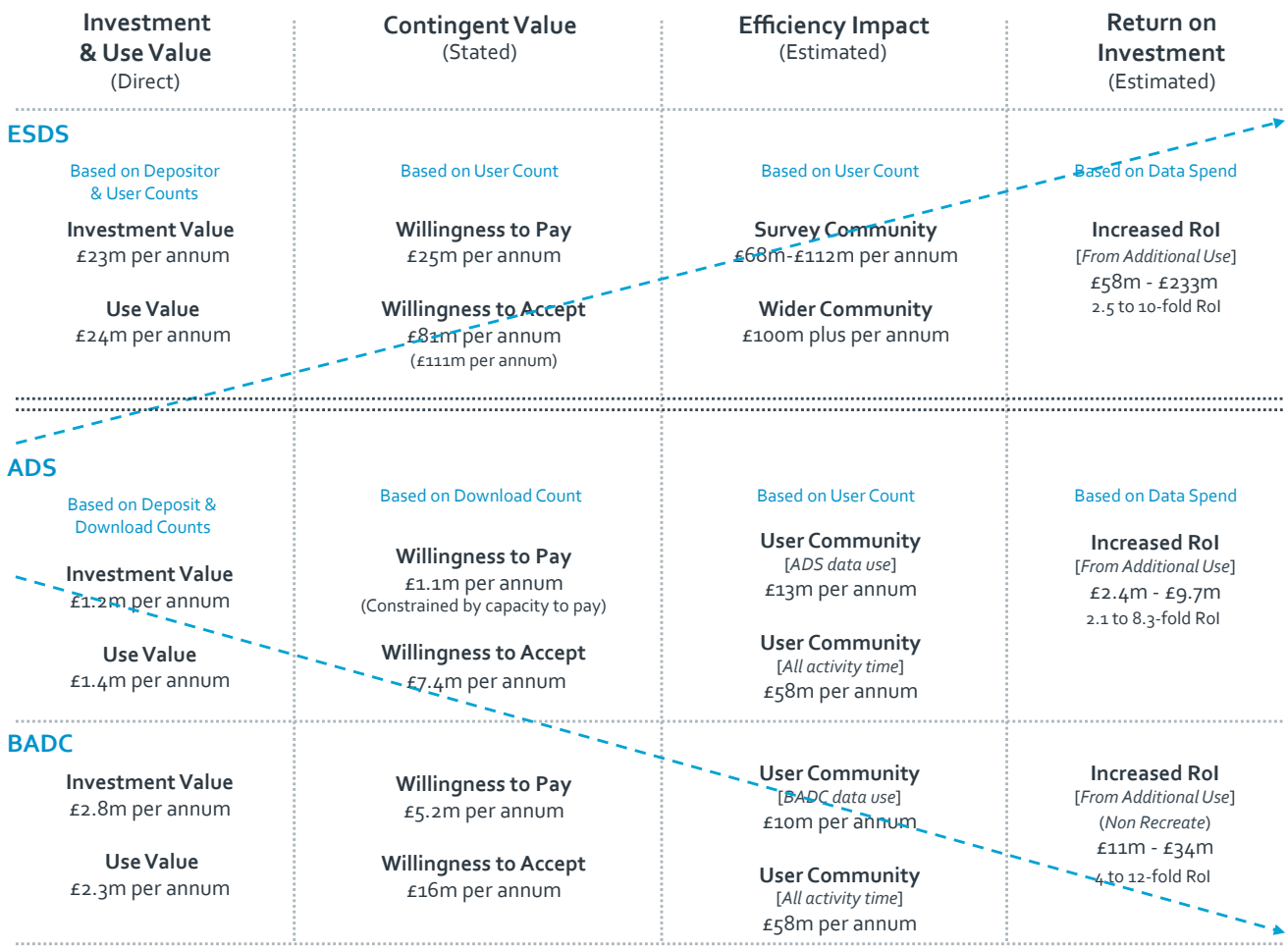
Source: *Authors' analysis based on data in the original studies.*

A key element of the research underlying the studies was to develop and refine the methods. Hence, in addition to contextual differences noted in Table 2, there are some methodological and implementational differences between the studies that make direct comparison difficult. In particular, the results from the ESDS study (the first in the series) are not comparable with the later studies, primarily due to two significant differences:

- ESDS time cost estimates were based on the UK higher education TRAC full economic cost (fEC) model with mark-up of 2.3 on salaries, rather than the UK Treasury Green Book method with mark-up of 1.3 on salaries, in part reflecting differences in data centre user communities; and
- ESDS activity cost estimates were based on depositor and user counts, rather than deposit and download counts, reflecting methodological choices at the time (See discussion in section 3.4).

