THE VALUE AND IMPACT OF THE BRITISH ATMOSPHERIC DATA CENTRE

Final Report, 30 September 2013

Neil Beagrie (Charles Beagrie Ltd) and

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

Charles Beagrie Ltd.
2 Helena Terrace
College Street
Salisbury
SP1 3AN

CONTENTS

Acknowledgements

We would like to thank: Daphne Charles for her input as a researcher to the study; Graham Parton, Sarah Callaghan, and colleagues at the British Atmospheric Data Centre for their collaboration and support; Simon Hodson, Neil Grindley, and Rachel Bruce at Jisc and Mark Thorley at NERC who provided feedback and guidance throughout; and finally, all the interviewees and survey respondents, who all gave valuable time and input to the study.

Neil Beagrie (Charles Beagrie Ltd) and John Houghton (CSES, Victoria University)

September 2013

Foreword - The Impact of Research Data Centres Report Series

This report forms part of a series of independent studies produced by the same authors on the value and impact of three UK research data centres. These reports cover the Economic and Social Research Data Service (ESDS), the Archaeology Data Service (ADS), and the British Atmospheric Data Centre (BADC). Each report was commissioned independently over a period of two years and there are therefore differences in these studies arising from varying call requirements and levels of funding. However, to enable ready comparison we have deliberately structured the reports from each study in a similar way. As independent studies each report is intended to be free-standing. Consequently, there is some commonality in sections of text across the reports, particularly when methods and approaches are discussed. To summarise and facilitate dissemination of key findings, a separate synthesis *The Value and Impact of Data Sharing and Curation: a synthesis of three recent studies on UK national data centres and services* is being prepared from all three studies for publication by Jisc.

BADC Impact Study Executive Summary

Background, Aims and Approaches

Jisc and other funders, together with Higher Education Institutions, are investing substantial resources in projects and services for the curation and long-term preservation of research data. It is a high priority area and there is strong interest in establishing the value and sustainability of these investments. The critical concept that determines how much or how little attention is paid to the long-term sustainability of digital content is how much value that content is perceived to have. This value is not solely economic, but in a tougher financial environment the economic arguments are increasingly important.

The proposal for this study of the British Atmospheric Data Centre (BADC) was put forward by the partners (Charles Beagrie Ltd and CSES Victoria University) to Jisc and the Natural Environment Research Council (NERC) following studies on the impact and value of the Economic and Social Science Data Service (ESDS) and the Archaeology Data Service (ADS). Hence, while free-standing, this study is part of a series on the value and impacts of research data services.

The aims of this study were to:

- Extend the feasibility testing and development of the partners' economic and survey collection methods to the BADC;
- Further refine the economic methods used in the studies of ESDS and ADS, testing and proving them in a different disciplinary area (science data); and
- Develop ways to assess and quantify the economic value and impact of such research data collections and services, with the ultimate objective of improving their prospects for sustainability.

For this study we have surveyed and analysed perceptions of the value of the digital collections held by the British Atmospheric Data Centre (BADC), and quantified the value and impact of those collections for BADC's user community using a range of economic approaches.

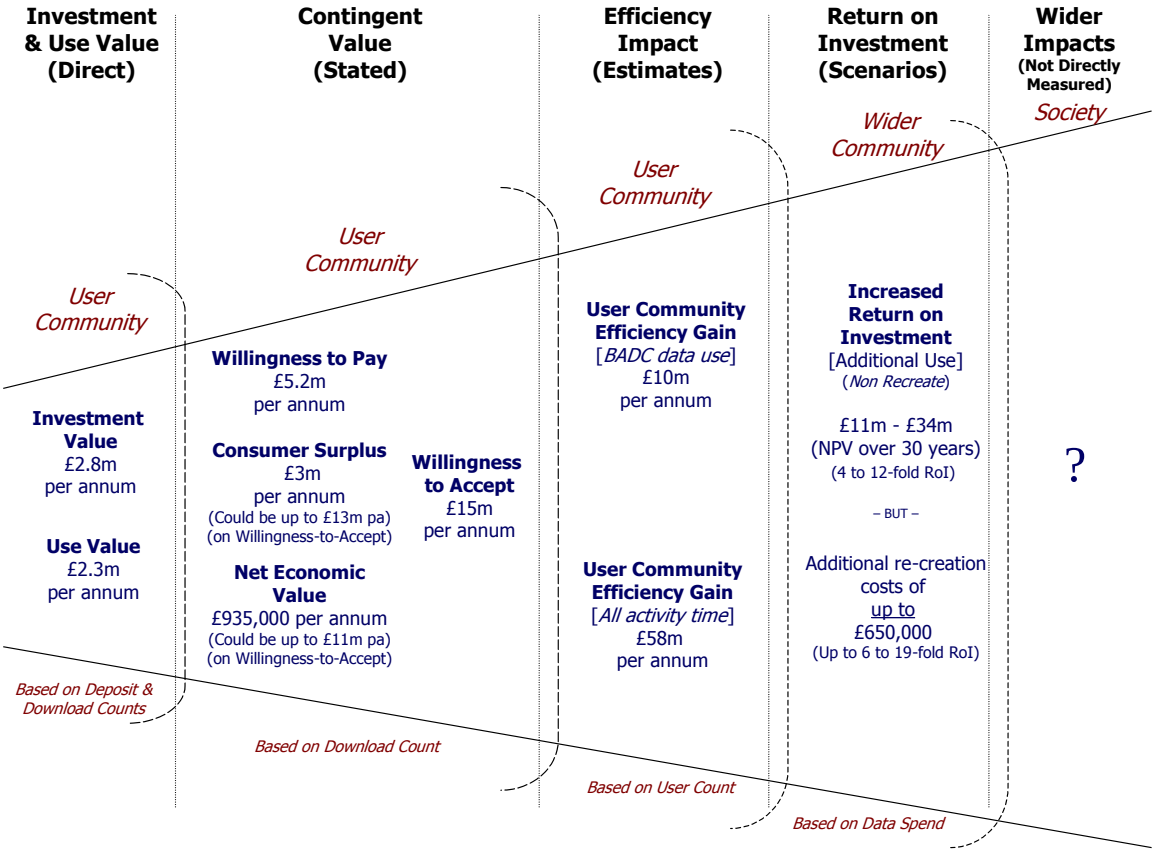
The study shows the benefits of integrating qualitative approaches exploring user perceptions and non-economic dimensions of value with quantitative economic approaches to measuring the value and impacts of research data services. Such a mix of methods is important in capturing and presenting the full range and dimensions of value. The approaches are complementary and mutually reinforcing, with stakeholder perceptions matching the economic findings. For example, both our qualitative and quantitative analyses highlight the important contribution of BADC data and services to research efficiency.

The Economic Value and Impact of BADC

While there were limitations in the data available, our *quantitative analysis* suggests that the economic benefits of BADC substantially exceed the operational costs. A very significant increase in research efficiency was reported by users as a result of their using the BADC, which we estimate to be worth at least £10 million per annum. We also estimate the value of the increase in return on

investment in data creation/collection resulting from the additional use facilitated by BADC to be between £11 million and £34 million over thirty years (net present value) from one-year's investment – effectively, a 4-fold to 12-fold return on investment.

Figure 1: The economic value and impact of the BADC



Source: Authors' analysis.

Overall, we estimate that:

- The direct investment value of BADC for data depositors is around £2.8 million per annum, and the use value of BADC to its user community around £2.3 million per annum at current prices and levels of use.
- Users' willingness to pay reveals that they value their access at around £5.2 million per annum – more than 2.5 times BADC's annual operating budget. Nevertheless, both quantitative and qualitative analysis shows that the free-to-access model is important in BADC achieving the impact it does, given resource constraints among its user community.
- When capacity to pay is limited, the amount that users would be willing to accept in return for giving up their access to BADC 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 £15 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, and we found that the efficiency impacts of BADC might be at least £10 million per annum, and could be as much as £58 million per annum – 2 to 11 times the sum of operational, depositor and user access time costs.

- Exploring scenarios relating to the impact of BADC data and services on returns to investment in the data held, we found that BADC facilitates additional use which may realise additional returns to the research and data creation/collection activities underpinning it that, to an approximation, may be worth between £11 million and £34 million over thirty years (net present value) from one-year's investment – effectively, a 4-fold to 12-fold return on investment.

- While different studies focus on different information services and content and use different methods and measures and are not strictly comparable, BADC's return on investment is comparable to and somewhat higher than reported returns to library and information services, which typically range from 3-fold to 6-fold.

The Qualitative Impact and Stakeholder Perceptions of BADC

Our *qualitative analysis* shows that interview and survey comments reveal strong support for the BADC, with many aware of the value of the services for them personally and for the wider user community.

For example, the user survey showed that 81% of the academic users who responded reported that BADC was very or extremely important for their academic research, and 53% of respondents reported that it would have a major or severe impact on their work if they could not access BADC data and services. In the depositor survey, depositors cited having the data preserved for the long-term and its dissemination being targeted to the academic community, as the most beneficial aspects of depositing data with BADC, both rated as a high or very high benefit by around 76% of respondents. Fifty-two per cent of depositor survey respondents suggested that being unable to deposit data with BADC would have a major or severe impact on them personally.

We have summarised qualitative benefits and value using approaches and formats developed by the Keeping Research Data Safe (KRDS) projects, particularly aspects of the KRDS Benefits Framework. The benefits identified were drawn from the interviews and surveys and the KRDS Benefits Analysis Toolkit and are shown in Table 6.2 of the report.

The quantitative and qualitative analyses independently show a similar picture of the value of BADC: they are complementary, reinforce each other, and lend credence to the findings. The qualitative findings illustrate individual user and depositor experience and benefits that can provide further insights into the economic analysis.

Recommendations

Recommendation 1: The unique combination of qualitative and quantitative approaches used in the BADC Impact Study 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. **There would be benefits from further research developing, refining and further exploring applications of the methods used in this study, as making the "business case" or funding case for data centres and services plays an increasingly important role in ensuring their sustainability.**

Recommendation 2: It is also clear in this and, to a lesser extent, the previous studies that different data services collect financial and operational data, such as user statistics, data deposit, access and download statistics, to varying levels of detail. More guidance is needed on the collection of such data to ensure a greater degree of standardisation of statistical records across data centres, as well as providing the basis for more comprehensive and reliable data for making the economic case for such data services. **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 case" or funding case.**

Recommendation 3: To date these approaches have only been applied to three UK national 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 at international, national and/or institutional levels as well as to the NERC data community.**

Recommendation 4: The study has looked at the aggregate value of BADC. There is also significant scope for more granular studies that focus on the value of specific collections or the economic value of BADC services to specific groups. There may also be some practical advantages to a narrower focus in simplifying some of the statistical 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 by the authors for the Archaeology Data Service (Beagrie and Houghton 2013), may also be beneficial. **These methods of valuation should also be applied at more granular levels than the overall collections or all stakeholders.**

Recommendation 5: Value and perceptions of value change over time. The BADC user community was previously surveyed in 2004, 2007 and 2010, and both BADC users and depositors were surveyed for this study in 2013. **BADC and funders should consider opportunities to repeat the BADC surveys and extend the available time series of comparative studies in future years. Ideally another survey of users and depositors should be considered within the next three to five years.**

Recommendation 6: 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 the uses to which the data are put after research use where substantial additional benefits can arise. To an extent, some of these impacts can be captured through the return on investment scenarios explored in our analysis. Nevertheless, there can be very substantial wider benefits. In the context of atmospheric data, 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 (see Box 4.1 in the main body of the report). **Consider research on the wider societal benefits and economic impacts that are generated by atmospheric research data sharing and curation, and the contribution to this made by BADC.**

1 INTRODUCTION

1.1 Background

Jisc, the Research Councils, and other funders together with HEIs are investing substantial resources in projects and services for research data. It is a high priority area for funders and institutions, and there is strong interest in establishing the value and sustainability of this investment.

Although a number of studies have looked at methods of determining cost-benefit and broad indicators of value, there remain significant challenges in establishing baseline data for measuring value and impact in a quantitative way and there are still only a relatively small number of socio-economic studies focussing specifically on the impact of research data services or research data infrastructure.

The Centre for Strategic Economic Studies at Victoria University and Charles Beagrie Ltd are currently working on a study for Jisc on the Impact of the Archaeology Data Service[1] and they have also recently completed a study for ESRC on the economic impact of the Economic and Social Research Data Service (Beagrie et al 2012). These studies have applied a unique combination of qualitative and quantitative methods to research data services in social science disciplines (ESDS) and to a humanities discipline (ADS). This study of the British Atmospheric Data Centre (BADC) applies the methods to a range of primarily science disciplines served by BADC. It is complementary to, and builds on, the existing impact studies, as well as earlier user surveys of the BADC itself (BADC 2007, RIN 2010, and Parton 2013).

This BADC value and impact study is funded primarily by Jisc, with additional funding from the Natural Environment Research Council (NERC), the BADC's core funder.

1.2 Aims of the Study

The aims of the study are set out in the project plan (Beagrie 2012). Namely: the project partners will extend feasibility testing and development of the economic and survey collection methods to the BADC, thus further refining the economic methods used in the study of ESDS and being used in the study of ADS, testing and proving them in a different disciplinary area. Developing ways to assess and quantify the economic value and impact of such research data collections and services has the ultimate objective of improving their prospects for sustainability.

In this study, we use a range of economic approaches and draw on baseline data gathered through desk research (e.g. BADC user statistics), online surveys and interviews. Information on the more direct benefits and impacts is derived partly from interview and survey responses from the producers and users of BADC data and services, draws on previous cost work at BADC and elsewhere, and on internal BADC reports and statistics. The work includes exploration of the costs and cost savings involved in using BADC data and services, the value to depositors and users, and impacts on the wider user and research communities.

[1] This study is also scheduled for release in late July/early August 2013. For further information see the ADS Impact project website: http://archaeologydataservice.ac.uk/research/impact

Although not within the formal aims of the study, it was also recognised by BADC staff that the online surveys of users and depositors we proposed for this study provided an opportunity in 2013 to complete another tri-annual customer survey in an existing longitudinal series, and could supplement previous BADC user surveys completed in 2007 and 2010. Where relevant, therefore, we designed specific sections of the surveys to provide comparable questions to the previous surveys (and in some cases to extend them to cover known gaps) as well as to provide other new data required for our study aims.

1.3 Brief Description of BADC

The BADC is the UK's Natural Environment Research Council's designated data centre for atmospheric sciences, delivering data curation and access services to the UK academic community. The BADC conducts this "National Capability" role within NERC's National Centre for Atmospheric Sciences (NCAS). This function is fulfilled by the BADC's role"… to assist UK researchers to locate, access, and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by Natural Environment Research Council (NERC) projects."[2] Given its role within the UK atmospheric academic community, it has also become the de facto point of contact for UK researchers needing access to the meteorological products of third party data providers including the UK Met Office and the European Centre for Medium Range Weather Forecasting (ECMWF). There is also considerable interest from the international research community in BADC data holdings, in particular the Met Office data.

The BADC is operated by the Centre for Environmental Data Archival (CEDA) group, within the RAL Space department at the Science and Technology Facility Council's Rutherford Appleton Laboratory (STFC RAL) in Oxfordshire. It is one of the three NERC data centres operated by CEDA, the others being the NERC Earth Observation Data Centre (NEODC) and the UK Solar System Data Centre (UKSSDC). CEDA also hosts the Intergovernmental Panel on Climate Change's Data Distribution Centre (IPCC-DDC).

The BADC came into existence in 1994 in response to NERC's requirement for a dedicated UK data centre for atmospheric research. Prior to this, the data holdings were managed by the Geophysical Data Facility (GDF) at the Rutherford Laboratory in Oxfordshire (later to become the STFC Rutherford Appleton Laboratory), from which the BADC evolved. Originally the GDF, operated by the Science and Engineering Research Council (SERC), served less than 200 registered users, from which the BADC's registered user community has now grown to more than 22,500 users (Parton 2013). During the intervening period, the BADC archive has grown to more than 1 Petabyte of accessible online data. It was amalgamated with the NERC Earth Observation Data Centre (NEODC) in 2005 to form CEDA. This included an amalgamation of hardware and user services across the BADC and NEODC. Since 2005, CEDA has grown further, incorporating the operation and systems of the UKSSDC and IPCC-DDC.

[2] BADC Website home page: http://badc.nerc.ac.uk/home/index.html

The data held at the BADC are of two types:

- Datasets produced by NERC-funded projects – these datasets are of high priority since the BADC may be the only long-term archive of the data, and indeed NERC has designated BADC to be the primary long-term archive of these data.
- Third party datasets that are required by the UK atmospheric research community and are most efficiently made available through one location (e.g. Met Office and ECMWF datasets). Here the BADC is primarily acting in a facilitation role to assist the wider research community to access data that would otherwise be prohibitive to them (e.g. on cost or technical grounds).

All BADC data are available on-line through the website or via an ftp service. Software is provided to assist in the manipulation of the data and extensive information is provided on the data collection procedures, formats, data quality, contact names and references to journal papers. Other specialist services include: the development of value added data products such as averaged and gridded data; visualisation services for viewing large datasets; resources to support project work, such as shared workspaces; and hosting processing resources and group workspaces alongside the archive.

Although there are some commonality of systems, operation, management and user community between the four data centres operated by CEDA, this report focus only on the BADC and its user community. However, the relative scales of BADC and NEODC operation and the closeness of the systems supporting these two data centres and their respective communities mean that the findings reported here for the BADC may also be indicative of a similar findings for the NEODC as well, however exploration of this lies beyond the scope of this report.

1.4 Report Contents

This report present a comprehensive account of the research undertaken for this study, as it both explores a range of qualitative and quantitative methods that can be used in assessing the value and impacts of a research data services such as BADC, and presents the results of the analysis.

Beginning with a brief introduction to the approaches and methods used to collect the data necessary for analysis (Section 2), this report then presents a brief review of previous studies of the BADC user population and BADC statistics (Section 3). While such BADC studies and statistics are often historic and not contemporary with ours, they provide some context for our work.

The report then presents a brief review of other studies focusing more narrowly on a range of related information and data services, which highlights the quantitative approaches used (Section 4). This provides the platform for an outline of the economic approaches used in this study.

Section 5 presents a detailed account of the quantitative economic analysis of the value and impacts of BADC undertaken for this study. Examining the practically and merits of the various economic approaches employed, it concludes by suggesting that the methods used can be successfully deployed and that they demonstrate the "business" or funding case for BADC.

Section 6 presents an account of the complementary qualitative approaches and analysis, and shows not only how the qualitative survey responses, interviews and stakeholder interactions are in

concordance with the quantitative economic analysis, but also how they provide insights into the nature and dimensions of value and impact.

This is followed by a brief summary of findings, conclusions and recommendation (Section 7).

Appendix 1 presents a full summary of the user and depositor survey results.

Appendix 2 presents additional data cuts by registered users, non-registered users and “unsure” for respondents to Questions 1-4 of the user survey.

2 APPROACH AND METHODS USED TO COLLECT DATA

In selecting conceptual approaches for this study, we have taken account of the practical limitations of collecting the necessary data through survey and interview techniques, and sought to maximize economy in data collection through commonality (i.e. the same data can be used to inform more than one of the approaches). We combined:

- Desk-based analysis of existing evaluation literature and reports, looking at both methods and findings;
- Existing data from KRDS and other studies of the costs and benefits of research data infrastructure and services;
- Existing management and internal data collected by BADC, such as user registration and access statistics, deposit records, internal operational and financial reports; and
- Original data collection in the form of an online survey of BADC users and depositors, and semi-structured interviews.

After a brief introduction to the data collection methods used, we present the economic analysis which is based primarily on the online surveys – the detailed results of which are described in Appendix 1.

2.1 Desk Research

Desk research included: analysis of existing evaluation literature; existing data from KRDS and other studies of the costs and benefits of research data infrastructure and services; reports of previous BADC user surveys in 2007 and 2010; and analysis of existing management and internal data collected by BADC, such as user registration and access statistics, deposit records, internal operational and financial reports.

2.2 Interviews

A list of potential interviewees was compiled in consultation with BADC to cover a representative range of users and depositors. Thirteen users and depositors nominated by BADC agreed to be interviewed by telephone. Of these, all were registered users (though not all had actually downloaded or used any data from the BADC in the last 12 months), and six were also depositors or generators of data hosted by BADC. Semi-structured interviews were conducted using a pre-defined questionnaire. Not only did these interviews provide valuable information, they also provided the opportunity to test and inform development of questions for the surveys.

2.3 Online Surveys

Two online surveys were conducted, aimed at depositors of data with BADC and BADC users, respectively. The survey questionnaires were developed iteratively by the project team with external review and input from BADC staff and a small group of BADC customers who helped test and refine pilot versions of the surveys. Given the nature of some of the economic approaches being explored, and the range of affiliations, roles and seniority of the survey populations, substantial effort was needed to design questionnaires suitable for an online survey.

Significant effort, therefore, was spent on trying to reduce the likely burden on recipients in terms of time to complete the questionnaires, wording of the invitations to participate, advance notice of the surveys via the CEDA websites, news feeds, social media and email shots, and offering five £50 Amazon vouchers in a draw for participants. These were followed up with similar reminders at appropriate intervals during the survey's open period, which solicited an additional ~30 % increase in the number of responses. As a result, the surveys enjoyed high response rates and good completion rates given the topics and number of non-mandatory questions.

The questionnaires used a range of standard survey approaches, including question logic in their design to steer respondents through the appropriate sections of the questionnaire, and use of "critical instances", such as the last data accessed/downloaded (for users) or last data deposited/updated (for depositors). A number of questions also sought specific information on the costs of creating and collecting the data, the time and cost involved in preparing it for deposit, the time and cost of access for users, the benefits and efficiency impacts of access, and contingent valuation (i.e. willingness to pay or accept) using stated preference techniques. Answers to these questions must be interpreted carefully, in the context of open-ended text comments in the surveys and other findings from the interviews and desk research, to ensure that protest and outlier answers are excluded from the economic analysis, or included with suitable caveats. These questions were supplemented by qualitative questions asking for views on the importance and impact of BADC for both depositors and users.

The question logic and the questions themselves are set out in Appendix 1 with the survey responses.

3 PREVIOUS BADC USER SURVEYS AND BADC STATISTICS

3.1 BADC User Survey 2007

The aim of the 2007 BADC User survey was threefold:

- To determine the skills base of the BADC user community;
- To determine the experience of the BADC user community of its datasets and supporting services at that time; and
- To identify those areas where the BADC should improve, those areas where the BADC was doing well, and those areas where the BADC could explore/develop.

Some 285 respondents completed the survey, of whom 87% were registered users, 7% not registered, and 6% did not know whether they were registered or not. Some 74% of respondents were UK based. Other demographic profiling is limited to areas of research.

The survey looked at the browser and operating system respondents were using, their familiarity with data formats and their proficiency with data analysis tools. In the analysis, a distinction was made between those in the BADC target user community (those from the atmospheric and climate change science communities) and those outside it.

The survey also investigated user satisfaction with the various BADC services, including ease of locating data, and concluded that the BADC was providing a good service that met the expectations of the user community. However, the survey did not examine the BADC's impact or related economic factors.

3.2 RIN Survey 2010

The RIN study on *Data centres: their use, value and impact* (RIN, 2011) included BADC as one of eight research data centres examined, five of which were surveyed. There were 759 respondents to the RIN survey from BADC users, of whom 67% were from the academic community.

The RIN survey was conducted jointly with CEDA to allow a parallel consultation of the NEODC user community alongside that of BADC. Parton (2011) provides additional information on the profile of the BADC respondents that is not available from the RIN report itself, namely: although the 2010 survey did not capture the location of respondents, institute names for the respondents were requested. Taking a conservative estimate from this information, Parton suggests that some 51% are identifiable as being UK based.

The RIN study notes that data centres had improved the culture of data sharing and re-use. Some 69% of BADC respondents indicated that the impact had been large, and a further 29% agreed that there was some degree of improvement (RIN 2011, p.30).

The most widely-agreed benefit of data centres was research efficiency. Data centres make research quicker, easier and cheaper, and ensure that work is not repeated unnecessarily. Research quality is another important benefit, although not rated quite as highly as efficiency. Data centres provide

services and support which are highly valued by researchers, including: user support; access to otherwise-unavailable datasets via reciprocal sharing arrangements; and curation, preservation and long-term access for datasets, both for their own research and for datasets created by others.

The RIN study found it proved more difficult to identify areas where research based upon data centre resources had gone on to have significant social, economic or environmental impacts. Nevertheless, a few cases did illustrate how research based on data centre resources has had a positive impact upon wider society and the economy through the development of new tools and methodologies, new policies and regulatory controls, and new products or services. A number of the examples cited involve data obtained from the BADC.

The RIN study surveyed users of five of the data centres and quantified the research benefits under three main categories. The figures for BADC have been extrapolated from three separate figures in their study report and presented in Table 3.1.

Table 3.1: BADC Research Benefits (extrapolated from RIN 2011)

Data Centre Research Benefits	*BADC N=*	*BADC %*
Research efficiency		
It has reduced the time required for data acquisition / processing	618	68%
It has improved the efficiency of research	622	62%
It has reduced the financial cost of data acquisition / processing	612	62%
It has reduced duplication of effort (i.e. unnecessary recreation of data)	609	57%
It has enabled me to undertake a greater quantity of research	614	42%
Research quality		
It has increased the use of data in my research	614	40%
It has improved the quality of the data I use within my research	613	47%
It has improved the evidence base of my research	624	46%
It has helped to improve the quality of my research outputs	620	47%
Research novelty		
It has created new intellectual opportunities (e.g. merging of several data sets to answer new questions)	600	33%
It has enabled research to go ahead that otherwise might not have done	636	48%
It has permitted more novel research questions to be answered / tackled	616	38%
It has enabled new types of research to be carried out	604	34%

Source: Data extrapolated from RIN 2011, Figures 23, 24, 25 pp 34-39.

3.3 BADC User Stats Report 2013

Over the course of the study, significant support was provided by Graham Parton from CEDA in the collation and interpretation of current and historic BADC user statistics. Towards the end of the study this work has been collated in a *BADC User Statistics Report* (Parton 2013). This report presents details of the current and historical user base where suitable information was available to the author. It discusses BADC user statistics for the following user groups: registered users; active

registered users; non-registered users; and profiles of users registered in a specific calendar year. The primary sources of information for the review were the user database maintained by CEDA, HTTP and FTP download logs, and BADC website access logs.

The report examines changes in the user community registering to use the BADC and reports the number of active registered users. The report also analyses the access statistics attributable to registered users compared with and other access types to estimate the number of active non-registered users. While the BADC's primary user community of university based atmospheric and climate change researchers in the UK has remained an important part of the user community, Parton (2013) noted that it is clear the BADC is increasingly being used by a geographically and demographically more diverse user community. The report concludes that these findings emphasise that data centres such as the BADC can both facilitate the research of the community it is dedicated to, but also enable research from a wider community too, demonstrating the wider impact of such facilities and the data they hold.

For the purposes of this study, the key user statistics groups and their characteristics discussed in the BADC user statistics report can be summarised as shown in Table 3.2.

Table 3.2 Key BADC User Statistics

Statistical Group	Comments
Registered Users	Statistics on registered users are taken from the CEDA user registration database. This has been in use for the BADC since 1996. The register is cumulative so it contains both current active users and historic (inactive) users. There are currently around 22,500 registered users. Registered users provide profile information at registration, but their updating of that information as it changes in subsequent years is known to be very uneven. However, as users are required to review their details when applying for restricted content, the most recently active registered users are known to keep key profiling information, such as their title and institute details, up to date. However, information for non-active users, and on user research domain for the recently active users, remains largely static.
Active Registered Users	This is a subset of registered users. Active registered users are defined as registered users who have accessed a restricted BADC dataset during the 12 months prior to the CEDA reporting period in April each year. There were 3,497 active registered users during 2012.
Non-Registered Users	Approximately 50% of BADC datasets are available publicly via the Web without registration. Statistics on these public accesses, including IP addresses, are available in BADC download logs. Using this information and comparing it with that for registered users BADC have estimated that there could have been around 2,462 active non-registered users during Jan-Dec 2012. Relatively little is known about this group of users.
Profiles of Users	Comparison of profiles of users registered in specific years provides

Registered in a Specific Calendar Year	information on trends and changes in the registered user base and overcomes issues of currency in the overall registered user database.

3.4 Relevance to the BADC Impact Study

The previous user surveys in 2007 and 2010 have areas of overlap with the surveys undertaken for this study that allow some broad comparisons over a six year period. Results from this study are presented in full in Appendix 1 and discussed in sections 5 and 6.

The user survey populations have increased from 285 in 2007, to 759 in 2010, and 1,141 in 2013 reflecting more effective recruitment to the surveys over time as well as some growth in the user population in this period. Although circumstantial evidence was available in the 2010 survey, the present study is the first to address BADC depositors specifically alongside its users.

BADC allows optional user registration. In the 2007 survey, 87% of respondents reported that they had registered, 7% not registered, and 6% were not sure. The respective figures in the 2013 survey were 78% registered, 6% had not, and the remaining 16% were not sure. No comparable figures are recorded for the 2010 survey. The 2013 figures are broadly comparable with those reported in the 2007 BADC User Survey, but with a somewhat lower number of respondents stating that they were registered in tandem with a somewhat higher number being unsure. This change may reflect the often historic nature of user registration and the passage of time between the surveys.

Where recorded, the sector and subject domains between the three surveys are similar with a gradual expansion of the range of users and disciplines over time: in the 2010 user survey 67% of users were from the academic sector and in the 2013 user survey 61% from the academic sector. There is a noticeable downward trend in the percentage of respondents from the UK over the three surveys with 74% being UK based in 2007, an estimated 51% in the 2010 survey, and 43% in the 2013 survey – although the absolute number continues to increase. The 2013 survey allows some additional profiling of the non-registered users that was not possible in previous surveys (see Appendix 2).

