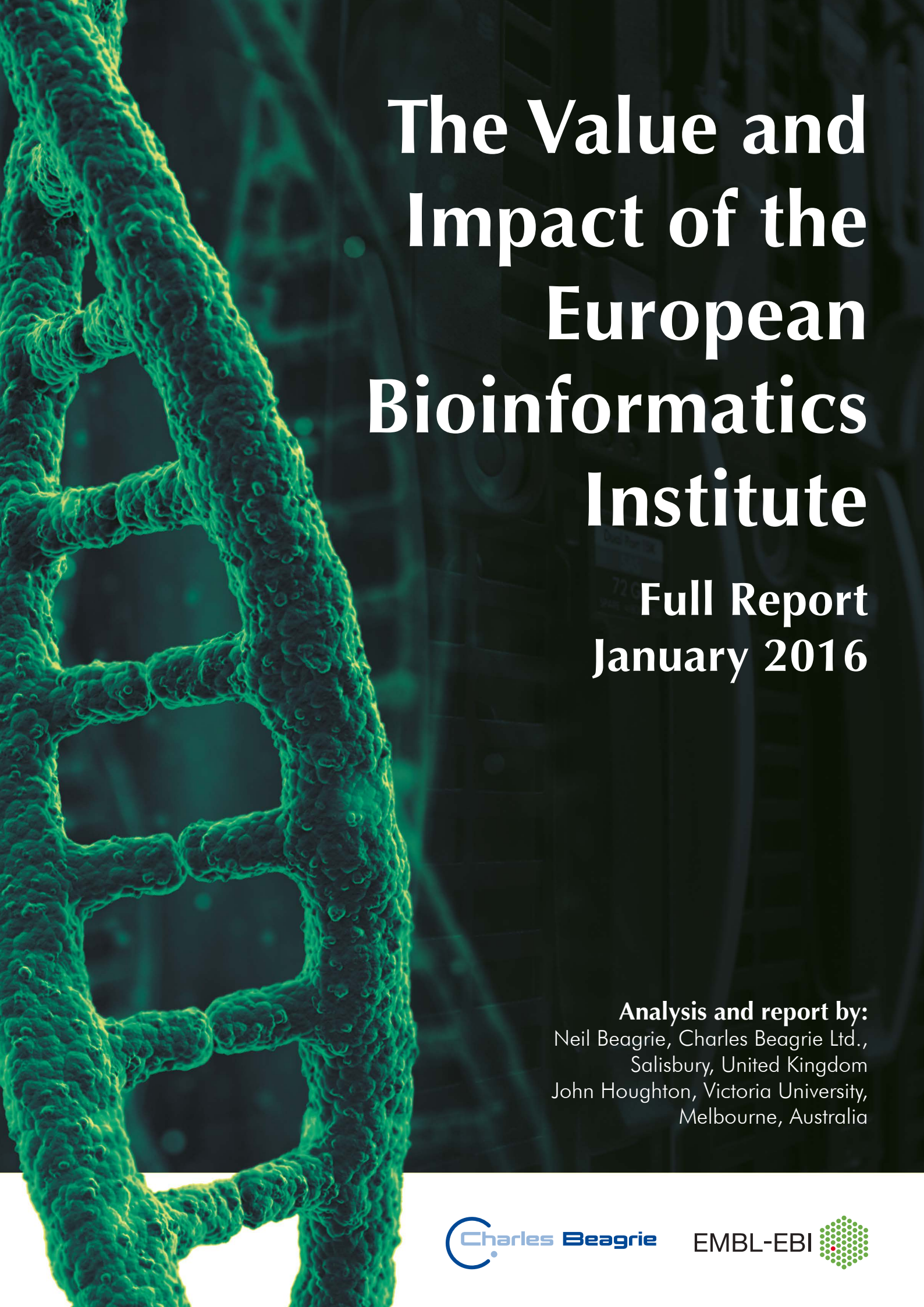


The Value and Impact of the European Bioinformatics Institute

Full Report
January 2016

Analysis and report by:
Neil Beagrie, Charles Beagrie Ltd., Salisbury, United Kingdom
John Houghton, Victoria University, Melbourne, Australia

Charles Beagrie

EMBL-EBI

About this Publication

The artwork in this report was taken from the EMBL-EBI Annual Scientific Report 2014. Cover image credit: Spencer Phillips, EMBL-EBI.

This report and a separate version of the executive summary are available online in printable format, at http://www.beagrie.com/EBI-impact-report.pdf and http://www.beagrie.com/EBI-impact-summary.pdf respectively.

Contents

The Authors

Neil Beagrie is director of consultancy at Charles Beagrie Ltd, an independent consultancy company specialising in the digital archive, library, science and research sectors. He has extensive experience in cost-benefit analysis, research data management and digital preservation. Neil has been the lead consultant or principal investigator for a wide range of research studies including the Keeping Research Data Safe (KRDS) projects, which investigated the costs and benefits of digital research data curated by UK universities. He is a Fellow of the RSA and was awarded the 2014 SMPTE Archival Technology Medal for his long-term contributions to digital preservation.

John Houghton is a Professorial Fellow at Victoria University's Centre for Strategic Economic Studies (CSES) and an Associate at Charles Beagrie Ltd. He is an economist specialising in applying economic assessment techniques to information technology policy, science and technology policy and in exploring the value and economic implications of open access scholarly publishing and open data. A number of John's studies have also focussed on the economics of research data services. In 1998, John was awarded a National Australia Day Council, Australia Day Medal for his contribution to IT industry policy development.

Neil and John have co-authored a series of four major studies on the value and impact of large research data services in the UK.

Acknowledgements

We would like to thank: EMBL-EBI staff who contributed to the study particularly Mary Barlow for her support and input throughout, Rodrigo Lopez Serrano, Chuck Cook and James Knoop (intern) for their work on EMBL-EBI web statistics, and Jenny Cham for assistance with previous EMBL-EBI surveys; Charles Beagrie staff and associates Daphne Charles and Peter Williams who assisted with interview transcription and notes for Phase 1 and Phase 2 case studies; the EMBL-EBI users who kindly participated in user pilot testing of the user survey for the report; 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** (Victoria University, Melbourne)
January 2016

Executive Summary

Introduction

The European Bioinformatics Institute (EMBL- EBI), located on the Wellcome Genome Campus in Hinxton, UK, manages public life science data on a very large scale, making a rich resource of information freely available to the global life science community. EMBL-EBI is one of a handful of organisations in the world involved in global efforts to exchange information, set standards, develop new methods, and curate complex genome information.

We present here the results of a quantitative and qualitative study of the Institute, examining the value and impact of its work. Our focus is the economic impact and can be seen as complementary to traditional academic measures, such as citation counts.

This study was conducted as part of an on-going programme, led by EMBL-EBI, to develop a framework and evidence base for demonstrating how the Institute captures and curates the increasingly vast output of genome research and allows it to be easily located, understood, and applied.

Key Findings

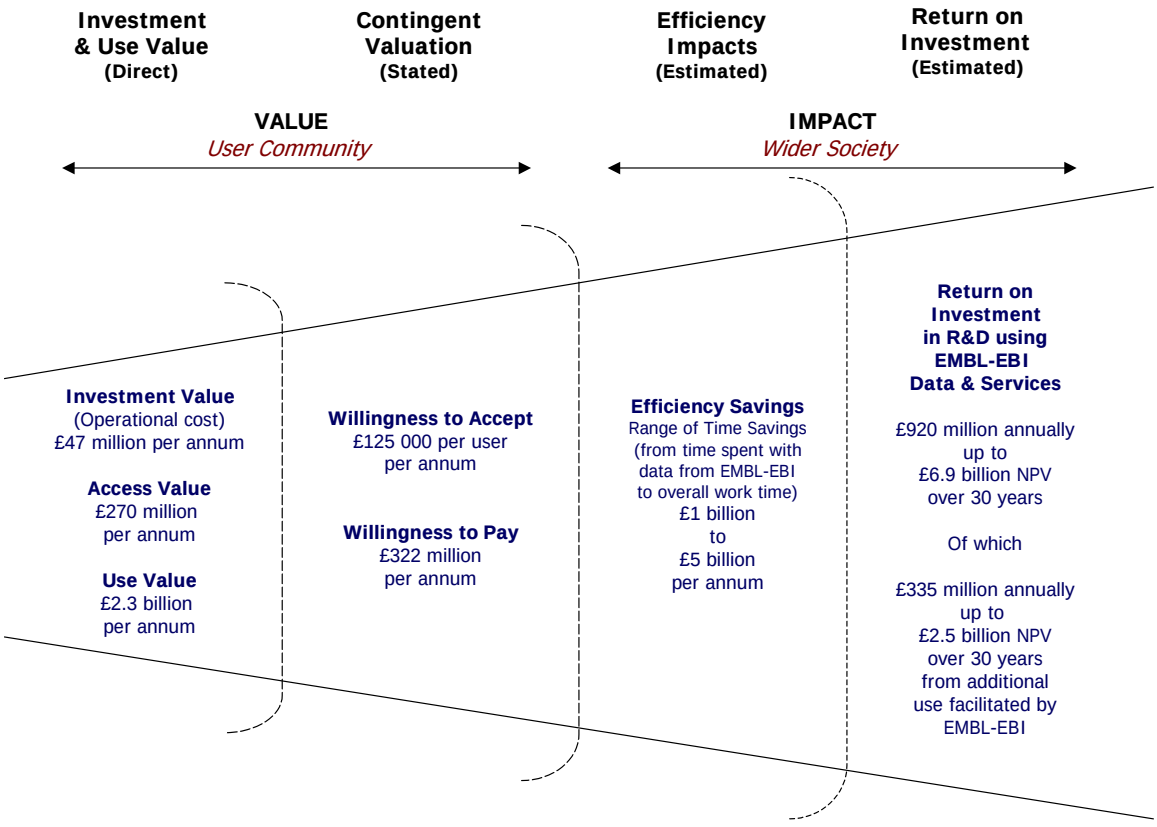
The qualitative and quantitative analyses reveal that EMBL-EBI services are utilised widely and valued highly by their user community.

The quantitative analysis, summarised in Figure S1 below, explores the value and impact of EMBL-EBI data and services, and shows:

- **Access (use) value:** The most direct measure of the value is the time and therefore costs users spend accessing EMBL-EBI data and services - an estimated £270 million during the year to May 2015.
- **Contingent valuation:** is an alternative approach to estimating what something is worth, measuring the value users place on a freely provided service, and is an estimated £322 million during the year to May 2015.
- These estimates give a sense of the minimum direct value of EMBL-EBI's data and services to its user community, and compare very favourably with the approximately £47 million annual operational expenditure, with a minimum direct value to users that is equivalent to around 6 times the direct operational cost.
- **Efficiency impacts**: Users reported that EMBL-EBI data and services made their research significantly more efficient. This benefit to users and their funders is estimated, at a minimum, to be worth £1 billion per annum worldwide - equivalent to more than 20 times the direct operational cost.
- **Return on Investment in R&D**: during the last year the use of EMBL-EBI services contributed to the wider realisation of future research impacts conservatively estimated to be worth some £920 million annually, or £6.9 billion over 30 years in net present value.
- A large number of survey respondents stated that they could neither have created/collected the last data they used themselves nor obtained it elsewhere. As a result it is estimated that

during the last year EMBL-EBI data and services underpinned future research impacts worth £335 million annually, or £2.5 billion over 30 years in net present value, that could not otherwise have been realised.

Figure S1: The value and impact of EMBL-EBI data and services



Source: Authors' analysis.

The qualitative analysis reveals a similar picture of the value and impact of EMBL-EBI data and services. The user survey, launched in May 2015, received 4 509 responses, providing an excellent foundation for analysis.

The survey found that users spend considerable time on research. The mean number of hours spent in research was 34 hours per week. More than half of this time (mean of 56 per cent) was spent working with data. One-fifth (20 per cent) of this was with data from EMBL-EBI. These numbers reflect the vital importance of genome data (and related services) and the intensity of its use among life sciences researchers.

More than half of all respondents (55 per cent) said that not having access to EMBL-EBI services and resources would have a "major" or "severe" impact on their work or study.

The quantitative assessment and the qualitative one independently revealed similar findings about the economic value and wider impact of EMBL-EBI data and services. This gives added credence to the findings.

Methodology

Quantifying the value and impact of EMBL-EBI is complicated due to many factors: its open and free provision of services, the collaborative nature of the work researchers undertake, the wide range of services it provides, and the diversity of communities it serves. Our approaches took account of the practical limitations inherent in collecting data through interview and survey techniques.

The quantitative economic approaches used included: estimates of access and use value, contingent valuation using stated preference techniques, an activity-costing approach to estimating the efficiency impacts of EMBL-EBI data and services, and a macro-economic approach that seeks to explore the impacts of EMBL-EBI use on returns to investment in research. These approaches allowed us to develop a picture, beginning with estimates of minimum direct values for the EMBL-EBI's user community and moving progressively toward approaches that measure wider social and economic value.

Knowledge of the EMBL-EBI user population numbers and levels of use underpin estimates of the value and impact of EMBL-EBI as a whole. EMBL-EBI data and services are almost entirely open resources. Users are not required to register and are not directly identified or recorded.