There were also differences in the survey questionnaires applied across the studies, leading to minor differences in the data collected and treatment of it in efficiency impact and return on investment estimates in the ESDS study. Figure 3 presents a summary of the findings from the three studies.

Figure 3: The value and impacts of the three UK data centres



Note: Due to contextual and methodological differences between the studies the results are not comparable.

Source: *Authors' analysis based on data in the original studies.*

The Economic and Social Data Service (ESDS)

With the above caveat regarding comparability in mind, our analysis of the ESDS showed a direct investment and use value to its user community of £23 million to £24 million per annum at 2009 prices and levels of activity. Willingness to pay is an expression of value by users, who revealed that they valued their access at around £25 million per annum.

The contribution of ESDS data and services to its user community can be seen in terms of its impact on their research and teaching efficiency (e.g. in terms of time saved). We found that the total estimated efficiency impacts of ESDS data and services among its non-student user community might be worth as much as £100+ million per annum.

Exploring the potential impacts of ESDS on returns to investment in the data hosted and delivered, we found that ESDS facilitates additional use which realises additional returns that could be worth some £58 million to

£230 million over 30 years (net present value) from one year's investment expenditure – effectively, a 2.5- to 10-fold return on investment.

The Archaeology Data Service (ADS)

Using the different costing and estimation methods noted, our analysis of the ADS shows a direct investment and use value to its user community of £1.2 million to £1.4 million per annum at 2011 prices and levels of activity. ADS users revealed that they value their access at around £1.1 million per annum, despite resource constraints and sometimes very limited capacity to pay. When capacity to pay is limited the amount that users would be willing to accept in return for giving up their access for a year can be a better indicator of the value they place on it (i.e. what they would sell it for, rather than what they can and would pay for it). Looked at this way, ADS data and services are worth around £7.4 million per annum to its user community.

The contribution of ADS to its user community can also be seen in terms of its impact on their research, teaching and studying efficiency. We found that the total estimated efficiency impacts of ADS among its user community might be as much as £58 million per annum at 2011 activity levels. However, some respondents may have interpreted the survey question as relating to the efficiency impact on their time spent with ADS data and/or data from all sources, rather than their total work time – which had been intended. Based on respondents' time spent with ADS data, the implied efficiency impacts would be worth around £13 million per annum. This is still a very substantial impact.

Exploring the potential impacts of ADS on returns to investment in the data hosted and delivered, we found that ADS facilitates additional use (i.e. by those who could neither obtain the data elsewhere nor create/collect it themselves) which realises potential additional returns which could be worth £2.4 million to £9.7 million over 30 years (net present value) from one year's investment expenditure – effectively, a 2- to 8-fold return on investment.

A unique feature of the ADS Impact Study was the inclusion of an analysis of the evolving, cumulative value of the centre. That analysis suggested that the value of ADS data and services has increased as the collection has grown and the service has developed.

The British Atmospheric Data Centre (BADC)

Analysis of the BADC, which is broadly similar to that of the ADS, shows a direct investment and use value to its user community of £2.3 million to £2.8 million per annum at 2012 prices and levels of activity. BADC users reveal that they value their access at, and would be willing to pay, around £5.2 million per annum, despite constraints on capacity to pay. As we've said before, when capacity to pay is limited the amount that users would be willing to accept in return for giving up their access for a year can be a better indicator of the value they place on it. Looked at this way, BADC data and services may be worth as much as £16 million per annum to its users.

The contribution of BADC to its user community can also be seen in terms of its impact on their research, teaching and studying efficiency. We found that the total estimated efficiency impacts of BADC among its user community might be as much as £58 million per annum at 2012 prices and levels of activity. However, it appears that in this study, too, some respondents may have interpreted the survey question as relating to the efficiency impact on their time spent with BADC data and/or data from all sources, rather than their total work time. Based on respondents' time spent with BADC data, the implied efficiency impacts would be worth around £10 million per annum.

Exploring the potential impacts of BADC on returns to investment in the data hosted and delivered, we found that BADC facilitates additional use (i.e. by those who could neither obtain the data elsewhere nor create/collect it themselves) which realises additional returns which could be worth some £11 million to £34 million (net present value) over 30 years from one year's investment expenditure – effectively, a 4- to 12-fold return on investment.