The statistics in the BADC User Statistics Report are important for comparison with the survey populations but, as noted in table 3.2, each has relative strengths and weaknesses. Information on non-registered users is particularly difficult and estimation methods are needed. No statistical group provides a direct comparison with the 2013 survey population, which contains known active registered users, known active non-registered users, and a relatively high 16% of active users who were unsure if they were registered or not.

The survey populations in the 2007 and 2010 surveys are seen as a statistically sound subset of the BADC user community and thus the conclusions from the surveys can be taken as indicative of the experiences of the BADC user community at that time. In terms of sectors and main subject domains, the 2013 survey is similar to those of 2007 and 2010 and the BADC registered users. On this basis the 2013 survey population may also be seen as representative, although the geographical split continues to diverge over time in the three surveys, as noted above.

Full details of the 2013 survey are presented in Appendix 1.

4 Approach and Methods to Measure Value and Impact

This section presents a brief review of past studies that seek to estimate the value of research-related and publicly-funded information and services, and then outlines the quantitative economic and qualitative approaches that we use for this study.

While there are many forms of value and many ways to estimate the economic costs, benefits and impacts of an activity, we focus on the more direct value of the BADC data service to its users. We examine and, where possible, seeking to quantify, the value of the data service to its users, rather than examine what those users subsequently do and the wider impact of their work. Consequently, there are wider economic and societal benefits and impacts of value that are not covered in the following economic analysis, but which we seek to show in more general qualitative approaches used in Section 6.

4.1 A Brief Review of Approaches to Valuing Research and Information Services

Much has been written about the costs and, to a lesser extent, the benefits of more open access to research publications. To date, somewhat less attention has been paid to the value of the open curation and sharing of research data, although a few studies are emerging. There are also related literatures on the value of the arts, museums, library and information services, and public sector information (PSI). Our purpose here is to briefly explore and draw ideas from these literatures, and to assess which approaches might be most usefully applied and adapted for this study. A key criterion in selection of approaches is the practicality of data collection using survey and interview techniques.

4.1.1 Library and information services and public sector information

Library and information services have been the focus of many studies estimating their value in quantitative and qualitative terms, with the former including numerous studies based on investment and use value and estimates of consumer surplus, and a number based on contingent valuation, leading to cost-benefit analyses (Noonan 2003; Missingham 2005; Imholz and Arns 2007; Svanhild 2009; Oakleaf 2010). Most focus on the more direct economic impacts or value and calculate a Return on Investment (RoI), but some explore wider impacts (e.g. the contribution of library spending to employment and gross domestic product) and outcomes (e.g. the relationship between library spending and successful grant applications (Tenopir et al. 2010)). One example, measuring the economic impact of the British Library, combined contingent valuation in the forms of willingness to pay and accept with investment value and estimates of the cost of alternatives (British Library 2004). Such an approach to library valuations has become a, if not *the*, standard.

Indeed, the use of contingent valuation has become increasingly common in a range of areas where there are intrinsic values associated with what is being valued, including the environment, the arts and museum services, as well as library and information services (Bakhshi et al. 2009). Based on public choice, contingent valuation in the form of willingness to pay and/or willingness to accept captures the preferences of people as they choose to spend their money and time on A rather than

B, thereby foregoing the latter. Contingent valuation is also increasingly widely used in marketing, and in that context is often the basis for conjoint analysis, which involves respondents being asked to make explicit, rather than implicit, trade-offs (Ramirez 2009).

Public sector information (PSI) has been the focus of a number of studies seeking to estimate its value and the benefits to be derived from making it freely available. Some forms of PSI are similar to social, scientific and observational data (e.g. national statistics, geospatial data, etc.).

PIRA (2000) combined measures of the investment cost (i.e. the amount spent on the collection/generation of the information) and expenditure on PSI by users and re-users, then, for final users, estimated the value as expenditure on PSI or, where the PSI was freely available, as the investment cost of its collection/generation. They estimated the investment value of PSI (i.e. what governments invest in the acquisition of PSI) and the economic value of PSI (i.e. the national income attributable to activities built on the exploitation of PSI). In the European Union, they put the former at around EUR 9.5 billion per annum in 1999, and the latter at around EUR 68 billion (equivalent to approximately 1.4% of EU GDP).

The PIRA report was popularised by Weiss (2001) in the influential report *Borders in Cyberspace*. Drawing on the PIRA report, Weiss highlighted the comparison between the US and Europe, noting that the US invested twice as much as Europe in PSI, but earned 40 times more from it. Weiss suggested that this was because the US had an open access model for PSI, whereas the EU countries used a cost recovery approach. Aside from the many difficulties in estimation and attribution, a potential problem with the PIRA approach is that it may overestimate the value of PSI because it does not account for the possible use of alternative information.

In the MEPSIR (Measuring European Public Sector Information Resources) study of Dekkers et al. (2006), demand and economic performance were measured in an extensive survey by directly asking both PSI holders and re-users for key economic data, such as total turnover against turnover related to PSI, total number of staff against the number of staff dedicated to handling PSI, and estimates of the domestic market size for particular types of PSI. The market value was then estimated from the average revenues multiplied by the average number of re-users per PSI domain, minus the cost of PSI collection/generation. This produced a much lower number than suggested by the PIRA study, despite it being market size rather than value added, and coming five years later. Making some adjustments to the MEPSIR estimates with the benefit of hindsight, te Velde (2009) suggested that the value might drop further – to between one-fifteenth to one-twentieth those reported in the original PIRA study.

In their report to the UK Office of Fair Trading, DotEcon (2006) and collaborator Pollock (2009) adopted a bottom-up approach to estimating the economic value of PSI products and services in the UK in an effort to overcome some of the limitations of the PIRA approach. DotEcon adopted a contingent valuation approach and estimated the net economic value of PSI from willingness to pay for PSI minus the cost of supplying it. Using a survey and published sources, the value of PSI was estimated from the net consumer surplus from PSI (i.e. the amount that customers might be prepared to pay over and above what they do pay), and the total producer surplus that arises from the provision of PSI (i.e. the extent to which revenues exceed the costs of supplying the product or

service). Adding these two estimates gave the net economic value of PSI in the UK - around £ 590 million per annum in 2005.

The DotEcon report also provided estimates of the value lost from not making PSI freely available, by looking at the consumer detriment resulting from: unduly high pricing; restriction of downstream competition, such as refusing to supply or discrimination; and failure to exploit PSI. This goes to the heart of important counterfactual issues by starting to look at the value of things that do not happen unless data is made freely available. They suggested that the net value of PSI in the UK could have been approximately doubled by resolving the problems identified. While much less subject to over-estimating the value of PSI, a potential weakness of this approach lay in estimating price elasticities of demand, especially where the PSI was supplied free of charge – rather throwing the necessary assumption of linearity into question.

These studies suggest the considerable potential of approaches based on investment and use value, and estimating net economic value from a survey-based approach to willingness to pay or accept (i.e. contingent valuation through stated preferences). They also suggest that the challenge of better integrating the economic value with other qualitative forms of value largely remains.

4.1.2 Research publications

Research publications have been the focus of quite extensive economic analysis, although most focus on the costs of creation, access and preservation, and few studies look at the value and benefits of access. Those addressing the value and benefits have done so from the supply-side, in the form of macro-economic analysis based on estimates of the impacts of changes in accessibility and efficiency on returns to R&D expenditure, and from the demand-side, in the form of estimates of the impact of research on innovation and the value of that innovation to firms (Mansfield 1991, 1998; Beise and Stahl 1998; Houghton et al. 2011). The latter approach is not readily applicable to many research data services as commercial users often make up a relatively small share of total users. While the former approach is relatively data intensive, some aspects of it could be applied to research data services.

Houghton et al. (2006) and Houghton et al. (2009a) were among the first studies to explore the costs *and* benefits of open access to research publications. Houghton et al. (2009a) outlined a detailed activity cost model, based on the IDEF0[3] modelling standard that is often used in business process reengineering, and used it as the basis for their analysis of the potential costs and cost savings throughout the scholarly communication lifecycle arising from alternative publishing models. The focus of this part of the work was activity-based costing, looking at the activity cost differences between the alternative models and efficiency gains in terms of research information search, discovery and access, research library negotiation, acquisition and handling, publisher handling, and use efficiencies.

Exploring the wider impacts of more open access to research findings, Houghton and Sheehan (2009) and Houghton et al. (2009a) developed a modified Solow-Swan model to estimate the impacts of changes in the accessibility of the information and efficiency in its access and use on

[3] http://en.wikipedia.org/wiki/IDEF0

returns to R&D expenditure. The standard Solow-Swan approach makes some key simplifying assumptions, including that:

- All R&D generates knowledge that is useful in economic or social terms (efficiency of R&D); and
- All knowledge is equally accessible to all entities that could make productive use of it (accessibility of knowledge).

Obviously, these assumptions are not realistic. In the real world, there are limits to efficiency and barriers to access. Addressing these real world limitations, they introduced *accessibility* and *efficiency* into the standard Solow-Swan model as negative or friction variables, then explored the impact on returns to R&D of reducing the friction by increasing accessibility and efficiency.

Houghton et al. (2009a) then put the costs, cost savings and returns to R&D together into a cost-benefit analysis. Because there is a lag between research expenditure and the realisation of economic and social returns to that research, the impact on returns to R&D was lagged and the value of those returns discounted accordingly. The cost-benefit comparisons were made over a 20 year transitional period, and suggested that the cost savings and additional returns to R&D resulting from enhanced accessibility and efficiency would exceed the costs of open access publishing models.

To date, this is one of the few methods to have explored both the direct efficiency impacts of more open access to information, in the form of activity costs and cost savings, and the wider economic and efficiency impacts, in the form of increased returns to R&D arising from increasing the accessibility of the information. It has been applied in a number of subsequent studies by the original authors and colleagues and by others (Houghton et al. 2009b; Houghton 2009; CEPA/RIN 2011).

There is an increasing number of research funding agencies mandating that the findings from the research they fund be made openly and freely available. In the United States, the National Institutes of Health (NIH) spends USD 28 billion on research annually, resulting in around 65,000 peer-reviewed articles. The NIH Public Access Policy "requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication."[4] Through the PubMed repository, these manuscripts and materials are made available to the public. *The Federal Research Public Access Act* (FRPAA) had sought to extend the NIH policy by requiring that US government agencies with annual extramural research expenditures more than USD 100 million make manuscripts of journal articles stemming from research funded by that agency publicly available via the Internet.[5]

Houghton et al. (2010) outlined one possible approach to measuring the impacts of the proposed US FRPAA on returns to public investment in R&D. The project involved further development and refinement of the modified Solow-Swan model discussed above (Houghton and Sheehan 2009, and Houghton et al. 2009a), particularly in relation to the most appropriate lag and distribution over time of returns to R&D, the most appropriate depreciation rate for the underlying stock of R&D knowledge arising from federally funded R&D, and metrics to measure potential changes in accessibility and efficiency.

[4] http://publicaccess.nih.gov/

[5] http://www.taxpayeraccess.org/issues/frpaa/index.shtml The FRPAA has now been superceded by the Fair Access to Science and Technology Research Act (FASTR).

Key data required for the modelling included: the implied archiving costs, the volume of federally funded research outputs (e.g. journal articles), and the levels of federal research funding and expenditure trends. The preliminary analysis used publicly available sources and published estimates. Data relating to federal research funding, activities and outputs were taken from the National Science Board (2010), and the analysis explored three sources for archiving costs: the LIFE^2 Project lifecycle costs (Ayris et al. 2008), and submission equivalent costings from arXiv (2010) and NIH (2008).

Preliminary modelling by Houghton et al. (2010) suggested that over a transitional period of 30 years, the potential incremental benefits (i.e. over and above the existing NIH mandate) of the proposed FRPAA archiving mandate for all federally funded R&D might be worth around:

- Four times the estimated cost using the higher end LIFE^2 lifecycle costing;
- Eight times the cost using the NIH costing, which it was suggested would probably be the best estimate; and
- Twenty-four times the cost using the historical arXiv costing.

Perhaps two-thirds of these benefits would accrue within the US, with the remainder spilling over to other countries. Hence, the US national benefits might be of the order of five times the costs, with the benefits from increased accessibility worth more than USD 1 billion over 30 years (Net Present Value).

These studies demonstrate that more open access to publicly funded research publications can bring cost savings and efficiency improvements for both the producers and users of the information, as well as increasing the return on investment in the research by making it more accessible. Crucially, they address the former through activity costing and the latter through a return on investment approach.

4.1.3 Research data

Research data repositories have been the topic of a number of studies over recent years, as detailed below. Efforts to understand the costs and benefits involved in research data curation and sharing typically mix quantitative and qualitative methods, and rely primarily on case studies and extrapolation. Some have provided templates for assessing costs and benefits, but few have tried to look at the value or benefits of open access to a wide range of research data types.

In a series of projects under the title Keeping Research Data Safe (KRDS), Beagrie et al. (2008; 2010) explored the costs and benefits of research data curation and sharing in the UK and elsewhere in Europe. The initial KRDS study investigated the medium to long-term costs of the preservation of research data to Higher Education Institutions (HEIs), and provided a brief overview of the potential benefits to HEIs from the preservation and sharing of research data. It developed a framework and guidance for determining costs consisting of: a list of key cost variables and potential units of record; an activity model divided into pre-archive, archive, and support services, and by duration of activity; and a resources template, including major cost categories.

A series of case studies from Cambridge University, King's College London, Southampton University, and the Archaeology Data Service at York University illustrated different aspects of costs for research data within HEIs. Selective illustrations of cost-benefits and costs over time were also provided.

Importantly, the study noted that the costs of a central data repository are an order of magnitude greater than that suggested for a typical institutional repository focused on e-publications alone. This was largely due to the additional infrastructure and staff (for documentation, metadata, validation, and support) required to promote access and re-use of datasets.

Fry et al. (2008) sought to identify the value of, and benefits arising from, the curation and open sharing of research data. They suggested that potential benefits include:

- Maximised return on investment in data collection;
- Broader access, where costs would be prohibitive for individual researchers/institutions;
- Potential for new discoveries from existing data, especially where data are aggregated and integrated;
- Reduced duplication of data collection costs and increased transparency of the scientific record;
- Increased research impact and reduced time-lag in realising those impacts; and
- New collaborations and new knowledge-based industries.

The Fry et al. (2008) study used a mixed-method approach, including a literature review and qualitative case studies, to inform the development of a model on which to build a business case for data sharing in UK Higher Education. This was based on extensions to the research data preservation cost model proposed by Beagrie et al. (2008, 2010), to allow estimation of costs and benefits to users depositing or accessing data. Based on the work of co-authors Houghton and Rasmussen, the report presented a simple example of cost-benefit analysis applicable to an individual dataset or repository, based on costs and potential cost savings. It described the data requirements and walked the reader through the process step-by-step. The approach was then extended to explore the more diffuse benefits of data curation and sharing at the institutional and disciplinary levels. Importantly, the report included an outline questionnaire and template to facilitate cost-benefit analysis.

Recognising that no single approach has dominated across the studies of research publications, research data and PSI that attempt to measure the value and economic impacts, Beagrie et al. (2012) drew on a number of approaches to explore the economic value and benefits of the UK Economic and Social Data Service (ESDS). They began with approaches that can be seen as estimating minimum values, and moved progressively toward approaches that can be seen as measuring some of the wider values. These included: investment and use value, contingent valuation using stated preference techniques, economic welfare in the form of consumer surplus and net economic value. Wider benefits and impacts were explored by looking at the efficiency gains enjoyed by users and assigning an economic value to them (e.g. an activity cost savings), and by estimating the impacts of increased data use on returns to investment in the data collection/creation and the related data infrastructure services necessary for hosting and sharing the data. As these latter impacts are recurring during the useful life of the data, Beagrie et al. (2012) used a simple Perpetual Inventory Method to estimate the overall value of the impacts over time. The analysis was based on extensive user and depositor surveys. In addition to the economic

analysis, Beagrie et al. (2012) drew on approaches, such as the KRDS Benefits Framework, and impact case studies to illustrate qualitative benefits defined in the surveys, and agency and user interviews that could not be reflected fully by economic analysis alone. The approaches used to explore the economic value and benefits of the ESDS have subsequently been applied in a study exploring the value and impact of the Archaeology Data Service (Beagrie and Houghton 2013). These studies explore the direct costs and benefits and wider economic impacts, and demonstrate the possibility of using a number of approaches to estimating the value and impacts of research data services.

Box 4.1: The value of atmospheric data and its use (US NOAA)

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 from the uses to which the data are put 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 analysis (below). Nevertheless, there can be very substantial wider benefits. In the context of atmospheric data, work by the US National Oceanic and Atmospheric Administration (NOAA) on the value of meteorological data is indicative.

Studies conducted or commissioned by NOAA have shown that the sum of US public and private sector meteorology expenditures is around USD 5.1 billion (including all operations and research). The total value of weather services to US households was estimated to be USD 286 per household and the net benefit USD 26 billion a year (i.e. USD 31 billion in benefits minus USD 5 billion in costs), a cost–benefit ratio of 6.2 (Lazo et al. 2009).

It has also been suggested that US electricity producers save USD 166 million annually using 24-hour temperature forecasts to improve the mix of generating units that are available to meet electricity demand. Incremental benefits are relevant in assessing the merits of investments that will improve forecast accuracy. The incremental benefit of an improvement in temperature forecast accuracy is estimated to be around USD 1.4 million per percentage point of improvement per year. For a 1°C improvement in accuracy, the benefit is about USD 59 million per year. It is estimated that a perfect forecast would add USD 75 million to these savings (Teisberg et al. 2010).

For example, for every US Dollar that railway companies spend in acquiring NOAA climate data, it is estimated that they receive a USD 13,140 savings in infrastructure costs that would be required to maintain their own climate database storage, archiving, and reporting system. After extrapolating these savings to the entire Class I freight railroad sector, the potential benefits are approximately USD 11.5 million (Centrec Consulting Group, 2005).

Source: NWS & NOAA (2011) *Value of a Weather Ready Nation*. Washington DC.

4.2 Approaches Used in This Study

A review of previous studies of the value of information services, research publications and data suggests that no single approach has dominated across related and yet diverse fields. Consequently, building on our studies of ESDS and ADS (Beagrie et al. 2012 and Beagrie and Houghton 2013), we propose to draw on a number of approaches to explore the economic value and impacts of BADC data and services, beginning with approaches that can be seen as estimating minimum values and moving progressively toward approaches that can be seen as measuring some of the wider value. These include:

- Investment and use value;
- Contingent valuation, using stated preference techniques;
- Welfare approaches to estimating consumer surplus and net economic value;
- An activity cost approach to exploring the efficiency impacts of BADC data and services; and
- An approach that seeks to explore the impacts of increased use on returns to investment in data creation/collection.

The Keeping Research Data Safe (KRDS) Benefits Framework is then used to present a summary of qualitative benefits and value identified in the interviews and surveys.

In selecting these approaches, we have taken account of the practical limitations of collecting the necessary data through survey techniques, and sought to maximize economy in data collection through commonality (i.e. the same data can be used to inform more than one of the approaches).

4.2.1 Investment and use value

The most direct indicators of value are the investment value (i.e. the amount of resources spent on the production of the good or service) and use value (i.e. the amount of resources spent by users in obtaining the good or service). Measures of the investment in access suggest the minimum amount that the good or service is worth to the consumers.

Both investment and use value can be established from user and depositor surveys through questions about the time and costs involved in the creation of the data, preparation and deposit of the data, and its discovery, access and use, together with usage statistics and financial information from the BADC.

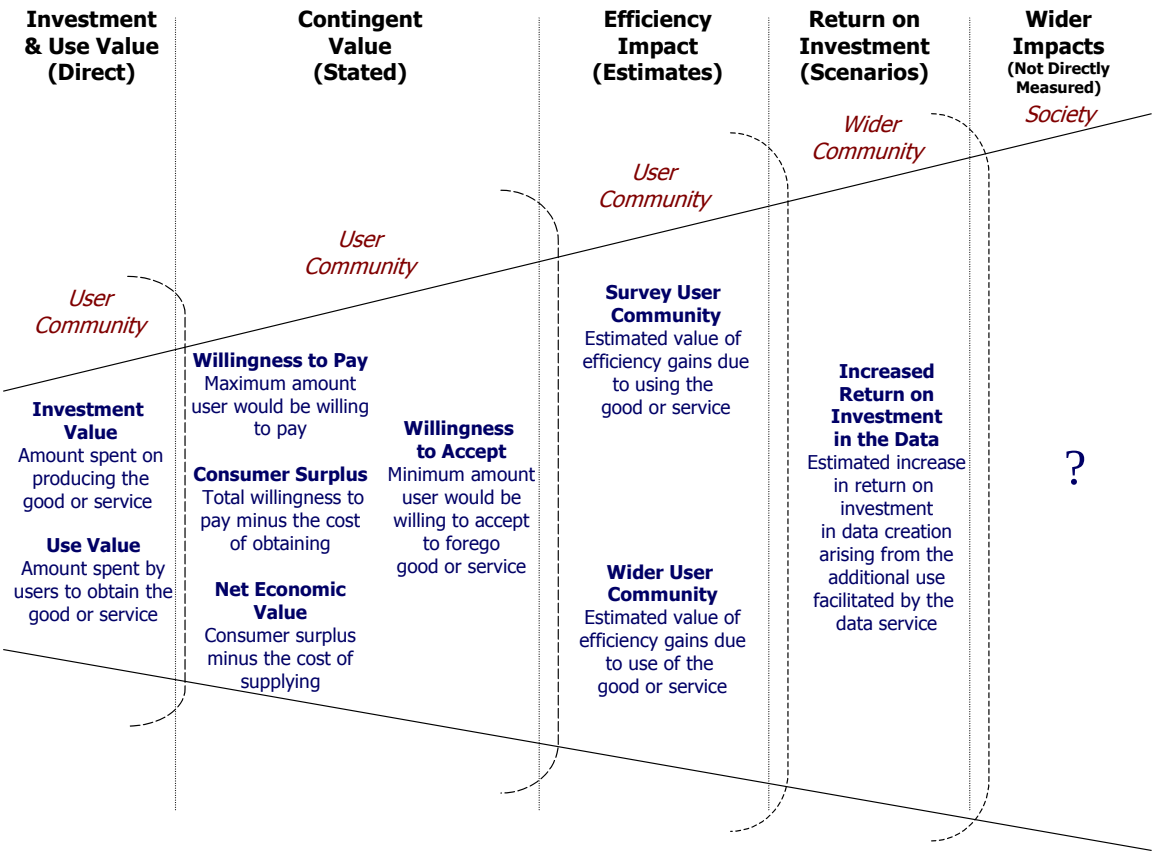
4.2.2 Contingent value

Contingent valuation involves the assignment of money values to non-market goods and services based on preferences (DTLR, 2002). If a good or service contributes to human welfare, it has economic value, and whether something contributes to an individual's welfare is determined by whether or not it satisfies that individual's preferences. An individual's welfare is higher in situation A than situation B, if the individual prefers A to B. Preferences are revealed by what an individual is willing to pay for a good or service and/or by the amount of time and other resources spent obtaining the preferred good or service. Where preferences are not revealed in the market,

individuals can be asked what they would be willing to pay or to accept in return for the good or service in a hypothetical market situation (i.e. stated preferences). For a public good the value is the sum of "willingnesses", as consumption is non-rivalrous (i.e. the same information can be consumed many times).

The key difference between willingness to pay and willingness to accept is that the former is constrained by the person's ability to pay (typically by disposable income), whereas the latter is not. Hence, willingness to pay directly measures the demand curve with a budgetary constraint and willingness to accept measures the demand curve without a budgetary constraint (British Library 2004). In the case of some research data services, where many users expect institutional support and where there is a relatively large number of student users and users who are private/unaffiliated individuals, retired persons and volunteers, willingness to pay will be highly constrained and willingness to accept can be the better indicator of the value users ascribe to the service.

Figure 4.1: Methods for exploring the economic value and impacts of BADC



Source: Authors' analysis.

Where there is a bundle of different goods and services these can be treated in the aggregate or dis-aggregated and re-aggregated in a way that reflects the bundling and/or use (e.g. multiplying the average willingness to pay expressed by users of specific types of research data by uses of those data types), thus weighting individual survey respondents' expressed preferences by the structure of the

bundle and its use. This can be particularly important where most users of a data service use just some part of the service and not all, and so the value that they express (i.e. would be willing to accept or pay) relates to just some parts of the service.

4.2.3 Consumer welfare and net economic value

The benefit or welfare impact of a good or service for a consumer is measured by the consumer surplus. In a market situation, willingness to pay is made up of what is actually paid and any excess willingness to pay over and above the price paid (i.e. consumer surplus). Hence, consumer surplus is the net gain derived by the consumer from the purchase of a marketed good or service. In a non-market context, all the willingness to pay is consumer surplus because there is no market price. In practice, however, some expenditure, be it in the form of time or money, will be incurred in obtaining the non-market good or service (e.g. time spent accessing the data service). In this case, the consumer surplus will be the net gain (i.e. willingness to pay minus the cost of obtaining). Net economic value is the consumer surplus minus the cost of supply.

Box 4.2: What value is and is not being captured?

Think of the example of pharmaceuticals. Imagine that a pharmaceutical company does research into and develops a new drug. They then sell the drug around the world for 10 - 20 years. If one did a return on investment calculation, one would look at the expenditure on R&D and the revenue from sales.

The wider value and benefit of the new drug is in the lives saved by the better drug, or the efficiency gain in hospitals through using a better drug, with shorter hospital stays, etc. A return on investment calculation does not directly measure these things, but it not true to say that they are not captured, to some extent, because the revenue from sales is an expression of the value of the drug. Doctors prescribe the new drug because it saves lives, governments, patients and doctors pay what they do for the new drug because it has the effects it does (e.g. saving lives, raising hospital efficiency, etc).

So the methods for economic valuation that we are using in this study can, to a limited extent and by proxy, capture the wider value and impacts even though we are not directly measuring them.

4.2.4 Efficiency impacts and returns on investment

Wider benefits and impacts can be explored: (i) by looking at the efficiency gains enjoyed by users and assigning an economic value to them; and (ii) by estimating the impacts of increased data use facilitated by the data service on returns to investment in the data collection/creation and the related data infrastructure services necessary for hosting and sharing the data. As these latter impacts are lagged and recurring during the useful life of the data, it is necessary to estimate the overall value of the impacts over time.

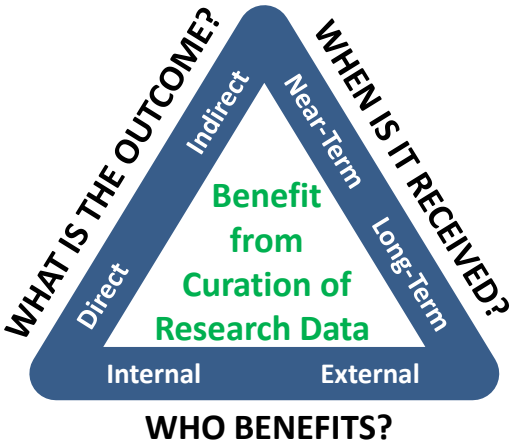
4.2.5 Keeping Research Data Safe (KRDS) Benefits Framework

Measuring benefits is often quite challenging, especially when these benefits do not easily lend themselves to expression in quantitative terms. Often a mixture of approaches will be required to analyse both qualitative and quantitative benefits and value, and present complete picture.

The conceptual framework used for the assessment of wider qualitative benefits and value in this study is the Keeping Research Data Safe (KRDS) Benefits Framework (Beagrie et al 2010). The KRDS Benefits Framework is a tool for identifying, assessing, and communicating the benefits from investing resources in the curation/long-term preservation of research data.

To assist institutions in applying the Framework, the KRDS project created a KRDS Benefits Analysis Toolkit (KRDS, 2011). This aims to help institutions identify the full scope of benefits from management and preservation of research data and to present them in a succinct way to a range of different stakeholders (e.g. when developing business cases or advocacy).

Figure 4.2: The KRDS Benefits Framework



Source: Beagrie et al 2010, and KRDS 2011.

The KRDS Benefits Framework uses three dimensions to illuminate the benefits investments potentially generate. These dimensions serve as a high-level framework within which thinking about benefits can be organised and then sharpened into more focused value propositions. The Framework can be customised and extended as needed to visualise and present benefits in different ways. Further information and guidance on its application is available in the Toolkit (KRDS 2011). For this study we have used the simplest form of presentation based on these three dimensions of a KRDS Benefit Framework. This has formed a model for application in section 6.3.

5 QUANTITATIVE ANALYSIS OF THE VALUE AND IMPACT OF BADC

The focus of the economic analysis is to try to shed light on the value of BADC data and services to its users (i.e. both the users of BADC data and services and depositors of data with BADC) and to the wider research and data community. Self-evidently, our survey respondents are a self-selected sub-group of users and depositors (e.g. because they responded to the survey). The response rates to both user and depositor surveys were good, and the brief analysis of respondents by the major categories suggests that they are reasonably representative of the BADC user community (see Section 3.4, above). However, on average, one might expect that those taking the time to respond to a survey are likely to use and value BADC data and services more than those who did not respond, and our user survey respondents' reported frequency of use suggests that they are among the more frequent users when compared with BADC usage statistics.