The various proxy measures of users and levels of use, such as Unique IP addresses or Web downloads, have significant limitations, which are amplified by the large number of individual EMBL-EBI services surveyed. In the absence of user registration, we estimated EMBL-EBI's user population by combining information from log data, a user survey, and external studies. The result was an estimate of 198 000 direct active users during the year to May 2015, accessing data 88 million times. This figure is felt to be conservative as it excludes what is known to be extensive secondary use. There are a number of observations in the conclusions which may help to validate and improve future knowledge of the user population and related research data metrics.

1 INTRODUCTION

1.1 BACKGROUND TO THE STUDY

Founded in 1974, the European Molecular Biology Laboratory (EMBL) now operates across five locations: Heidelbery, Hamburg, Grenoble, Monterotondo, and EMBL-EBI in Hinxton.

Established in 1992, the European Bioinformatics Institute (EBI) gathers, curates, and maintains a vast library of data resulting from life-science experiments, covering the full spectrum of molecular biology. EMBL-EBI supplies this data, without charge, to researchers.

While it is best known for its bioinformatics services, about 20 per cent of the institute is devoted to basic research. EMBL-EBI also has an extensive training programme that helps researchers in academia and industry make the most of the vast amount of data produced in life science experiments every day. This Industry Programme promotes pre-competitive collaboration and ensures that the institute's public offerings align with companies' needs. As of October 2015 EMBL-EBI employed a staff of 570 drawn from 57 nations.

The core databases of this non-profit organisation are produced in collaboration with other leading organisations, including the National Centre for Biotechnology Information in the US, The National Institute of Genetics in Japan, the SIB Swiss Institute of Bioinformatics, and the neighbouring Wellcome Trust Sanger Institute.

The largest part of EMBL-EBI funding comes from the governments of the 21 EMBL member states and two associate member states. Significant funding is also supplied by the European Commission, the US National Institutes of Health, The Wellcome Trust, Research Councils UK, and EMBL-EBI's industry partners. Led by the Biotechnology and Biological Science Research Council (BBSRC), the UK government has provided substantial capital funding through the Large Facilities Capital Fund (LFCF) to further develop EMBL-EBI's technical infrastructure and expand their activities within a new building.

EMBL-EBI is now applying a range of methods for exploring the value, benefits, and impacts of research data and services, including those developed by Charles Beagrie Limited and Prof John Houghton and applied in previous studies of the economic impact of research data centres (Beagrie and Houghton 2014).

Over the last four years Charles Beagrie Ltd and John Houghton have completed studies assessing the economic value and impact of the Archaeology Data Service (Beagrie and Houghton 2013a), the British Atmospheric Data Centre (Beagrie and Houghton 2013b), and the Economic and Social Research Data Service (Beagrie, Houghton et al 2012). These studies combined qualitative and quantitative methodologies to measure the value and impact of research data and associated services and tools.

These methods have broken ground in measuring the value and impact of major research data services. In a recent review of the international state of the art as regards the relationships between large-scale science facilities and innovation performance, this work was one of three studies highlighted to UK Department of Business, Innovation and Skills as being particularly good examples

of 'good practice' in the measurement of economic impacts (Technopolis 2013, p31-2).[1] The desire to apply leading best practice in assessing impact was a major reason for their selection and application by the EMBL-EBI.

EMBL-EBI and the authors have worked together in this study, applying these methods not only to evaluate the institute's work but also to develop potential synergies with components of the EMBL-EBI's own emerging evaluation programme, such as its annual user survey and a growing set of case studies.

1.2 AIMS AND SCOPE

This study employs a range of economic approaches, using baseline data gathered through desk research with EMBL-EBI internal reports and statistics, annual reports, online survey answers, and interviews with users and staff. This study includes exploration of the costs and cost savings involved in using EMBL-EBI data and services, its value to users, and its impacts on the wider commercial, healthcare, and research communities.

The main project outputs are:

- This study report on the value and impact of the EMBL-EBI's data and services; and
- Three case studies illustrating individual EMBL-EBI services direct impact that are included in Appendix 4 but also available separately as part of the EMBL-EBI impact case studies series.

1.3 LAYOUT OF THIS REPORT

Section 2 presents a brief description of the services provided by EMBL-EBI as well as a review of previous user surveys and impact studies. Section 3 describes the techniques used to collect the data necessary for analysis, and the approaches and methods used for measuring value and assessing economic impact.

The quantitative analysis of the value and impact of EMBL-EBI is presented in Section 4, followed by a brief presentation of the main qualitative findings from the user survey and case studies (Section 5). These are followed by some concluding remarks (Section 6).

Appendices include a full description of the results of the user survey conducted for this study (Appendix 1) and a summary of the free-text comments from the survey, which provide rich insight into how users see EMBL-EBI data and services (Appendix 2), a summary overview of the modified Solow Swan economic model used to estimate the potential contribution of EMBL-EBI to returns to R&D (Appendix 3) and the three case studies (Appendix 4).

[1] Other studies and their approaches and methodologies are explored in Appendices D, E, and F of the Technopolis report.

2 EMBL-EBI

2.1 A BRIEF DESCRIPTION OF EMBL-EBI

EMBL-EBI maintains a comprehensive range of freely available and up-to-date databases, which collectively cover the full range of molecular biology, from nucleotide sequences to full systems. It has a mandate to make its tools and infrastructure freely available to the global scientific community. EMBL-EBI is a central partner in global efforts to exchange information, set standards, develop new methods, and curate complex information.

Quantifying the impact of EMBL-EBI's work is complicated due to many factors: its open and free provision of services, the collaborative nature of the work it does, the range of services it provides, and the diversity of communities it serves. The ubiquitous requirement of reference biological data for many areas of research mean that acknowledgement can be rare, in the same way that basic services are rarely highlighted in the achievements of everyday life.

EMBL-EBI gathers biological information from both published literature and directly from experimental research. This data is then processed, incorporated into all relevant databases, classified, annotated, and aligned with existing data to become a value-added resource. These resources are provided as defined services. The most strategically important databases are accompanied by comprehensive tools and training. This allows scientists to share data, perform complex queries, and analyse results in different ways. Scientists can work locally by downloading EMBL-EBI data and software or use web services to access EMBL-EBI resources programmatically.

2.2 PREVIOUS IMPACT STUDIES AND USER SURVEYS OF EMBL-EBI

EMBL-EBI has an ongoing impact evaluation programme, which includes user surveys and externally conducted studies. It is measuring both the outcomes and the benefits through a mixture of assessment of the three vital effects of discovery, understanding and application. This includes use of counter-factual models. The evaluation programme is supplemented and illustrated by relevant case studies. The programme is based on statistics/user data gathered both internally (annual report) and through surveys (general user and training). In 2010, the EMBL-EBI launched its first user survey, which was repeated in 2013 and 2014. It will continue to conduct these surveys annually.

The EMBL-EBI was one of eight data centres included in a 2010-2011 study, conducted by the Research Information Network, of the use, value, and impact of research data centres. It was not included in an associated Research Information Network (RIN) user survey because an independent EMBL-EBI survey was already underway. Because of this EMBL-EBI was not a major focus of the RIN's resulting report. EMBL-EBI did, however, contribute *Combining databases to accelerate drug discovery*, one of the two very short case studies on new products and services in the report (Research Information Network 2011, pp 52-3).

More recently, as part of the EU-funded Evaluation of Research Infrastructures in Open innovation and research systems (EvaRIO) project set up to develop an evaluation model of European Research Infrastructures, EMBL-EBI impacts were explored through a case study of the methodology it has

developed. The EvaRIO case study[2] report showed that EMBL-EBI provides real and lasting benefit to the users of its services. It noted that EMBL-EBI had a positive influence on scientific working practices; it adapts quickly to emerging fields of science, having increased from five research fields (as outlined by web of science) in 1996 to 35 in 2012; and that the outputs of its research activities increase year on year with the publication rate of the institute growing on average 14 per cent each year since 2006.

As part of the EvaRIO study EMBL-EBI worked with a large industrial user and estimated that the worth of its activities dependant on EMBL-EBI services was in excess of 150 million Euros since joining the EMBL-EBI Industry programme and 20 million Euros in 2013. In addition it established that between 15 and 20 patent applications submitted by the company in the last five years could not have been done without data provided by EMBL-EBI. An SME company was also evaluated. It stated that one-third of the data used in its drug design services provided by the company was obtained from EMBL-EBI services.

[2] http://evario.u-strasbg.fr/uploads/autres-docs-BETA/EvaRIO_CAse_Study_EMBL-EBI.pdf

3 APPROACHES AND METHODS

This section presents a brief description of the approaches and methods used to measure the value and impact of science facilities, and then outlines the methods used for collecting data and assessing the economic value and impact of EMBL-EBI data and services. It is important to emphasise that the focus is on measuring value and impacts in economic terms.

3.1 A REVIEW OF METHODS

Assessing the value and impacts of research data and related services is a relatively new field and no single approach dominates. There is a growing body of literature on the value and impact of science facilities, but the emphasis tends to be on 'Big Science' facilities rather than on data repositories and related infrastructure and services. Methodologically, these studies fall into three main groups: those using various forms of Input-Output (IO) analysis; those featuring case studies and examples; and various forms of cost-benefit analysis, typically using activity costing and/or contingent valuation to underpin the analysis. These methods can be combined, with complementary use of qualitative and quantitative approaches highlighting the various dimensions and mechanisms through which value and impact can be determined.

In their review of approaches to measuring the effects and impacts of research infrastructures, EvaRIO (2013b, p14) noted:

> *Quite surprisingly, the use of standard methods such as cost-benefit analysis (in the strict sense of the term, i.e. using the whole apparatus of hypotheses and tools such as price estimation, consumer surplus, welfare, discount rate etc.) is apparently relatively rare. Similarly, the production function based approach, so widely used in the case of the estimation of the return on R&D expenditure at micro as well as meso (e.g. regional) or macro level is apparently relatively neglected.*

Our study aims, in part, to fill this void.

Input-Output analysis

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

'Big Science' facilities are often single-site facilities (e.g., a synchrotron), but may also be distributed facilities (e.g., the Square Kilometre Array), networked facilities (e.g., CERN's Large Hadron Collider Computing Grid), or virtual collections (e.g., the Consortium of European Social Science Data Archives). Input-Output analysis is best suited to single site facilities because they have a more directly measurable impact on the local and/or national economy. Nevertheless, it is possible to use an Input-Output approach for more distributed activities, albeit within a single country. Tripp and

Gueber (2011), for example, used an Input-Output approach to explore the economic impact of the Human Genome Project on the US economy. It is worth noting that this study suggested a very high multiplier of 141 (i.e., benefits of $141 for every $1 of US Federal Government funding).

Case studies

Another and often complementary approach involves case studies, which typically follow the innovation impacts on suppliers and users through surveys and/or through tracing the development of spin-off firms and the use of information derived from the science facilities. Such case studies are widely used in the evaluation of research facilities and activities, and can focus on the scientific, economic, and/or wider social impacts. Among studies of larger facilities, those of CERN have reported the value of supplier contracts and the ways in which these have facilitated the development of new products or processes, and NASA's Spin-off Database reports on the number and revenue of spin-off firms emerging from the space agency's research work (Technopolis 2013, p47).

While case studies provide concrete examples and often highlight the mechanisms through which impacts can be realised, they are limited because it is not possible to scale up a case study to estimate overall impacts. Consequently, case studies are often combined with broader economic estimates and/or formal frameworks for analysis.

Mixed method approaches to cost-benefit analysis

Among previous studies adopting a more formal framework are a series of projects named Keeping Research Data Safe (KRDS), undertaken by Charles Beagrie Ltd. The initial KRDS study investigated the medium- to long-term costs to UK higher education institutions of the preservation of research data, and provided a brief overview of the potential benefits from such preservation (Beagrie et al. 2008).

A second phase project (KRDS2) further developed the previous, activity-based cost model and developed a benefits framework, illustrated with two benefit case studies from the National Crystallography Service at Southampton University and the UK Data Archive at the University of Essex. The study found that research data curation and sharing can bring significant benefits to current researchers in the short-term as well as long-term benefits for future research (Beagrie et al. 2010).

Fry et al. (2008) also sought to identify the benefits arising from the curation and sharing of research data. Their study also used a mixed-method approach, including qualitative case studies to inform the development of a model on which to build a business case for data sharing in UK Higher Education. This was based on extensions of the research data preservation cost model proposed by Beagrie et al. (2008), to allow estimation of the benefit/cost to users depositing or accessing data.

Based on the work of co-authors John Houghton and Bruce Rasmussen of Victoria University (Australia), the report presented a simple example of cost-benefit analysis applicable to an individual dataset or repository, based on costs and potential cost savings. The approach was then extended to explore the more diffuse benefits of data curation and sharing at the institutional and disciplinary levels. Unfortunately, due to limited data availability, the study provided a framework for analysis without presenting a detailed analysis of the case studies.

Research infrastructure metrics

Most research data infrastructure and services are almost entirely open resources requiring no registration or direct identification and recording of the user population. Alternative metrics (altmetrics) for users and levels of use, such as Unique IP addresses or Web downloads, are often measured but are known to have significant limitations. The synthesis of our previous studies (Beagrie and Houghton 2014) noted the variability of research infrastructure metrics and the implications for economic analysis, and made the following recommendation:

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

Recently Jisc reported ongoing work underway in a NISO[3] altmetrics working group and its implications for research data metrics, including the potential use of COUNTER[4]. This work appears potentially relevant to the EMBL-EBI impact evaluation framework and a watching brief on its progress or participation should be considered.