3.2 Qualitative analysis and findings

The interviews underlined the value seen by users and depositors in the data centres and the overall feedback was very positive, although there were also some suggestions about where improvements could be made.

Academic users reported that use of the centres was very or extremely important for their academic research, and a majority of respondents (between 53% and 61% across the three surveys) reported that it would have a major or severe impact on their work if they could not access the data and services. In the depositor surveys, respondents reported that having the data preserved for the long-term and its dissemination being targeted to the academic community were the most beneficial aspects of depositing data with the centres.

In addition to the surveys, three impact case studies were undertaken for the ESDS study to assess the policy and practice impact of research based on data accessed via the ESDS. As with previous ESRC research on impact, we found it difficult to identify case studies that could conclusively show direct impact on policy and practice, because of the widely acknowledged difficulties associated with attribution and time-lags. Nevertheless, all three case studies were valuable in demonstrating how research based on ESDS has had significant impact in terms of debate and media coverage of major social issues.

The KRDS Framework proved an effective and straightforward way to summarise the qualitative findings. The simplest form of presentation of results in the six areas of the framework was used for ESDS (Beagrie et al 2012, p 43-4) and BADC (Beagrie and Houghton 2013b, p51-2). Several generic benefits are shared across the two data centres, (for example, data preserved for the long-term, service targeted at academic community and supporting their needs), but there are also more discipline and community specific additions and variations, and different relative prioritisation and ordering of benefits in the framework for each. For the ADS, the focus and resourcing of the impact study allowed the use of a more detailed version of the framework, including a stakeholder benefit analysis (Beagrie and Houghton 2013a, p59-61).

The ADS stakeholder focus group revealed that our impact study had changed their perceptions, increasing recognition of the value of the ADS and digital archiving and data sharing generally. Most stakeholders already valued ADS highly, but felt that the study had extended their understanding of the scope of that value and the degree of its value to other stakeholders. They were positive about seeing value expressed in economic terms, as this was something they had not previously considered or seen presented, but they also felt that it was important not to dwell exclusively on economic measures of value.

Indeed, the three studies show the benefits of integrating a range of quantitative economic approaches to measuring the value and impacts of research data archiving and sharing, with qualitative approaches exploring user perceptions and wider dimensions of value.