Box 5.1: BADC User and Access Statistics

There are two key data elements used in the following economic analysis to scale the activities reported by survey respondents to overall use of the BADC. Namely, counts of the number of users accessing BADC data and of the number of accesses made during 2012 (i.e. in the 12 months immediately prior to the surveys). The most important for analysis is the latter (i.e. the number of accesses).

Users: While access to much of the data at BADC is unrestricted and open to non-registered users, some resources require user registration and specific application to the restricted resource. There are also some users who voluntarily register as BADC users without accessing restricted resources. Users are defined as 'active' if they accessed BADC data during the previous 12 months. During 2012, there were 3,497 active registered users using HTTP or FTP to access BADC. BADC download logs also contain information about 'public' access (i.e. occasions where the user has not logged in with a BADC account id), including the session's IP address. As Parton (2013) reports, if internet usage is assumed to be similar between the registered and non-registered BADC user communities (i.e. taking into account factors such as a user accessing from multiple IP addresses over time or within a defined IP subnet – e.g. via university networks), then comparing ratios of the number of IP addresses per registered user to the number of IP addresses for non-registered users presents a way to obtain an estimate for the number of active non-registered users. Taking this approach Parton (2013) estimated that 2,462 active, un-registered users accessed BADC by HTTP or FTP during 2012. Thus, the total number of active BADC users during 2012 was estimated to be 5,959.

Accesses: While there is no direct indicator of accesses that matches the survey respondents' reported access events, removing activity attributed to scripts and web crawlers, the BADC session and download logs suggest that there were some 4.8 million accesses during 2012 over 22,608 use days. For the purpose of estimation we take these use-days to be the closest that available log data come to measuring uses in the sense reported by survey respondents.

It is also important to note that few users use all of BADC, but rather experience just part of it, and they can only express costs and value relating to the part(s) they use. In fact, just 29 (3%) of user survey respondents reported using all of the data types delivered by BADC during the last three years.[6] Consequently, it is important to consider weighting the survey responses to better reflect the totality of deposits with, and uses of, BADC data and services.

Unfortunately, we have been unable to source reliable data on the total number of data deposits with BADC by data type that match the deposit events reported by depositor survey respondents. Consequently, we treat the depositor survey respondents as a representative population. Moreover, due to some limitations in the survey question about frequency of access in the face of widely varying access frequencies, and to uncertainty in regard to the equivalence of users' reported "access events" and user "use days" recorded in BADC log files, we have not been able to source sufficient data to unpack and (re)weight respondent users use data. As a result, the following estimates are no more than indicative.

5.1 The Value and Impacts of BADC Data and Services

This section explores the value of BADC data and services circa 2012 based on reported current levels of activity and use. In view of the limitations in available deposit and access data, the following should be treated as no more than indicative estimates.

5.1.1 Investment and Use Value

The most direct indicators of value are investment value (i.e. the amount of resources spent on the production of the good or service) and use value (i.e. the amount of resources spent by users in obtaining the good or service). Measures of the investment in access suggest the minimum amount that the good or service is worth to the consumers.

Investment value includes annual BADC operational funding, the costs that depositors face in preparing data for deposit and in making those data deposits. For simplicity, each can be expressed as an annual cost in current prices and at current levels of activity by focusing on a single year snapshot for 2012. Activity times have been converted to costs using average annual salaries for academic staff and equivalents for non-academic staff, and average annual school leaver and graduate salaries for undergraduate and postgraduate students, respectively, based on Times Higher Education Surveys and the Green Book method of costing (see Box 5.2 for details).

As we are unable to source reliable data on the total number of data deposits by data type, we treat the depositor survey respondents as a representative population. BADC staff agree that this is a reasonable assumption. However, there were just 42 respondents to the depositor survey, of which, when asked about their last deposit as a critical incident, 18 reported costs relating to one-off deposits, 2 reported costs relating to initial deposits in a series, and 14 reported costs relating to subsequent deposits in a series. As a result, the following depositor costs should be interpreted as no more than indicative.

---

[6] Note, however, that although a user may use data from all the data-types archived by the BADC, this does not necessarily equate to them having made use of all datasets therein.

Box 5.2: Costing Activity Times

Activity times were converted to costs by assigning each respondent to a salary group based on the Times Higher Education Salary Survey for 2012, then scaling to include non-wage labour costs using a 30% uplift, based on the Green Book method (Green Book 2011). For students, we used the school leaver and graduate average salaries reported in the Telegraph and Guardian Higher Education pages, for undergraduates (£19,000 pa) and postgraduates (£22,800 pa), to reflected the opportunity cost of earnings forgone. Across the respondents, this resulted in a costing of around £37 per hour for staff and £15 to £18 per hour for students - £36 per hour for those saying that their duties included teaching, £34 for those researching (including postgraduate students), and £24 for those studying. These categories are not exclusive.

It should be noted that slightly more than half of the user respondents were based outside the UK, and while many of those were in comparable developed countries, some were not. For developing country respondents, we are likely overestimating their actual costs. Hence, the costings presented should be thought of as "UK equivalent costs." As depositors were predominantly UK-based, depositor costs are not affected in this way.

Source: Authors' analysis.

With the above caveat in mind, based on the times and costs reported by respondents to the depositor survey (Depositor questions 4 to 11, Appendix 1), depositor preparation and deposit costs amount to an estimated £771,000 per annum circa 2012. The operating budget for BADC is around £2 million per annum. Hence, treating data creation/collection costs as sunk costs (i.e. assuming that the data would have been collected whether or not BADC existed), *investment value* amounts to around £2.8 million per annum.

Use value includes BADC user access costs. Again costing is done on the basis of average annual salaries (Box 5.2). Multiplying the mean of user critical incident access costs reported by user survey respondents (questions 10 to 13, Appendix 1) by the estimated number of accesses from BADC access statistics (i.e. approximately 22,600 use-days during the last 12 months according to BADC use logs) suggests user access costs or *use value* of around £2.3 million per annum.[7] Again, however, due to limitations in the survey question about frequency of access and uncertainty in regard to the equivalence of users' reported "access events" and user "use days" recorded in BADC log files these estimates are no more than indicative.

[7] To this one might add the cost of the time spent using the data accessed via BADC. However, we have not included this as the time spent using the data is the use value of the data, rather than that of the data service per se.

Table 5.1: Investment and Use Value

Investment Value	=	BADC Operational Budget + Data Preparation & Deposit Costs *Respondents' Annual Deposit Costs*	2,049,600 + 771,009	=	£ 2,820,609 *per annum*
Use Value	=	BADC User Access Costs *Mean Cost of Last Access* *x* *Total Number of Use-days*	100 x 22,608	=	£ 2,257,170 *per annum*

Source: Authors' analysis.

5.1.2 Contingent Value

The contingent value of a non-market good or service is the amount users are "willing to pay" for it and/or "willing to accept" in return for it. For a public good the value is the sum of "willingnesses", as consumption is non-rivalrous (e.g. the same information can be consumed many times). The key difference between willingness to pay and willingness to accept is that the former is constrained by the person's ability to pay (e.g. by disposable income), whereas the latter is not.

Willingness to pay (WTP)

User survey respondents were asked to express their *willingness to pay* in terms of: (i) an annual subscription fee; and (ii) on a pay-per-access basis (question 35, Appendix 1). Removing the 30 "protest answers" that are typical of this technique, we had 532 WTA (willingness to accept) responses and 575 WTP (willingness to pay) responses. The mean of the individual willingnesses to pay was £1,120 per annum. The mean on a pay-per-access basis was £418. Of course, what a person is willing to pay for a data services depends on the nature and extent of their use of that data service. Hence, it is necessary to 'weight' responses to reflect overall use/access patterns, as indicated by the proxy of logged use days.

Thus, dividing the mean annual willingness to pay by the mean frequency of use, and then multiplying this and the mean pay-per-view willingness to pay by the estimated number of total accesses suggests a willingness to pay for access to BADC data and services of around £5.2 million per annum. Again, due to limitations in the survey and uncertainty in regard to the equivalence of reported "access events" and user "use days" recorded in BADC log files these estimates are no more than indicative.

When the individual pay-per-access answers are multiplied by the individual frequency of access answers it amounts to a mean of £1,430 per annum (i.e. 1.3 times the mean annual willingness to pay reported by respondents). This is not surprising, as one would expect there to be a premium for the pay-per-access model because there is no longer-term commitment to pay involved.

Willingness to accept (WTA)

Willingness to accept is not constrained by capacity to pay, which may be an important limitation to willingness to pay among research data users. Moreover, some respondents expressed a willingness to accept nothing in return for giving up their access to BADC because they believe that the data should free, rather than it being of no value to them.[8] Excluding these non-protest zero responses, the mean of the individual willingnesses to accept was £7,072 per annum (question 34, Appendix 1). Dividing the individual willingnesses to accept per annum by the individual frequencies of access per annum and multiplying by the total number of use days recorded by BADC suggests a willingness –to accept of around £15 million per annum.

Table 5.2: Contingent Value based on Stated Preferences

Willingness to Pay	=	(((Willingness to Pay Per Annum ÷ Frequency of Access Per Annum) x Total Number of Accesses) + (Willingness to Pay Per Access x Total Number of Accesses)) ÷ 2	(((1,120 ÷ 25) x 22,608) + (418 x 22,608)) ÷ 2	=	£ 5,241,466 *per annum*
Willingness to Accept	=	(Willingness to Accept Per Annum ÷ Frequency of Access Per Annum) x Total Number of Accesses	686 x 22,608	=	£ 15,513,782 *per annum*

Source: Authors' analysis.

5.1.3 Consumer Surplus and Net Economic Value

The welfare impact or *consumer surplus* for a non-market good or service with public good characteristics is indicated by the total willingness to pay minus the cost of obtaining or accessing the good or service. From the data above, this suggests a net consumer surplus of around £3 million per annum.

The *net economic value* of a good or service is indicated by the consumer surplus (i.e. the willingness to pay minus the users' cost of obtaining access) minus the cost of supplying the good or service (i.e.

[8] Some also expressed a zero willingness to pay, for the same reason.

BADC's annual operating budget). This implies that the net economic value of BADC data and services to its user community is around £935,000 per annum.

Table 5.3: Consumer Surplus and Net Economic Value

Consumer Surplus	=	Willingness to Pay - Cost of Obtaining *BADC User Access Costs*	5,241,466 - 2,257,170	=	£ 2,984,296 *per annum*
Net Economic Value	=	Consumer Surplus - Cost of Provision *BADC Operational Budget*	2,984,296 - 2,049,600	=	£ 934,696 *per annum*

Source: Authors' analysis.

However, as noted, willingness to pay is constrained by capacity to pay, which may be a significant factor among research data users. If willingness to accept, which is not constrained by capacity to pay, were taken as an alternative expression of value, the implied consumer surplus could be around £13 million per annum and the net economic value £11 million per annum (more than 5 times BADC's operational budget).

5.1.4 Efficiency Impacts

Respondents to the user survey were streamed into those studying, teaching and researching, and asked a series of questions about their activities (questions 19 to 33, Appendix 1), including:

- How many hours per week, on average over the last twelve months, they had spent studying/teaching/researching;
- What share of their studying/teaching/researching time they spent with data from BADC and data from all other sources; and
- What impact BADC data and services had on their studying/teaching/researching efficiency?

Following feedback during pilot testing of the questionnaire, the questions asking users to estimate the share of their studying/teaching/researching time spent with data from BADC and all other sources offered percentage scales, which were in 10 percentage point increments from 0% to >90%. While not ideal (e.g. it involves interpreting categorical data as continuous), these scales provided an approximate guide.

The questions asking users to estimate what impact BADC data and services has on their studying/teaching/researching efficiency also offered percentage scales, for the same reasons noted

above. Beginning from "negative change" and "no change", respondents were offered 5%, 10% and then 10 percentage point increments to 90%, and then >90%. Again this is not ideal, but did provide some guide as to impacts on efficiency. It is worth noting that the “negative impact” response was reported by one student, one teacher and seven researchers – just nine out of 756 responses (1.2%).

- **Student** respondents spent a mean 18% of their time with data from BADC. They reported spending a further mean of 45% of their study time using data from other sources. Combined with their study time, this suggests that student respondents spent an average of around 5 hours a week with BADC data over the last 12 months (median 3 hours), and a total mean of 21 hours a week with data from all sources (median 17 hours). Asked to what extent their use of BADC data and services had changed their study/learning efficiency, ignoring the single negative response the reported mean was a 34% efficiency gain, which translates to an efficiency gain equivalent to around 12 hours a week at current activity times.

- **Research** respondents spent around 18% of their time with data from BADC, and a further mean of 50% of their research time using data from other sources. Combined with their research time, this suggests that research respondents spent an average of around 6 hours a week with BADC data over the last 12 months (median 3 hours), and a total average of 21 hours a week with data from all sources (median 18 hours). Asked to what extent their use of BADC data and services had changed their research efficiency, ignoring negative responses, the reported mean was a 28% efficiency gain, which translates to an efficiency gain equivalent to around 8 hours a week at current activity times.

- **Teaching** respondents spent an average of 7% of their time with data from BADC, and a further mean of 29% of their time using data from other sources. Combined with their teaching and preparation time, this suggests that they spent an average of around 1 hour a week with BADC data over the last 12 months, and a total average of 4 hours a week with data from all sources (median 1 hour). Asked to what extent their use of BADC data and services had changed their efficiency, ignoring the single negative response, the reported mean was a 15% efficiency gain, which translates to an efficiency gain equivalent to 2 hours a week at current activity times.

Each of the respondents was allocated to a salary scale according to their role and affiliation and an activity cost was calculated using a 30% uplift based on the Green Book method (see Box 5.2). This enabled us to estimate the approximate value of the reported efficiency impacts. In the absence of detailed BADC user data records (see Section 3.4), we took the activity share of responses as indicative of the user population. On that basis, we found that the efficiency impact of BADC data and services could be worth some £4.3 million per annum among regular teaching users, £3.2 million per annum among student users, and as much as £51 million per annum among research users. Hence the total estimated efficiency impacts of BADC data and services among its user community might be as much as £58 million per annum at current activity times, or £55 million per annum if undergraduate and school students are excluded – more than 11 times the sum of operational, depositor and user access time costs. This is likely to be an upper bound estimate for the reasons discussed below.

However, these impacts seem high and it seems that some respondents may have interpreted the question as relating to the efficiency impact on their studying/teaching/researching time spent with BADC data and/or data from all sources, rather than their total studying/teaching/researching time – which had been intended.[9] If this were so, then the implied efficiency time saving impacts would be lower, and the overall value of the efficiency impacts of BADC data and services among its regular users could be around £10 million per annum at current BADC use activity times – double the sum of operational, depositor and user access time costs. This is still a very substantial impact.

Table 5.4: Efficiency Impacts

Efficiency Gains from BADC Use (*All Activity Time*)	=	Estimated Number of Users *(During the last year)* x Cost Per Hour x Activity Time x Per Cent Efficiency Impact	5,959 x (24 to 36) x (2 to 12) x (15% to 34%)	=	£ 58,455,837 *per annum*
Efficiency Gains from BADC Use (*BADC Data Time*)	=	Estimated Number of Users *(During the last year)* x Cost Per Hour x Activity Time x Per Cent Efficiency Impact	5,959 x (24 to 36) x (0.1 to 2) x (15% to 34%)	=	£ 10,311,956 *per annum*

Source: Authors' analysis.

It is important to note that these estimates are likely to be upper bound values. First, because it might be expected that those responding to the survey would be more intensive data users than non-responding users and non-users – although responses about their impression of data use by others in their field does not reflect this (questions 22, 27 and 32, Appendix 1). Second, because the few negative responses are excluded as non-quantifiable. Third, because it expresses the impact in time and money equivalents, rather than quality impacts, such as completeness, appropriateness, etc., thereby focusing on one dimension of efficiency.[10] In the end, the efficiency impacts are likely to be more about how much gets done, and how well, in a given time, rather than the amount of time spent. We are using hours and pounds as proxy measures of the value of the efficiency impacts.

[9] There may also have been a tendency to pick a central answer from the efficiency percentage range offered.

[10] Conversely, estimates are based on current activity times, rather than on the implied activity times before the efficiency impact. Current activity times are used because they better reflect the 'doing more and better in a given time' impacts.

5.1.5 Return on Investment in Atmospheric Data and Services

It is also possible to explore some scenarios relating to the potential impacts of BADC data and services on returns to investment in the data. There are a number of data elements required for such an analysis, including: annual investment in the research data and services; average returns to that investment; and the level of, and increase in, access and use resulting from BADC activities. The user and depositor surveys and BADC operational data and reports provide a foundation for estimates:

- Treating depositor survey respondents as representing the population of depositors, suggests annual data creation costs of £11 million (questions 12 to 15, Appendix 1), depositor preparation and deposit costs of around £770,000 per annum (questions 4 to 11, Appendix 1), and BADC operation costs of some £2 million per annum, suggesting total data and services investment costs of around £14 million per annum.

- There is an extensive literature on returns to R&D, which, while varied, suggests that returns are high - typically in the region of 20% to 60% per annum (Bernstein and Nadiri 1991; Griliches 1995; Industry Commission 1995; Salter and Martin 2001; Scott et al. 2002; Dowrick 2003; Shanks and Zheng 2006; Martin and Tang 2007; Sveikauskas 2007; Hall et al. 2009). Much of this literature relates to the natural, biological and medical sciences and one might expect average returns in such fields to be relatively high. BADC data holdings are quite varied as are the fields in which the data users operate (Parton 2013). Work by the US National Weather Service (NWS) and National Oceanic and Atmospheric Administration (NOAA) on the value of atmospheric data suggest potentially high returns (NWS & NOAA 2011) and it is easy to see areas in which returns to atmospheric data might be high (e.g. in informing decisions that may lead to a reduction in the impacts of climate change). Nevertheless, to be conservative, we explore the range of returns characteristically identified in the literature (i.e. 20% to 60%).

- The other issue is what impact BADC data and services have on access and use of the data hosted and delivered. Some 59% of respondents to the user survey indicated that they could not have obtained the data in any other way if BADC had not existed (question 15, Appendix 1).[11] All of these represent additional use. However, responses to question 16 (Appendix 1) suggested that, of these, 89% thought the data was beyond their scope to create or collect themselves. This suggests that at least 59% of BADC use may be additional use, of which 89% could not have created/collected the data themselves. Hence, 54% of BADC use is effectively additional use, and the remaining 5% saved data re-collection costs (e.g. would otherwise have re-collected all the data they thought they were able to).

We proceed on the basis of these data, estimating the increase in annual return on investment in the data collections due to additional use. As these returns are recurring during the useful life of the data we use a simple Perpetual Inventory Method to estimate the overall value of the impacts. Drawing on preliminary work on the UK R&D Satellite Account (Evans et al. 2008) we depreciate publicly-funded research data at 5% per annum, and following the lead of the US R&D Satellite

[11] Comparable, though somewhat higher than the response to a similar question relating to CEDA data centres (Parton 2011).

Account (Sveikauskas 2007) we set the average useful life of the data/knowledge created each year at the accepted average of 30 years – although, of course, the useful life of data can be much longer and/or much shorter. For preliminary estimation we distribute the returns normally over year 1 through year 9. Applying a 3.5% discount rate to estimate net present value (Green Book 2011), we then model the recurring returns outlined above as follows.

Additional use by those who could not (re)collect the data

At the lower-bound average 20% return on data investment, if 54% of BADC use is additional use (i.e. the share of respondent users who could not have got the data elsewhere or recreated/recollected it themselves) the implied increase in returns on one-year's data and services investment would be £11 million over 30 years in net present value (NPV), and at the upper bound 60% average return the implied increase in returns would be £34 million (NPV). Given non-sunk data services costs of around £2.8 million per annum, this suggests a 4-fold to 12-fold return on investment arising from the additional use of the data facilitated by BADC services.

Table 5.5: Return on Investment to Additional Use Facilitated by BADC

Return on Investment (*Additional Use*)	=	((Mean of Total Costs *Creation + Deposit + Operation* ÷ (Number of Additional Uses x Average Return))	13,912,371 ÷ (22,608 x 54%) x (20% to 60%)	=	£ 11m to £ 34m *(from one year's investment over 30 years NPV)*
		÷	÷	=	=
		Non-Sunk Data Services Costs *BADC Operational + Deposit Costs*	2,820,609		4-fold to 12-fold RoI

Source: Authors' analysis.

Of course, if the other element of additional use, namely that by users who could not have got the data elsewhere but could have (re)created it themselves, is taken into account, then the implied cost saving of not (re)creating it should be added to the implied additional returns to investment. That would suggest the addition of up to £650,000 per annum in saved (re)creation costs (Box 5.3).

Hence, BADC data and services facilitate additional use which realises additional returns to the research and data creation/collection activities underpinning the data it hosts, that, to an approximation, could be worth between £11 million and £34 million (NPV) over 30 years from a single year's investment. As such, BADC exhibits a 4 to 12 fold return on investment.

Box 5.3: Proxy indicators of value to users: re-creation costs

Indicators of value to users include the implied cost savings resulting from reuse (i.e. the costs of data collection/creation that are saved multiplied by the number of additional uses). Of course, this is an upper bound estimate of what the users may be saving as they may alternatively forego use. Nevertheless, just to explore a possible reuse related cost savings: 11% of user respondents said they could not have obtained the data they downloaded in any other way, but could have re-collected or re-created the data.

If this were characteristic, and 11% of all users were able to re-collect the data and had done so instead of accessing BADC during the last year, at the average of depositor reported data creation costs, less BADC operation, user access and deposit costs, it would have cost them around £650,000.

Of course, some of the accesses are repeat accesses to the same data, it is more likely that they could have re-collected the smaller- and cheaper-to-create data collections than the larger and more expensive-to-create data collections, and some would have foregone the data if faced with re-creation costs, so this is likely to be very much an upper bound cost saving.

Source: Authors' analysis.

While different studies focus on different information services and content and use different methods and measures, BADC's return on investment is comparable to, and somewhat higher than, reported returns to library and information services. For example:

- British Library (2004) concluded that: "The British Library generates value around 4.4 times the level of its public funding."
- King (2010) summarized findings relating to library services and concluded that: special libraries exhibit a RoI ratio of 2.9 to 1, academic libraries 3.4 to 1 (for staff), and public libraries 5.8 to 1.
- Imholz et al. (2007) summarized a number of studies, finding, inter alia, that Ohio public libraries showed a RoI ratio of 3.8 to 1 and the Carnegie Library in Pittsburgh 3 to 1.
- Houghton (2011) estimated the benefit/cost ratio of the Australian Bureau of Statistics making data and publications freely available online and using Creative Commons licensing at 5.3 to 1.

Although of course, each study focuses on a different mix of information/data and related services, so none are strictly comparable.

5.1.6 Summary of the quantitative economic analysis

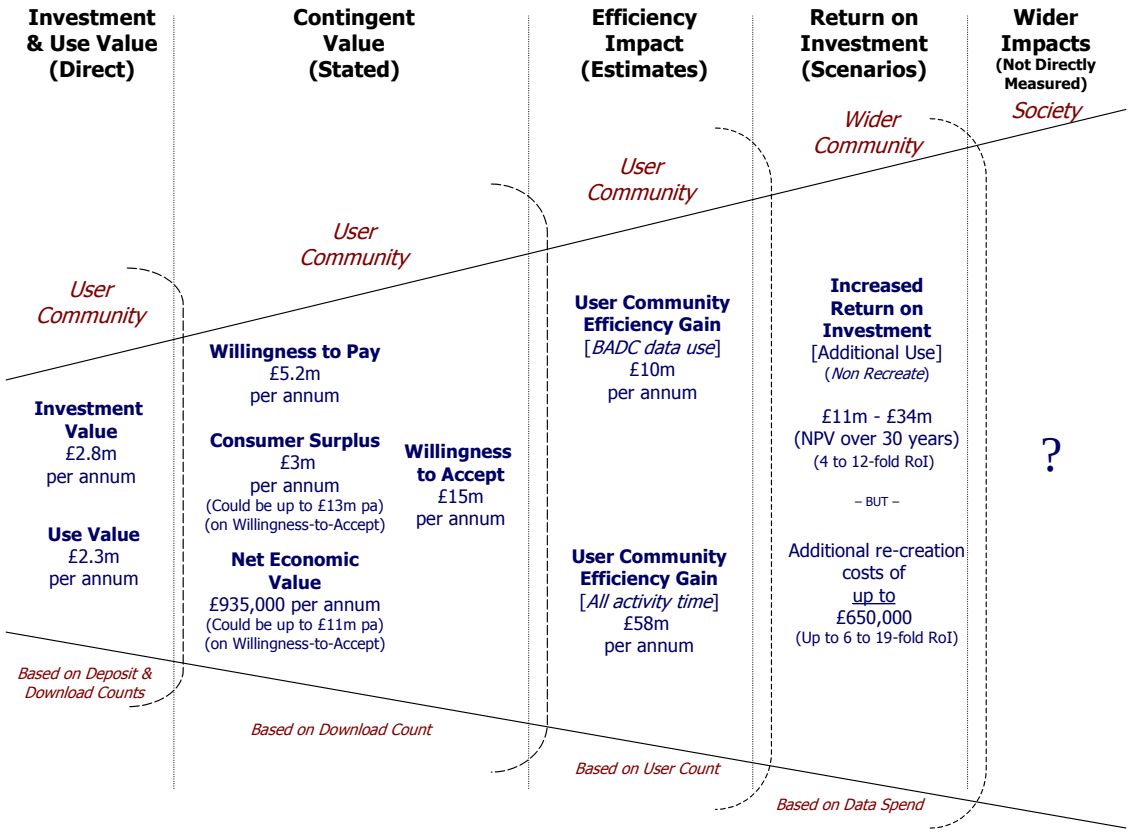
While there are many limitations in the data, our survey results and BADC operational data provide some basis for estimation. Figure 5.1 summaries these results. It shows a direct investment and use value to the BADC user community of £2.3 million to £2.8 million per annum at current prices and levels of activity. Willingness to pay is an expression of value by the users, who reveal that they value their access at around £5.2 million per annum, despite resource constraints and limitations on

capacity to pay. If one adds the cost of access to what users are willing to pay it indicates what they do and are willing to pay, which is around £7.5 million per annum.

When capacity to pay is limited the amount that users would be willing to accept in return for giving up their access to BADC for a year can be a better indicator of the value they place on it, as it is not constrained by their capacity to pay. Looked at this way, BADC data and services are worth around £15 million per annum.

The contribution of BADC data and services to its user community can be seen in terms of its impact on their research, teaching and studying efficiency. We found that the total estimated efficiency impacts of BADC data and services among its user community might be as much as £58 million per annum at current activity times – more than 11 times the sum of operational, depositor and user access time costs. However, these impacts appear high and it seems that some respondents may have interpreted the question as relating to the efficiency impact on their studying / teaching/ researching time spent with BADC data and/or data from all sources alone, rather than their total studying/teaching/researching time – which had been intended. If this were so, then the implied efficiency time saving impacts might be around £10 million per annum (Figure 5.1). This is still a very substantial impact – double the sum of operational, depositor and user access time costs.

Figure 5.1: The value and impacts of the BADC data service



Source: Authors' analysis.

Exploring some scenarios relating to the potential impacts of BADC data and services on returns to investment in the data, we found that BADC data and services facilitate additional use that may realise additional returns to the research and data creation/collection activities underpinning the data it hosts, that, to an approximation, could be worth between £11 million and £34 million over 30 years in Net Present Value from one year's investment – a 4-fold to 12-fold return on investment.

6 QUALITATIVE APPROACHES AND ANALYSIS

In addition to the quantitative economic analysis, we undertook qualitative analysis based on user and depositor surveys and interviews. This provides the opportunity to check that the quantitative and qualitative findings are in accord and sheds light on the dimensions and nature of the value of BADC data and services.

6.1 Interviews

Thirteen users and depositors agreed to be interviewed by telephone. Of these, all were registered users and six were also depositors or generators of data hosted by BADC. Semi-structured interviews were conducted using a pre-defined questionnaire. Table 6.1 presents a synthesis of key messages that emerged from responses by interviewees. Interviews were in confidence and key messages have been aggregated and anonymised for presentation below. Where issues are mentioned by more than one user or depositor, the number of interviewees is given in the text - e.g. (2/13 for users) or (3/6 for depositors), respectively. Note the interviews are a very small sample of 13 and should be seen as supplementing the information gathered via the much larger range of respondents to the surveys discussed in the next section.

Table 6.1: Key messages from the interviews

	Key Messages	**Details**
1	Data Volumes	Often measured in multiple terabytes, BADC datasets can be large. For users this may be the significant driver for a different pattern of use, whereby data is downloaded once, stored locally and then reused multiple times (2/13). One user felt he was pushing the limits of data transfer and his institution is putting in fibre optic cable to help him get the data from BADC.
2	BADC Data Storage for Data Sharing	The availability of storage space for data sharing is also important for some BADC users (3/13) due to the issue of large datasets.
3	Pattern of Use	The pattern of use by individual users may be intermittent; a user may not need to download fresh data for months at a time, but BADC is no less important for that. Interviewees (6/13) were at pains to make the point that when people need it, BADC is critical, but then they might not need it again for a long period.
3	Importance of Automated Data	For some depositors, data volumes raise other significant issues. Automation of the deposit process is necessary. Some interviewees (3/6) had written the upload/generation procedures

	Deposit	themselves others used tools provided by BADC. Even so, the process can be time-consuming and speed of ingest is an issue. Comments included “Transfers are going on 24 hours a day" and "currently one deposit is taking two months to deposit (in chunks)". One major depositor would like to see improvements such as more automation and verification of data transfer and expressed concern that without this BADC may not have the capacity to handle increasing volumes of both data and metadata in the future.
4	BADC is a Unique Source of Data in the Required Format	For most interviewees (7/13) BADC was the only source of suitable data. All seven said that if they had not had access to BADC they would have had to delay or change their research in a fundamental way, or abandon it altogether. Some interviewees (5/13) are able to get data from other sources, having contacts or special arrangements with the originators.
5	Supply from Alternative Sources (when available) can Involve Significant Overhead	The time and effort required to get data from alternative sources was also mentioned; estimates and experience indicated this could add anything from a week or several weeks, to over a year to the timescale, and entail hours of negotiation and emailing. One user calculated that it would cost him about £50,000 per year in equipment, staff, travel and subsistence if he could not get the data via BADC.
6	Quality of BADC Support and Documentation	Some interviewees (4/13) specifically commented that BADC has better user support and background documentation than alternative data centres; "BADC and their people have been very helpful though in support and supplying data in usable format, so they are my preferred source". One had used a dataset from a European data centre which was comparable to BADC but concluded that BADC data quality, operational service and Help Desk were much better. BADC staff in general were praised by depositors (2/6) as well as users.
7	Feedback on User Interface	For some, the user interface can be regarded as "clunky" and it can be difficult to find something unless you already know where it is. A major depositor estimated that more than 50% of the customers he has directed to BADC will come back to say they cannot find the data. A user commented that the interface looks like it replicates the data storage file structure rather than being organised for the prospective user.