3.2 METHODS USED TO COLLECT DATA

Desk research

Desk research included analysis of: existing evaluation literature; existing data from Keeping Research Data Safe and other cost benefit studies of research infrastructures; web pages and EMBL-EBI service publications; direct interviews with EMBL-EBI users; and existing EMBL-EBI management and internal data, such as access statistics, previous user survey results, and internal operational and financial reports.

Information was compiled by EMBL-EBI on web and programmatic accesses and unique hosts recorded over the previous two years for the 39 EMBL-EBI services included in the user survey. These are discussed further in section 4.1.

Interviews and case studies

Twenty-nine interviewees were selected to provide a cross-section of staff from key EMBL-EBI services and functions and their external users. In total, 29 people (17 EMBL-EBI staff and 12 external users) were interviewed via 22 individual or group interviews. A semi-structured interview was conducted using a pre-defined questionnaire.

[4] http://researchdata.jiscinvolve.org/wp/2015/06/18/a-standard-for-research-data-metrics/

Three impact case studies were also undertaken , published in Appendix 4, as part of the EMBL-EBI impact case study series . All are intended to be concise (maximum 2 pages) and to illustrate pathways to impact. Their main findings and the implications for this study are summarised briefly in Section 5.4.

A user survey

An online survey was conducted aimed at measuring the impact of EMBL-EBI services for its users. The survey questionnaire was developed iteratively by the project team, with external review and input from EMBL-EBI staff, and included observed pilot testing by external users. Given the nature of some of the economic valuation approaches being explored, the range and number (39) of EMBL-EBI services, along with the affiliations, roles and seniority of the survey populations, substantial effort was needed to design a questionnaire suitable for an online survey.

Significant effort was put into reducing the amount of time it would take recipients to complete the questionnaire, and five Raspberry Pi 2 with Pibow Cases were offered in a draw as an incentive for participation. As a result, the survey enjoyed a high response rate and good completion rates, especially given the topics and number of non-mandatory questions.

The questionnaire used a range of standard survey approaches, including the use of "critical instances", such as the last data accessed/downloaded (for users). A number of questions sought specific information on: 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 were interpreted carefully, in the context of open-ended text comments in the survey and other findings from the interviews as well as desk research, to ensure that protest and outlier answers were excluded from the economic analysis or were included with suitable caveats. These quantitative questions were supplemented by qualitative questions asking for views on the importance and impact of EMBL-EBI for users, to ensure that the quantitative and qualitative findings were in accord. Full details of the survey and responses are available in Appendix 1.

3.3 METHODS USED TO MEASURE VALUE AND IMPACT

Building on the experience of previous collaborative studies, a number of approaches to exploring the value and impact of EMBL-EBI data and services were pursued in parallel. In doing so, quantitative and qualitative approaches are combined, with an emphasis on the former.

Quantitative approaches

The quantitative approaches used include: estimates of investment and use value, contingent valuation using stated preference techniques, an activity-costing approach to estimating the efficiency impacts of EMBL-EBI data and services, and a macro-economic approach that seeks to explore the impacts of EMBL-EBI use on returns to investment in research. Thus we begin with approaches that can be seen as estimating minimum direct values for the user community and move progressively toward approaches that can be seen as measuring the wider value and impacts for the economy and society (Figure 1).

In selecting these approaches, the practical limitations of collecting the necessary data through interview and survey techniques have been taken into account with commonality of data sought where possible (i.e., the same data can be used to inform more than one of the approaches).

Investment and use value

The most direct indicators of value are investment value (i.e., the amount of time and money spent on the production and delivery of the good or service) and use value (i.e., the amount of time and money spent by users on obtaining and using the good or service). Measures of the investment made by users in access and use suggest the minimum amount that the good or service is worth to them. Both investment and use value can be established from user interviews and surveys, through questions about the time and costs involved in the discovery, access, and use of EMBL-EBI data and services.

Contingent valuation

Contingent valuation involves the assignment of monetary values to non-market goods and services based on preferences (i.e., Preference Theory). 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 (DTLR 2002).

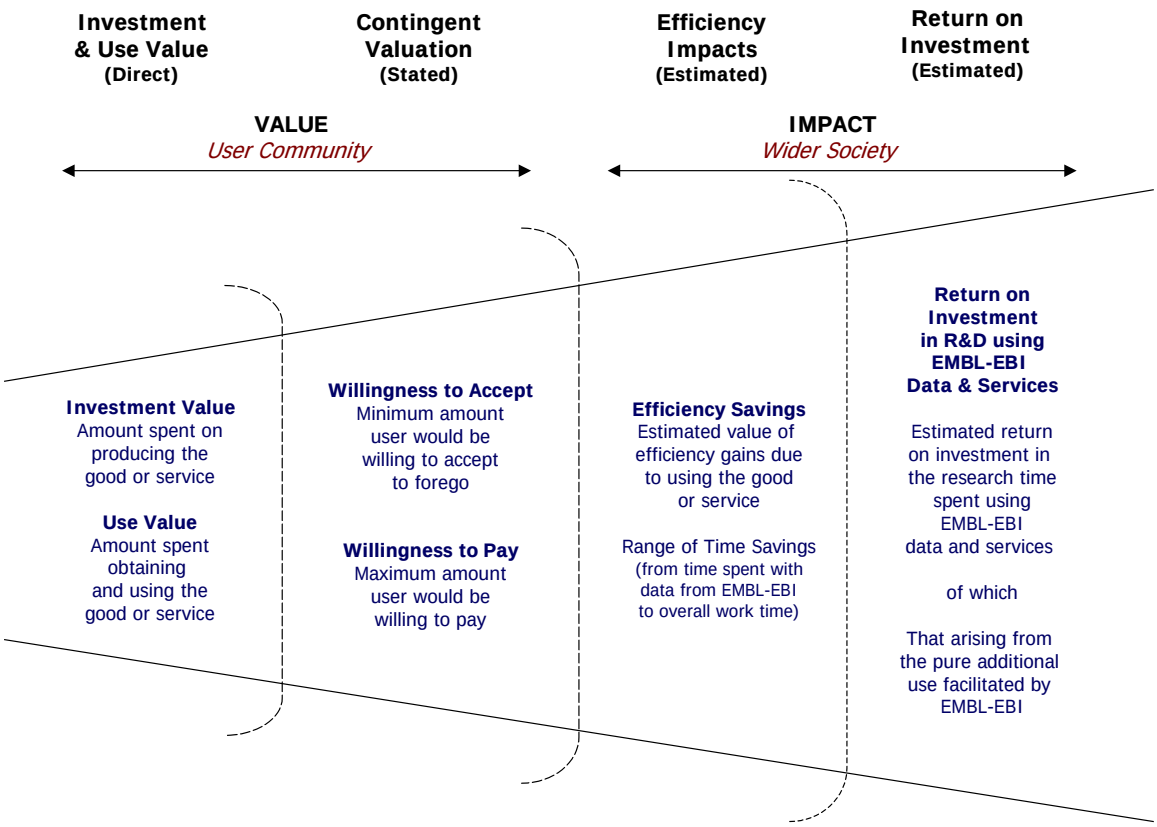
Preferences are revealed by what an individual is willing to pay for a good or service and/or by the amount of time and other resources spent obtaining the preferred good or service. Where preferences are not revealed in the market, individuals can be asked what they would be willing to pay for, or to accept in return for being without, the good or service in a hypothetical market situation (i.e., stated preference). 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 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, while willingness to accept measures the demand curve without a budgetary constraint (British Library 2004).

Efficiency impacts

Wider benefits and impacts can be explored by looking at the efficiency gains enjoyed by users and assigning an economic value to them, such as the value of time savings (productivity), and the avoidance of costs for users that would otherwise be involved in the creation/collection of the data for themselves or obtaining it elsewhere. For this we combine user survey questions about perceived efficiency impacts with activity costing.

Figure 1: Methods used for exploring the value and impacts of EMBL-EBI



Source: Authors' analysis

Return on investment

A sense of the scale of the value and impacts of EMBL-EBI data and services can be derived from an exploration of the potential return on investment in the research time spent using EMBL-EBI data and services, using a modified Solow-Swan model outlined in Appendix 3 (Houghton and Sheehan 2009, Houghton et al. 2009). A subset of this value will be the return from the pure additional use of the data facilitated by EMBL-EBI data and services (i.e., that by users who could neither obtain the data elsewhere nor create/collect it for themselves). As these impacts are recurring through the useful life of the data, it is necessary to use a simple Perpetual Inventory Method (Dey-Chowdhury 2008) to estimate the overall value of the impacts over time (see below).

Box 1: What value is and is not being captured?

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

Source: Authors' analysis

Qualitative approaches

Qualitative approaches included analysis of desk research, interview summaries, and user survey responses. Likert scaling was used in selected questions of the User Survey and word clouds composed of key themes that emerged from comments sections on the questionnaire were used for presentation in the report. The case studies used "pathways to impact" to explore and illustrate the impact that research based on EMBL-EBI data and services has. EMBL-EBI, as part of its ongoing impact assessment, has developed a logic model which outlines frequently noted impact pathways.

Likert-type scaling

A Likert scale is a psychometric scale in which responses are scored along a range. It is the most widely used approach for scaling responses in survey research. Named after its inventor, psychologist Rensis Likert, the scale allows people to give varying degrees of positive or negative response to a question or statement. Thus the scale captures the intensity of feelings for a given item. We used a 5-point and a 6-point Likert-type scale in Questions 6 and 27, respectively, of the user survey (see Appendix 1).

Pathways to Impact

In recent years, the UK government has placed increasing emphasis on the need for evidence of economic and social returns on its investment in research. The UK Research Councils (RCUK) introduced "Pathways to Impact" guidelines to encourage researchers to think about what they might do to ensure their research makes a difference, including considering the activities and timescales to deliver societal and economic impact and the wider communities these outcomes impact. These guidelines influenced the approaches employed in the case studies conducted for this study.

4 QUANTITATIVE ANALYSIS OF THE VALUE AND IMPACT OF EMBL-EBI

When combined with operational data from EMBL-EBI, the user survey responses provide a foundation for a number of economic estimates of value and impact. Details of the survey responses can be found in Appendix 1.

4.1 DATA LIMITATIONS AND ESTIMATES

Key operational data include financial and technical elements, including the overall annual expenditure on the operational provision of EMBL-EBI data and services; user numbers; and use levels based on statistics, such as the total number of data requests made on, and unique IP addresses (hosts) accessing, the various data and services.

There are number of important limitations and caveats:

- *Expenditure*: It is difficult to apportion overall operational expenditure to any particular subset of activities, such as the hosting and delivery of the data and services covered in the user survey. However, approximately 20 per cent of EMBL-EBI core funding is spent on basic research, rather than data and services. There are also short-term collaborative projects and external funding for them.
- *Data requests:* The total number of requests to a service can only be seen as indicative of the number of user access/download events. This is because (1) not all requests are successfully completed (e.g., when an FTP fails due to network interruption and is restarted), and (2) a sometimes significant share of requests are automated and/or robot/spider activities that are not related (directly) to a researcher's access and use.
- *Unique hosts:* The number of unique hosts (unique IP addresses) visiting the EMBL-EBI's data services is used as an indicator of the number of users although there can be significant differences between unique host counts and users over time (Fomitchev 2010). For example, there may be a number of users behind a single IP address (e.g., a number of research users from a single university), and a single user may appear as multiple IP addresses (e.g., when accessing from work and from home, when using a range of connected devices, and when their ISP uses dynamic IP addressing). The unique hosts counts may also include automated and robot/spider activities.
- *Users and data requests*: A further issue is that it is quite common for people to curate a local copy of EMBL-EBI data for use by local users. In our user survey 12 per cent of respondents reported curating a version of EMBL-EBI resources or tools for local use within their organisation; 18 per cent reported curating a free public resource from many sources including EMBL-EBI; and 7 per cent reported curating a subscription resource from many sources including EMBL-EBI. While those downloading for curation will be captured in EMBL-EBI use statistics, none of the secondary users or uses will be counted. As such, the host and request counts recorded by EMBL-EBI logs are necessarily understating the full extent of use.

There are also limitations in terms of the completeness of the access request and unique host data available from EMBL-EBI:

- Some of the data services are, in whole or in part, hosted by external agencies, and some of those external hosts collect different and/or more limited use statistics.
- Some of the data services statistics are available only for previous years or for more limited time frames (e.g., Reactome and Ensembl statistics are available for just one month during 2013).
- Access statistics are collected and reported on a monthly basis. The number of unique hosts accessing EMBL-EBI across the year is estimated from monthly statistics. EMBL-EBI does not track unique hosts at the individual data service level. Consequently, estimates for the number of unique hosts visiting during the last year across the data services and months can only be indicative.

Operational expenditure

Taking EMBL-EBI average annual expenditure over the three years 2012 through 2014, subtracting expenditures relating to research activities, and converting from Euros to British Pounds at average annual exchange rates (OECD 2015), gives an estimated average annual operational expenditure on data and services of around £47 million per annum over the last three years.