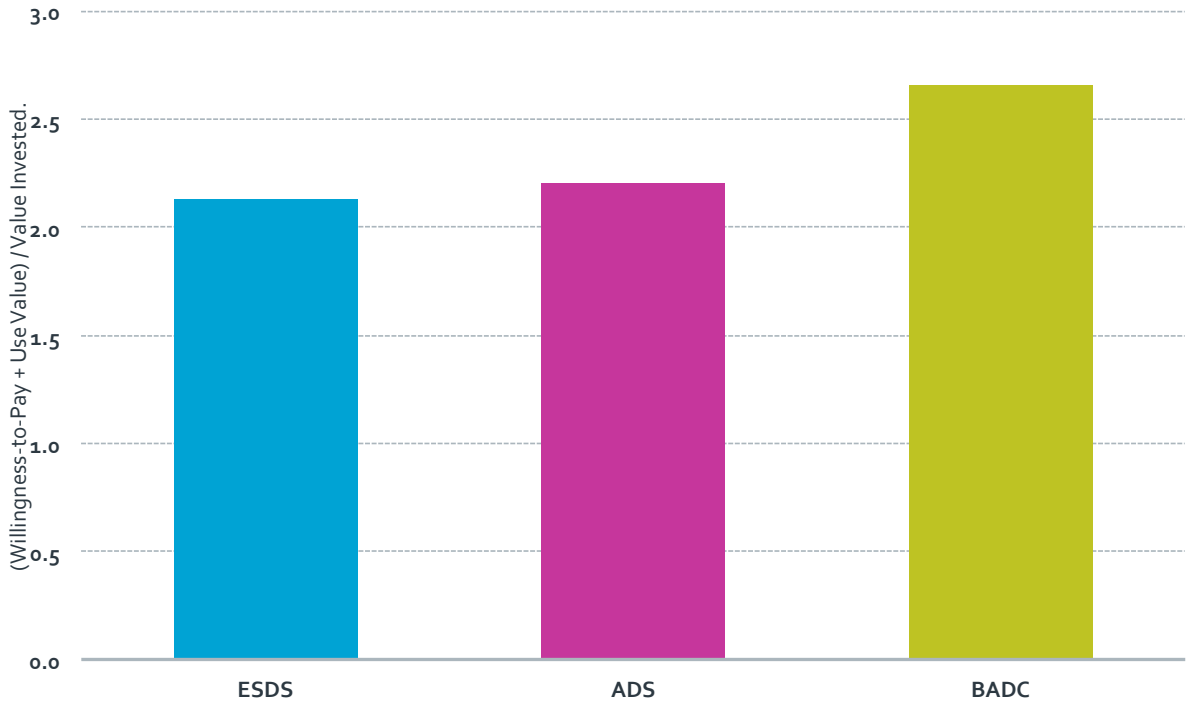
3.3 Summary of key findings

The economic analysis indicates that data sharing and data curation via the centres studied has a substantial and measurable positive return on investment and, by facilitating additional use, increases the return on investment in the original creation/collection of the data hosted.

We were able to estimate, for the first time, the indirect investment made in the data centres by depositors (and their funders) in terms of the time and other costs involved in preparing material for deposit. These indirect investments are substantial.

Nevertheless, what users pay, in terms of their access time, and what they would be willing to pay for access, is 2.2 to 2.7 times greater than the value invested in the centres, in terms of operational costs plus depositor costs (Figure 4).

Figure 4: Ratio of use value plus willingness to pay to value invested in the data centres

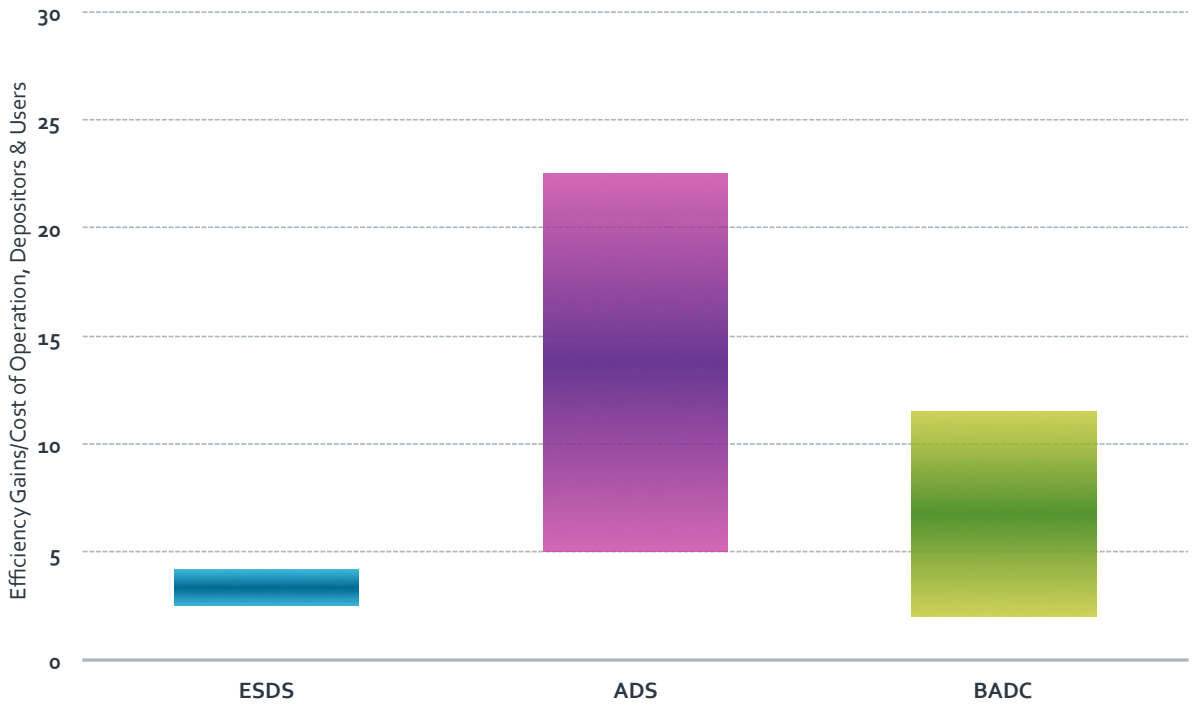


Note: Due to contextual and methodological differences between studies the results are not directly comparable.