8	Benefits for Depositors	Depositors (3/6) value the fact that BADC saves them having to set up and maintain an infrastructure for public access. Savings were cited in terms of storage capacity and also in personnel to answer customer queries. Estimated cost savings from having to provide a public interface run from £10-15,000 to £100,000's p.a., though for some depositors (2/6) it would not be feasible for them or their institutions to undertake this themselves at all. Depositors appreciate that while BADC cannot answer every customer query themselves they do act as an important filter for straightforward enquiries.
9	Benefits for Users	Benefits mentioned by the interviewees included network efficiencies and computer storage space, saving time and improving data quality for others. Having a single central data source is cost effective, as the alternative would require duplication of resources across all the institutions that use them. Time saved in setting up licencing contracts alone is substantial. It took one customer about 30 hours emailing to negotiate the extra permission required from a number of bodies to get data from the US for which licences were already in place.
10	Views on Data Standards	Where standards were mentioned (4/13), most felt that BADC has done a good job trying to set standards and formats and that this is a good thing overall, but a view was also expressed that atmospheric data still lags behind other sectors. One felt that success as regards metadata for resource discovery is perhaps not as strong as in formatting standards - users cannot always find the data. Some felt that having standards saves time overall, but can also cost time in terms of preparing standards-compliant data.
11	BADC impacts in other areas	BADC data is used in teaching (4/13); outreach to industry e.g. insurance, and to feed local models for river catchment and agriculture where it improves local models and modelling processes. It sustains national and international collaboration (2/13). The work from BADC data (mostly Met Office derived) is one of many sources feeding into government climate change policy. BADC also has an impact on capacity-building with regard to student development. Students apply for a data set and learn how to use it ethically, developing analytical tools to turn data into meaningful results. This helps to build skills and produces better scientists. Some depositors mentioned a multiplier effect on citations and international re-use of data; BADC generates more users and more feedback for their data, and may have

		helped widen usage of their data to non-traditional sectors.
12	Importance of access to Met Office Data via BADC	Finally, Met Office data particularly radiosonde /MIDAS data are the most widely used BADC data (9/13). Four interviewees stressed that they used it frequently (e.g. weekly) or regarded it as particularly important.

6.2 Surveys

6.2.1 Users of BADC data and services

Results from the user survey and a detailed commentary are presented in Appendix 1. This section provides a short overview of the qualitative questions and responses and a selection of user comments from those presented in full in the Appendix.

There are currently around 22,500 BADC registered users of whom BADC estimate 3,497 were active users during 2012. However approximately half of all datasets are open access – with no need to be registered to download and use. BADC estimates it has approximately an additional 2,462 active non-registered users, hence around 5,959 active users in total (Parton 2013).

Email invitations to participate were sent to all active registered BADC users, and additionally active non-registered users were invited to participate in the survey via the CEDA news feed, CEDA and BADC websites, Facebook page, and Twitter feed. Some 1,141 respondents completed the user survey in total (18% response rate from the total estimated active user population).

Questions in the user survey were primarily focussed on obtaining data for the economic analysis, but a number focussed on exploring broader qualitative benefits and impacts. The responses echo those from the interviews and provide supporting data from a larger and broader range of respondents. They also support the quantitative analysis for the significant economic benefits and value of BADC presented in section 5.

Question 7 of the user survey asked how important BADC data and services were to their work, on a 5-point scale from “not at all important” to “extremely important”:

- 81% reported that BADC was very or extremely important for their academic research;
- 42% that BADC was very or extremely important for learning or skills development;
- 35% that BADC was very or extremely important for government policy research.;
- Across responses, the highest importance ratings on a scale of 1 to 5 were for academic research (with an average score of 4.12), learning and skills development (2.78), and government policy research (2.19).

Question 8 asked what impact it would have on their work or study if they could not access data and services from BADC:

- 53% reported that it would have a major or severe impact on their work if they could not access BADC data and services:
- A further 34% said it would have a moderate impact;
- Less than 3% said it would have no impact.

Some 112 users also provided free-text comments to explain or amplify their responses to this question. In Appendix 1 comments have been categorised by common themes and selected comments included. Selected comments are copied for illustration below.

Selected Comments from free-text Q8

'BADC is a one-stop dataset portal which houses key datasets covering primarily the UK/European region. Without the BADC, it would be time-consuming for me to search for the datasets at individual hosts.'

'BADC is fundamental to the way that academic atmospheric science is undertaken.'

'CRU TS dataset distributed by BADC is used both for historical climate impact modelling as well as bias-correction of climate scenario data. Therefore, it is crucially important to our work.'

'Would stop or at very least hold back many millions of pounds worth of academic research. Would force direct acquisition of datasets from Met Office which in my experience has been fraught with difficulties and costs.'

'Without the assistance and services provided by BADC the … project … would have been many orders of magnitude more difficult to implement and could well have failed.'

'To have data that is of high quality based on the reputation of the scientists who have prepared it saves me time in my work and that is a big benefit.'

'Not only does BADC act as a repository for general data, it is vital for hosting and archiving data from collaborative projects, which form the majority of the work I do.'

Source: Appendix 1 User Survey Question 8.

To explore possible valuations through counterfactual means and establish the extent to which BADC use is additional use, respondents were asked if they could have obtained the data they used in another way had BADC not existed (questions 15 and 16).

Question 15 asked if BADC had not existed would they have been able to obtain the data they last used from another source:

- 59% said they could not have obtained the data elsewhere, while 41% could.

Question 16 followed up by asking the respondents if they could not have obtained the data elsewhere, would they have been able to collect/recreate the data themselves:

- 89% said they could not have collected/recreated the data themselves;
- Of those answering “no” to question15 (i.e. could not have obtained the data elsewhere), 91% could not have collected/recreated the data themselves. Hence, around 54% of responding users could neither obtain the data elsewhere nor collect/recreate it themselves, and their use can be considered additional use arising as a result of the availability of BADC data and services.

Among those who could have collected/recreated that data themselves, comments alluded to the additional time, effort and costs involved in doing so, and to a concern that the alternative data would not be as good.

In Appendix 1 comments to question 15 have been categorised by common themes and selected comments included. Selected comments are copied for illustration below.

Selected comments from free-text Q15

‘Possibly from printed reports in the Public Record Office, otherwise it would have required travel to UK Met Office in England costing around £500.’

‘The data are available elsewhere but I am familiar with BADC and how to use it and it conveniently has many datasets under one banner so this saves time and effort in going round many different sources.’

‘We have direct contacts with most of the international climate research centres. If BADC did not exist we would still be able to access data - in some cases this would be easier than accessing the data at BADC and in other cases harder. The Time and cost would increase without BADC access as there would be more data sources to deal with. Hard to quantify - but probably significant (more than double, less than 10x).’

‘I would have to write new download scripts. The BADC service is quite good for scripting -- other services are not so good. It would take a few days of work to switch sources.’

‘Perhaps yes, but the time, effort and cost involvement would have surely jeopardise my project for all practical purposes.’

‘I am unsure whether I would have found the data, but it would have taken more time to find a source (if it exists). Having multiple data sets stored in one location is extremely convenient (& efficient)’

‘Could have gotten the model data from xxxxx, but that portal is absolutely horrid and no sign of becoming easier to use. Your [BADC’s] portal is much easier to use.’

‘It would take significantly longer to obtain this data, and considerably more if it involved a web interface to get the data rather than script-based ftp. Also, the BADC service is incredibly reliable and intermittent provision often experienced with other providers incurs a time cost to projects.’

Source: Appendix 1 User Survey Q15.

Question 17 asked if the respondents considered that they had benefited in a list of areas as a result of using data and/or services from BADC. Respondents suggested that they had benefited from using data from BADC in a variety of ways including:

- 71% saying they benefited from discovery of required data through single site;
- 59% from gaining access to multiple datasets/portals through one licence/account; and
- 52% from the long-term preservation of data offered by BADC.

Note that the responses to this question primarily reflect a researcher perspective of the benefits to them personally and their own research. It will not necessarily reflect relative benefits to other stakeholders, such as funders or institutions, from activities, such as reduced hosting or licensing costs.

In a similar related theme, question 18 asked to what extent had the respondents benefited, in a variety of ways suggested to them, from BADC. Responses included:

- Almost 50% reported a high or very high benefit from dataset documentation;
- 34% a high or very high benefit from online help guides;
- 32% a high or very high benefit from data discovery tools.

It was clear that users saw documentation, user support such as the online help guides, and data discovery tools as major benefits.

6.2.2 Depositors of data with BADC

Emails to participate to the Depositor survey were sent to 143 BADC users identified as past and present depositors. Around 20 of these users were no longer reachable through their contact details held by BADC. A link to the Depositor survey was also posted on the CEDA website and news feeds. Overall 42 responses were received by the close of the survey.

Results from the depositor survey and a detailed commentary are presented in Appendix 1. This section provides a short overview of the qualitative questions and responses and a selection of depositor comments from those presented in full in the Appendix.

Questions in the depositor survey were primarily focussed on obtaining data for the economic analysis, but a number focussed on exploring broader qualitative benefits and impacts. The responses echo those from the interviews and provide supporting data from a larger and broader range of respondents. They also support the quantitative analysis for the significant economic benefits and value of BADC presented in section 5.

The final two depositor survey questions explored the nature and extent of benefits derived by depositors, in qualitative terms.

Question 16 asked depositors to select the level of perceived benefit to them as a result of depositing/providing their data to the BADC, for a range of possible outcomes:

- Depositors cited that the data are preserved for the long-term and dissemination being targeted to the academic community as the most beneficial aspects of depositing data with BADC, both rated as a high or very high benefit by ~76% of respondents;
- These were followed by fulfilling grant obligations or organisational mandate (48%) and the fact that BADC handles user licensing and access (42%) as other important benefits.

The final question, number 17, explored the impact of being unable to deposit data with BADC on the respondents, their work groups and organisations:

- 52% suggested that being unable to deposit data with BADC would have a major or severe impact on them personally;
- 42% said it would have a major or severe impact on their work group;
- 27% said it would have a major or severe impact on their organisation.

6.3 Summary of Findings

We used the KRDS Benefits Framework (see section 4.2.5) as an underlying conceptual framework for summarising and illustrating the qualitative value and benefits arising more generally from the BADC. To do this we have drawn on the KRDS Benefits Analysis Toolkit (KRDS 2011 - Benefit Framework Tool), the interviews, and relevant parts of the user and depositor surveys, and their observations on the benefits of BADC. Relevant findings on benefits from these sources are described in sections 6.1 - 6.2 and Appendix 1. The outcomes from a KRDS benefits analysis can be presented in a simple table structured around the dimensions of the KRDS Framework. It provides a simple visual summary of the key qualitative benefits from BADC in an easily digestible form.

Table 6.2: Summary of qualitative benefits from BADC arranged in a KRDS Benefits Framework

KRDS Benefits Summary for BADC	
Direct Benefits	**Indirect Benefits (Costs/Risks Avoided)**
Access to data provides new research opportunities International sharing and re-use of data on topics of shared interest with the UK Re-purposing and re-use of data maximises the value of data holdings Increasing research and teaching productivity: time and resource savings for researchers and teachers Verification of research through increased data citation and access Assisting in provision of a skills base Customer service ethos of BADC staff Availability of storage space for collaborative project data	Protecting returns on earlier investments No re-creation of data and reducing potential duplication of effort Existing data available for new audiences and non-traditional sectors Avoidance of higher costs associated with getting the data from other sources
Near Term Benefits	**Long-Term Benefits**
Value to current researchers and students Single point of access: increasing speed and ease of use and access to data for researchers and students The reliability of the service and the quality of data, documentation, and support: often makes BADC the preferred supplier even when other sources are available Secure storage for data intensive research Availability of data underpinning journal articles	Data preserved for the long-term: secures value of high quality data for future researchers and students Value added over time as collection grows and develops critical mass Impact on wider profession as a centre of excellence, and in setting and mandating data format standards Promoting quantitative methods, skills and use across students and early career researchers Fostering innovation in research and data management practice
Private Benefits	**Public Benefits**
Service targeted at academic community and supporting their needs Source of high-quality and often unique historical data Motivating new research that otherwise could not be undertaken Increasing visibility/citation of research and underlying data products Time saved in setting up licensing contracts for data use Articulating user needs to data providers Fulfilling grant obligations for deposit and open data Fulfilling EU/UK Inspire requirements for geospatial data Aggregator of data for data providers – providing them with one point of access to UK customers Higher usage and profile of their data for providers Removing user enquiries burden from data providers Removes archive infrastructure burden from providers	Sustaining national and international research collaboration Feeding into government climate change policy Data use in industry, e.g. insurance and agriculture Research integrity since others can check the outcomes and interpretation of research Fostering transferable skills in analysis of large datasets Single central data source is cost-effective

Source: KRDS 2011 and authors' analysis

7 CONCLUSIONS AND RECOMMENDATIONS

The approaches used for this study applied a unique combination of quantitative and qualitative methods to provide the full picture of the nature and dimensions of value and explore the full range of impacts from data sharing and data curation at BADC.

The economic analysis shows that data sharing and data curation via the BADC has a very high value relative to its operational costs. A very significant increase in research efficiency was reported by users as a result of their using the BADC, which we estimate to be worth at least £10 million per annum. We also estimated the value of the increase in return on investment in data creation/collection resulting from the additional use facilitated by BADC to be between £11 million and £34 million over thirty years (net present value) from one-year's investment – effectively, a 4-fold to 12-fold return on investment.

The qualitative analysis in the user survey showed that 81% of the academic users who responded reported that BADC was very or extremely important for their academic research, and 53% of respondents reported that it would have a major or severe impact on their work if they could not access BADC data and services. In the depositor survey, depositors cited that having the data preserved for the long-term and its dissemination being targeted to the academic community as the most beneficial aspects of depositing data with BADC, both rated as a high or very high benefit by around 76% of respondents. Fifty-two per cent suggested that being unable to deposit data with BADC would have a major or severe impact on them personally.

The quantitative and qualitative analyses independently show a similar picture of the value of BADC: they are complementary, reinforce each other, and lend credence to the findings. The qualitative findings illustrate individual user and depositor experience and can personalise and provide further insights into the economic analysis.

The approaches used to assess impact across all three data service studies (ESDS, ADS, and BADC) show a similar pattern of findings, with data sharing via the data centres having a big measurable impact on research efficiency and on return on investment in the data and services. These findings are important for funders and for making the economic case for investment in data curation and sharing and research infrastructure to funders and other stakeholders.

The approaches used to explore value and impact are 'doable' and transferable between different disciplinary contexts and collections. However, relatively little can be transferred in terms of implementation, 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. 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 for each survey.

This study, and others in the series, has considered the impact of the data centre as a whole across all of its collections. This in itself can pose some challenges as the data centres have developed a great diversity of data collections that may be accessed and used very differently. This may mean that usage statistics have inherent variation and limitations. Similarly, often very little may be known

about the profile of public (or non-registered) users of collections in 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. We confronted data difficulties about user/depositor counts and deposit and download/use counts in this study, which has limited the scope and reliability of the economic estimates.

Contingent valuation is a method that typically generates some protest responses, and did so in this and the other studies. Despite assurances in the survey questionnaire that the service would remain free at the point of use, users were naturally fearful, leading to protest responses and comments. At the same time, the fear generated, comments made, and the protest zero answers and comments made because users believe the service should be free at point of use are still an important message for funders and services to note.

RECOMMENDATIONS

Recommendation 1: The unique combination of qualitative and quantitative approaches used in the BADC Impact Study 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. **There would be benefits from further research developing, refining and further exploring applications of the methods used in this study, as making the "business case" or funding case for data centres and services plays an increasingly important role in ensuring their sustainability.**

Recommendation 2: It is also clear in this and, to a lesser extent, the previous studies that different data services collect financial and operational data, such as user statistics, data deposit, access and download statistics, to varying levels of detail. More guidance is needed on the collection of such data. Doing so would help to ensure a greater degree of standardisation of statistical records across data centres as well as providing the basis for 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 case" or funding case.**

Recommendation 3: To date these approaches have only been applied to three UK national 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 at international, national and/or institutional levels as well as to the NERC data community.**

Recommendation 4: The study has looked at the aggregate value of BADC. There is also significant scope for more granular studies that focus on the value of specific collections or the economic value of BADC 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 by the authors for the Archaeology Data Service

(Beagrie and Houghton 2013) may also be beneficial. **These methods of valuation should also be applied at more granular levels than the overall collections or all stakeholders.**

Recommendation 5: Value and perceptions of value change over time. The BADC user community was previously surveyed in 2004, 2007 and 2010, and both BADC users and depositors were surveyed for this study in 2013. **BADC and funders should consider opportunities to repeat the BADC surveys and extend the available time series of comparative studies in future years. Ideally another survey of users and depositors should be considered within the next three to five years.**

Recommendation 6: 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 the uses to which the data are put after research use where 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 analysis. Nevertheless, there can be very substantial wider benefits. In the context of atmospheric data, 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 (see Box 4.1 in the main body of the report). **Consider research on the wider societal benefits and economic impacts that are generated by atmospheric research data sharing and curation, and the contribution to this made by BADC.**

REFERENCES

arXiv.org (2010) *arXiv Business Model White Paper*, Cornell University. Available http://arxiv.org/help/support/whitepaper

Ayris, P. et al. (2008) *The LIFE2 Final Report*, London & Bristol: The Joint Information System Committee (Jisc). Available http://www.life.ac.uk/2/documentation.shtml

BADC (2007), *2007 User Survey Analysis*. Available http://badc.nerc.ac.uk/community/news/BADC_survey_2007.pdf

Bakhshi, H., Freeman, A. and Hitchen, G. (2009) Measuring intrinsic value, MPRA Paper No 14902, University of Munich. Available http://mpra.ub.uni-muenchen.de/14902/

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

Beagrie, N. (2012) *The Value and Impact of the British Atmospheric Data Centre (BADC) – JISC Project Plan*. Available http://www.beagrie.com/BADC_project_plan.pdf

Beagrie, N., Chruszcz, J. and Lavoie, B. (2008) *Keeping Research Data Safe*, Jisc, London and Bristol. Available http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx.

Beagrie, N., Lavoie, B., &Woollard, M. (2010) *Keeping Research Data Safe 2 Final Report.* London, Jisc. Available http://www.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 http://www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf

Beise, M., and Stahl, H. (1998) *Public Research and Industrial Innovations in Germany*, ZEW Discussion Paper No. 98-37. Available http://ideas.repec.org/p/zbw/zewdip/5209.html

Bernstein, J.I. and Nadiri, M.I. (1991) 'Product demand, cost of production, spillovers and the social rate of return to R&D,' NBER Working paper 3526.

British Library.(2004) *Measuring the Economic Impact of the British Library,* Power point presentation on unpublished report. Available http://www.bl.uk/aboutus/stratpolprog/increasingvalue/publicvalue/confpres/pungelwesmarks.pdf

CEPA and Ware, M. (2010) *Transition dynamics of improving access to research*, RIN, London. Available http://www.cepa.co.uk/publications.php.

CEPA/RIN (2008) *Activities, costs and funding flows in the scholarly communications system*, RIN, London. Available http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/activities-costs-and-funding-flows-scholarly-commu.

CEPA/RIN (2011) *Heading for the open road: costs and benefits of transitions in scholarly communications*, RIN London. Available http://www.rin.ac.uk/our-work/communicating-and-disseminating-research/heading-open-road-costs-and-benefits-transitions-s.

Dekkers, M., Polman, F., teVelde, R., & de Vries, M. (2006) MEPSIR: Measuring European Public Sector Information Resources, Brussels: European Commission. Available http://www.epsiplus.net/psi_library/reports/mepsir_measuring_european_public_sector_resources_report

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

Dowrick, S. (2003) A Review of the Evidence on Science, R&D and Productivity. Canberra: Department of Education, Science and Training.

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

Evans, P., Hatcher, M. and Whittard, D. (2008) The preliminary satellite account for the UK: a sensitivity analysis, *Economic & Labour Market Review*, 2(9), September 2008, 37-43.

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

Green Book (2011) *Appraisal and Evaluation in Central Government*, HM Treasury, London. Available http://greenbook.treasury.gov.uk/index.htm. Available http://www.hm-treasury.gov.uk/d/green_book_complete.pdf

Griliches, Z. (1995) R&D and productivity: Econometric Results and Measurement Issues, In Stoneman, P. (Ed.) *Handbook of The Economics of Innovation and Technological Change*. Oxford: Blackwell, 52–89.

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

Houghton, J.W. (2009) *Costs and Benefits of Alternative Publishing Models*: *Denmark*, DEFF, Copenhagen (September 2009). Available http://www.knowledge-

exchange.info/Admin/Public/DWSDownload.aspx?File=%2fFiles%2fFiler%2fdownloads%2fDK_Costs_and_benefits_of_alternative_publishing_models.pdf.

Houghton, J.W. (2010) 'Costs and benefits of alternative scholarly publishing models: Lessons and developments,' in *Publishing in the networked world: Transforming the Nature of Communication*, Eds. Turid Hedlund and Yasar Tonta. ISBN 978-952-232-086-5, pp.385-403.Available http://hdl.handle.net/10227/599.

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

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

Houghton, J.W., Rasmussen, B., Sheehan, P.J., Oppenheim, C., Morris, A., Creaser, C., et al. (2009a) *Economic Implications of Alternative Scholarly Publishing Models: Exploring the Costs and Benefits*, Report to The Joint Information Systems Committee (Jisc) by Victoria University & Loughborough University. Available http://ie-repository.jisc.ac.uk/278/. Addendum http://www.cfses.com/EI-ASPM/Jisc%20EI-ASPM%20Report%20%28Addendum%20April%2009%29.pdf

Houghton, J.W., de Jonge, J. and van Oploo, M. (2009b) *Costs and Benefits of Research Communication: The Dutch Situation, S*URFfoudation, Utrecht (June 2009). Available http://www.surffoundation.nl/wiki/display/economicstudyOA/Home.

Houghton, J.W., Rasmussen, B. and Sheehan, P. (2010) *Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs*, The Scholarly Publishing and Academic Resources Coalition (SPARC), Washington DC. Available http://www.arl.org/sparc/publications/papers/vuFRPAA/index.shtml

Houghton, J.W., Steele, C. and Sheehan, P. (2006) *Research Communication Costs in Australia: Emerging Opportunities and Benefits*, Report to the Department of Education and Training, Canberra. Available http://hdl.handle.net/1885/44485.

Houghton, J.W., Swan, A., & Brown, S. (2011). *Access to Research and Technical Information in Denmark*, Forsknings- ogInnovationsstyrelsen and Denmark's Electronic Research Library (April 2011). Available http://www.fi.dk/publikationer/2011/adgang-til-forskningsresultater-og-teknisk-information-i-danmark

Imholz, S., and Arns, J. (2007) *Worth Their Weight: An Assessment of the Evolving Field of Library Valuation*, New York: Americans for Libraries Council. Available http://www.ala.org/ala/research/librarystats/worththeirweight.pdf

Industry Commission (1995) Research and Development, Report No 44, Industry Commission, Canberra.

King, D.W. (2010). Presentation on Library Value. Available http://libvalue.cci.utk.edu/sites/default/files/King2010June.pptx

KRDS (2011) *I2S2/KRDS Digital Preservation Benefits Analysis Toolkit*. Charles Beagrie Ltd, Salisbury. Available http://beagrie.com/krds-i2s2.php

Mansfield, E. (1991). Academic research and industrial innovation. *Research Policy* 20(1), 1991, 1-12.

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

Martin, B.R., & Tang, P. (2007) *The benefits from publicly funded research*. SWEPS Paper No. 161, Science Policy Research Unit, Brighton: University of Sussex. Available http://www.erawatch-network.eu/reports/sewp161.pdf

Missingham, R. (2005). *Libraries and economic value*, National Library of Australia Staff Paper, Canberra. Available https://www.nla.gov.au/openpublish/index.php/nlasp/article/viewArticle/1213/1498

National Institutes of Health (2008). *Analysis of Comments and Implementation of the NIH public access policy*, Washington DC.: NIH. Available http://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf.

National Science Board (2010) *Science and Engineering Indicators 2010*, National Science Foundation (NSB 10-01). Available http://www.nsf.gov/statistics/

Noonan, D.S. (2003) 'Contingent valuation and cultural resources: a meta-analytic review of the literature', *Journal of Cultural Economics*, Vol 27, pp159-176.

Oakleaf, M. (2010) *Value of Academic Libraries: A Comprehensive Research Review and Report*, Chicago: Association of College and Research Libraries (ACRL). Available http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/value/val_bibliography.pdf

Parton G.A (2011) *2010 CEDA User Survey Report – Benefits of the CEDA data centres*. Available http://cedadocs.badc.rl.ac.uk/897/

Parton, G. A. (2013) *BADC User Statistics Report 2013*, May 2013. Available http://cedadocs.badc.rl.ac.uk/947/

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

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

Ramirez, J.M. (2009) "Measuring: from Conjoint Analysis to Integrated Conjoint Experiments". *Journal of Quantitative Methods for Economics and Business Administration***9**: pp28-43. See http://en.wikipedia.org/wiki/Conjoint_analysis

Research Information Network (2011) *Data centres: their use, value and impact*. Research Information Network. Available http://rinarchive.jisc-collections.ac.uk/our-work/data-management-and-curation/benefits-research-data-centres

Royal Society (2012)*Science as an open enterprise*. Availablehttp://royalsociety.org/policy/projects/science-public-enterprise/report/

Salter, A.J. & Martin, B.R. (2001) The economic benefits of publicly funded basic research: a critical review, *Research Policy* 30(3), 509-532.

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

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

Svanhild, A. (2009) 'Libraries and return on investment (ROI): a meta-analysis', New Library World, 110(7), pp311-324.

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

TeVelde, R. (2009) 'Public Sector Information: Why Bother?,' in Paul Uhlir (ed.) The socioeconomic effects of public sector information on digital networks: Towards a better understanding of different access and reuse policies, National Academies Press, Washington DC. Available http://www.nap.edu/openbook.php?record_id=12687&page=40.

Tenopir, C., Love, A., Park, J., Wu, L., Baer, A., & Mays, R. (2010) *University Investment in the Library, Phase II: An International Study of the Library's Value to the Grants Process*, Elsevier Connect White Paper. Available http://libraryconnect.elsevier.com/whitepapers/roi2/lcwp021001.html

Weiss, P. (2002) *Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts,* National Oceanic and Atmospheric Administration. Available http://epsiplatform.eu/reports/borders-in-cyberspace

APPENDIX 1: SURVEY RESPONSES

Two online surveys, directed at: (i) users of BADC data; and (ii) depositors of data with BADC, were launched on 16th January 2013, and closed on 5th February 2013. There were 1,141 responses to the user survey and 42 responses to the depositor survey.

The survey questionnaires were developed iteratively by the project team with external review and input from BADC staff and others. Given the nature of some of the economic approaches being explored, and the range of affiliations, roles and seniority of the survey populations, substantial effort was needed to design questionnaires suitable for an online survey.

The questionnaires used a range of survey approaches, including question logic in their design to steer respondents through the appropriate sections of the questionnaire, and use of "critical instances", such as the last data accessed/downloaded (for users) or last data deposited/updated (for depositors). These questions were supplemented by qualitative questions asking for views on the importance and impact of BADC for both depositors and users. A small number of questions also sought specific information on the costs of creating and collecting the data, the time and cost involved in preparing it for deposit, the time and cost involved in accessing data for users, the benefits and efficiency impacts of access, and contingent valuation (i.e. willingness to pay or accept) using stated preference techniques. Answers to these questions must be interpreted carefully, in the context of open-ended text comments in the surveys and other findings from the interviews and desk research, to ensure that protest and outlier answers are either excluded from the economic analysis, or included with suitable caveats.

The following sections describe the survey results.