User population

There are many difficulties involved in attempting to estimate the number of users. These, together with data limitations, force an attempt to triangulate towards a plausible and conservative estimate of the EMBL-EBI user population. Three adjustments have been used in this triangulation:

- EMBL-EBI log data report data requests from around 10.8 million unique hosts during the most recent year for which data can be compiled. It is widely accepted that there are limitations to such unique host counts, with sometimes substantial differences between the number of unique users and reported hosts. For example, Fomitchev (2010) reports the results of an analysis of web site traffic logs and argues that both unique IP address and unique cookie counts overestimate unique visitors by a constant factor that grows linearly with sampling time, resulting in 6 to 8 times unique visitor overestimation at the end of a 28 day sampling period, due to the dynamic nature of IP addresses, the number of devices used by individual users, and the multiple locations they use to access the Internet. Noting that EMBL-EBI log data are compiled monthly, a first adjustment is to divide the reported 10.8 million unique hosts by 6 to 8 (mid-range 7) to estimate the number of unique visitors.
- As noted, there may be a number of users behind a single IP address (e.g., a number of research users from a single university) and a single user may appear as multiple IP addresses (e.g., when accessing from both work and home and when using a range of connected devices). The balance of these will depend on the balance of institutional and home use. After data cleaning, the user survey received 4 185 responses from 3 622 unique IP addresses (i.e., 1.16 people per unique host), suggesting a second adjustment to account for the balance of multiple users behind unique hosts.

- The 10.8 million unique hosts reported in EMBL-EBI log data is unique month-to-month across the 12 months, but it is not also unique across the 39 services that EMBL-EBI provides. For example, if a host visited the same data service 10 times during the year it would be counted once in the 10.8 million, but if a host visited 10 different services once each during the year it would be counted 10 times. The user survey respondents reported accessing an average (median) of 9 services during the last twelve months, suggesting a second adjustment to account for the use of multiple services.

These three adjustments suggest that the 10.8 million unique hosts reported in EMBL-EBI log data might represent around 198 000 unique direct users across the 39 services during the last 12 months.

By way of confirmation we attempted to estimate the global population of life-sciences researchers from independent sources, realising that life-sciences itself is not a perfectly defined category. We noted two data elements:

- UNESCO (2010) reported just over 7.2 million full-time equivalent researchers worldwide circa 2007. Few countries report detailed statistics by field of research, Australia among them. The Australian Bureau of Statistics (ABS 2010) reported gross expenditure on R&D of almost AUD 28 billion during 2008-09, of which life-sciences (i.e., biological sciences, medical & health sciences, and agricultural & veterinary sciences) accounted for some AUD 6.4 billion, or 23 per cent.
- While publication rates vary significantly between research fields, we note that UNESCO (2010) reported a total of 986 009 scientific publications worldwide during 2008, of which some 207 418 or 21 per cent were in the biology and biomedical fields.

If these proportions were approximated worldwide and the ratios of staffing approximately matched funding and publication ratios, then there might be around 1.5 million to 1.7 million full-time equivalent life-sciences researchers worldwide who, at a maximum, would be *potential* users of EMBL-EBI data and services. These estimates relate to full-time equivalent and now somewhat dated researcher counts. Nevertheless, they suggest that our estimate of 198 000 EMBL-EBI users would be equivalent to around 13 per cent of the maximum possible potential worldwide life-sciences FTE researcher population (no doubt use would be higher in some regions and lower in others).[5]

These estimates are believed to be conservative for two reasons. First, because EMBL-EBI users are predominantly institutional users (i.e., from universities and firms) rather than predominantly home users), suggesting lower than average dynamic IP addresses. Second, because the estimated user number includes only direct users of EMBL-EBI data and services and does not include secondary users of curated data, which we know to be widespread as around 30 per cent of respondents to our user survey reported undertaking some form of curation for secondary use.

[5] This implies that our survey reached almost 14 per cent of the user population and we received responses from 2.3 per cent (or 17 per cent of those contacted).

Requests and access

EMBL-EBI log data report receiving around 5.5 billion data requests during the most recent year for which data can be compiled, equivalent to some 513 requests per unique host. The user survey recorded an estimated 547 accesses/downloads per user during the last 12 months, approximately 631 per unique host—with considerable uncertainty due to difficulties interpreting the numbers involved when users reported using EMBL-EBI data and services more than once a day. One would expect the survey respondents to be among the more frequent users, as they would have a greater motivation to respond to the survey and are more likely to be known to and contactable by EMBL-EBI. These ratios support adjustments for both differences between survey respondents and the overall user population and for differences between unique hosts and users.

Hence, for the purposes of estimation, we suggest that users may access and use EMBL-EBI data and services an average of 445 times a year (i.e., a total of around 88 million access and use events if there were 198 000 users).

4.2 WEIGHTING AND SCALING

It is common to weight survey data to better reflect the overall population (Crockett 2011). In this case, there are two main reasons that might encourage weighting the EMBL-EBI user survey responses. First, the use of convenience sampling[6] and discretionary responses suggest that the survey respondents cannot be assumed to be representative of the overall user population. Second, the scope and complexity of the data and services available from EMBL-EBI suggest that users will use different parts of the overall service, rather than the whole, and do so with different levels of intensity. Ideally, therefore, the process of weighting the survey data better reflects the nature and intensity of overall use of EMBL-EBI data and services.

The user survey included a number of elements that need to be treated differently:

- Questions about the number of hours spent working per week along with the share of that time spent working with data in general, and from EMBL-EBI in particular, cannot be scaled/weighted in any simple way due to the multiple dimensions upon which differences may depend. However, respondents were also asked what use patterns they thought typical of others in their field, and responses suggested that they believe they are typical of others. These responses show that the survey respondents believe themselves to be representative of the user population.
- Questions about the time and costs involved in accessing the data last used, obtaining it elsewhere, and collecting/creating it themselves are all critical incident questions. As such, the responses are randomised and do not require weighting.

[6] The subjects were selected because they are possible to recruit easily for the study from existing EMBL-EBI email contact lists.

- Questions about the efficiency impacts and contingent valuation relate to the intensity of use of EMBL-EBI data and services (e.g., what someone is willing to pay for the data services depends on the nature and intensity of their use of those data and services) and should be scaled/weighted according to the respondent's relative intensity of use compared with that of the overall user population.

There are, however, some data constraints that limit the extent to which weighting is practicable. As a result, the extent of weighting/scaling is limited, with un-weighted results presented and used in the analysis, and weighted results mentioned where applicable to give a sense of the possible range of estimation.

4.3 ESTIMATING THE VALUE AND IMPACT OF EMBL-EBI

Knowledge of the EMBL-EBI user population numbers and levels of use underpin estimates of value and impact, but lack of solid data has forced us to conservatively estimate both user numbers and levels of use. Consequently the following estimates of value and impact can only be interpreted as indicative. Nevertheless, within bounds, most scale in an approximately linear fashion. So, for example, if there are twice as many users, then the value estimates can be doubled (and vice-versa) to give an approximate sense of value.

Activity times are converted to costs by assigning each respondent to a salary group based on the UK Times Higher Education Salary Survey and information from the UK Department of Education for 2014-15, then scaling to include non-wage labour costs using a 30 per cent uplift, based on the HM Treasury Green Book method (Green Book 2011, p59). For students, we use the school leaver and graduate average salaries reported in the UK Complete University Guide, to reflect the opportunity cost of earnings forgone. Non-academic respondents are allocated to a comparable academic staff level and salary. Across the respondents, this resulted in an average costing of around £37 per hour, including both staff at all levels and students.[7]

It should be noted that the majority of user survey respondents (more than 80 per cent) were based outside the UK, and while many were in comparable developed countries, some were not. For developing country respondents, there is likely to have been an overestimation of actual costs. Hence the costings presented in the following analysis should be thought of as UK equivalent costs.[8]

Investment and use value

The most direct indicators of value are investment value (i.e., the amount of resources spent on the production and delivery 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 made by users accessing the data services suggest the minimum amount they are worth to them. For simplicity, both investment

[7] A simple confirmation of this is to divide UK Gross Expenditure on R&D (GERD) by the number of full-time equivalent R&D personnel, to get an approximate full cost per researcher. On standard working hours that is around £48 per hour. When the inclusion in our survey of students and teaching staff is taken into account, this suggests that £37 per hour is a reasonable estimate.

[8] Using the mean hourly cost for costing also assumes that the survey respondents' job levels (i.e. role within affiliation) are broadly representative of the overall user population.

and use values can be expressed as an annual cost in current prices and at current levels of activity, by focusing on a single year snapshot (circa 2014).

Investment value

Investment value includes annual operational spending, the costs that data creators and contributors face in collecting/creating the data, preparing data for deposition and making data contributions, as well as the costs of collaborators in annotating and adding value to the data, etc. As we are unable to source information on quantitative data collection/creation costs, data contributions to EMBL-EBI, or collaborative contributions, we are unable to estimate the investment value of EMBL-EBI data and services[9]. However, there remains the annual average £47 million operational expenditure (i.e., the estimated annual average expenditure over the last three years).

Use value

There are two ways to look at use value. First, at the lower-bound, the value of using EMBL-EBI data services is reflected in the time/cost of accessing and obtaining the data and services. Second, at the upper-bound, the value of using EMBL-EBI data and services is reflected in the time/cost of accessing, obtaining *and* using the data and services.

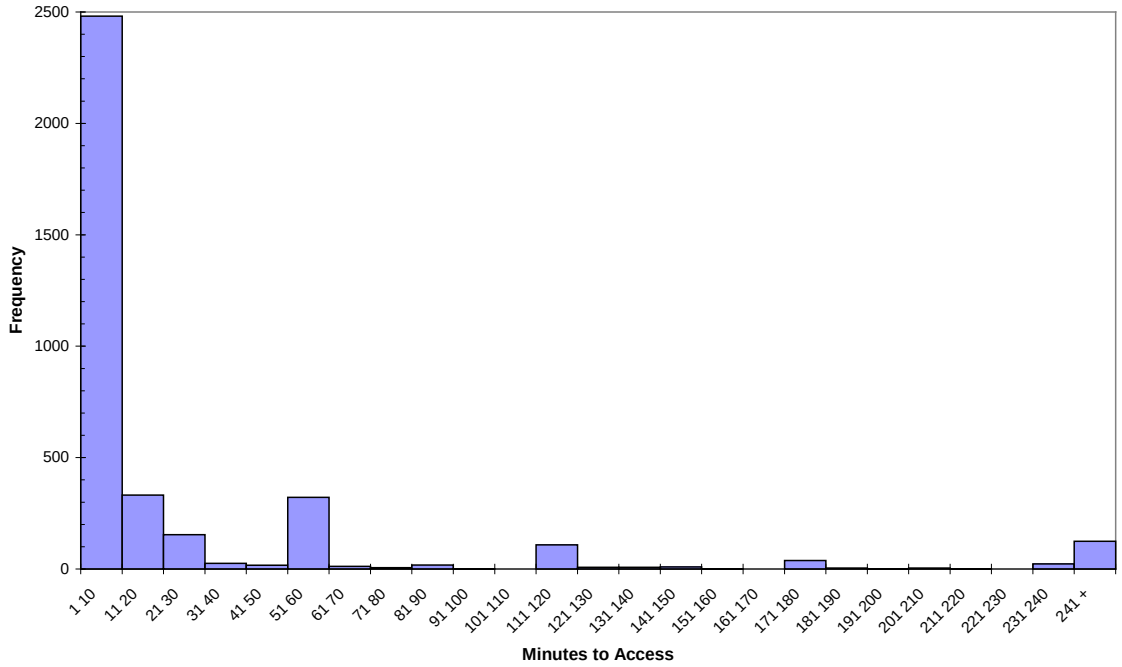
Access time/cost

Responses to the question about time to access and obtain the last data used varied widely, with a mean of 51 minutes reported.[10] However, there was a large number of low times reported and relatively few very long times reported, with the median reported being 5 minutes. This reflects the different methods of access. Many found access easy, with 506 respondents saying it took 1 minute or less, with comments such as: “it’s bookmarked”, “I know where to go”, “it takes no time at all”, and so on. At the other end of the scale, 51 respondents reported access times of 24 hours or more. Such cases were accompanied by comments about download speeds using FTP, difficulties setting up a programmatic access, and so on. Examining the distribution of reported access times suggests that the median would be a better indicator of central tendency (Figure 2). Hence, subsequent analysis is based on the median time to access and obtain the last data used.

[9] However the value of making both the narrative and the data from publicly funded research openly available is broadly recognised. For example, Elixir reports for the Protein Data Bank storing and making all structural data available for broad reuse costs less than 1 per cent of regenerating one year’s new depositions. https://www.elixir-europe.org/system/files/elixir_scientific_programme_1.pdf

[10] It is important to note that there are limitations to use value as a metric, and it can only be used as a snap-shot. However, one should not interpret it too simply. It is an expression of value, which does not change in direct proportion with the ease or difficulty of use. If it were easier (cost less time) to use EMBL-EBI, there would be more use; and if it cost more time to use EMBL-EBI, it would be used less, with the impact on (use) value unlikely to be large - depending on whether expenditure on EBI data and services was realising increasing or decreasing marginal returns. It will also be affected by the elasticities of demand and supply (e.g. driven by changes in expenditure on life-sciences research).

Figure 2: The distribution of reported access times–far from a normal distribution



Source: Authors' analysis

Simply multiplying the reported median time of last access by the mean hourly cost and the estimated number of data accesses during the last year suggests a total worldwide user access cost of around £270 million per annum (an average of around £1 675 per person across the survey respondents) at UK equivalent costs.

*((median time of last access * mean hourly cost) * estimated annual accesses)*
=
£270 million pa

Use time/cost

Use value is expressed by the amount of money and time people are willing to spend accessing and using a product or services. An estimate of the wider use value of EMBL-EBI data and services can be derived from user survey responses about the average hours spent on research each week and the share of that time spent with data, including that from EMBL-EBI. Converting the reported times spent with data from EMBL-EBI to pounds at the average hourly cost suggests an approximate worldwide use value of £2.3 billion per annum (an average of £11 625 per person across survey respondents) at UK equivalent costs.