Source: *Authors' analysis based on data in the original studies.*

A very significant increase in research efficiency was reported by users as a result of their using the data centres. These estimated efficiency gains ranged from 2 and up to more than 20 times the costs – including operational, depositor and user costs (Figure 5). Note that the ESDS results relate to time spent working with data from ESDS (as the minimum) and all data (as the maximum), whereas ADS and BADC results relate to time spent with ADS/BADC data (minimum) and total working time (maximum). Obviously, the latter produces wider ranges.

Figure 5: Ratio of efficiency gains from using the data centres to data centre costs

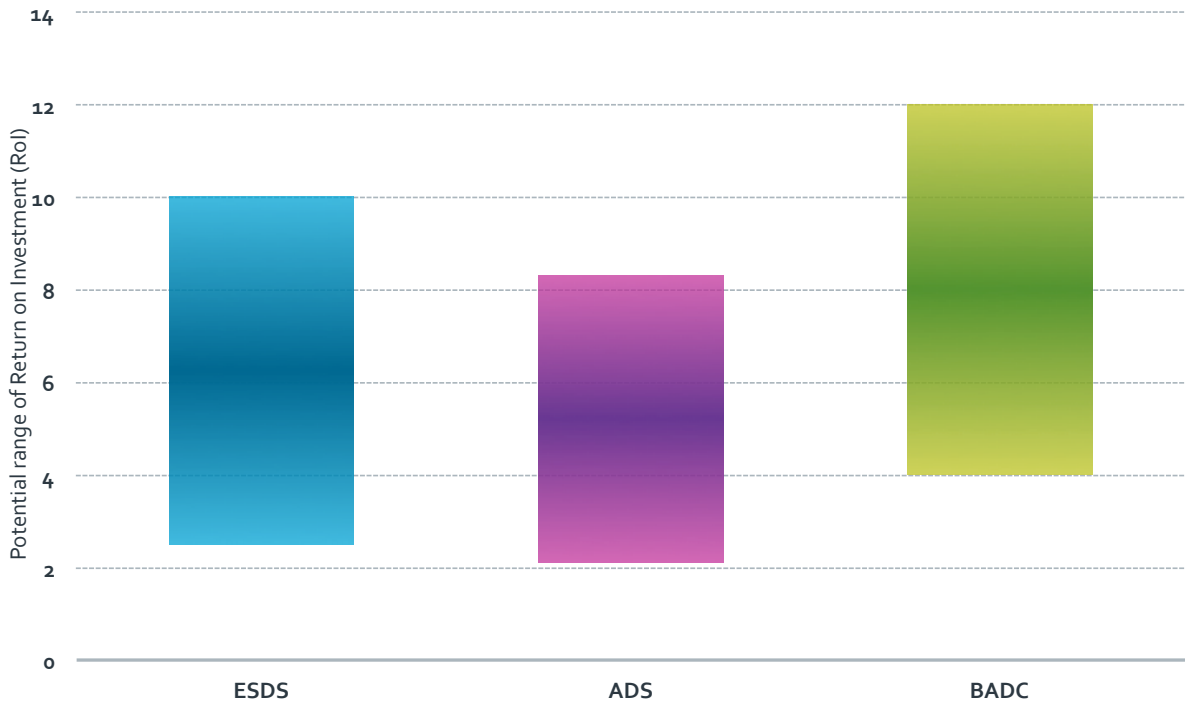


Note: Due to contextual and methodological differences between studies the results are not directly comparable.

Source: *Authors' analysis based on data in the original studies.*

We also estimated the value of the increase in return on the original investment in the creation/collection of the data hosted resulting from the additional use facilitated by the centres, and found returns ranging from twice and potentially up to 12 times the investment in the data centres (Figure 6). However, difficulties in estimating original data creation/collection costs from depositor survey responses (due to depositor population sizes and difficulties framing comparable questions across widely divergent data types and depositor practices) should be taken into account in interpreting these results.

Figure 6: Ratio of value of the additional (re-)use of the data hosted to data centre costs



Note: Due to contextual and methodological differences between studies the results are not directly comparable.

Source: *Authors' analysis based on data in the original studies.*

The qualitative analysis showed that many of the academic users see the centres as very or extremely important for their academic research, and a majority of respondents reported that it would have a major or severe impact on their work if they could not access the data and services. There was also a strong belief expressed that such services should remain free to users at point of use. Having the data preserved for the long-term and its dissemination being targeted to the academic community were seen as the two most beneficial aspects of depositing data with the centres. The KRDS Framework proved an effective and straightforward way to summarise the qualitative findings in each of the studies.