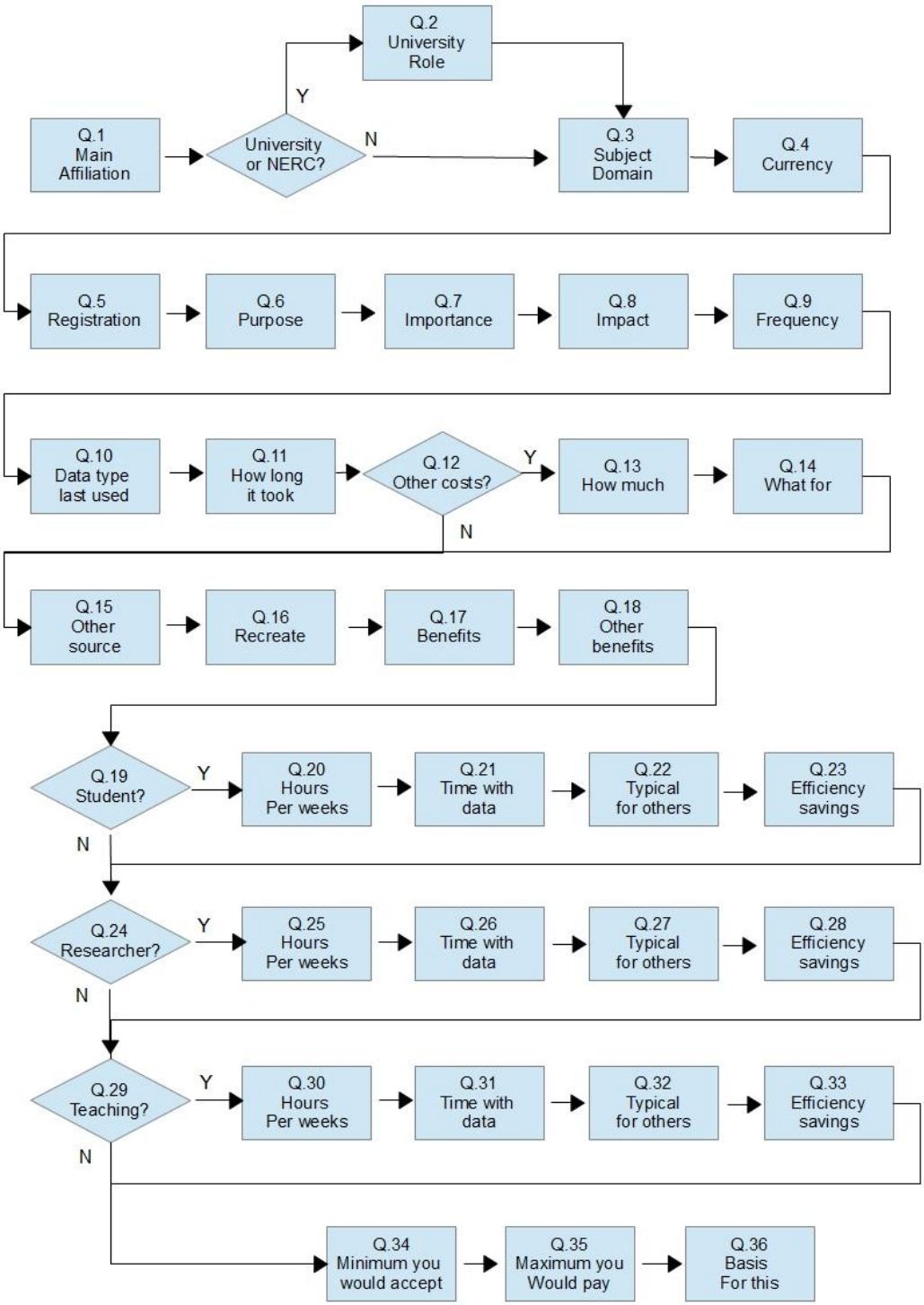
User survey

There are currently around 22,500 BADC registered users of whom BADC estimate 3,497 were active users during 2012. However approximately half of all datasets are open access – with no need to be registered to download and use. BADC estimates it has approximately an additional 2,462 active non-registered users, hence around 5,959 active users in total (Parton 2013).

Email invitations to participate were sent to all active registered BADC users, and additionally active non-registered users were invited to participate in the survey via the BADC website, Facebook page, and Twitter feed. Some 1,141 respondents completed the user survey in total (18% response rate from the total estimated active user population).

The User Survey consisted of 36 questions and followed the following logic in the sequence of questions:

Figure A1: BADC User Survey Question Logic



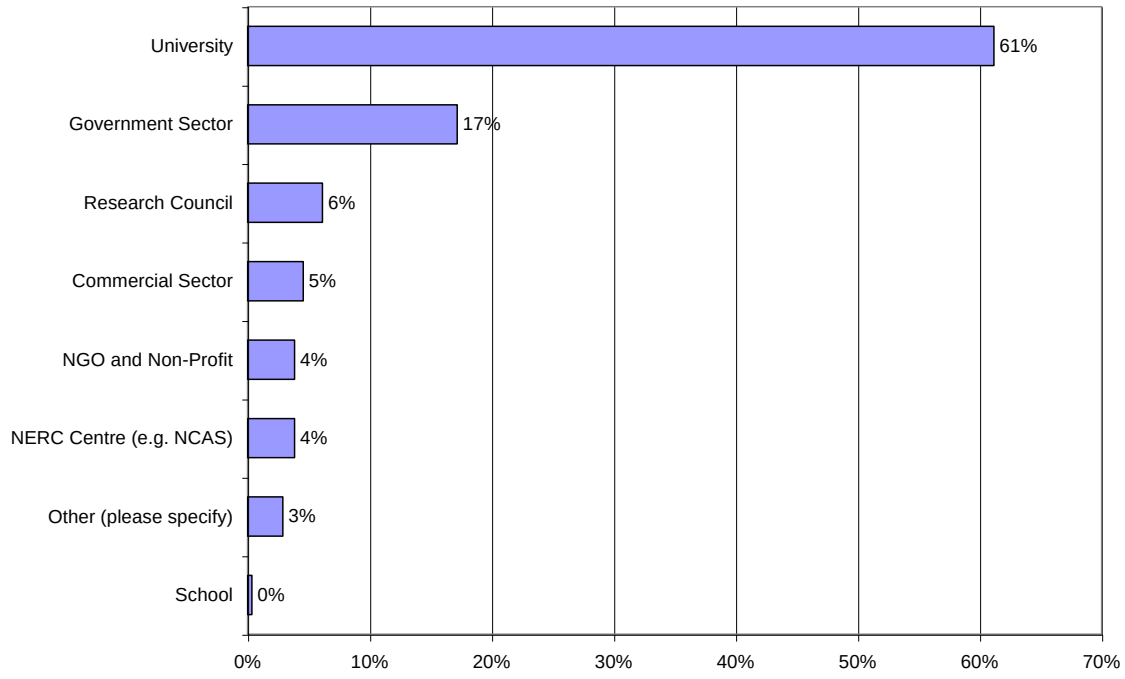
User demographics

The first part of the BADC users questionnaire explored demographics, including the respondents' main role, affiliation and country location and currency.

Q1: Main affiliation?

Reflecting BADC's user community, the majority of respondents were based at universities (61%), with 17% in the government sector. All other categories accounted for less than 7% each.

Figure A2: Main affiliation of BADC users (N=1141)

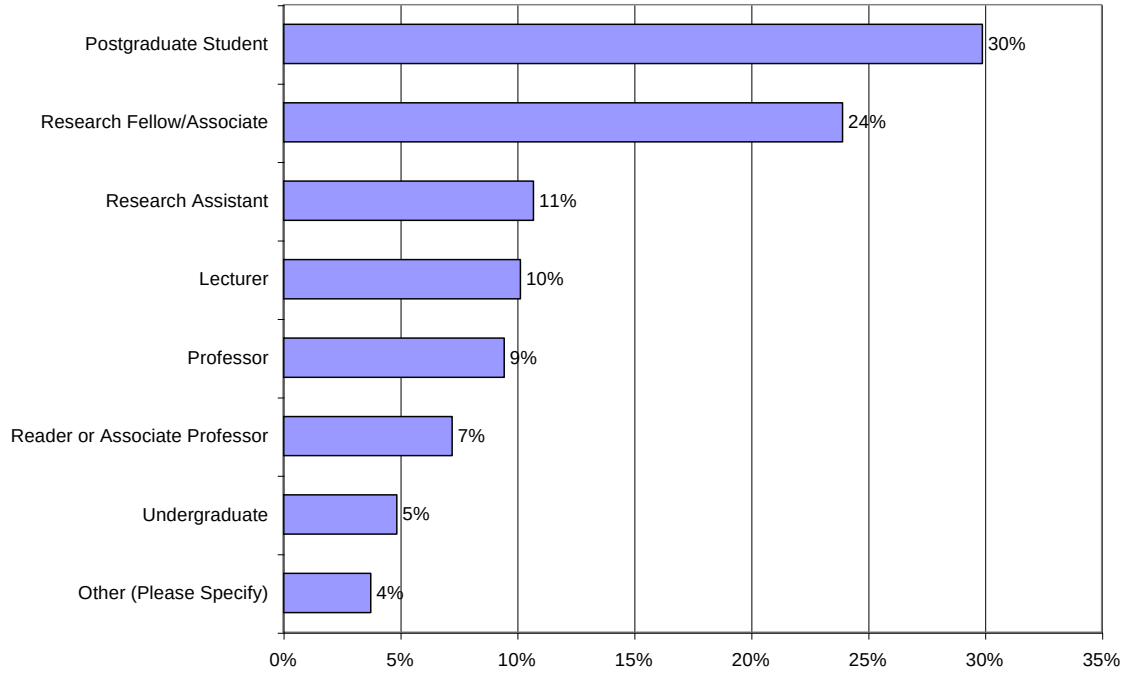


Source: BADC survey, Authors' analysis.

Q2: What is your main role within the university?

As might be expected, postgraduate students (30%) and research fellows/associates (24%) were the largest user categories. Undergraduate students accounted for 5% of university-based user respondents. Hence, students accounted for 35% of respondents.

Figure A3: Users role within the university (N=719)



Source: BADC survey, Authors' analysis.

A direct comparison with the BADC registered user statistics cannot be easily made due to the difference between user information held BADC and the user survey. The *BADC User Statistics Report* 2013 (Parton 2013) has attempted to infer the academic hierarchy of its registered users through inference between available information, but is unable to give a close enough match for most of the above categories. Even so, it is apparent that there is a lower proportion of respondents that are undergraduates for the user survey (~20% in the BADC user statistics compared with 5% survey respondents) than registered BADC users, but there are comparable levels of postgraduate users (~25%).

BADC staff believe the difference in undergraduate numbers can be accounted for through the fact that these undergraduate users have a shorter "user lifetime" with the BADC than those further up the academic ladder and a higher turn-over, but persist in this state or as users in the BADC registered user statistics as accounts remain open. It is also well known by the BADC that those who remain active registered users will not keep their BADC profile information up to date, thus further skewing the BADC user statistics to the lower levels of academia.

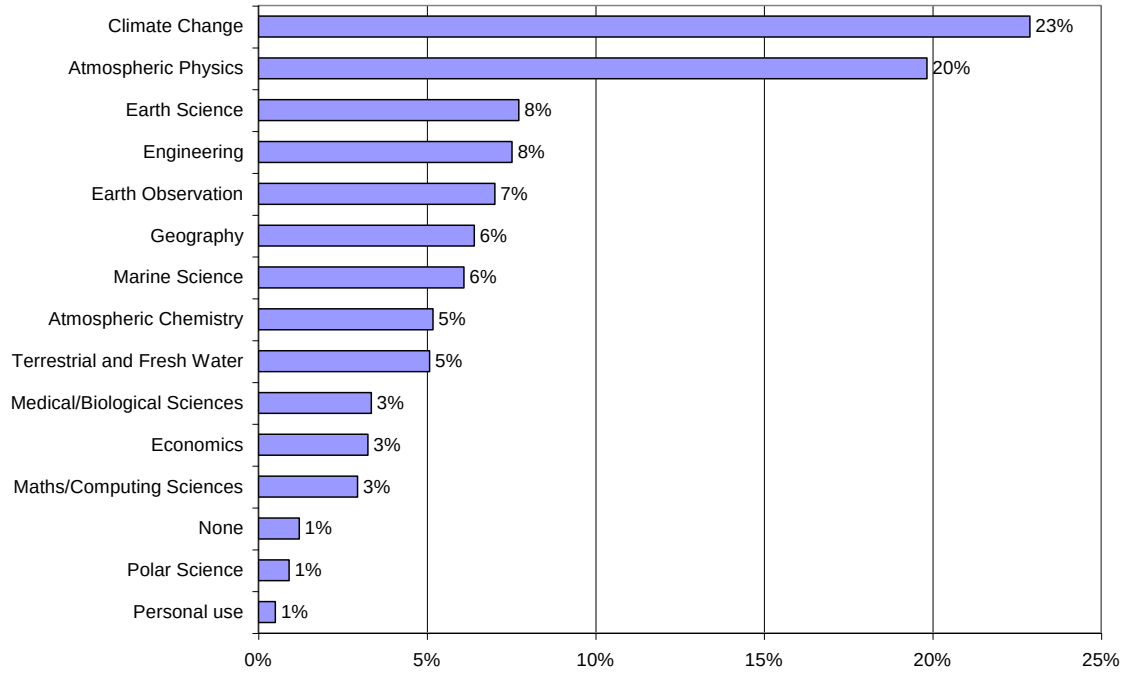
Q3: What is your main subject domain?

While a wide range of subject domains was reported, the largest groups of user respondents cited climate change (23%) and atmospheric physics (20%) as their main subject domains.

According to the BADC User Statistics report 2013 (Parton 2013), this is in strong agreement with the split of recently registered BADC users, especially when 2012 registrants are reviewed where the

Climate Change category was first offered as a research category in its own right (previously such users classified themselves under other available categories).

Figure A4: Main subject domain of BADC users (N=983)



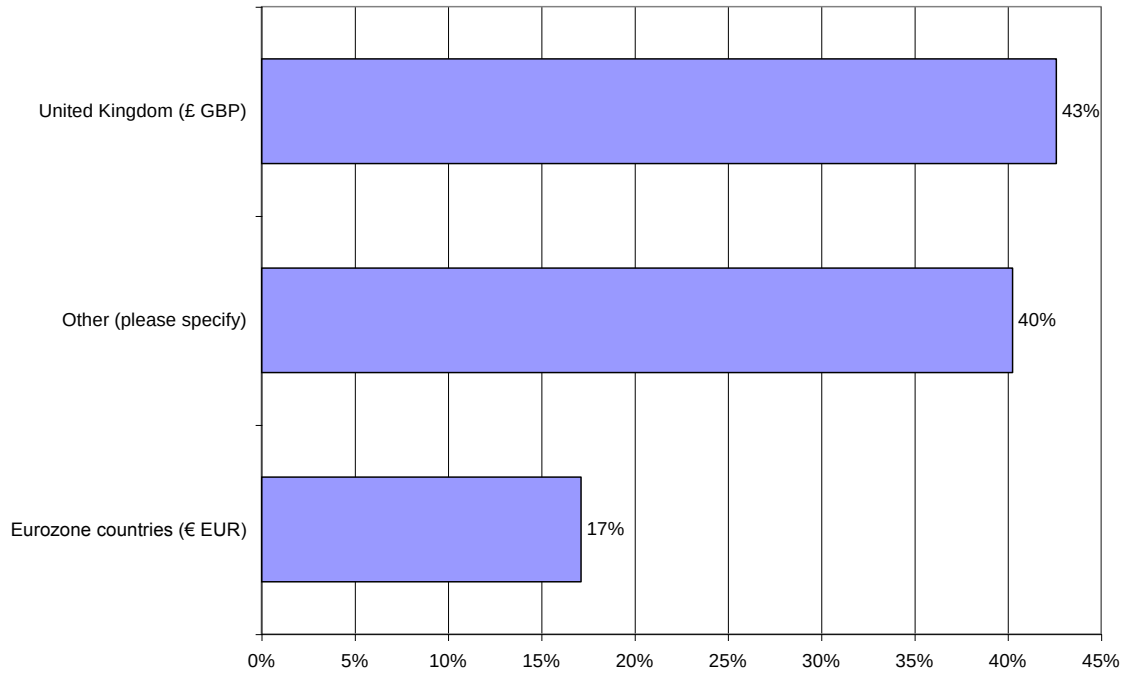
Source: BADC survey, Authors' analysis.

Q4: Country and currency?

Respondents were asked for their country and currency, both to explore the spread of BADC's user community and to ascertain their currency to enable activity costs to be estimated. Forty-three per cent of respondents were UK based, 17% from Eurozone countries and a further 40% from other countries. While the US, Canada, Australia, etc. were among these, there was a very wide range of developed and developing countries cited.

All subsequent cost responses were converted to British Pounds using spot exchange rates at 15th February 2013 (i.e. close to the survey closing date).

Figure A5: Country and currency of BADC users (N=1068)



Source: BADC survey, Authors' analysis.

The above geographical splitting highlights a significant departure from the BADC annual user registration statistics which indicate that while there has been a decreasing proportion of new registrants from the UK compared to those from abroad, dropping from 71% in 2007, and 68% in 2010, this still remained at over 60% per annum, with European, US and BRIC users at around 10% each in 2012 (Parton 2013 figure 6). However, the geographical splitting from the 2013 survey is closer to that reported in the RIN Survey in 2010, which was estimated at 51% UK based (Parton 2011). This divergence remains an anomaly across both the 2010 and 2013 surveys. For the 2013 survey, the geographical differences seem partly attributable to active registered users responding to the survey being drawn from more than the most recent year's registrations. The BADC statistics show higher levels of EU user registration in other recent years and sharp increases, with a doubling of BRIC user registrations in the last 12 month period (Parton 2013 Figure 6). There is also a higher than average number of UK based users in 2013 in the "unsure" category (see Appendix 2).

Data from previous user surveys and BADC User Statistics are described further in section 3 and the potential representativeness and comparability of the survey sample and BADC statistics are discussed further in section 3.2.

Q5: User registration?

BADC allows optional user registration. 78% of respondents reported that they had registered, and 6% had not (N=1058). The remaining 16% were not sure whether they had registered or not.

These figures are broadly comparable with those reported in the 2007 BADC User Survey but with a somewhat lower number of respondents stating that they were registered and in tandem a

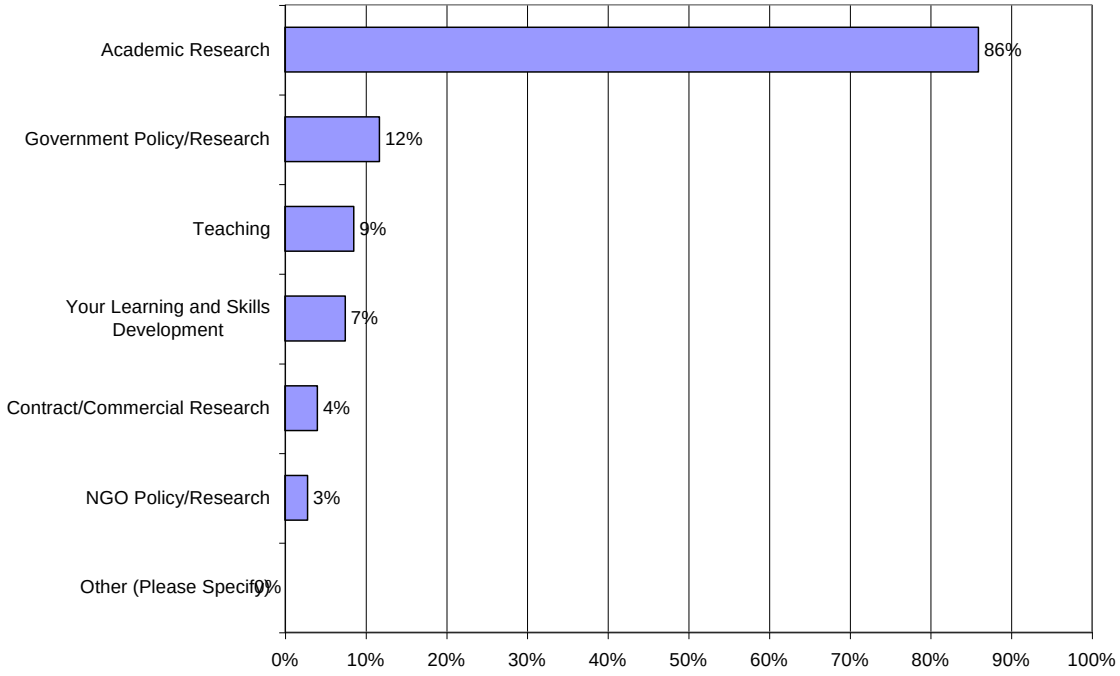
somewhat higher number being unsure. This change may reflect the often historic nature of user registration and the passage of time between the surveys. The respective figures in 2007 were 87% registered, 7% not registered and 6% not sure (BADC 2007).

Nature and frequency of use

The next section of the BADC user questionnaire explored the nature and frequency of use, looking at the main purpose of respondents' use, the impacts of use, the frequency of use and the types of data used.

Q6: What is the main purpose of your use of BADC data and services?
Reflecting the user and respondent communities, academic research was the most often cited main purpose of use (86%). Government policy research (12%), teaching (9%) and learning and skills development (7%) were the other categories most widely cited.

Figure A6: Main purpose of use of BADC data and services (N=1068)



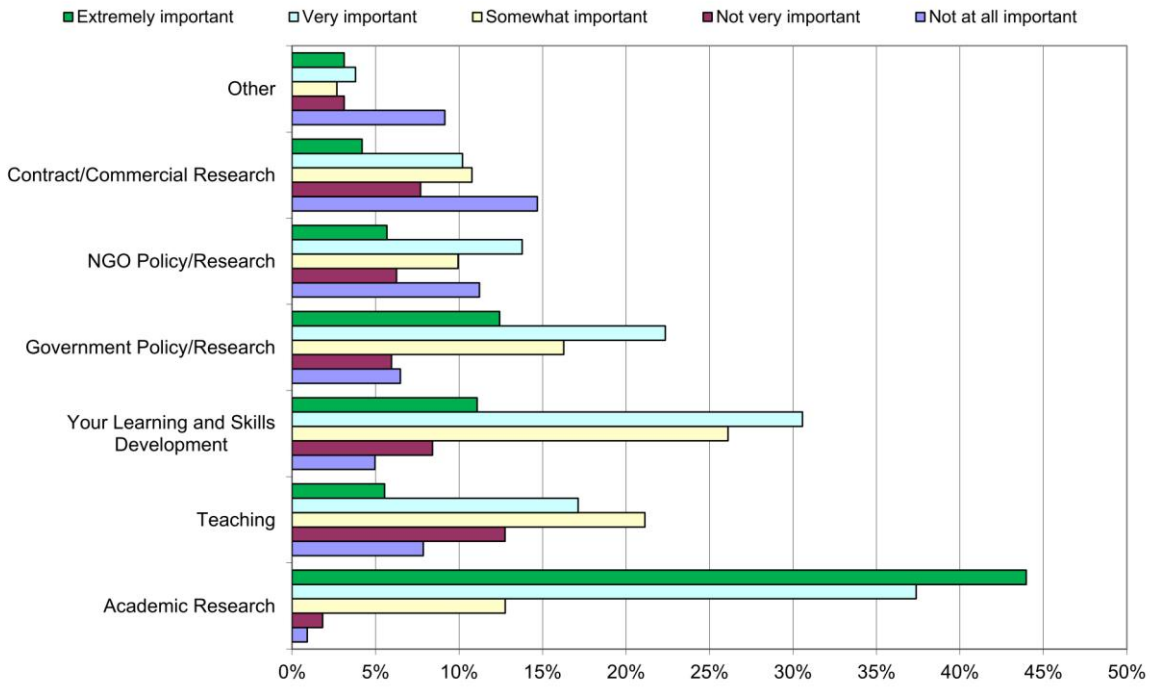
Source: BADC survey, Authors' analysis.

These results are very close to figures reported in the 2010 RIN/CEDA User Survey, where respondents were asked about their target audiences for their research outputs (RIN 2010, Parton 2011).

Q7: How important are data and services from BADC for the following areas of your work?
Asked how important BADC data and services were to their work, on a 5-point scale from "not at all

important" to "extremely important", 81% reported that it was very or extremely important for their academic research, 42% for learning or skills development, and 35% for government policy research (N=1019). Across responses, the highest importance ratings were for academic research (4.12), learning and skills development (2.78), and government policy research (2.19).

Figure A7: Importance of BADC data and services for users (N=1019)

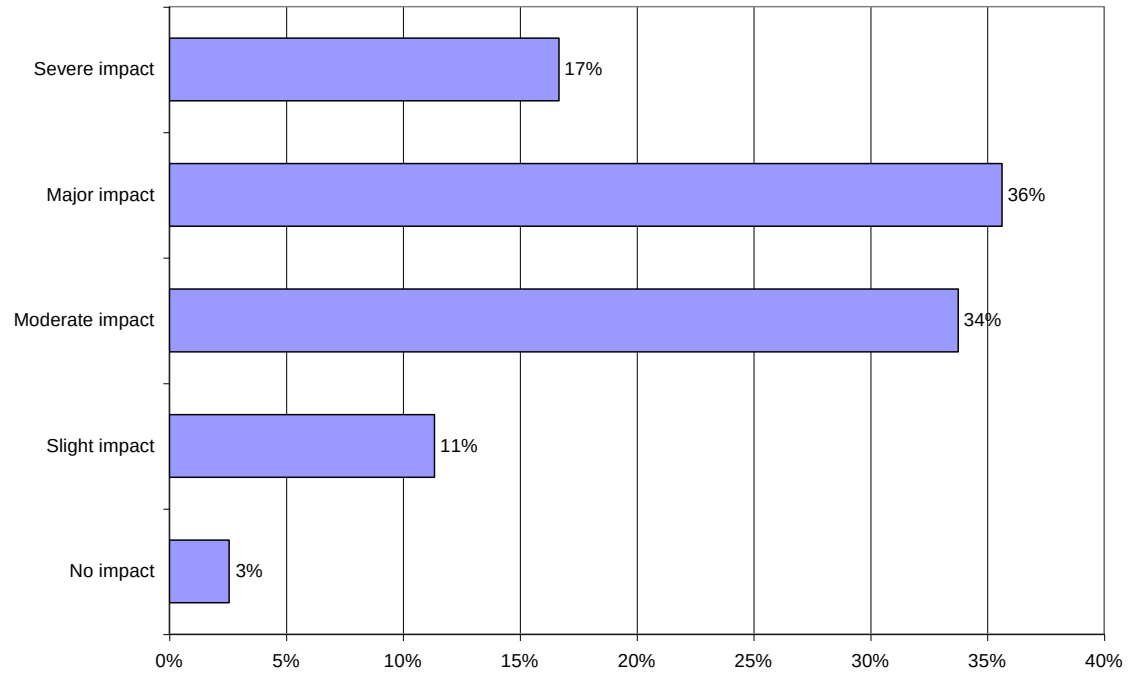


Source: BADC survey, Authors' analysis.

Q8: What impact would it have on your work or study if you could not access data and services from BADC?

53% reported that it would have a major or severe impact on their work if they could not access BADC data and services, and a further 34% said it would have a moderate impact. Less than 3% said it would have no impact.

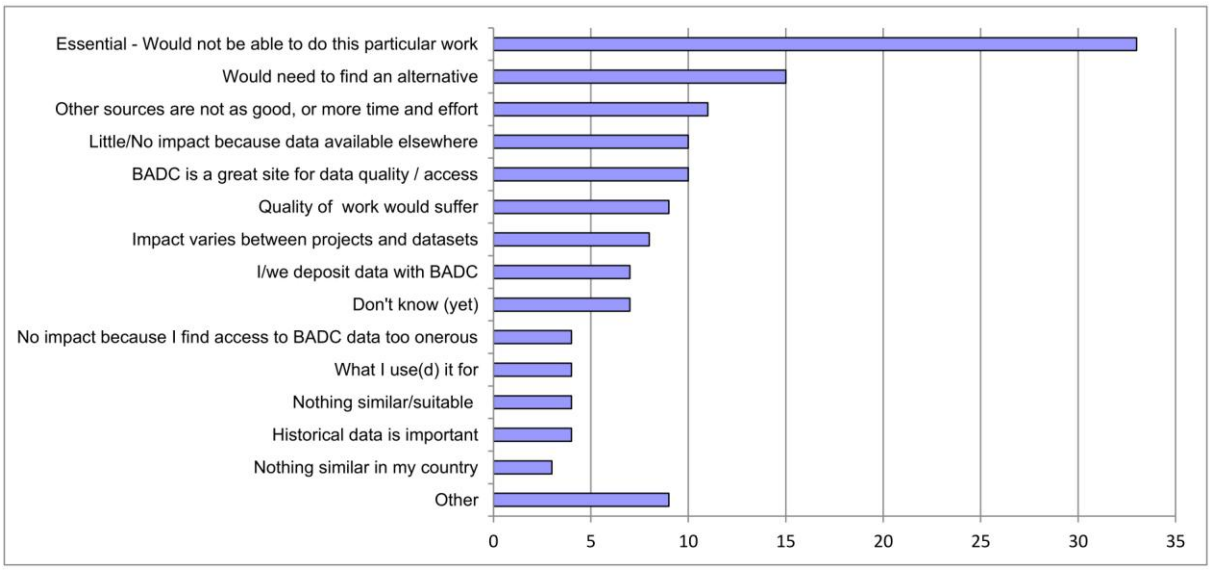
Figure A8: Impact on users of not being able to access BADC (N=1013)



Source: BADC survey, Authors' analysis.

Some 112 users also provided free-text comments to explain or amplify their responses to this question. Comments have been categorised by common themes in Figure A9 below and selected comments included for illustration of major themes.

Figure A9: Analysis of User Survey Question 8 free-text comments (N=112)



Selected Comments

'BADC is a one-stop dataset portal which houses key datasets covering primarily the UK/European region. Without the BADC, it would be time-consuming for me to search for the datasets at individual hosts.'

'BADC is fundamental to the way that academic atmospheric science is undertaken.'