*(((mean time with EMBL-EBI data per week * mean hourly cost) * weeks pa) * estimated users)*
=
£2.3 billion pa

Contingent valuation

The contingent value of a non-market good or service is the amount users are willing to pay for it and/or are willing to accept in return for giving it up. 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 is that the amount that users are willing to accept in return for giving up access is typically higher than the amount they would be willing to pay, primarily because the latter is constrained by what they can afford (e.g., by disposable income, limited research grants, etc.).

The method requires specific wording of the questions and an opportunity for open ended comments to enable analysis of the thinking behind responses and the identification of protest answers (DTLR 2002). Respondents' comments as to the rationale for their answers to these questions provide invaluable insights into their thinking about the value of such services. Among the reasons reported for being willing to accept only very high amounts in return for giving up access is the belief that the resource is invaluable, with respondents entering amounts in the millions of pounds. Another group of respondents thought through the implications of not having access, suggesting that they could not do their research without it, and putting in amounts equivalent to their annual or sometimes multi-year salary or research grants, with amounts ranging from around £100 000 to £1 million. Others did a range of "back of the envelope" calculations, such as the amount it would cost to obtain the data elsewhere or create/collect it themselves. The mean amount respondents reported being willing to pay for an annual subscription to access EMBL-EBI data and services was £1 628 (median £126), equivalent to £322 million per annum across the worldwide user population (at UK equivalent costs). This is equivalent to 6.8 times the EMBL-EBI annual operational cost.

*(mean willingness to pay * estimated users)*
=
£322 million pa

Of course, what a person is willing to pay for data services depends on the nature and extent of their use of those data services. Hence it is preferable to weight survey responses to reflect overall use/access patterns, as indicated by the relative frequency of accesses during the last year. Among user survey respondents there were an estimated 631 access events per unique host during the last year, compared to an average 513 requests per unique host for EMBL-EBI data services overall. Scaling the reported mean willingness to pay on that basis suggests an overall willingness to pay around £262 million per annum (at UK equivalent costs)—equivalent to some 5.5 times operational costs. However, due to difficulties in interpreting survey responses reporting a frequency of use of greater than once a day, we use the un-weighted willingness to pay in subsequent estimations.

Efficiency impacts

The value of the efficiency impacts of EMBL-EBI data and services among their user community can be estimated from questions about the time spent on research and the share of that time working with data, and estimates of the efficiency time savings experienced by survey respondents. To estimate the overall efficiency impacts from user survey responses, these mean time savings can be weighted to reflect the overall frequency of access.

Excluding the 18 responses reporting negative but un-quantified impacts,[11] estimated efficiency impacts could be worth as much as £26 000 on average per person per annum, or £5 billion per annum across the worldwide user population (at UK equivalent costs). However, it is possible that some respondents may have misinterpreted the question, thinking that the efficiency impact referred to time with data or time with data from EMBL-EBI, rather than overall working time. If that were the case for all respondents, then the efficiency impacts would still, at a minimum, be worth an average £5 380 per person per annum, or just over £1 billion per annum across the worldwide user population (at UK equivalent costs). It would nonetheless represent a benefit to users, and their funders, that is more than 20 times the direct operational cost.

*(((estimated users * mean hourly cost) * share of time with data) * efficiency impact)*
=
£1 billion and possibly up to £5 billion pa

Weighted to reflect the estimated overall frequency of use, these impacts are worth between £865 million and £4.2 billion per annum worldwide (at UK equivalent costs).

Additional use facilitated by EMBL-EBI

The user survey asked whether respondents could have obtained the data they last used elsewhere and, if not, whether they could have created/collected it themselves. Answers to these critical incidence questions allow us to estimate the value of the additional use facilitated by EMBL-EBI data services (i.e., pure additional use that could not otherwise have occurred).

Some 45 per cent of user survey respondents indicated that they could not have obtained the last data they used elsewhere, nor could they have created or collected it themselves. If this were true of all users and uses, then some 45 per cent of use is effectively additional use. A further 21 per cent may have saved data creation/collection costs (i.e., would otherwise have re-created or re-collected the data they thought they were able to).

There is an extensive literature in economics on returns to R&D, which, while varied, suggests that returns are high—typically in the region of 20 per cent to 60 per cent 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

[11] For example, some respondents to our user survey reported having undertaken training but not yet made use of EMBL-EBI data and services. Thus experiencing a negative impact of their time saving at that point in time.

average returns in such fields to be relatively high. Nevertheless, to be conservative, we explore the mid-range of returns characteristically identified in the literature (i.e., 40 per cent).

Box 2: Estimating returns to research activities

Robert Solow won the Nobel Prize in 1987 for his work on a macro-economic model that subsequently formed the basis for Growth Accounting.[12] Using a modified Solow-Swan model developed by Houghton and Sheehan (2009), we can explore the likely return on investment in research activity time. 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 what is largely publicly-funded research data at 5 per cent per annum, and following the lead of the US R&D Satellite Account (Sveikauskas 2007), we set the useful life of the data/knowledge created each year at an 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 (Sveikauskas 2007). Applying a 3.5 per cent discount rate to estimate net present value (HM Treasury Green Book 2011, p99), we then model the recurring returns.

Source: Authors' analysis.

For those who could neither have obtained the data they last used elsewhere nor created/collected it themselves, we divide the mean hours spent working with EMBL-EBI data by their mean frequency of use to estimate a time/cost per use. This is multiplied by the average hourly cost (i.e., the cost per hour of the research activity), the average returns to R&D expenditure, and the estimated share of total data accesses that are additional use. This suggests an increase in returns due to the additional use facilitated worth around £335 million annually and possibly as much as £2.5 billion over 30 years in net present value (at UK equivalent costs).

*((((time with EMBL-EBI data pa / frequency of use) * hourly cost) *
average returns to R&D) * additional share of total use)*
=
£335 million annually
£2.5 billion over 30 years NPV

[12] See http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1987/solow-bio.html and https://en.wikipedia.org/wiki/Solow%E2%80%93Swan_model

Savings from not having to obtain the data elsewhere or create/collect themselves

Of course, if the other element of additional use, namely that by users who could not have obtained the data elsewhere but could have (re)created it themselves, is taken into account, then the implied cost savings could be added to the implied additional returns. Net of access costs incurred using EMBL-EBI, the potential savings from not having to obtain the data elsewhere or create/collect it themselves could be as much as £1.4 billion per annum. Of course, these savings are a part of the overall efficiency savings reported by survey respondents.

*(hours to obtain * hourly cost) * (estimated users * share who could)*
+
*(hours to create * hourly cost) * (estimated users * share who could)*
=
£1.4 billion pa

Potential wider and longer term impacts of EMBL-EBI data and services

One indicator of the potential wider and longer term impacts of EMBL-EBI data and services is the impact of the research to which they contribute. While it can be no more than indicative, it is possible to use the modified Solow-Swan model developed by Houghton and Sheehan (2009) to estimate the potential wider and longer term impacts based on data collected through the user survey about research time spent with data from EMBL-EBI, from EMBL-EBI use logs, and assumptions about various modelling parameters, such as the average return to R&D, and so on (described above and in Box 2).

Assuming an average 40 per cent return to expenditure on R&D and basing analysis on the estimated use value of EMBL-EBI data, modelling suggests that the wider value of the research done using EMBL-EBI data and services during the last year might be worth as much as £920 million annually or perhaps £6.9 billion over 30 years in net present value (at UK equivalent costs).

*((((mean time with EMBL-EBI data per week * mean hourly cost) * weeks pa) *
estimated users) * average return to R&D)*
=
£920 million annually
£6.9 billion over 30 years NPV

From the analysis presented above we can see that, of this, some £335 million annually or £2.5 billion over 30 years in net present value (at UK equivalent costs) might be from research conducted by those who could neither have obtained the data elsewhere nor created/collected it themselves, and so depends directly upon EMBL-EBI data and services.

Of course, these values arise from the research and its subsequent use, and cannot be directly attributed to EMBL-EBI data and services. The latter are but a contributing element. Nevertheless, these estimates give a sense of the scale and importance of the activities to which EMBL-EBI data and services make an important and widely appreciated contribution.

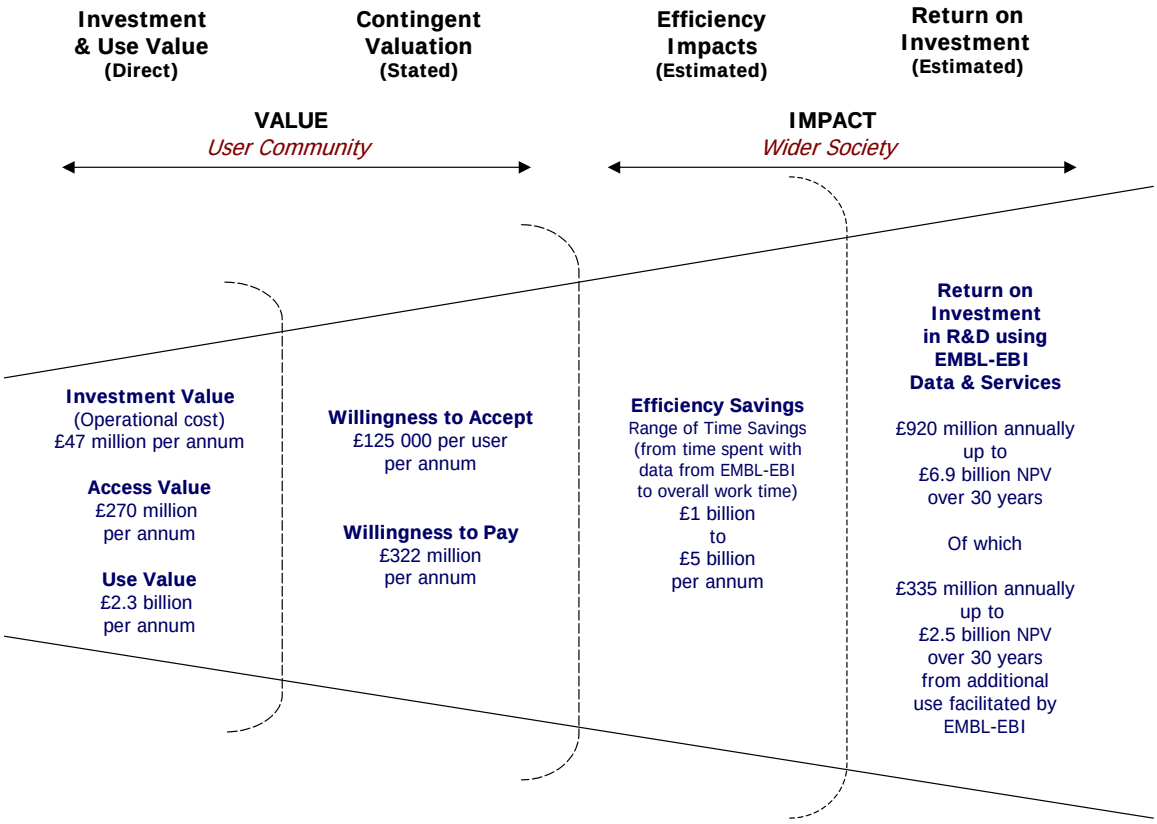
4.4 SUMMARISING THE ECONOMIC IMPACTS

The measures of impact below, as stated previously, are based on estimated values of EMBL-EBI user population and levels of use. These conservative estimates, within bounds, scale the values below in an approximately linear fashion and therefore twice the users or double the intensive of use would double the value (and vice versa) and the reported values should be treated as indicative.

Value of EMBL-EBI data and services

The most direct measure of the value of something is the amount of time and/or money people are willing to spend obtaining it (i.e., the activity-cost). Hence, the time users spend accessing EMBL-EBI data and services reflects what it is worth to them: an estimated £270 million over the last year. Contingent valuation is an alternative approach to estimating what something is worth, measuring what users say rather than what they do, and we estimate that EMBL-EBI users would have been willing to pay £322 million for their access during the last year. These estimates give a sense of the (minimum) direct value of EMBL-EBI's data and services to its user community, and compare very favourably with the approximately £47 million annual operational expenditure (i.e., a minimum direct value to users that is equivalent to around 6 to 7 times the direct operational cost).

Figure 3: The value and impacts of EMBL-EBI data and services



Source: Authors' analysis.

Impact of EMBL-EBI data and services

Users report that their use of EMBL-EBI data and services contributes to the efficiency with which they can perform their work, with a range of time savings (from time spent with data from EMBL-EBI to overall work time) worth an estimated £1 billion and possibly up to £5 billion during the last year across the worldwide user community—a benefit to users and their funders that, at a minimum, is equivalent to more than 20 times the direct operational cost. These efficiency gains could be realised through users working shorter hours (i.e., doing the same work in less time), working the same hours (i.e., doing more work in the same time), or a combination of both. In reality, of course, researchers use the time saved to do additional research which brings additional returns, to the benefit of research funders who get more ”bang for their buck“ and of society as a whole, which benefits from the discoveries and innovations arising from the additional research.

As a part of their research activities, users spent time worth an estimated £2.3 billion using data obtained from EMBL-EBI during the last year (i.e., obtaining, manipulating, and analysing data from EMBL-EBI). At average returns to R&D that would be worth some £920 million annually, or £6.9 billion over 30 years in net present value. So, during the last year, EMBL-EBI data and services contributed to the realisation of future research impacts conservatively worth £6.9 billion, and possibly more because bio-medical research is often characterised by higher than average returns to R&D.