Overall, the three studies show the benefits of integrating a range of quantitative economic approaches to measuring the value and impacts of research data archiving and sharing, with qualitative approaches exploring user perceptions and wider dimensions of value.

While, taken individually all have limitations, the approaches used to assess value, impacts and benefits across all three data centre studies show a similar pattern of findings, with data sharing via the data centres having a large measurable impact on research efficiency and on return on investment in the data and services. These findings are important for funders, both for making the economic case for investment in data curation and sharing and research data infrastructure, and for ensuring the sustainability of such research data centres.

3.4 Data limitations and interpretation

When implementing the economic approaches there are inevitably some limitations in the data to consider. What is the best way to undertake the economic estimations often depends on the availability of data and/or the confidence one has in the data available. It can also be important to be clear that key data variables are independent of each other, which can be an issue in interpreting data centre web logs and access statistics (e.g. in the independence of user and use counts).

For example, estimates of the number of users of the data centre may be based on registration with the service and subsequent recognition through tracking cookies or IP addresses, or may simply be an estimate of the number of unique IP addresses using the services. The latter may be less than ideal if, for example, a user works from an office and from home, thus appearing as multiple unique IP addresses. The former may be less than ideal if, for example, there is no process for updating registrations or for de-registration of those no longer using the service.

The concept of 'active user' is used by all three data centres to distinguish recent users, but with very different timeframes defined for inclusion (i.e. those who have either registered or renewed their registration in the preceding three years for ESDS; those using the service once a week or more during the last year for ADS; and those using the service once or more during the last year for BADC). While many active users are registered, there are also active users who are not registered. Much less is known about this group.

Similarly, data centre logs of data deposits and accesses/downloads need to identify a deposit or access activity that matches the depositors'/users' perceptions of the act of deposit or access. This can be difficult if, for example, users are able to browse parts of the data collection online, perhaps even interrogate it (e.g. using a tool such as Beyond 20/20[9]), without appearing to download data. There may also be 'uses' recorded in the logs, such as accidental visits in passing, automated web crawlers, etc. that can be difficult to eliminate.

Hence, implementation of the economic methods must take account of the availability and reliability of the data, as different data may be used for some estimations if the alternative data are judged to be more reliable or complete. For example, an estimate of the overall user communities' willingness to pay for a data centre might ideally be based on factoring the survey respondents' accesses/downloads as a share of total accesses/downloads (as was the case for ADS and BADC). However, if access/download data are deemed less reliable, an alternative, second-best approach might be based on factoring survey respondents as a share of total data centre users (as was the case for ESDS). Such choices and judgements are an integral part of the analysis.

The focus of the economic analysis is on the value of a data centre to its user community. Self-evidently, survey respondents are a self-selected sub-group (because they responded to the survey). While response rates to our surveys (around 13%-18% for users and 30%-34% for depositors) were good, on average, one might expect that those taking the time to respond to a survey are likely to use and value the data and services more than those who did not respond; and reported frequencies of use suggest that our respondents were among the more frequent users.

It is also important to note that few users use all of the data or services provided by a centre, but rather experience just part of it, and they can only express costs and value relating to the parts they use. Consequently, wherever possible, it is necessary to weight the survey responses to reflect the wider depositor and user communities, and the deposits with, and uses of, the data and services. While the results from ESDS and ADS were weighted, due to data limitations, those from BADC were not.

9 See http://beyond2020.com/index.php/data-solutions/products/professional-browser

4. Experience gained and lessons learned

The three data centre studies applied a unique combination of quantitative and qualitative methods to provide the full picture of the nature and dimensions of value and to explore the full range of impacts from data sharing and data curation.

The quantitative findings express the value and impacts in economic terms. The qualitative findings illustrate individual user and depositor experience and can personalise and provide further insights into the value of data curation and sharing.

Moreover, as there are inevitably some limitations in any individual approach, a further advantage of using multiple approaches is the potential for confirmation of findings. Independently, both quantitative and qualitative analyses show a similar picture: they are complementary, they reinforce each other, and lend credence to the findings.

Our work shows that the methods used to explore economic value and impact are 'doable' and transferable between different disciplinary contexts and collections. However, relatively little can be transferred in terms of implementation between different studies as there is significant tailoring for specific contexts and communities involved, and this has resource implications. The data collection and economic analysis are time consuming and need to be tailored to the specific nature of operation and use of each data centre. As discussed in previous sections, this has implications for the extent to which the results arising from independent studies can be compared.