'CRU TS dataset distributed by BADC is used both for historical climate impact modelling as well as bias-correction of climate scenario data. Therefore, it is crucially important to our work.'

'Would stop or at very least hold back many millions of pounds worth of academic research. Would force direct acquisition of datasets from Met Office which in my experience has been fraught with difficulties and costs.'

'Without the assistance and services provided by BADC the … project … would have been many orders of magnitude more difficult to implement and could well have failed.'

'To have data that is of high quality based on the reputation of the scientists who have prepared it saves me time in my work and that is a big benefit.'

'Not only does BADC act as a repository for general data, it is vital for hosting and archiving data from collaborative projects, which form the majority of the work I do.'

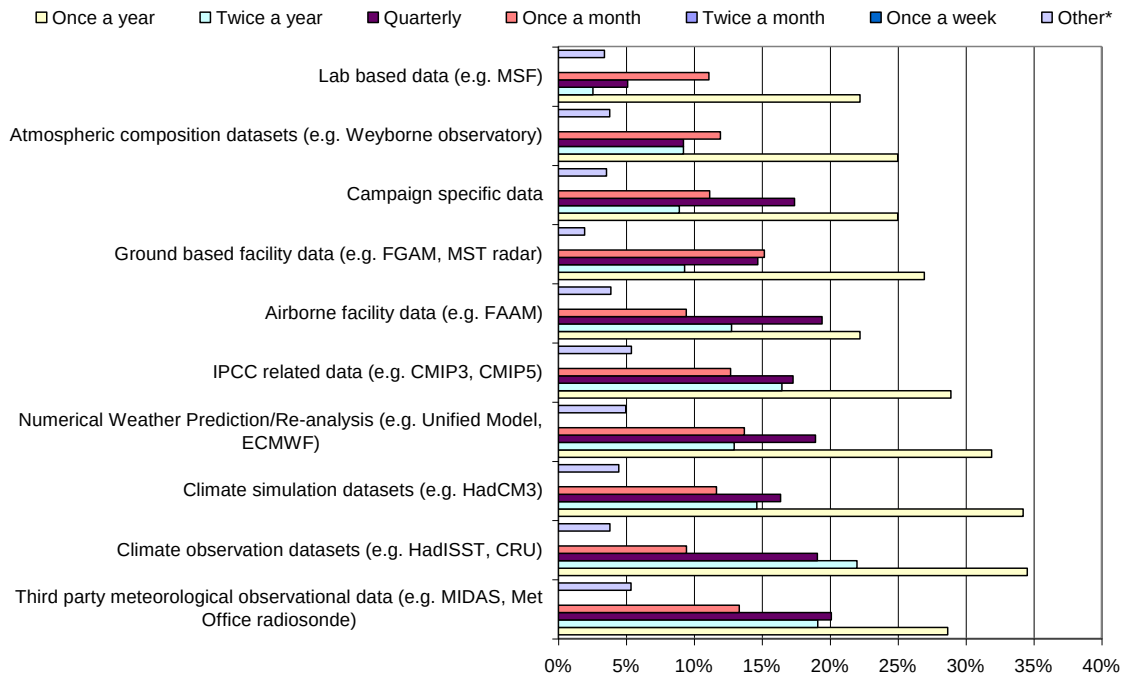
Q9: How frequently do you access/download data from BADC?

Asked how frequently they accessed various types of data hosted by BADC, answers varied widely by type of data and nature of use. Numerical weather prediction/re-analysis (e.g. Unified Model, ECMWF), third party meteorological observational data (e.g. MIDAS, Met Office radiosonde), and climate simulation datasets (e.g. HadCM3) were among the most frequently used data types.

Categories were converted to days per year, based on 44 working weeks and 220 working days per year. Overall, ignoring the "unable to say" responses, respondents reported a mean of around 25 accesses/downloads per annum (median 6).

The spread across the different data categories is in part due to number of datasets varying in each category, with some categories containing datasets that support a much broader user base than others. For example, there are very few users of the lab based data, which included only a handful of datasets, while the 3rd party meteorological data support the broadest user base and contain a large number of such datasets.

Figure A10: Frequency of BADC access/download (N=848)



Source: BADC survey, Authors' analysis.

Time and costs involved in access

The next section of the users questionnaire explored the time and costs involved in accessing data and services from BADC.

Q10: Which data type(s) did you last use?

To get a better idea of current use and to explore the time and costs involved, we posed a critical incident question about the data type last used. As respondents were allowed to nominate more than one data type (i.e. may have downloaded more than one data type in the last access session), results are expressed as a share of data types, rather than respondents. Climate observation datasets (e.g. HadISST, CRU) were most frequently cited as the last data type used (27% of uses), followed by third party meteorological observational data (e.g. MIDAS, Met Office radiosonde) 22%, and numerical weather prediction/re-analysis (e.g. Unified Model, ECMWF) and IPCC related data (e.g. CMIP3, CMIP5) both 13%.

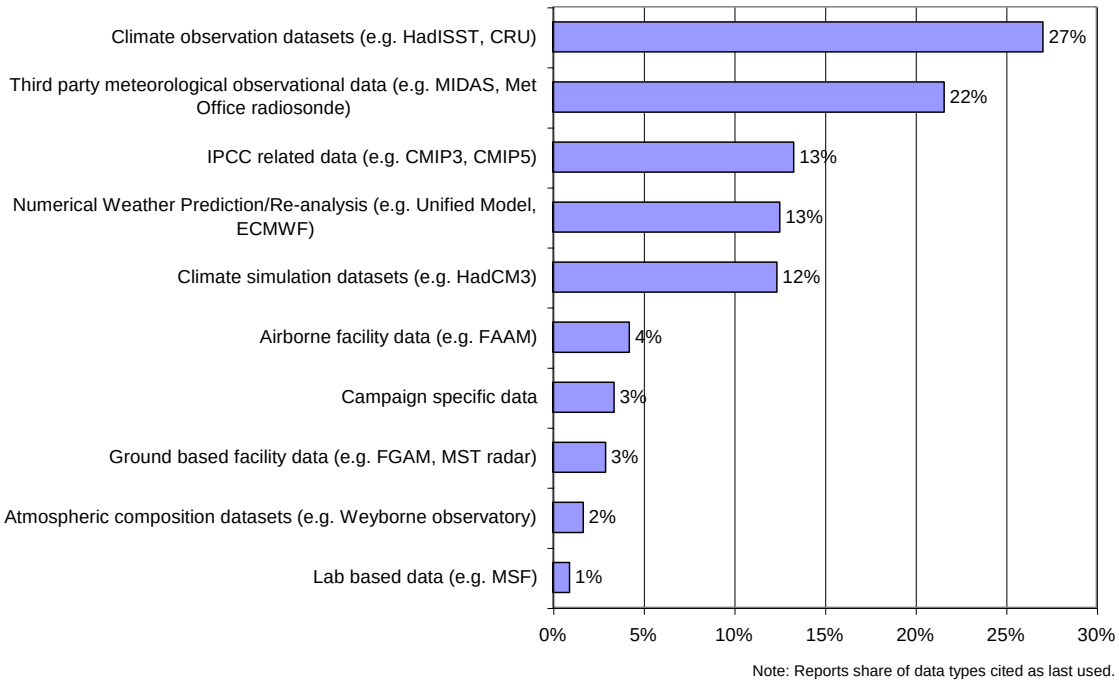


Figure A11: Last data type used (N=907, with 1301 type responses)

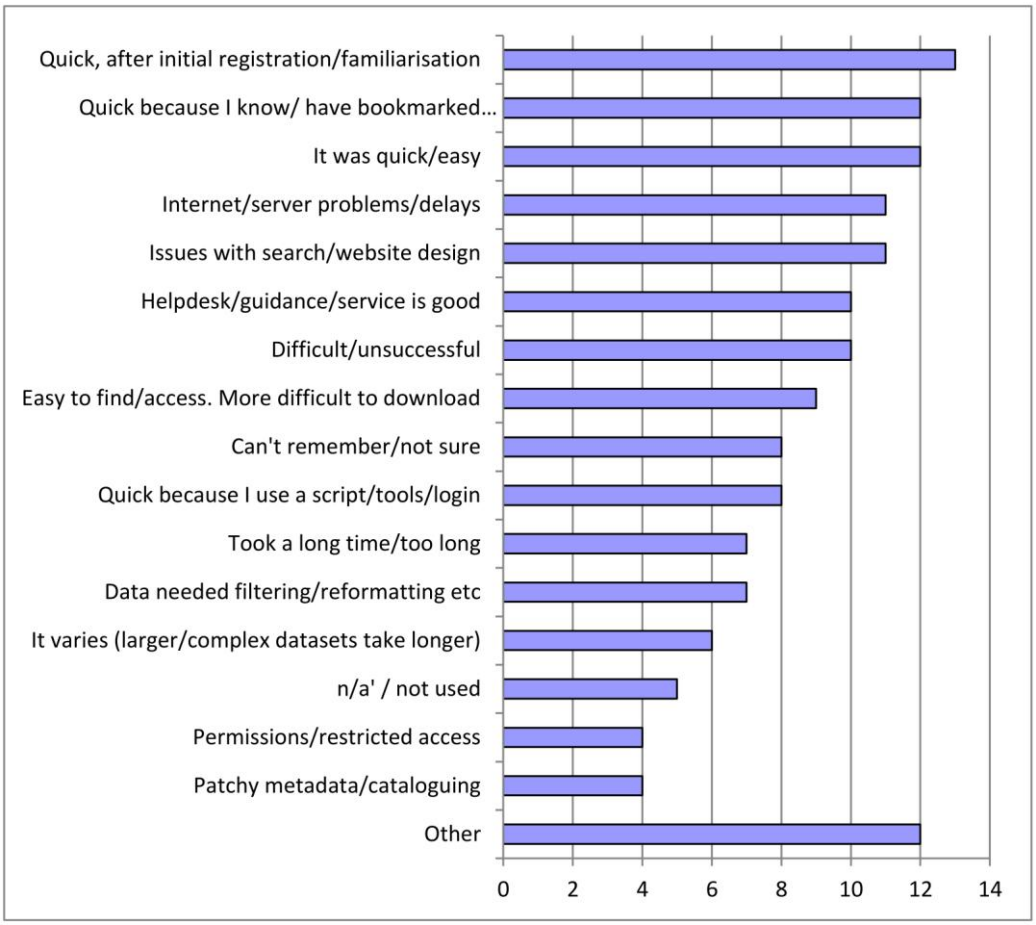
Source: BADC survey, Authors' analysis.

The breakdown of dataset categories is as expected given the interest in climate change data for work following the Intergovernmental Panel on Climate Change (IPCC) reports and wider press coverage of these data. In addition, as noted earlier, the third party meteorological data are a set of supporting data widely used as supplementary data in a width range of research disciplines and account for a large number of registered users.

Q11: How long did it take you to find and access the data you last used from BADC?
Respondents reported spending and average of around 2 ½ hours accessing the last data they used (N=870), with a maximum of 25 hours reported and a median of 1 hour.

Some 122 users also provided free-text comments to explain or amplify their responses to this question. Comments have been categorised by common themes in Figure A12 below and selected comments included for illustration of major themes. Comments suggested a range of experiences, with some noting how quick and easy it was to access BADC data and others reporting difficulties. One factor to note is the breakdown of time to search and access versus the time to download large datasets, with the latter sometimes more significant than the former.

Figure A12: Analysis of User Survey Question 11 free-text comments



Selected comments

'I was able to access the data with 15 minutes, but it took me more than an hour to download the data due to my internet connectivity.'

'I first looked at the BADC dataset Index. I was a bit overwhelmed by the quantity of datasets. When I performed a search I was not really satisfied with the results and was not sure that some datasets had not been missed. Finally it is a search on the Met Office website that led me to the MIDAS data set on BADC'

'It took a while to understand what was available. Then, it took me some time to find what I actually need. The data download was slow too.'

'Reflecting experiences related by colleagues and end users. xxx datasets are easy to find - and quick to download for small amounts - harder to access extensive datasets without significant investment in script development. Extremely hard to ensure that a local copy of a dataset is complete and up to date with the latest version. …it can be hard to find some datasets - metadata quality and completeness is patchy and in a few cases non-existent...'

'Thanks for an excellent data service.'

'It would be helpful to have an overview/list of some data (e.g. for radiosonde) in a plain ascii file including location in order to make a plot, selection according to coordinates etc.'

'…when searching for datasets the first time, I took me hours to find them.'

'I have a perl script that ftp's the data to my local computer in a matter of seconds.'

'It depends on my internet connexion, it can take just a few minutes if the network is correct.'

Q12: Were there any other costs involved?
Just 3% (30 respondents) reported that there were other costs involved, in addition to the time reported above (N=952).

Q13: If there were other costs involved, how much were they?
Fourteen respondents reported other costs, with a mean of £1,176. However, a number of comments suggested that some of the additional costs reported might be for their time (i.e. doubling up on their answer to the previous question) or for things that might be consider overhead and already included in activity costing costs (e.g. internet access). There was also one protest response, which has been excluded.

The time taken to access data was translated to an activity cost and added to these other costs to estimate the total costs involved in accessing BADC data and services. The costing method is outlined in Box A1. On that basis, the mean total cost of the last access was £100, with a median cost of £25.

Box A1: Approach to costing user time

User time was costed using average academic and graduate salaries reported in the Times Higher Education Supplement surveys, based on responses to Q2 (i.e. main role with the university). Each respondent was costed at the UK average salary reported for that academic level. Postgraduates were costed at the average graduate salary foregone, and undergraduates at the average school leaver salary foregone.

For non-university respondents, we used the overall average academic salary, as representative of the functions performed. In all cases, salaries were translated to an hourly rate based on 220 working days per annum and 7.5 hours per day. An uplift of 30% was applied to capture non-wage labour costs, using the method employed across UK government (Green Book 2011).

It should be noted that somewhat more than half of the user respondents were based outside the UK, and while many of those were in comparable developed countries, some were not. For developing country respondents, we are likely overestimating their actual costs. Hence, the costings presented should be thought of as "UK equivalent costs."

Additionality and benefits

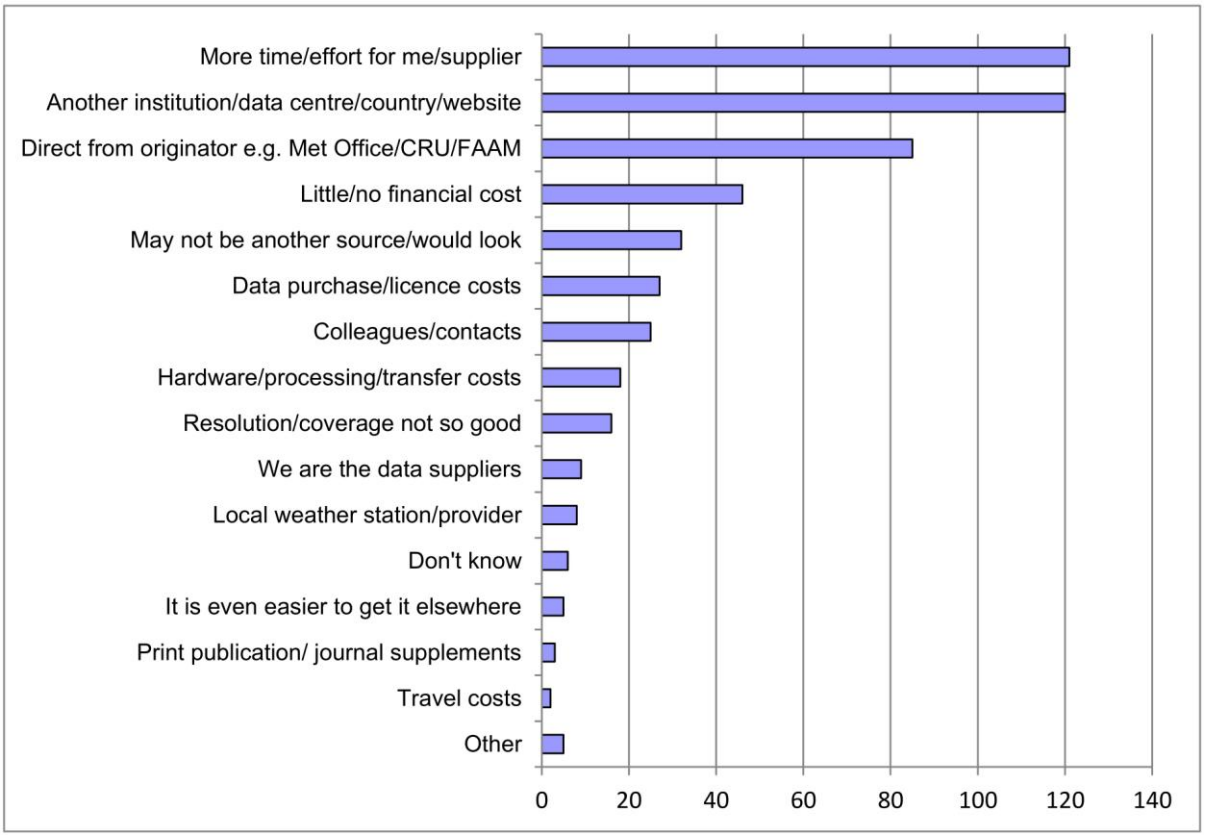
The next section of the questionnaire explored users' perceived benefits from using BADC data and services and the extent to which their use was additional use arising from the existence of BADC.

Q15: If BADC had not existed would you have been able to obtain the data you last used from another source?

Fifty-nine per cent said they could not have obtained the data elsewhere, while 41% could (N=892).

Some 296 users also provided free-text comments to explain or amplify their responses to this question. Comments have been categorised by common themes in Figure A13 below and selected comments included for illustration of major themes. Comments suggested that often additional time, effort and costs would be involved in obtaining the data from elsewhere.

Figure A13: Analysis of User Survey Question 15 free-text comments



Selected comments

'Possibly from printed reports in the Public Record Office, otherwise it would have required travel to UK Met Office in England costing around £500.'

'The data are available elsewhere but I am familiar with BADC and how to use it and it conveniently has many datasets under one banner so this saves time and effort in going round many different sources.'

'We have direct contacts with most of the international climate research centres. If BADC did not exist we would still be able to access data - in some cases this would be easier than accessing the data at BADC and in other cases harder. The Time and cost would increase without BADC access as there would be more data sources to deal with. Hard to quantify - but probably significant (more than double, less than 10x).'

'I would have to write new download scripts. The BADC service is quite good for scripting -- other services are not so good. It would take a few days of work to switch sources.'

'Perhaps yes, but the time, effort and cost involvement would have surely jeopardise my project for all practical purposes.'

'I am unsure whether I would have found the data, but it would have taken more time to find a source (if it exists). Having multiple data sets stored in one location is extremely convenient (& efficient)'

'Could have gotten the model data from xxxxx, but that portal is absolutely horrid and no sign of becoming easier to use. Your [BADC's] portal is much easier to use.'

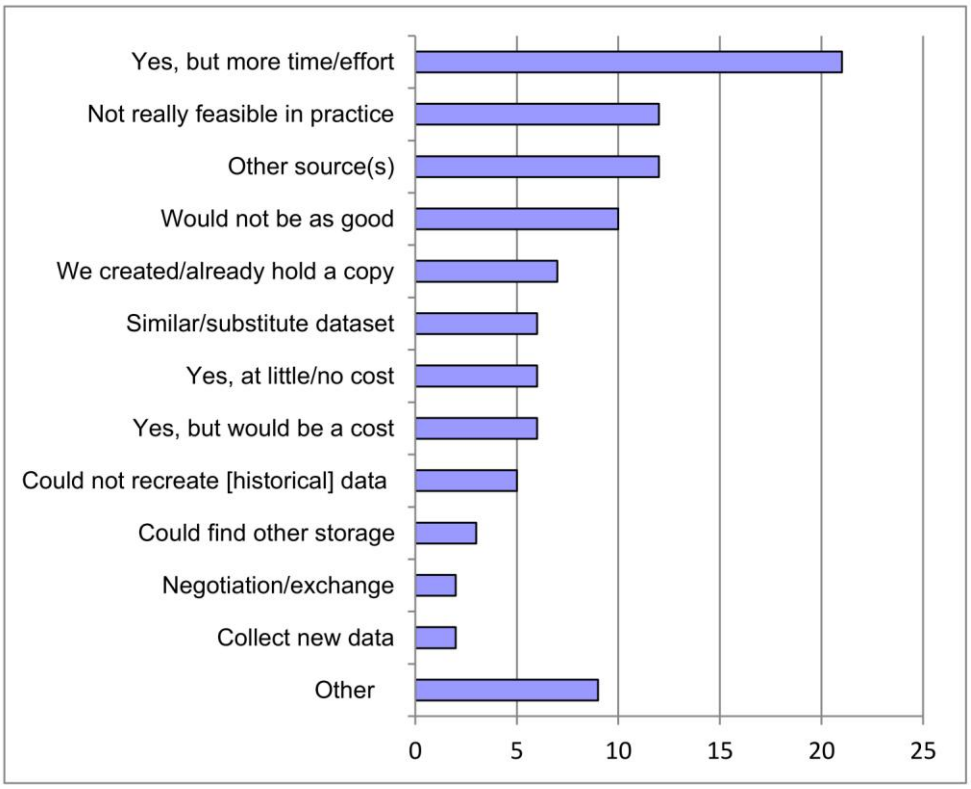
'It would take significantly longer to obtain this data, and considerably more if it involved a web interface to get the data rather than script-based ftp. Also, the BADC service is incredibly reliable and intermittent provision often experienced with other providers incurs a time cost to projects.'

Q16: If you could not have obtained the data elsewhere, would you have been able to collect/recreate the data yourself?

Eighty-nine per cent said they could not have collected/recreated the data themselves (N=868). However, some answered this question who had not answered "no" to Q15. Of those answering "no" to Q15 (i.e. could not have obtained the data elsewhere), 481 (91%) could not have collected/recreated the data themselves. Hence, around 54% of responding users could neither obtain the data elsewhere nor collect/recreate it themselves, and their use can be considered additional use arising as a result of the availability of BADC data and services.

Some 81 users also provided free-text comments to explain or amplify their responses to this question. Comments have been categorised by common themes in Figure A14 below. Among those who could have collected/recreated that data themselves, comments alluded to the additional time, effort and costs involved in doing so, and to a concern that the alternative data would not be as good.

Figure A14: Analysis of User Survey Question 16 free-text comments(N=81)

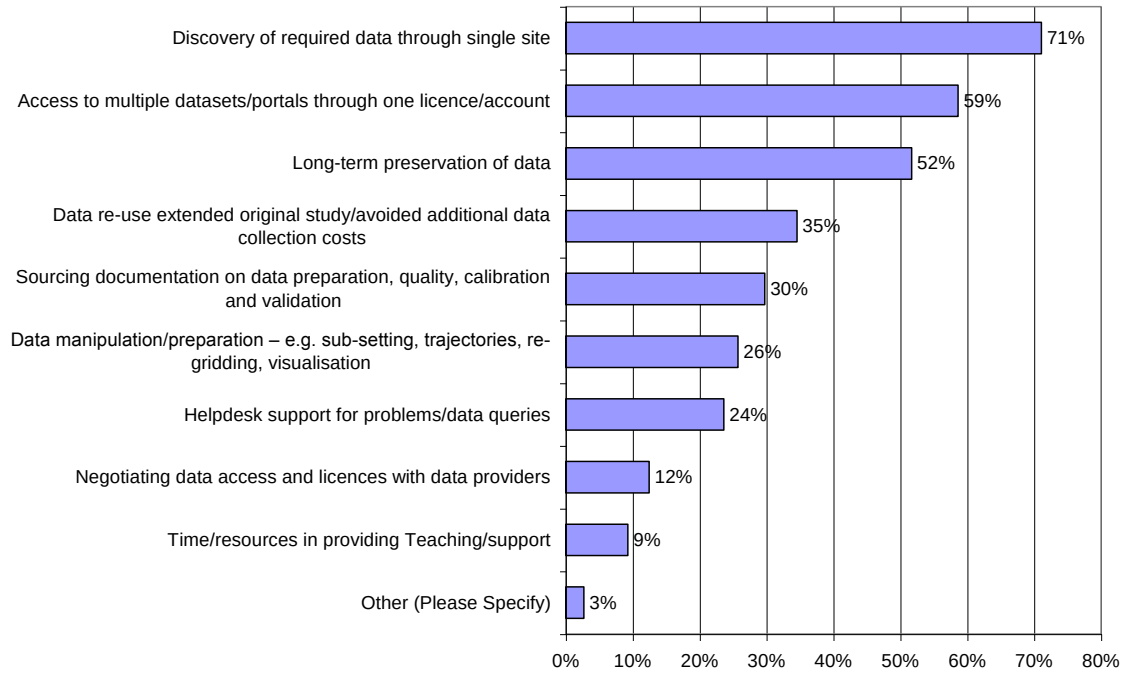


Q 17: Do you consider that you have benefited in any of the following areas as a result of using data and/or services from BADC?

Respondents suggested that they had benefited for using data from BADC in a variety of ways, with 71% saying they benefited from discovery of required data through single site, 59% from gaining access to multiple datasets/portals through one licence/account, and 52% from the long-term preservation of data offered by BADC.

While there were few additional comments, they included: 'Access to data expanded my horizons, opened up new research questions and generated new findings.'

Figure A15: Benefits from using BADC (N=852)

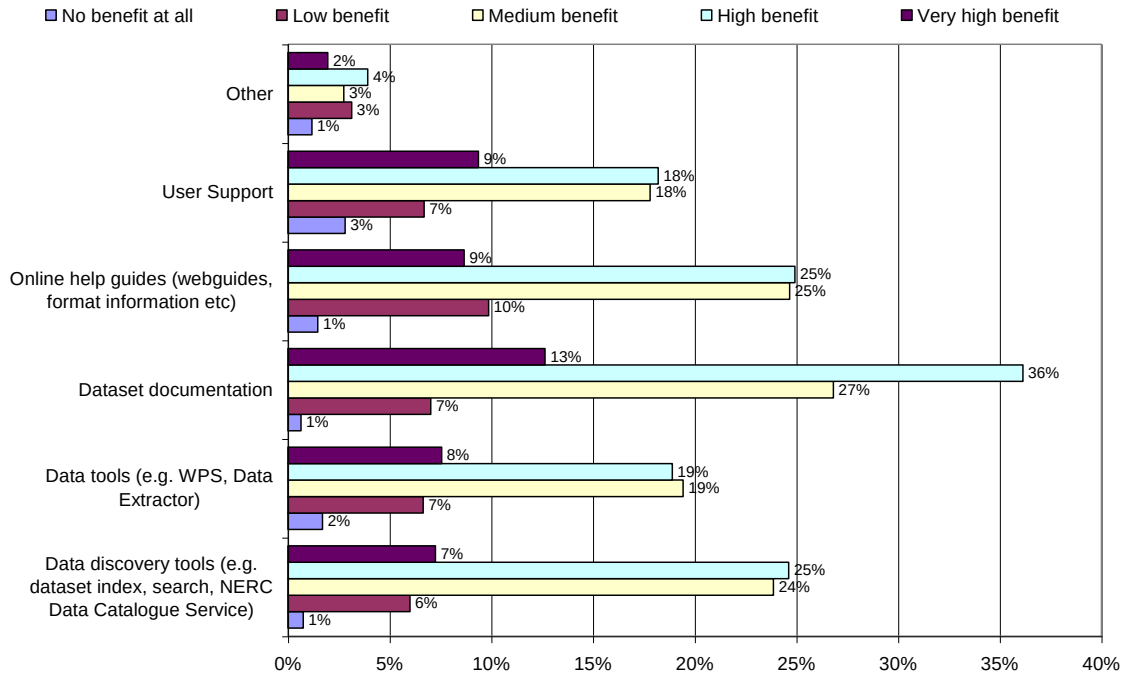


Source: BADC survey, Authors' analysis.

Q18: To what extent have you benefited from BADC in the following ways?

Almost 50% reported a high or very high benefit from dataset documentation, 34% from online help guides, and 32% from data discovery tools.

Figure A16: Benefits from using BADC (N=850)



Source: BADC survey, Authors' analysis.

Efficiency impacts

The next section of the user questionnaire explored the efficiency impacts of using BADC data and services for students, researchers and teachers/lecturers. The same series of questions was put to each group, although the primary focus of analysis is on research impacts. Moreover, one would expect that, due to limited experience, students might be less able to form mature judgements about their efficiency. Nevertheless, all results are included for completeness.

While the questions on share of time spent with data sources referred to overall studying/research/teaching time, it seems possible that a number of respondents incorrectly interpreted it as relating to time spent with data and/or thought that the time shares should sum to 100%. As a result, responses have an upward bias, and must be interpreted with caution.

Students

Q19: Are you an undergraduate or school student?
Six per cent of respondents said that they were students (N=49).

Q20: Over the last 12 months, on average how many hours per working week did you spend studying and learning?
Asked how many hours a week they spent studying/learning, the students reported some very high numbers. A maximum average of 100 hours per week was reported by one respondent, while a number reported studying for an average of 40 hours per week and more. Two erroneous answers with data use time shares adding to more than 100% were deleted. The overall mean was 35 hours per week, median 40 hours per week (N=40).

Q21: Can you estimate the approximate share of your TOTAL study/learning time spent with data during the last 12 months?
Students estimated spending an average of around 18% of their time with data from BADC (N=41) and a further 45% of their time with data from other sources (N=40). That translates to an average of around 5 hours a week with data from BADC and 21 hours a week with data from all sources (including BADC).