Of course, it is important to consider the counterfactual and ask: how much of this impact might have been realised without EMBL-EBI, through alternative means? Some 45 per cent of EMBL-EBI user survey respondents reported that they could neither have obtained the last data they used elsewhere nor created/collected it for themselves. If that were true for all users, then during the last year EMBL-EBI would have facilitated future research impacts worth an estimated £335 million annually, or £2.5 billion over 30 years in net present value, that could not otherwise have been done (i.e., is directly attributable to EMBL-EBI). It is important to note that these estimates of value and impact focus on direct users and do not fully capture the contribution of collaborators or the value to both parties of that collaboration. Moreover, a number of users are known to obtain data from EMBL-EBI and curate it locally, making it available to other users in-house and/or incorporating it into data-based products and services used by others. Hence, the effective number of people using EMBL-EBI data and services is greater than the number of direct users, and our estimates do not capture the value for these secondary users of EMBL-EBI’s data and services.

5 QUALITATIVE ANALYSIS OF THE VALUE AND IMPACT OF EMBL-EBI

In this section we explore some of the qualitative information gathered during the study, primarily from the user survey (details of which can be found in Appendix 1) and case studies.

5.1 USER DEMOGRAPHICS

The survey was based on a convenience sample with an invitation to participate in the survey reaching 26 487 email addresses that did not bounce. There were 4 509 responses to the survey, a 17 per cent response rate. Due to the number and variety of EMBL-EBI data services, the survey was demanding, and a number of respondents did not complete enough of the questions to be included. After data cleaning, 4 185 responses were included.

Responses to our user survey were received from across the Eurozone and 91 non-Eurozone countries, from Albania to Zimbabwe. The bulk of respondents were from Eurozone countries (1 532), the United Kingdom (620), United States (354), and India (306). More than 100 responses were also received from Switzerland (109) and Brazil (127).[13]

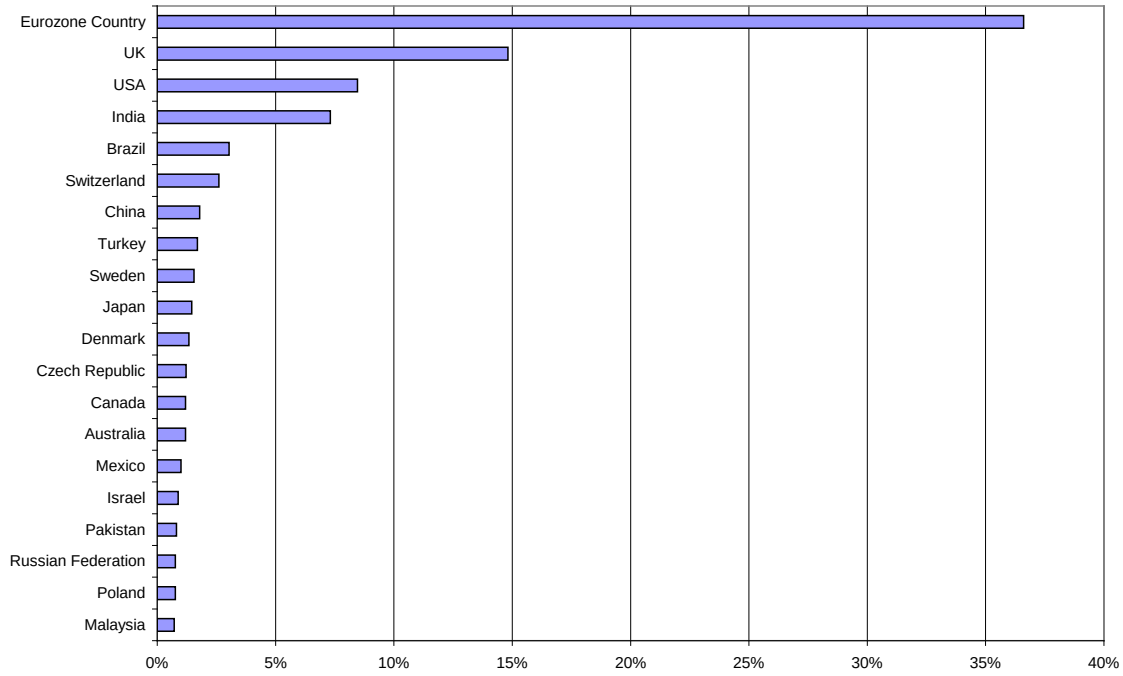
The overwhelming majority of respondents were in the academic sector (83 per cent), with almost 9 per cent coming from the corporate sector.[14] Forty-eight per cent of respondents described their work as mostly ”wet” laboratory (e.g., a scientist performing experiments in a lab or manager/overseeing ”wet” research), with 42 per cent saying mostly ”dry” working with computers (e.g., a bioinformatician). A further 10 per cent reported non-scientific and other working environments. Research was a part of the role for 93 per cent of respondents (N = 4040).

More than 17 per cent of respondents reported most recently using Ensembl, 15 per cent most recently using UniProtKB, 6.3 per cent Europe PubMed Central, 5.9 per cent GO (Gene Ontology), 5.6 per cent NCBI taxonomy, 5.4 per cent PDBe (Protein Databank in Europe), and just less than 5 per cent Pfam (Figure 4). The vast majority (86 per cent) of respondents use a web browser to access EMBL-EBI resources, with just 7 per cent downloading using FTP, and a further 6.5 per cent using programmatic access.

[13] Two non-answers were completed using IPLocation.

[14] Six ‘other’ responses were re-coded based on comments given.

Figure 4: The Top 20 services most recently used (share of respondents, per cent) N = 3941



Source: Authors' analysis.

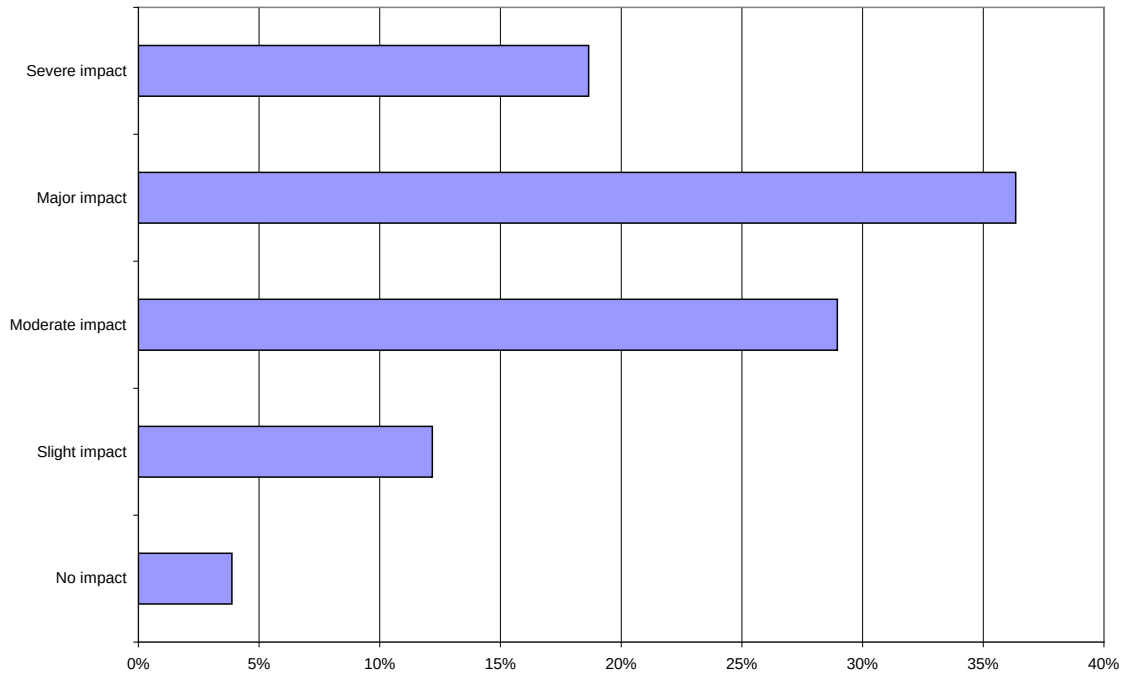
5.2 THE IMPACT OF ACCESS TO EMBL-EBI

More than half of all respondents (55 per cent) said that not having access to EMBL-EBI services and resources would have a major or severe impact on their work or study, with a further 29 per cent saying it would have a moderate impact (Figure 5). The weighted average score was 2.54 out of 5, reflecting the balance of responses towards the importance and impact of EMBL-EBI services and resources.

Some 463 respondents also added free text comments on this question. Major themes included:

- EMBL-EBI services are important, essential, fundamental, or difficult to replace. They rely/depend on them for their own work (N=125);
- Use them daily, regularly, frequently, or routinely (N=63);
- EBL-EBI services are high quality or superior to other sources, have unique elements, or are important to the scientific community as a whole (N=52);
- Use multiple EMBL-EBI tools and services (N=52);
- Could find alternatives for some things (e.g. NCBI, UCSC, Galaxy, Genbank) (N=50);
- Value the workshops, training, and courses (N=34);
- Use them in teaching (N=23);
- Free, Open, Sharing (N=22);
- Loss of EMBL-EBI would impact time, efficiency, or integrated workflows (N=20);
- EMBL-EBI services are useful, valuable, or helpful (N=18); and
- EMBL-EBI helps them keep skills up to date (N=14).

Figure 5: The impact of not having access to EMBL-EBI (share of respondents, per cent) N = 4155



Source: Authors' analysis.

When thinking about the value and impact of EMBL-EBI data and services, it is important to explore the counterfactual (i.e., what would users have done if EMBL-EBI did not exist). The majority of respondents (55 per cent) said they could not have obtained the last resource they used from another source, with 45 per cent saying they could have (N = 3887). Almost 80 per cent of those who could not have obtained elsewhere the data they last used said that they could not have created/collected it themselves. Just 21 per cent said they could have done so (N = 3691).

Combining these answers, of the 3 636 answering both questions, 1 643 (45 per cent) could neither have obtained elsewhere nor created/collected the last data they used themselves. This gives an indication of the proportion of use of EMBL-EBI data that is additional use, which could not otherwise have occurred.

Asked the extent to which they had benefited from using EMBL-EBI, responses suggested that most had benefited from data and tools, with training, user support, and collaborations also widely beneficial (Table 1). Among users, data and tools received a weighted score of 3.14 from 5 (N = 3734), with training scoring 2.48 (N = 2533), user support 2.19 (N = 2228), and collaborations 2.18 (N = 1930).

Table 1: Extent benefited from using EMBL-EBI (share of respondents, per cent)

	Data and tools	User support	Training	Collaborations	Other
haven't used	3.2%	38.1%	31.4%	46.0%	80.4%
no benefit	0.4%	4.2%	3.5%	5.5%	3.8%
low benefit	2.8%	12.1%	9.6%	11.0%	3.1%
medium benefit	14.7%	20.2%	19.2%	14.5%	4.7%
high benefit	43.7%	18.7%	22.8%	14.4%	4.7%
very high benefit	35.1%	6.7%	13.4%	8.6%	3.2%

Source: Authors' analysis.

To explore the impact of EMBL-EBI services on its user community, respondents were asked to estimate any resulting change in research efficiency. As in similar surveys that we have run in different disciplines, the reported efficiency impacts are considerable. Among those reporting a zero or positive impact, the mean was 46 per cent (N = 3239), with a median of 50 per cent. Just 18 respondents reported a negative impact, which could not be included in analysis as it is not numerical. A few comments by these respondents suggested that they had spent time on training and familiarisation, but had not yet used EMBL-EBI data and services enough to gain a benefit.

5.3 How users value their access to EMBL-EBI

Contingent valuation as explained in section 4.3 was used to estimate EMBL-EBI "non-market" value, based on asking what users would be willing to accept in return for giving up their access to EMBL-EBI and what they would be willing to pay for access in a hypothetical market situation. As outlined in section 4.3, high valuations were given by users who stated that they could not do their research without EMBL-EBI services, or calculation of the cost of recreating the data. This was counterbalanced with some users giving a zero value. The principal (and, perhaps, principle) reasons given for saying they would not be willing to accept anything in return for giving up access included such comments as "it's priceless", as well as statements that individuals believed science and research data should be open and free and thus they would not accept anything in return for it. There were 46 responses explicitly referencing Open Access/Open Data principles.

As a result, the range of amounts that users would be willing to accept was wide. The mean of values respondents would be willing to accept in return for giving up their access for a year was more than £125 000 (N = 2079).

Box 3: One respondent's comment on Ensembl / Ensembl Genomes

> I f****** love everything about ensembl. It is probably THE BEST tool out there. No one else has managed to put together that much information with such flexibility and capability for analysis, all while maintaining such a high degree of user friendliness. As someone who doesn't work with humans or mice, ensembl is invaluable. No other resource has taken the time to do such a great job for the other mammalian species out there. Seriously if it weren't for ensembl, many of the projects I have worked on and [am] working on would have been much more difficult and would have had lower quality results and performance. If there wasn't ensembl/ensembl genomes and I had to wade through the sluggish, inconsistent and painful XXXX stuff instead, I'd probably find a different job... Fund ensembl till the sun explodes!

Source: User survey. Authors' analysis. Some words and organisations redacted.