Implementation difficulties included survey design (e.g. in fashioning questions about what is a use and what is a deposit, and in quantifying efficiency and costs), and it requires a good deal of time to customise the questions and pilot test each survey.

We also confronted data difficulties in relation to user/depositor counts and deposit and download/use counts in all of the studies, to varying degrees. Data centre usage statistics have inherent variations and limitations. Relatively little may be known about the profile of public (or non-registered) users of data centres that require no registration for access. Even profile data on registered users can be historic and out of date, if periodic re-registration and updating is not required.

These three data centre studies focused at the level of the centre, but a more disaggregated focus would be possible (e.g. on individual data collections). However, the level of aggregation affects the appropriateness of methods. For example, a return on investment in the creation/collection of the data hosted approach, such as that used for these studies, should only be used at an aggregate level. Even then there are judgements and estimates involved (e.g. differing average returns to research in humanities and sciences). A number of the other methods used can work at lower levels of aggregation (e.g. investment and use value, and contingent valuation), and can be easier to implement at lower levels of aggregation.

Contingent valuation is a method that typically generates some protest responses, and did so in these studies. Despite assurances in the survey questionnaires that the service would remain free at the point of use, users were naturally fearful, leading to protest responses and comments. This fear, and the protests it brought us, are an important message for funders and data centres to note.

5. Recommendations

Our recommendations are relevant as appropriate to funders and depositors, data centres and repositories, and other future studies.

Recommendation 1: Continue support for the data centres studied

The studies show the three data centres are generating benefits that justify the investment in them made by funders, users and depositors. We assessed the aggregate value and impact of the individual data centres, therefore there is still scope for funders to look at how investment can be best targeted, or whether greater impact could be made with additional investment. For example, further research and analysis might focus on the impact of additional expenditure, looking in more detail at what 'causes' additional (re-)use that could increase the centre's return on investment or enhances user efficiency in such a way as to increase the benefits more than the costs.

Recommendation 2: Further develop the methods

The unique combination of qualitative and quantitative approaches used in these impact studies has now been applied to three UK data centres spanning very different disciplinary domains. The experience suggests that the approaches are complementary and mutually reinforcing, and while they are transferable they require significant customisation to fit disciplinary and service differences – somewhat limiting cross-study comparisons. There would be benefits from further research developing, refining and exploring applications of the methods used in these studies, as making the business or funding case for data centres plays an increasingly important role in ensuring their sustainability.

Recommendation 3: Promote standardisation of usage statistics

It is also clear from these studies that different data centres collect financial and operational data, such as user statistics, data deposit, access and download statistics, to varying levels of detail and using different definitions. More guidance is needed on the collection of such data. Doing so would help to ensure a greater degree of standardisation of operational records across data centres. This would be of greatest benefit to funders investing in a range of data centres, and would provide more comprehensive and reliable data for economic analysis. There would be considerable advantage to providing guidance regarding the collection of such data as it is fundamental to the economic analysis and in making the business or funding case.

Recommendation 4: Undertake analysis of the value and impact of other parts of the research data curation infrastructure

To date these approaches have only been applied to three UK-based data centres. However, they should be equally applicable to other international, national, or institutional repositories holding research data. We should consider applying these methods of valuation to a wider range of data centres and repositories at international, national and/or institutional levels, to assist investment decisions and assess their impact. There will be some significant differences in terms of levels of service, scale of use and collections, and depth of disciplinary expertise and coverage between them, which may require modified implementation of approaches. Any comparisons should aim to be "like for like".

Recommendation 5: Conduct more granular analysis

The studies have looked at the aggregate value of data centres. There is also significant scope for more granular studies that focus on the value of specific collections, or the economic value of services to specific groups. There may also be some practical advantages to a narrower focus in simplifying some of the statistics and the analysis of different usage patterns across collections and user groups. For the qualitative analysis, a more detailed KRDS analysis by specific stakeholder groups, similar to that undertaken for the Archaeology Data Service (Beagrie and Houghton 2013a), would also be beneficial.

Recommendation 6: Track changes over time

Value and perceptions of value change over time. Data centres and their funders should consider opportunities to repeat the surveys and extend the available series of comparative studies in future years. Ideally another survey of users and depositors with these data centres should be considered within the next three to five years.