Q22: Do you have any impression of what might be typical of other students at your level?
While these questions are particularly difficult to answer and the experience of students may be limited, they estimated that other students at a similar level to themselves might spend an average of around 19% of their time with data from BADC (N=32) and 38% of their time with data from all other sources (N=31). While high, these are similar to their own time shares.

Q23: How much has your use of data and services from BADC changed your overall study/learning efficiency?
Asked how much their use of BADC data and services had changed their overall study/learning efficiency, the average for those citing a zero or positive efficiency impact was 34% (median 30%) (N=37). One respondent suggested that the impact had been negative.

Selected comments

'Accessing, understanding and manipulating BADC data has greatly improved my skills in Excel such that I can complete operations and understand metadata (etc) far quicker than I'd be able to if I'd never had to use the data and get accustomed to it.'

'Through the extensive use of the data, through downloading, unzipping and subsequent usage, my computational skills and understanding of the topics the data were used for have both dramatically improved in terms of the efficiency upon which I conduct operations, process large volumes of data, organise data and am able to understand the data which is being used, thus saving time during like operations and otherwise through the transferable skills gained.'

Researchers

Q24: Is research part of your role?
Ninety-five per cent or respondents (N=784) said that research was a part of their role.

Q25: Over the last 12 months, on average how many hours per working week did you spend on research?
Asked how many hours a week they spent on research, some very high numbers were reported. A maximum average of 100 hours per week was reported by 11 respondents, while a number reported averaging 40 hours per week and more. The overall mean was 30 hours per week (median 30 hours per week) (N=738).

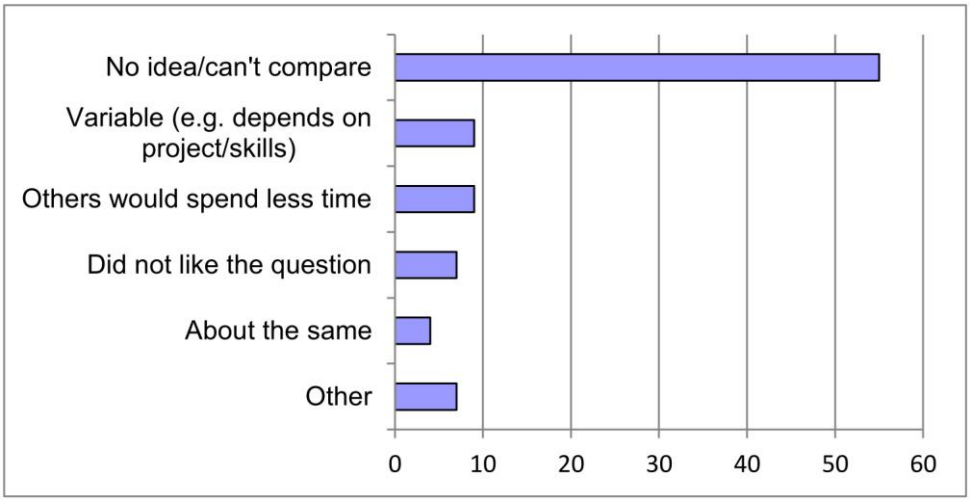
Q26: Can you estimate the approximate share of your TOTAL research time spent with data during the last 12 months?
Due to difficulties of interpretation nine erroneous responses were deleted. As noted, a number appear to have misinterpreted the question to refer to share of data use time, rather than overall research time. Responses must be interpreted with that caveat in mind.

Research respondents estimated spending an average of around 18% of their time with data from BADC (N=706) and a further 50% of their time with data from other sources (N=698). That translates to an average of around 6 hours a week with data from BADC and 21 hours a week with data from all sources (including BADC).

Q27: Do you have any impression of what might be typical of other researchers in the same sector as you?
For the purposes of subsequent economic analysis respondents were asked for their impression of the share of time spent with data that might be typical of others in their field, in order to judge the respondents relative data use intensity. Research respondents estimated that other researchers in their sector might have spent an average of around 21% of their time with data from BADC (N=400) and further 48% of their time with data from all other sources (N=394). These are similar to their own time shares.

Some 89 users also provided free-text comments to explain or amplify their responses to this question. Comments have been categorised by common themes in Figure A17 below.

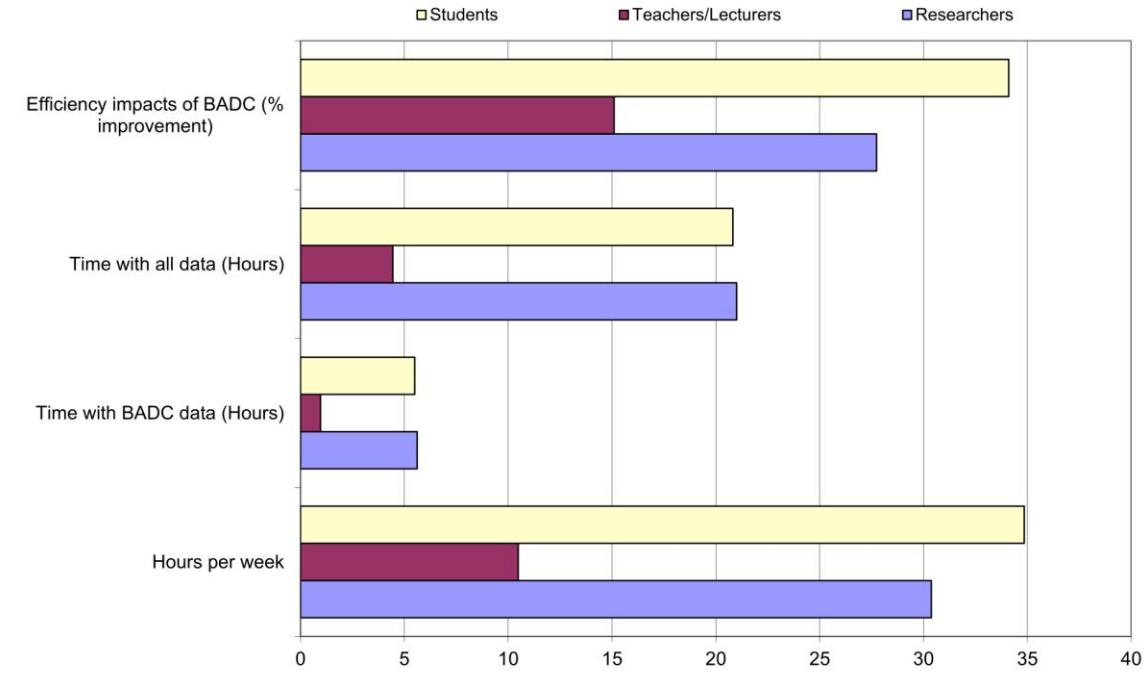
Figure A17: Analysis of User Survey Question 27 free-text comments (N=89)



Q28: How much has your use of data and services from BADC changed your overall research efficiency?

Asked how much their use of BADC data and services had changed their overall research efficiency, the average for those citing a zero or positive efficiency impact was a 28% positive improvement (median 20%) (N=498). Just 8 respondents suggested that the impact had been negative: they commented on the difficulties they had experienced using BADC and, hence, the time wasted.

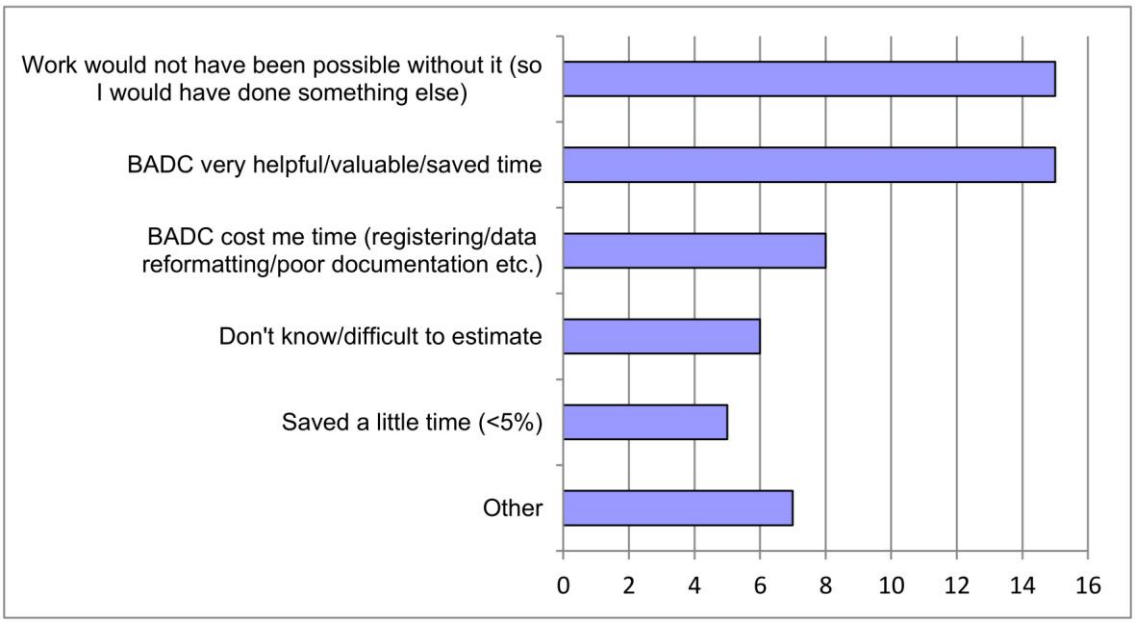
Figure A18: Time and efficiency impacts of BADC data and services



Source: BADC survey, Authors' analysis.

Some 53 users also provided free-text comments to explain or amplify their responses to this question. Comments have been categorised by common themes in Figure A19 below.

Figure A19: Analysis of Use Survey Question 28 free-text comments (N=53)



Teachers/Lecturers

Q29: Is teaching/lecturing part of your role?
Forty-one per cent of respondents (N=335) said that teaching/lecturing was a part of their role.

Q30: Over the last 12 months, on average how many hours per working week did you spend teaching and preparing learning materials?
Respondents reported spending an average of 10 hours per week teaching and preparing learning materials (median 8 hours per week) (N=308).

Q31: Can you estimate the approximate share of your total teaching and preparation time spent with data during the last 12 months?
Bearing the caveats noted above in mind, teaching respondents estimated spending an average of around 7% of their time with data from BADC (N=282) and a further 29% of their time with data from other sources (N=272). That translates to an average of around 1 hour a week with data from BADC and 4 hours a week with data from all sources (including BADC). One protest response was excluded.

Q32: Do you have any impression of what might be typical of other teachers/lecturers in the same sector?
Respondents estimated that other teachers/lecturers in their sector might have spent an average of around 12% of their time with data from BADC (N=151) and further 30% of their time with data from all other sources (N=143). One protest response was excluded. Again these estimates are similar to their own time shares.

Some 29 respondents provided free-text comments, with 27 simply saying they felt unable to estimate what might be typical.

Q33: How much has your use of data and services from BADC changed your overall teaching/lecturing efficiency?
Asked how much their use of BADC data and services had changed their overall teaching efficiency, one respondent suggested that the impact had been negative. The average for those citing a zero or positive efficiency impact was 15% (N=221).

Contingent valuation

The final three questions were addressed to all user types and involved the use of a contingent valuation approach using stated preferences. In the absence of a market price, respondents are asked how much they would be *willing to accept* in return for giving up their access to BADC and how much they would be *willing to pay* for access. A description of this method can be found in Section 4.2.

As is common with this method, there were a number of protest answers. There was also a small number of answers that betrayed the fact that respondents had not carefully read or understood the questions. However, an important issue emerging from the open-ended comments was that some respondents had said that they were willing to accept nothing in return for giving up access to BADC

because they believed that such data should be free, rather than because they did not value it. These 30 free data access and the protest responses were excluded from the subsequent analysis.

It should not be surprising that individual responses vary widely, as different users use different data and services, and use them with very different frequency. Some regular users may base their work on BADC data and services, while others may be occasional users who dip into BADC data as a matter of passing curiosity. Consequently, they are likely to value BADC data and services quite differently.

Q34: Imagine that the BADC was no longer open to new users. You have the option to either carry on using the BADC or to sell your rights as an existing user to a third party. If you sold your use of the BADC, what is the MINIMUM amount that you would be willing to ACCEPT as an annual payment in return for giving up ALL of your use of the BADC?
Respondents reported that they would be willing to accept a mean of £7,072 in return for giving up their access to BADC for a year (N=532). The highest response was £1 million, with £645,000 the second highest, and a further four responses of £100,000 or more. No fewer than 106 respondents said they were willing to accept £0 in return for giving up their access. When combined with their reported individual frequencies of access, this implies an average willingness to accept of around £686 per access.

Q35: The BADC receives funding from a number of bodies and is committed to providing free access. For this question, however, please imagine that this funding ceased to be provided. In this hypothetical case, what is the MAXIMUM amount you would be willing to PAY for your individual access to data and services from the BADC?
Respondents were then asked, what was the maximum amount that they would be willing to pay for access to BADC data and services: (i) per year, as an annual subscription, and (ii) per use, on a pay-per-view basis. Respondents said they would be willing to pay an average of £1,120 per annum (N=575), or an average of £418 per view (N=543).

One check on the willingness to pay answers is to compare the annual amount with the pay-per-use amount multiplied by the frequency of use per annum. This generally reflects a willingness to pay more per use, for the convenience and lack of longer term commitment to paying. Doing so suggests that the implied average pay-per-view willingness to pay would be equivalent to £1,430 per annum. This is a good match with the reported annual willingness to pay of £1,120 per annum when the tendency to attach a premium to pay-per-view is taken into account, suggesting some confidence in the answers.

What people are willing to accept in return for giving up access is not constrained by their ability to pay, whereas what they are willing to pay is constrained. Therefore, it is normal to see a willingness to accept that is higher than willingness to pay. The more constrained is capacity to pay, the larger the gap.

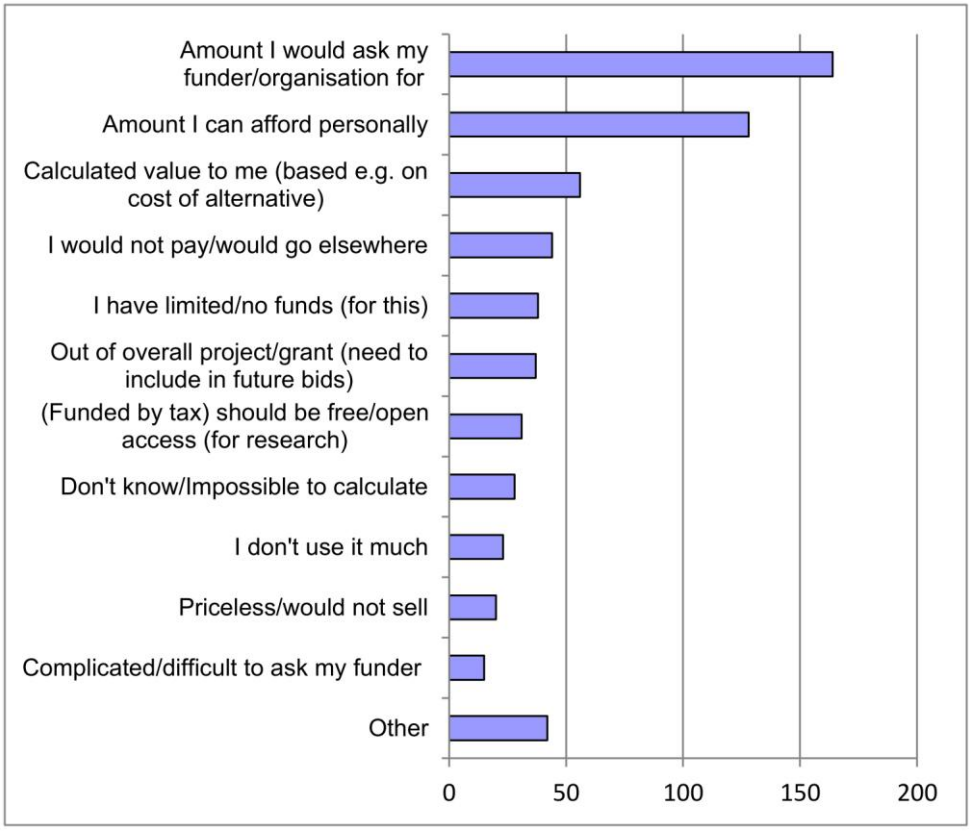
Q36: What was the basis for your answer to the previous two questions (e.g. amount I can afford personally, amount I would ask my funder or organisation to pay for me as a single user etc.)?
Respondents were given the opportunity to comments on the contingent valuation questions, as required by the method.

Some 522 users provided free-text comments to explain or amplify their responses to Q34 and Q35 as requested. Comments have been categorised by common themes in Figure A20 below. Note individual responses could span more than one of the identified overall themes.

Some 164 stated the basis of their calculation was the amount they felt they could ask their research funder or their institution to pay. A further 127 stated that it was the amount that they could personally afford to pay.

Some 61 comments indicated that the respondent had based their calculation of value on a measurable comparator, such as time saved, costs of other services (e.g. journal subscription) or of alternative data sources, or the commercial value of the data to them. Another 35 comments argued that data that has been funded by tax should be free for research purposes, although they were comfortable with the idea of charging for commercial use of that data.

Figure A20: Analysis of Use Survey Question 36 free-text comments (N=522)



Despite assurances in the questionnaire, some users found it hard to believe that this was just an intellectual exercise and responded to the prospect of the service becoming subscription-based. For example, 31 discussed payment models. Some infrequent users would only consider 'pay per use', though some of those who indicated that they did not use it often also said it was essential when they did. Several people felt 'pay per use' would be impractical due to: the automated nature of their downloads; variable size of datasets; uncertainty of the value of the data until you have

examined it (could you only include 'successful' use?); the hit and miss nature of the service/internet connection/data availability; what constitutes a dataset if this is a series; and would ‘pay per use’ be for a single file or for a day's downloading. Non-UK participants were concerned about reciprocity of national access.

Selected comments

‘Selling my rights to a public (and free) service is a strange idea to me, but I'm sure I would regret it so I put a high price ...’

‘the access to BADC data is invaluable, so the answer to 24. On 25, the values are on one hand a maximum I would personally pay and on the other hand some reasonable number for my funder.’

‘The amount I would be willing to pay personally, as I have no sources of funding for this aspect of my research. This results in a lower use than I would like to have of the BADC datasets - which is an excellent resource’

‘Amount of time I could guarantee saving by purchasing data… and therefore the amount I could take to my employer (PI for research grant).’

‘I cannot afford to pay for this data, especially when other sources could exist. I do hold BADC and their data holdings in the highest regard.’

‘It's an amount I am willing to pay myself. If I were to ask my employer (government), it would be so long and complicated at the administration level that I would give up!!!’

‘The extra time and effort required to get the same data elsewhere and the reduced coverage would mean that minimum amount I would be willing to accept is quite high although manageable. How much I could afford to pay is more about how much is assigned to the project and the consideration of buying in other data (which is not free). Per-use is low as sometimes data sets can only be found to be insufficient once downloaded which means that often 3 or 4 data sets are downloaded to create a usable one’

‘I treat it similarly to the costs of software licensing’

‘Research funds are tight! Although big headline research projects will always fund data access, but I believe a major value of BADC is that it allows quick checks, "pet" research mini-projects, "what-if" scenarios to be developed. These are very valuable services, but don't have much monetary value attached to them. For this reason, I would like to see BADC access remaining free.’

‘Cost of equipment to gather the necessary data’

‘I can't answer this. I wouldn't sell my user rights, since I don't consider that I own them. Our institution is unlikely to pay for anything so if funding ceased then that would be that.’

‘estimated funder contribution (pay-per-use would be difficult to legislate through a funding agency) note, consistency in availability would need to be improve drastically if this was to be offered as a "pay-per-access" service’

‘max: price of ECMWF wholesale data catalogue, min: a random figure, as "per use" could (should) mean a regulars batch job requesting any size data set(s)’

‘Reasonable maximum yearly amount for institution to pay for single user access with fairly heavy usage patterns. However if this were to happen I would expect the actual amount to be significantly less as it would make more sense to pay for access per institution or research group as some people will only use the service occasionally.’

‘In my opinion public bodies should provide all the data they collect and process for free (at least for research and teaching). I am not willing to pay any fees for accessing BADC data in the current format (sometimes scripting is needed to visualize the data)’

‘I don't use it very much, but when I do, it is invaluable. I wouldn't know how else to get the data I need, and wasted a lot of time before I found BADC. I'm a researcher doing my PhD in a small charity, and it would be hard to justify an annual subscription. I would have to pay on a single use subscription as needed. I greatly appreciate this service. I would not sell the use of my access to BADC, for any amount of money. What would I do instead?’

‘The amount I would ask my funder to pay. In the end it's a silly question because either a research council pays or some other organisation(s) pay as at present. If it were coming out of my own pocket, I would not pay anything. By the way, I don't believe that these are hypothetical questions...’

‘…Pay per access should be either (or both): (a) per successful download, or (b) per access to an entire single data set (as many times as required)’

‘… all publicly funded research is public research and no one should have to pay more for the results. At the same time I have a worry that paying for information/knowledge is a slippery slope for only the rich will benefit … though I admit there are costs to providing services so the question is, is the service worth it? If it is, then some fee for that service is justifiable but it must be priced accordingly and recognize that a scientist from Haiti should not have to pay the same as one from Stockholm.’

‘To give up BADC access to be given a yearly amount that would compensate for no longer being able to do my job. I think this question is very silly as this value is not the value I think I get from a year of using the BADC...’

‘…if you start charging for the basic climate data sets, people will find other sources and process the data themselves. I do not have objections to charging for those who repackage and then resell the data. If you are making a profit from the data then you are likely to be willing to pay for it. I am sure it is hard to keep track of this as there are many grey areas (e.g., non profit private research organizations).’

Depositor survey

Emails to participate to the Depositor survey were sent to 143 BADC users identified as past and present depositors. Around 20 of these users were no longer reachable through their contact details held by BADCA link to the Depositor survey was also posted on the CEDA website and news feeds. Overall 42 responses were received by the close of the survey, and given that no depositors were expected to participate that had not been emailed successfully, an estimated 34% response rate was obtained.

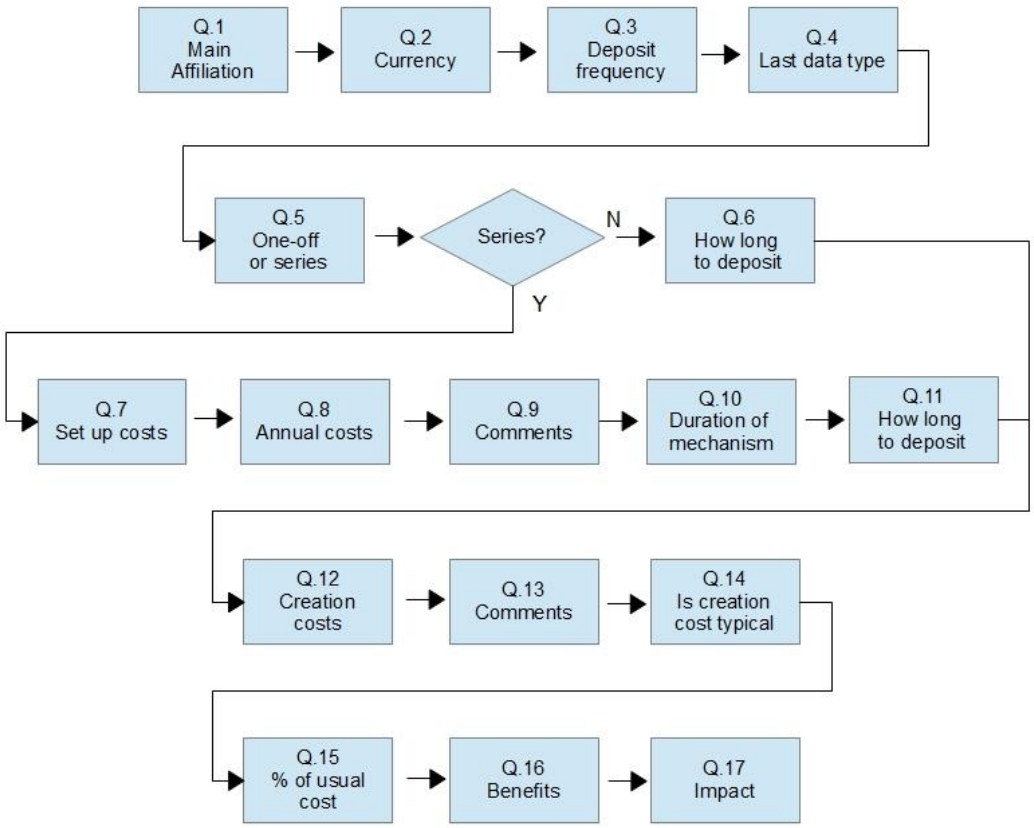
The Depositor survey was particularly difficult to construct and for users to complete, given the high level of deposit heterogeneity. Deposits to the BADC cover one-off deposits at the end of a project,

continual deposits in a limited series (e.g. during a field campaign) and on-going deposits to continually growing datasets. Within these latter two categories the frequency of deposit ranges from as frequent as every 5 minutes from Met Office operational network data; to monthly or yearly updates; through to ad-hoc updates. In addition, the size and complexity of the deposits ranges from dataset to dataset when number of files, overall data volume (plus individual file size), spread of products and access control requirements are considered. All of these factors are important when considering the costs involved with setting up deposits to the BADC. Likewise, the fact that the BADC has to deal with this heterogeneity and yet still maintain delivery of services and data to the end user community through just one data centre, is worthy of note.

As a consequence of the breadth of deposit types and complexity noted above the survey results presented here must be interpreted in combination with initial and follow-up interview findings.

The Depositor Survey consisted of 17 questions and followed the following logic in the sequence of questions.

Figure A21: BADC Depositor Survey Question Logic



Depositor demographics

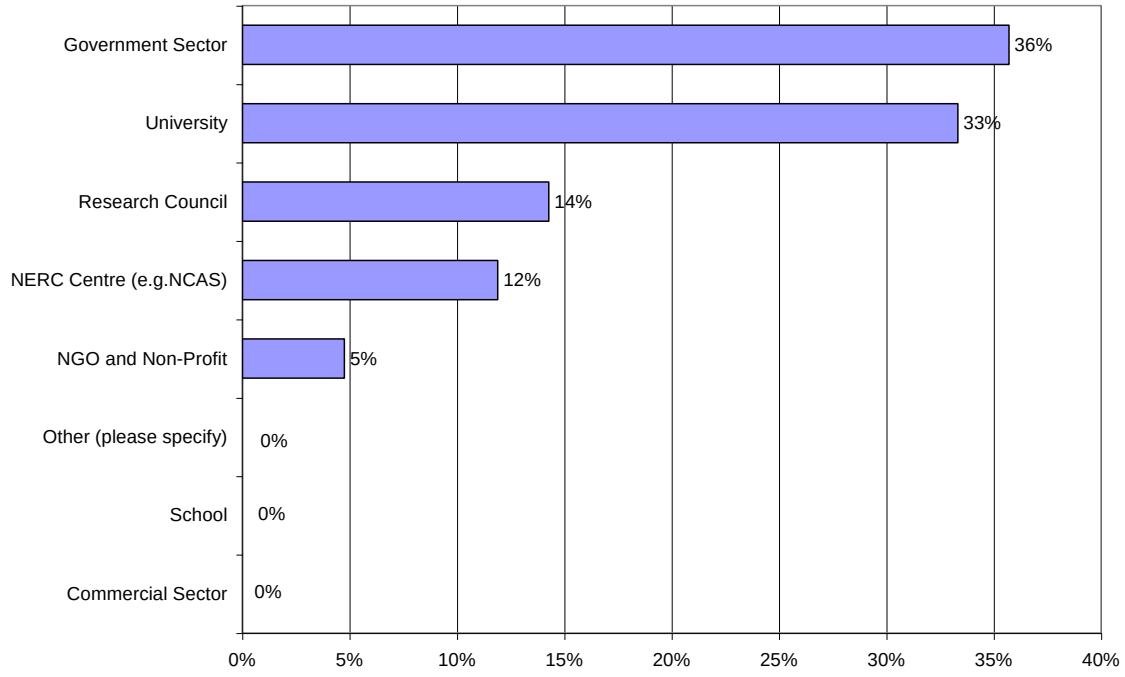
The BADC presently does not hold consistent way to identify its deposit community as demographics of this community have not historically been of interest to them. However a review of the results with BADC staff has indicated that these responses are considered a good reflection of depositors in general.

The first two questions in the depositor survey explored the affiliation and country location (currency) or respondents.

Q1: Main affiliation?

While there was a spread of affiliations reported, most depositor respondents were in the government sector (36%) or universities (33%).