Similar reasons inform what respondents would be willing to pay as an annual subscription for access to EMBL-EBI data and services, with what they or their organisation could afford to pay being an oft-cited limitation. There were also a number of respondents unwilling to pay anything because they believe science should be open and data free. The mean of the value respondents reported being willing to pay for access for a year was £1 628 (N = 2191).

It is clear from the qualitative answers and comments that users value highly their access to EMBL-EBI data and services, and that their access has a significant impact on their research efficiency and capacity to do their research.

5.4 THE CASE STUDIES

The full case studies are included within the report as appendix 4.

ChEMBL

ChEMBL is a unique public knowledgebase of chemical compounds and small molecules with their biological targets. It pulls together high-value information on compounds and their effects on biological systems from the available academic literature in a structured database. In this case study we cite four external users of the ChEMBL database, showing the range of applications of ChEMBL data alongside facts about the database, statistics derived from the impact survey, and public information. The case study demonstrates that ChEMBL's focus on enabling all aspects of discovery is utilised by universities and companies of all sizes, strengthening innovation from new research and the discovery of new treatments and drugs benefiting human and animal welfare as well as agriculture.

The Gene Ontology (GO)

GO has become part of the scientific infrastructure that supports much work in academia and industry. In this impact case study we present short extracts from five interviews conducted with external users of the EMBL-EBI GO service. We use these extracts to illustrate how GO is underpinning scientific investment, improving R&D, and increasing productivity and performance. The case study illustrates the scale of usage that lies behind individual EMBL-EBI web statistics and unique visitors for a single EMBL-EBI service such as GO. We also illustrate some of the efficiencies and qualitative benefits from sharing EMBL-EBI infrastructure with other members of the GO consortium..In terms of efficiency savings alone, we estimate this has a benefit to cost ratio of 3:1.

The Variant Effect Predictor Tool

Many organisms' genomes, including humans', have been fully sequenced, and the reference data sets are in Ensembl's databases. We can now start to understand what is happening in individual genomes by comparing them with reference genomes and then combining that information with knowledge from all other fields of molecular biology. Ensembl's Variant Effect Predictor (VEP) is a powerful open software tool that can analyse most types of variation data. It uses the extensive annotation in Ensembl to provide detailed functional predictions and annotation on the effects of variants. VEP is deployed in many critical areas of research such as cancer and rare diseases, where strong links have been established between changes in the genome and disease development. We present finding from interviews with three diverse users of the VEP tool demonstrating the research it supports.

6 CONCLUSIONS AND OBSERVATIONS

6.1 OBSERVATIONS CONCLUSIONS

The approaches used in this study apply a unique combination of quantitative and qualitative methods to provide a full picture of the nature and dimensions of EMBL-EBI's value and explore the range of impacts of its provision of data and services.

Both qualitative and quantitative analyses show that EMBL-EBI data and services are widely used, appreciated, and valued by their user community:

- The economic analysis shows that EMBL-EBI data and services are highly valued by users. A very significant increase in research efficiency was reported by users as a result of their use of EMBL-EBI data and services, which we estimate to be worth at least £1 billion per annum worldwide.

- The average time respondents reported spending on research varied widely. The mean was 34 hours per week, with a number reporting long hours. They reported spending a mean of 56 per cent of their research time working with data and 20 per cent of their research time with data obtained from EMBL-EBI. This reflects the intensity and importance of data and related services to current research in the life sciences.

- Showing the extent to which EMBL-EBI data and services facilitate research that could not otherwise have been done, 45 per cent of respondents also said they could neither have created/collected the last data they used nor obtained it elsewhere. It is estimated that the value of the future impacts of research that could not have been done without access to EMBL-EBI data and services is some £330 million annually and up to £2.5 billion over 30 years (net present value).

- The qualitative analysis in the user survey shows that 55 per cent of respondents reported that it would have a major or severe impact on their work if they could not access EMBL-EBI data and service. A further 29 per cent said that not having access would have a moderate impact.

- The case studies, conducted as a part of this study, reported in appendix 4, also demonstrate the value and impact of EMBL-EBI data and services and the pathways through which these impacts are achieved.

The quantitative and qualitative analyses independently show a similar picture of the value and impact of EMBL-EBI data and services: they are complementary, reinforce each other, and lend credence to the findings.

6.2 OBSERVATIONS

During the course of our research we have noted a number of things that could help all parties involved in funding and operating research data infrastructure.

1. **User population**. Economic estimates of the value and impact have been somewhat constrained due to limited information on the number of "real-world" users and uses of EMBL-EBI data and services, to which survey populations and results can be expanded. As pressure to demonstrate economic value and impact is likely to increase, EMBL-EBI may wish to develop ways to better understand user population numbers. This might take the form of a voluntary or compulsory user registration system. Such a system can be designed to minimise user impacts and concerns. It could potentially be built out from the existing voluntary registration scheme for *Train online* at EMBL-EBI, which has approximately 5 000 registrations. While far from perfect in terms of recording the actual number of active users, such a system would be a big step forward.
2. **Research data metrics**. It is also clear from a range of studies that different data services collect financial and operational data, such as user statistics, access, and download statistics, to varying levels of detail. More central guidance and/or widespread cooperation and coordination by the community is needed on the collection of such data. This would help to ensure a greater degree of standardisation of statistical records across data centres and services, as well as providing the basis for more comprehensive and reliable data for economic analysis. We have noted in section 3.1 a current NISO initiative on research data metrics that may be of interest to EMBL-EBI. Participation in this or similar initiatives, and developing further support for capturing relevant usage statistics, should be considered.
3. **Developing impact evaluation approaches**. The combination of qualitative and quantitative approaches used in this study has now been applied to a range of data centres and services spanning very different disciplinary domains. The experience suggests that the approaches are complementary and mutually reinforcing. 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" for data centres and services plays an increasingly important role in ensuring their sustainability.
4. **Granular studies**. This study looked at the aggregate value of EMBL-EBI data and services. There is also significant scope for more granular studies that focus on the value of specific data collections or services, or the economic value of EMBL-EBI data and services to specific groups and sub-sets of users and collaborators. 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 analysis by specific stakeholder groups may also be beneficial. This is the ongoing intention of EMBL-EBI
5. **Email lists**. EMBL-EBI is a large organisation and encompasses a diverse and numerous group of services. Communication with users is through individual services, and organisation-wide communication with users, for purposes such as this or previous surveys, has been through individual lists. Given the overlap in service use, there can be a significant overlap and

duplication in email addresses. A merged and de-duplicated list for organisation-wide activities was created by EMBL-EBI during the course of this study. Maintaining this list will be invaluable for future uses such as surveys.

6. **Survey Design**. Due to the large number and variety of EMBL-EBI data services, the survey for this study was demanding and a significant time was spent in survey design and testing prior to its launch. This study benefited from and built on approaches in previous EMBL-EBI user surveys and our own previous impact surveys. The design and communication of the survey produced excellent response and completion rates. There are new elements of the survey design (e.g., the use of the 39 services), which are closely aligned with the annual EMBL- EBI user survey and may be useful innovations for future EMBL-EBI work.

REFERENCES

ABS (2010) *Research and Experimental Development, All Sector Summary, 2008-09*, Cat No 8112.0. http://www.abs.gov.au/AUSSTATS/abs@.nsf/Latestproducts/8112.0Main%20Features32008-09?opendocument&tabname=Summary&prodno=8112.0&issue=2008-09&num=&view=

Beagrie, N., Chruszcz, J., and Lavoie, B. (2008) *Keeping Research Data Safe : a cost model and guidance for UK Universities*, JISC, London and Bristol. http://www.webarchive.org.uk/wayback/archive/20140615221657/http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf

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

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

Beagrie, N., Houghton, J.W., Palaiologk, A., and Williams, P. (2012) *Economic Impact Evaluation of the Economic and Social Data Service (ESDS)*, Economic and Social Research Council, London. http://www.esrc.ac.uk/files/research/evaluation-and-impact/economic-impact-evaluation-of-the-economic-and-social-data-service/

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

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

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

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. http://www.bl.uk/aboutus/stratpolprog/increasingvalue/publicvalue/confpres/pungelwesmarks.pdf

Crockett, A. (2011) *Weighting the Social Surveys*, UK Data Archive and Institute for Social and Economic Research, May 2011.

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

Dey-Chowdhury, S. (2008) Perpetual inventory method, Economic & Labour Market Review 2(9), September 2008, pp48-52. Office of National Statistics. http://www.ons.gov.uk/ons/rel/elmr/economic-and-labour-market-review/no--9--september-2008/methods-explained--perpetual-inventory-method--pim-.pdf

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. 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.

EvaRIO (2013a) *EvaRIO Case Study: EMBL-EBI* Deliverable D5.1. December 2013 http://evario.u-strasbg.fr/uploads/autres-docs-BETA/EvaRIO_CAse_Study_EMBL-EBI.pdf

EvaRIO (2013b) *Core study: Adapting the BETA method to the case of the evaluation of the impact of Research Infrastructures*, Final report: Results and Policy Recommendations - PART I.

Fomitchev, M.I. (2010) *How Google Analytics and conventional cookie tracking techniques overestimate unique visitors*, Academia. http://www.academia.edu/4492061/How_google_analytics_and_conventional_cookie_tracking_techniques_overestimate_unique_visitors

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

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. and Mohnen, P. (2009) Measuring the returns to R&D, NBER Working Paper 15622, NBER, Cambridge MA.

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

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

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

Kinman, G. and Jones, F. (2004) *Working to the Limit*. London: AUT Publications.

Kinman, G. and Wray, S. (2013) *Higher Stress: A Survey of Stress and Well-Being among Staff in Higher Education*. London: University and College Union.

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

Martin, B.R., and 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

OECD (2015) Exchange rates (indicator). doi: 10.1787/037ed317-en (Accessed on 23 July 2015).

Research Information Network (2011) *Data centres: their use, value and impact*, September 2011. http://www.rin.ac.uk/system/files/attachments/Data_Centres_Report.pdf

Salter, A.J. and 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., and Steinmueller, E. (2002) *The Economic Returns to Basic Research and the Benefits of University-Industry Relationships*. Report to the Office of Science and Technology, London.

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

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

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

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

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

Tripp, S. and Grueber, M. (2011) *The Economic Impact of the Human Genome Project*, Battelle Memorial Institute. https://www.genome.gov/27544383

Trounson, A. (2015) Staff 'donate' $1.7bn in unpaid work, *The Australian*, 22 July 2015, p29.

UNESCO (2010) *UNESCO Science Report 2010: The Current Status of Science Around the World*, UNESCO. http://www.unesco.org/new/en/natural-sciences/science-technology/prospective-studies/unesco-science-report/

Appendix 1: A Summary of the User Survey

This section provides a summary of the results from the user survey conducted as a part of this study. Details of the free text comments are presented in Appendix 2.

The survey was based on a convenience sample with an invitation to participate in the survey reaching 26 487 email addresses that did not bounce. The survey invitation email list was compiled from the EMBL-EBI training list, EMBL-EBI SME forum list, direct user download dispatch email list for 2015 (excluding @ebi.ac.uk addresses), the Biocuration society mailing list, and the EMBL-EBI Group Team Leader named contact lists. The survey was opened on 8th April 2015 and closed on 18th May 2015. A reminder email was sent on 7th May 2015. Five Raspberry Pi 2 with Pibow Cases were offered in a draw for participants.

We received 4 509 responses to the survey, a 17 per cent response rate. Due to the number and variety of EMBL-EBI data services, the survey was demanding and a number of respondents did not complete enough of the questions to be included in the results. After data cleaning, 4 185 responses were included.

Details of the data cleaning are discussed in the summary of responses below although the principal element of this process was the deletion of insufficiently complete responses. Primarily, this involved deleting 324 responses that did not get beyond Question 19 on frequency of use of various data services. A further 199 provided incomplete answers on frequency of use but did answer subsequent questions.

Demographics

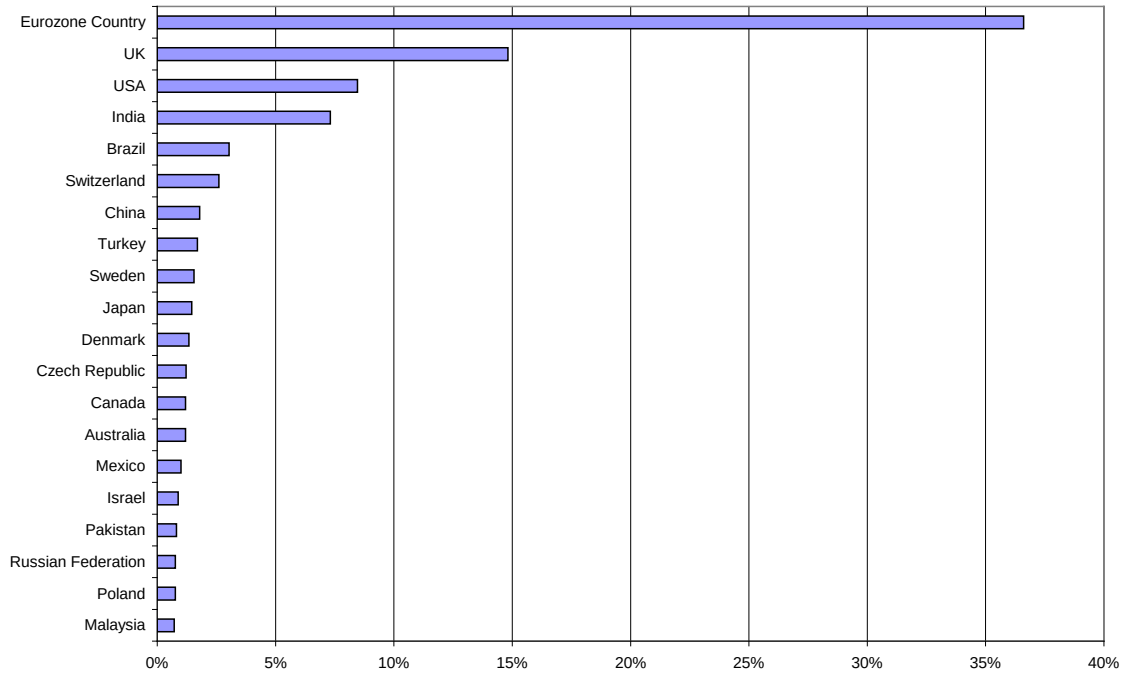
The first few questions sought to understand the characteristics of respondents, their country and currency, the sector in which they work, their affiliation and role within that sector, and the nature of their work.