Recommendation 7: Study the wider value and impact of collections

While the ready availability of data can have a significant impact on the efficiency of research users and, through increased use of the data, increase the return on investment in the data creation/collection, curation, and sharing involved, it is in the uses to which the data are put after research use that substantial additional benefits can arise. To an extent, some of these impacts can be captured through the efficiency impacts and return on investment scenarios explored in our analyses, and in the qualitative findings. Nevertheless, there can be very substantial wider benefits. In the context of atmospheric data, for example, work by the US National Oceanic and Atmospheric Administration (NOAA) on the value of meteorological data may be indicative of additional lines of research on these wider benefits (Beagrie and Houghton 2013b). Funders should consider research on the wider societal benefits and economic impacts that are generated by research data sharing and curation, and the contribution to this made by data centres such as those studied in this series of reports.

References

Beagrie, N., Lavoie, B., & Woollard, M. (2010) *Keeping Research Data Safe 2 Final Report* London: JISC. Available: **jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads**.

Beagrie, N., Houghton, J.W., Palaiologk, A. and Williams, P. (2012) *Economic Evaluation of Research Data Service Infrastructure: A Study for The ESRC*, Economic And Social Research Council. Available: **esrc.ac.uk/research/evaluation-impact/impact-evaluation/economic-impact-evaluation.aspx**.

Beagrie, N. and Houghton J.W. (2013a) *The Value and Impact of the Archaeology Data Service: a study and methods for enhancing sustainability*. Available: **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*. Available: **jisc.ac.uk/whatwedo/programmes/di_directions/strategicdirections/badc.aspx**

DTLR (2002) *Economic Valuation with Stated Preference Techniques*, London: Department of Transport, Local Government and the Regions. Available: **http://webarchive.nationalarchives.gov.uk/20120919132719/http://www.communities.gov.uk/documents/corporate/pdf/146871.pdf**

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

KRDS (2011) *Keeping Research Data Safe (KRDS) project website*. Available: **beagrie.com/krds.php**

Annex I: Basis of quantitative findings from the studies

	ESDS	ADS	BADC
Variable	*Basis of estimation*	*Basis of estimation*	*Basis of estimation*
Creation costs	Weighted cost of last deposit x total new deposits (i.e. not including updates).	Weighted cost of last deposit x total new deposits (i.e. not including updates).	Sum of depositor survey respondents' costs (i.e. survey respondents treated as population).
Depositor costs	Weighted annual deposit costs x total number of deposits.	Weighted last deposit costs x total number of deposit events.	Treating respondents' deposits as the total deposits.
Operational budget	Average per annum over last 5 years.	Per annum circa 2011.	Per annum circa 2012.
Investment value (excl creation costs which are treated as sunk costs)	Operation budget + Weighted deposit costs based on total number of deposits.	Operation budget + weighted deposit costs based on total number of deposits.	Operation budget + Deposits reported in depositor survey.
Use value (excludes depositor costs)	Users' + depositors' costs (based on user numbers and total deposits).	Weighted mean cost of last access x total user visits.	Mean cost of last access x total user access days.
Willingness to accept (WTA)	Users x weighted WTA.	(Weighted WTA / frequency of download) x total downloads. Frequency is the mean of the individualised frequencies.	(Individual mean WTA / individual frequency) x total downloads.
Willingness to pay (WTP)	Users x weighted WTP (mean of per annum and per view).	Total downloads x weighted mean WTP (mean of per annum and per view).	Total downloads x mean WTP (mean of per annum and per view).
Consumer surplus	Based on users' WTP (with download-based use value).	Based on downloads. Negative due to very limited capacity to pay among archaeological users.	Based on downloads.
Net economic value	Based users.	Based on downloads. Negative due to very limited capacity to pay among archaeological users.	Based on downloads.

	ESDS	ADS	BADC
Variable	***Basis of estimation***	***Basis of estimation***	***Basis of estimation***
		On a willingness to accept basis, as the archaeology community has limited capacity to pay.	
Efficiency Overall time data centre time	Based on active users (research and teaching). ESDS calculations were done differently, and were based on time spent with data rather than overall working time.	Based on active users (research, teaching and studying).	Based on active users (research, teaching and studying).
Return on investment in data creation Additional use (non re-create)	Returns from additional use (at average returns of 5% to 20%). Based on total downloads/uses.	Returns from additional use (at average returns of 5% to 20%). Based on total downloads/uses.	Returns from additional use (at average returns of 20% to 60%). Based on total downloads/uses.
Data re-creation costs	Not included	Up to max of.	Up to max of.

Source: *Authors' analysis based on the data from the individual studies.*