Figure A22: Main affiliation of BADC depositors (N=42)

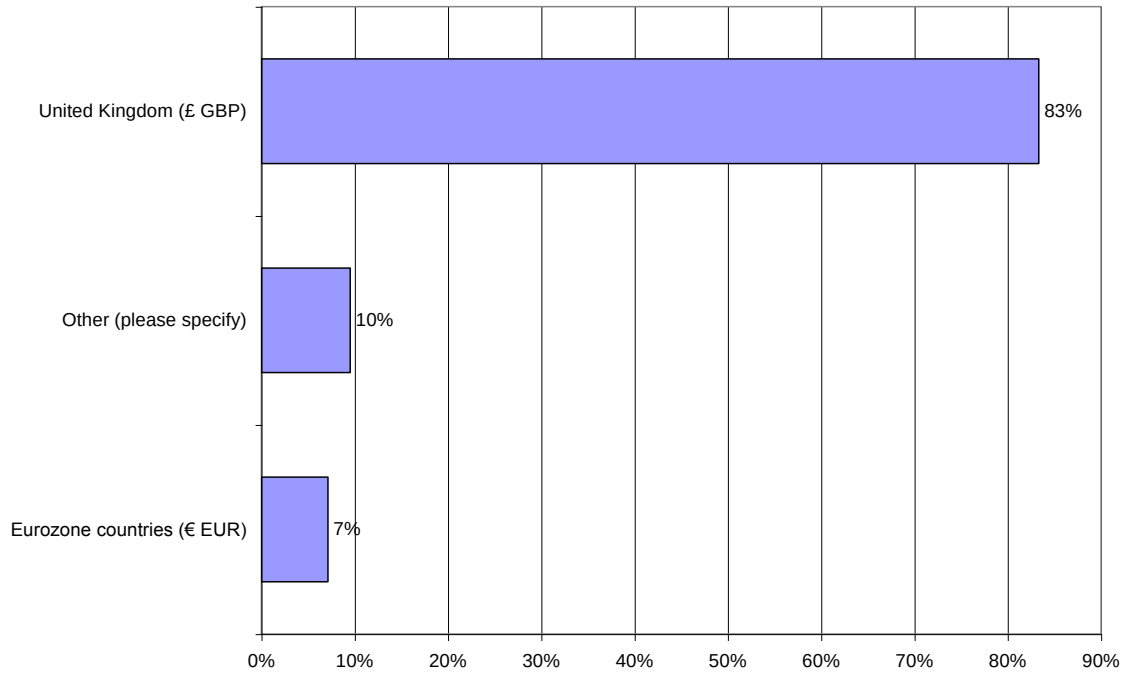


Source: BADC survey, Authors' analysis.

Q2: Country and currency?

The majority of depositor respondents were from the UK (83%), with 7% from Eurozone countries and 10% from a range of other countries.

Figure A23: Country and currency of BADC depositors (N=42)



Source: BADC survey, Authors' analysis.

Deposit activities and costs

The main section of the depositors' questionnaire explored the activities and costs involved in creating the data, preparing it for deposit, and depositing it with BADC. Questions were different for one-off deposits and those involving an initial deposit with subsequent updates.

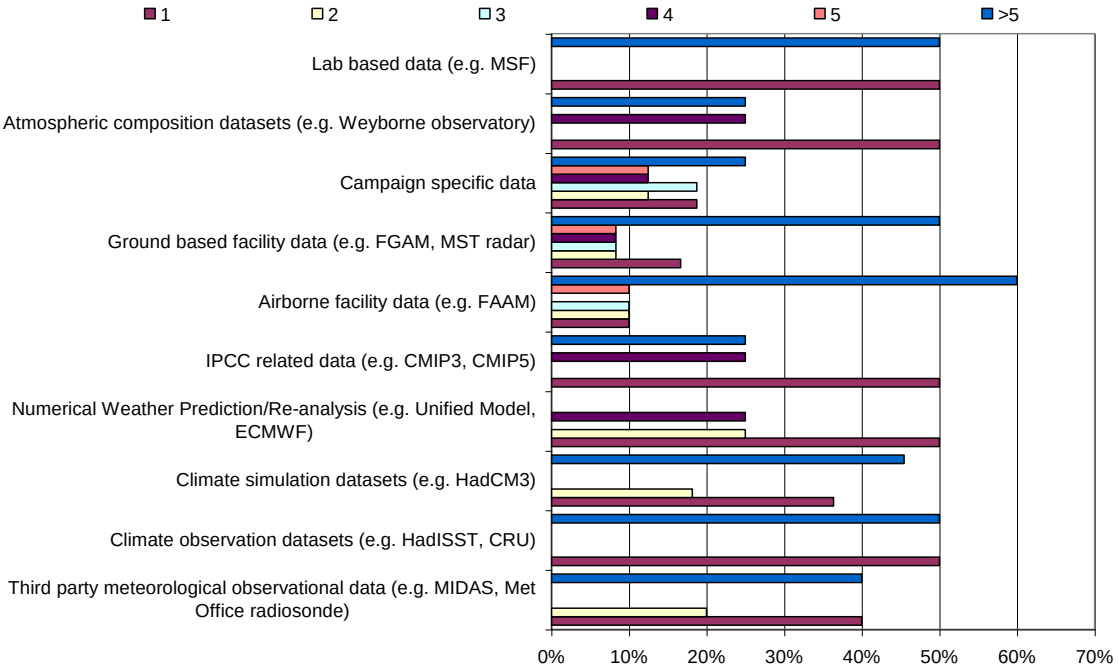
Q3: Approximately how many times did you deposit/update data to BADC over the last three years?

Because of the wide variety of data types and the nature of initial deposits and updates (e.g. multi-year automated deposits), this question proved difficult to interpret. Hence, these results should be treated with caution. Nevertheless, using information provided in comments to set a value and interpreting >5 to equal 6 where comments gave no further information, we are able to create an approximate frequency estimate.

Among depositor respondents, ground based facility data (e.g. FGAM, MST radar) was the most frequently deposited, accounting for 69% of reported deposits. Third party meteorological observational data (e.g. MIDAS, Met Office radiosonde) accounted for a further 31% of reported deposits.

Across data types there was a mean deposit frequency of 1,043 per annum over the last three years. However, the variation was high, with a standard deviation of 4,952 and a median deposit frequency of 2 per annum. This reflects the heterogeneous nature of deposits of different data types.

Figure A24: Frequency of deposit to BADC (N=41)

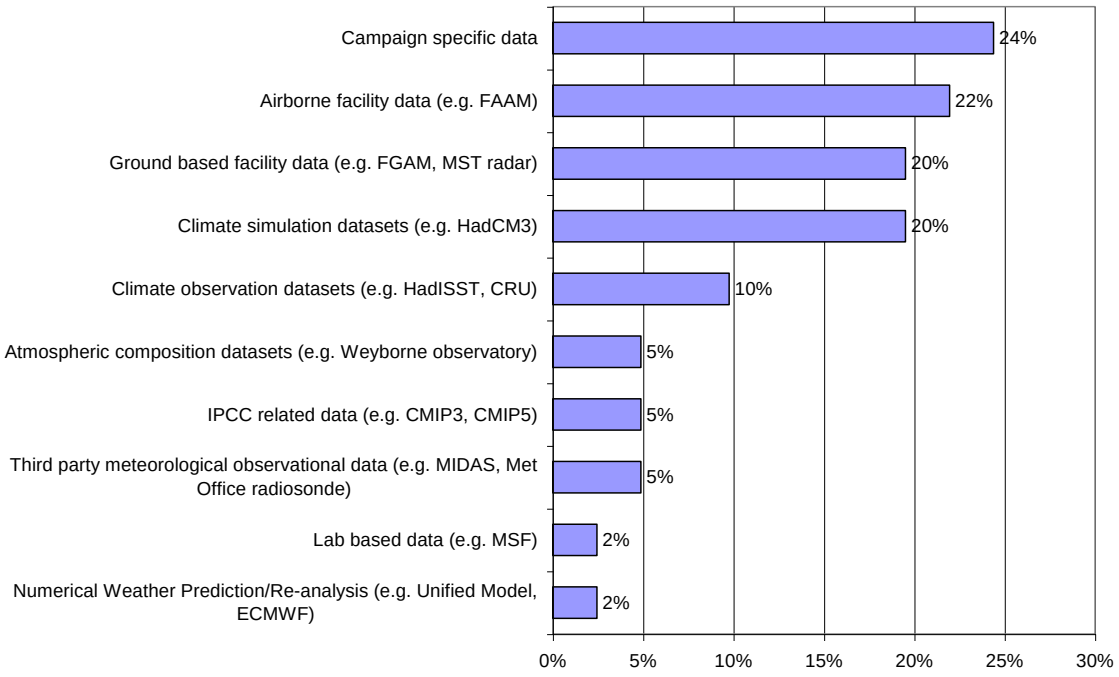


Source: BADC survey, Authors' analysis.

Q4: What was the last data type(s) you deposited?

To create a random response on deposits, we used a critical incident approach: asking what was the last data type respondents deposited. The most common data type responses were campaign specific data (24%), airborne facility data (e.g. FAAM) (22%), and ground based facility data (e.g. FGAM, MST radar) and climate simulation datasets (e.g. HadCM3) (20%).

Figure A25: Last data type deposited (N=41)

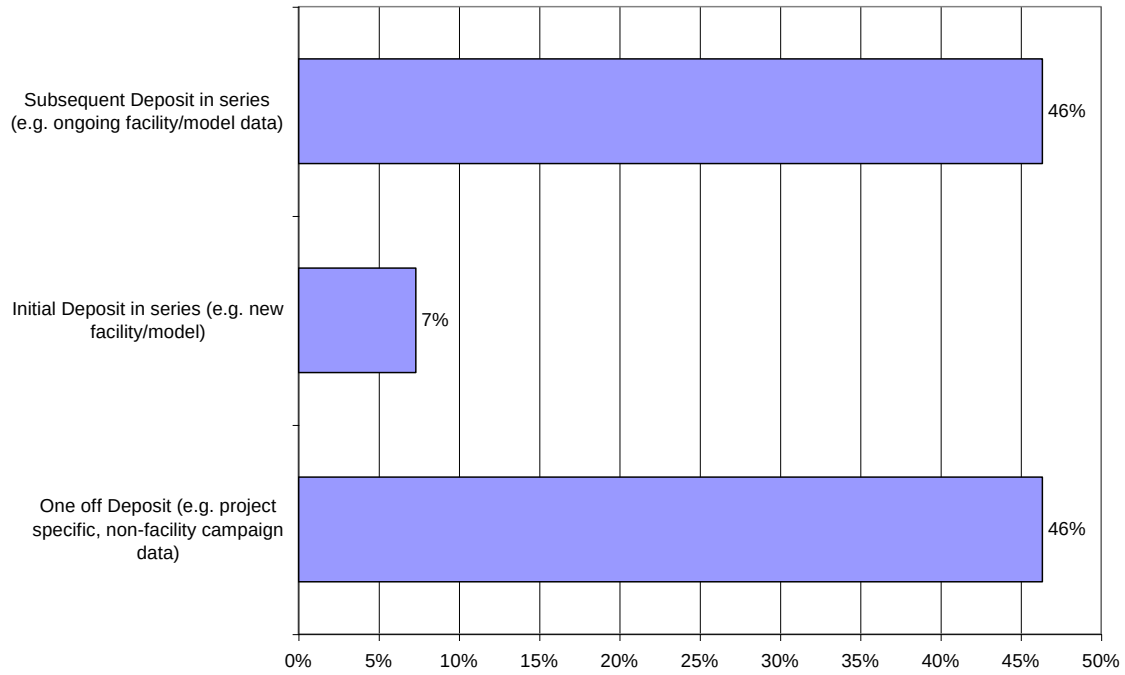


Source: BADC survey, Authors' analysis.

Q5: What type of deposit was it?

One-off deposits and subsequent deposits in a series accounted for 46% each, with the remaining 7% being initial deposits in a series. In the following questions, respondents were streamed into those making one-off deposits and those making deposits as a part of a series.

Figure A26: Type of deposit (N=41)



Source: BADC survey, Authors' analysis.

One-off deposits

Q6: Considering the last data you deposited, how long did it take to prepare them for deposit and submit them to BADC?

Categories of months, days and hours were converted to hours on the basis of 7.5 hours per working day and 18.3 working days per month (220 per year). On that basis, one-off depositor respondents reported an average data preparation and deposit time of 73 hours, with a reported maximum of 412 hours and a minimum of 3 hours (median 23 hours) (N=18).

Translating that time at average academic salaries plus non-wage labour costs of £37 per hour (See Box A1), suggests a mean data preparation and deposit cost of £2,713 (median £838) for those making one-off deposits. Taking frequency of deposit into account, this is equivalent to a mean cost of £4,209 per annum.

Series deposits

Q7: How much did it cost to set up the deposit mechanism for the series?

Depositors making initial or subsequent deposits in a series, were asked about the set-up and annual operating costs of any automated or semi-automated deposit processes they use.

For the series set-up, average reported staff costs were £74,588 (N=11) and average equipment costs £6,778 (N=9).

Q8: What are the annual (subsequent) costs of the series deposit?
For ongoing operation of the series deposit, average reported staff costs were £23,424 per annum (N=11) and average equipment costs £2,035 per annum (N=10).

These figures suggest mean additional costs of around £30,167 per annum - amortising the deposit system costs over the expected life of the systems (mean 5 years) (N=17).

Q9: Do you have any comments or important factors to note about your answer to the previous question?

Some 13 depositors provided free text comments to explain their responses to Q8.

Selected Comments

'The datasets are downloaded only once upon their initial official release. There were no further downloads unless there is a newer version (or amendments) of the dataset.'

'WHEN NEEDED. COULD BE DAILY, MONTHLY or YEARLY.'

'Depends on the project - when doing a project involving the data it's several times a month! In between, not at all. Overall, at least 10 times in a year.'

'We use the data, typically, for 3-5 year research projects, so download large data sets as needed, usually once or twice within the project period.'

'I work with data on JASMIN on a daily basis.'

Q10: How long do you expect the current automated system for the series deposit to operate before major system changes/completion (including time already elapsed)?
As noted, the mean expected operational life of deposit systems was 5 years (N=17).

Q11: Considering the LAST DATA you deposited, approximately how much time did it take to prepare them specifically for deposit and then to submit them to the BADC?
For those making series deposits, it took a mean of 71 hours to prepare the data and deposit it with BADC (median 8 hours) (N=15). This would imply a mean activity cost of £2,658 per series deposit.

Accounting for the reported frequency of deposit suggests mean total preparation and series deposit related costs of £35,597 per annum, with some series not involving set-up and deposit automation costs.

Data creation/collection costs

All respondents completing the survey were then asked questions regarding the original collection and creation costs of the data they most recently submitted to the BADC.

Q12: Considering the LAST DATA you deposited, can you estimate the TOTAL original financial cost of COLLECTING/CREATING them?
Some 20 respondents responded to question 12 giving a mean cost of £183,933 (median £27,500) for the collection/creation costs of the data they last deposited.

Q14: Was the creation cost of the data you last deposited typical for your deposits of that type of data?
As an additional check to support the critical incident responses, depositors were asked if the creation/collection costs were typical. Fifty per cent said the costs were average, 9% said they were less than average and 12% said more than average (N=34).

Q15: If the creation cost of your last data deposit was not average, can you estimate how much more or less than average it was?
Of the 7 indicating a variation from average these respondents suggested that the

creation/collection costs they reported were an average 89% of what they estimated would be typical. Such an adjustment would suggest mean data creation/collection costs of £178,933 per dataset deposited.

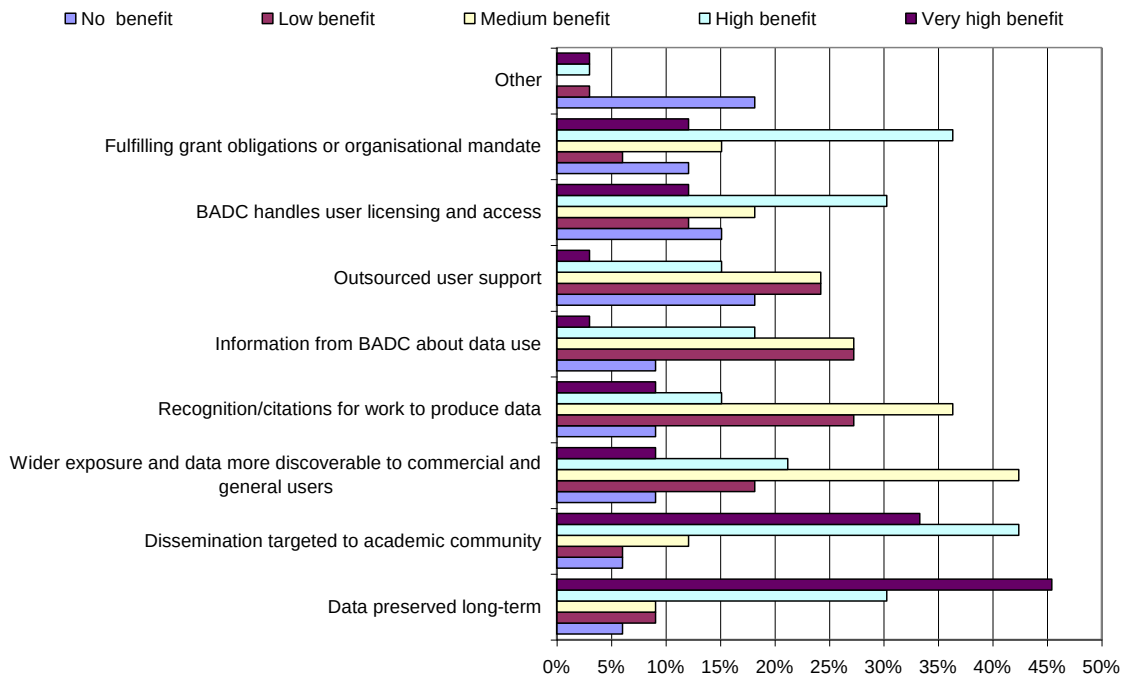
Benefits of depositing data with BADC

The final two depositor survey questions explored the nature and extent of benefits derived by depositors, in qualitative terms.

Q16: Please select the level of perceived benefit to YOU as a result of your depositing/providing data to the BADC, for each of the following possible outcomes?

Depositors cited that the data are preserved for the long-term and dissemination being targeted to the academic community as the most beneficial aspects of depositing data with BADC, both rated as a high or very high benefit by ~76% of respondents. These were followed by fulfilling grant obligations or organisational mandate (48%) and the fact that BADC handles user licensing and access (42%) as other important benefits (N=33).

Figure A27: Benefits of depositing data with BADC (N=33)



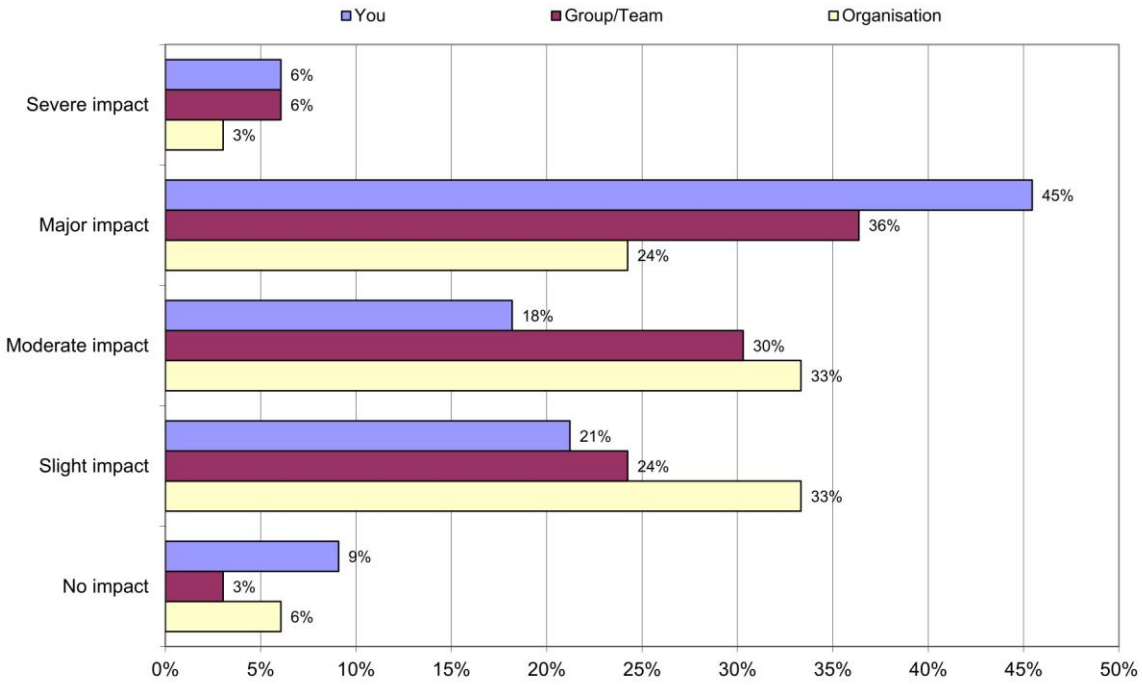
Source: BADC survey, Authors' analysis.

Q17: What impact would it have on you, your group or organisation if you could not deposit/provide data to the BADC?

The final question explored how being unable to deposit data with BADC would impact respondents, their work groups and organisations. Fifty-two per cent suggested that being unable to deposit data with BADC would have a major or severe impact on them personally, 42% said it would have a major

or severe impact on their work group, while 27% said it would have a major or severe impact on their organisation (N=33).

Figure A28: Impact of not being able to deposit data with BADC (N=33)

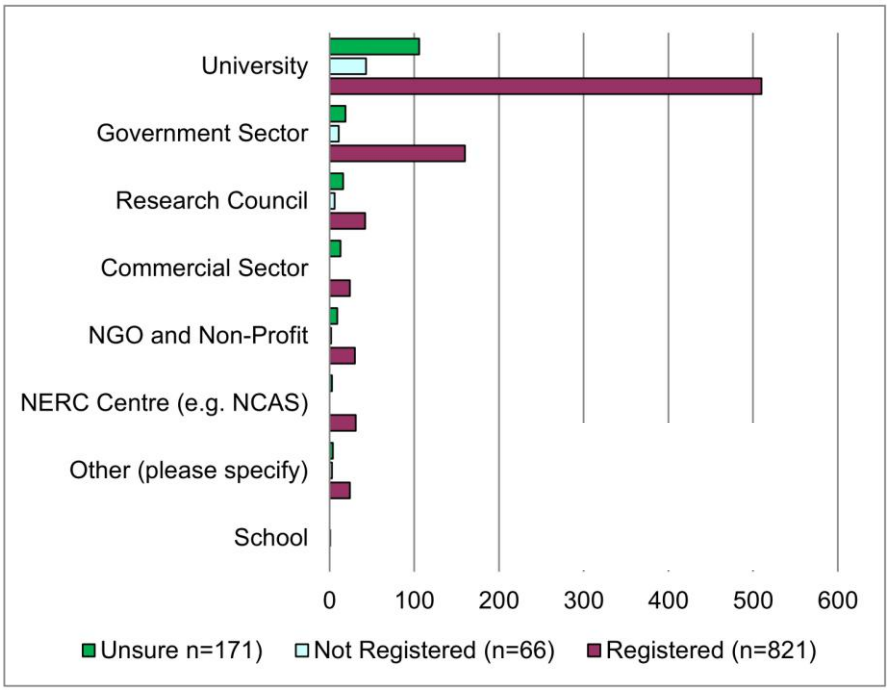


Source: BADC survey, Authors' analysis.

Appendix 2: Registered Users, Non-Registered Users and "Unsure" in the User Survey

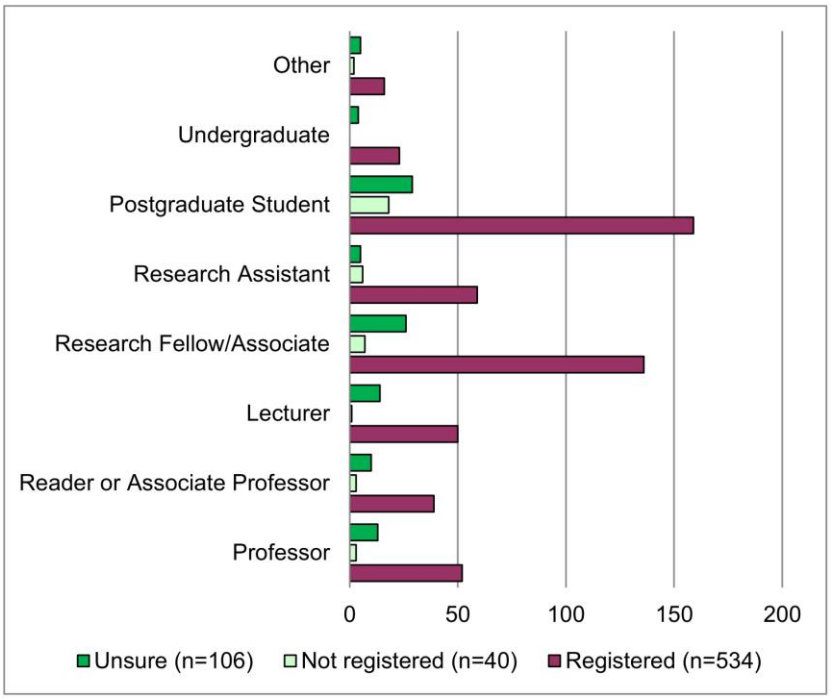
Appendix 2 presents additional data cuts by registered users, non-registered users and "unsure" for respondents to Questions 1-4 of the user survey. The 2013 survey allows some additional profiling of the non-registered active users that was not possible in previous surveys. Although not within the formal aims of the study, it was recognised that the online surveys of users and depositors we proposed for this study provided an opportunity in 2013 to complete another tri-annual customer survey in an existing longitudinal series, and could supplement previous BADC user surveys completed in 2007 and 2010 and existing BADC statistics. Note as discussed in the main body of the report, the percentage of respondents in the 2013 survey who were unsure of their registration status was a relatively high 16%.

Figure A29: Number of responses to User Survey Question 1 - Main affiliation of BADC users: sub-divided by registration status



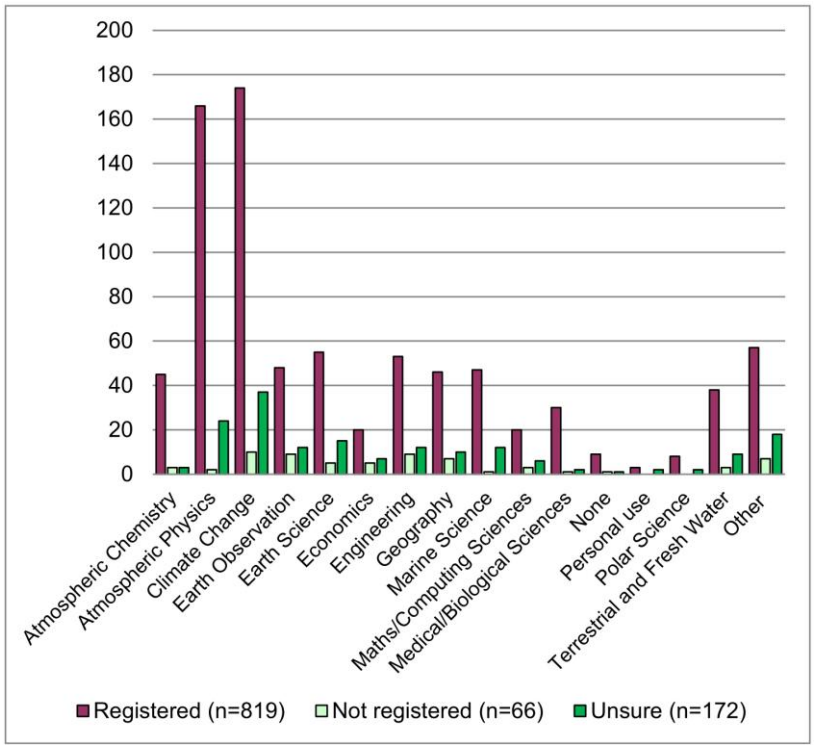
Source: BADC survey, Authors' analysis.

Figure A30: Number of responses to User Survey Question 2 - BADC academic users by main role within the university: sub-divided by registration status



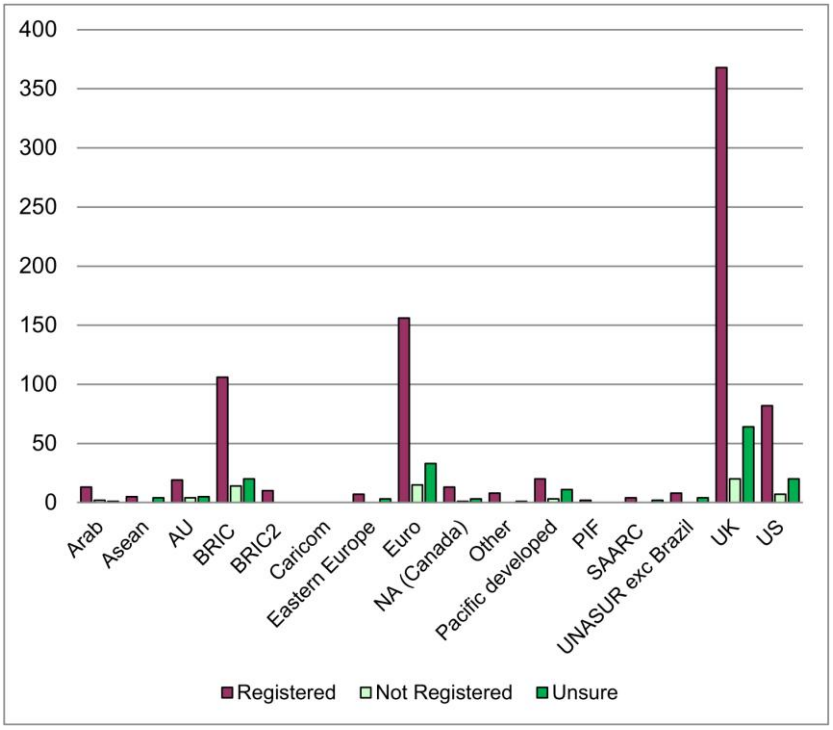
Source: BADC survey, Authors' analysis.

Figure A31: Number of responses to User Survey Question 3 - Main subject domain of BADC users: sub-divided by registration status



Source: BADC survey, Authors' analysis.

Figure A32: Number of responses to User Survey Question 4 Country and currency of BADC users: sub-divided by BADC economic areas and registration status



Source: BADC survey, Authors' analysis.