Q1. Country/Currency Zone in which you work or study [If you are in the Euro currency zone please select Eurozone Country]?

Responses were received from across the Eurozone and 91 non-Eurozone countries, from Albania to Zimbabwe. The bulk of respondents were from Eurozone countries (1 532), the United Kingdom (620), United States (354), and India (306). More than 100 responses were also received from Switzerland (109) and Brazil (127).[15] The Top 20 countries are shown in Figure A1.1.

[15] Two non-answers were completed using IPLocation.

Figure A1.1 Country base (share of respondents, per cent) N = 4185



Source: Authors' analysis.

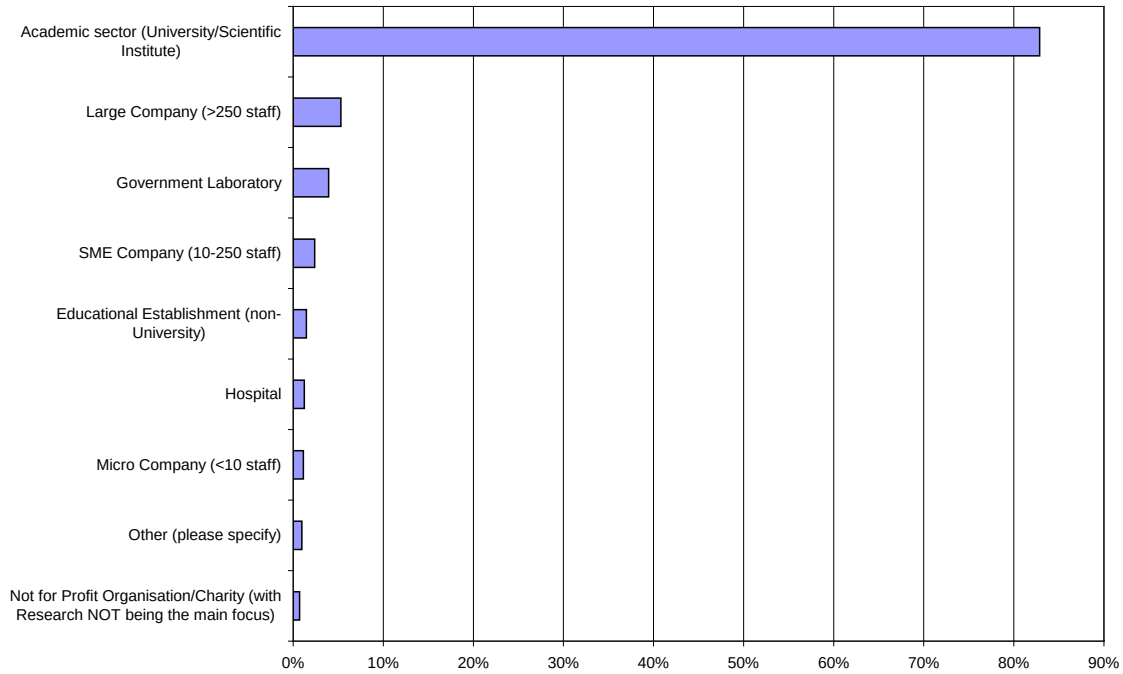
The respondents' currency was derived from their answers to Q1, and their value responses converted the British Pounds (GBP) at the spot exchange rate reported by Google Finance on 22nd May 2015 (i.e., as close to the time of the survey as possible).

Q2. Your Main Affiliation & Sector?

The overwhelming majority of respondents were in the academic sector (83 per cent), with almost 9 per cent coming from the corporate sector (Figure A1.2).[16]

[16] Six 'other' responses were re-coded based on comments given.

Figure A1.2 Affiliation and sector (share of respondents, per cent) N = 4185

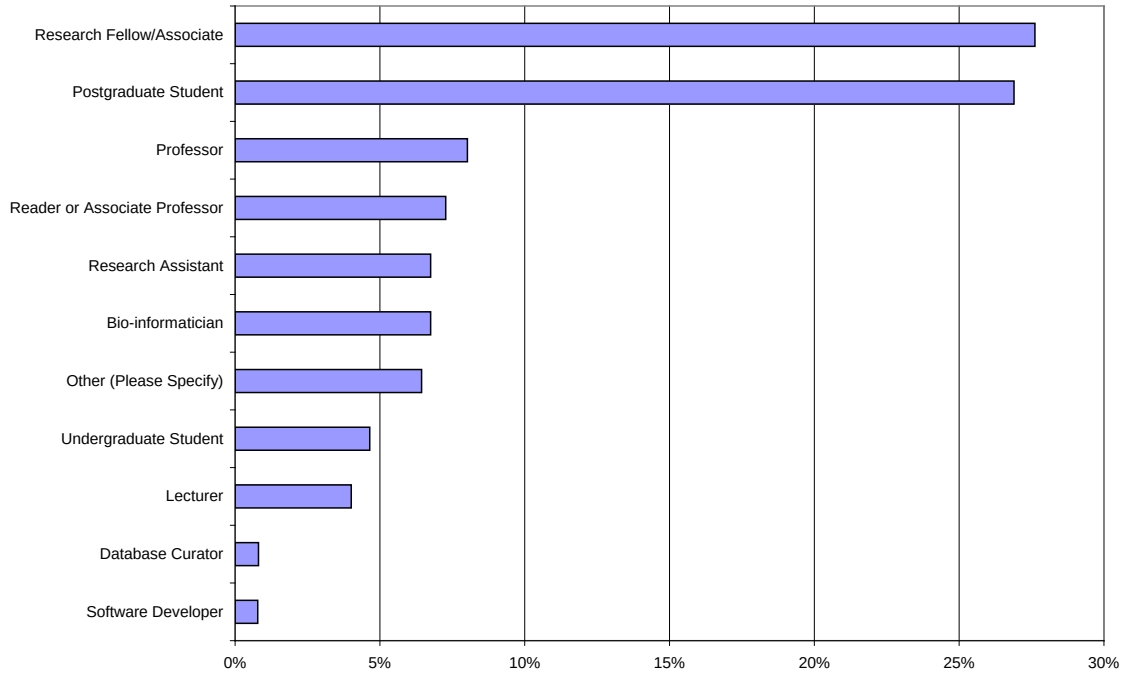


Source: Authors' analysis.

Q3 & Q4. What is your main role within this affiliation?

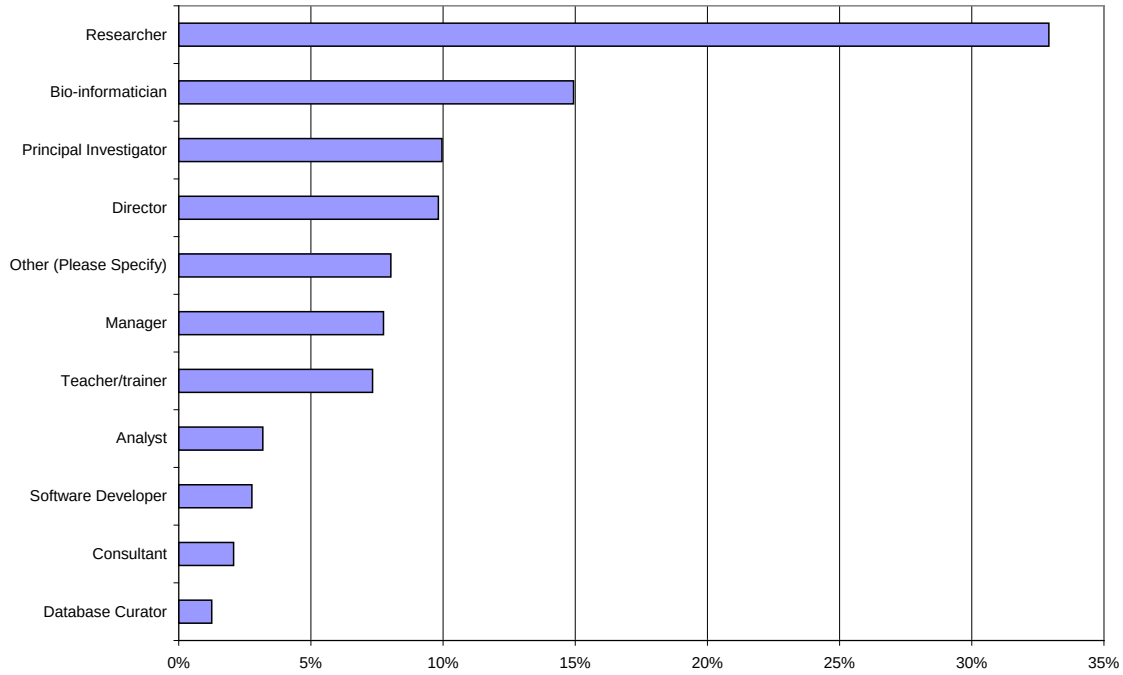
Respondents were taken to different versions of Q3/Q4 depending on their sector. Among the 3 465 academics, more than 25 per cent reported being a research fellow/associate or postgraduate student (Figure A1.3). Around 33 per cent of the academic respondents were students, of which 27 per cent were post-graduates and almost 5 per cent under-graduates. Some 33 per cent of the 723 non-academic respondents reported their role as researcher, 15 per cent bio-informatician, and 10 per cent principal investigator (Figure A1.4).

Figure A1.3 Main role in academic sector (share of respondents, per cent) N = 3465



Source: Authors' analysis.

Figure A1.4 Main role in non-academic sector (share of respondents, per cent) N = 723

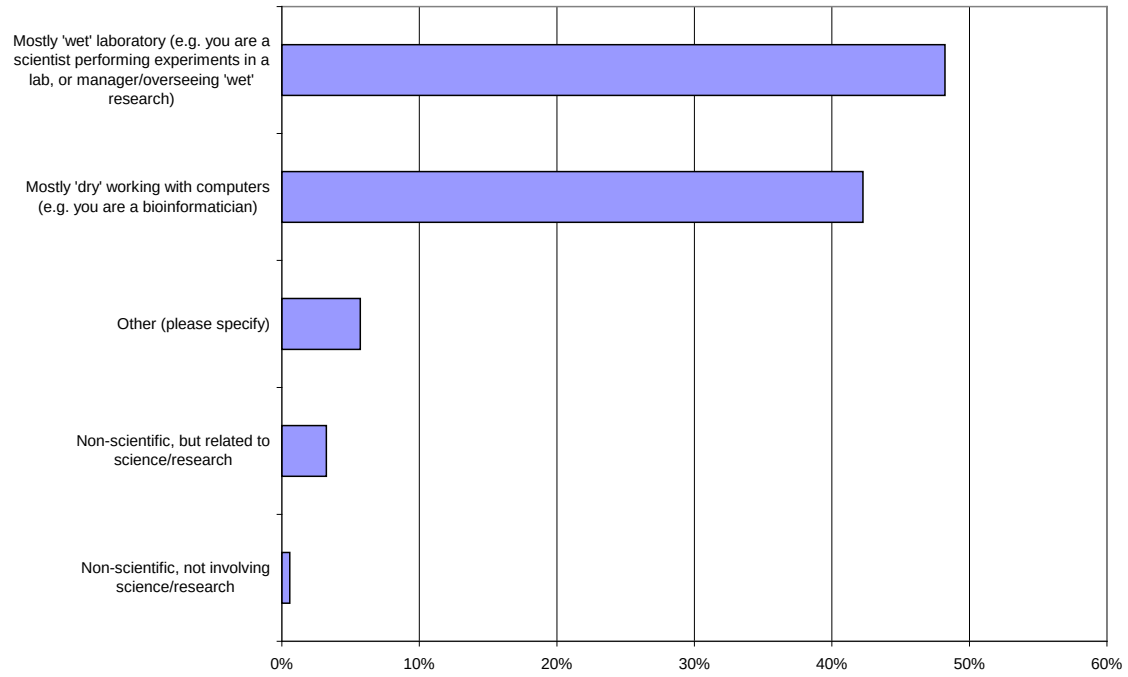


Source: Authors' analysis.

Q5. Which of the following most closely describes the nature of your work or study?

Forty-eight per cent of respondents described their work as mostly 'wet' laboratory (e.g., a scientist performing experiments in a lab or a manager overseeing 'wet' research), with 42 per cent saying mostly 'dry', working with computers (e.g., a bioinformatician). A further 10 per cent reported non-scientific (but related to science/research), non-scientific (not involving science/research), and other working environments.

Figure A1.5 Nature of work or study (share of respondents, per cent) N = 4172

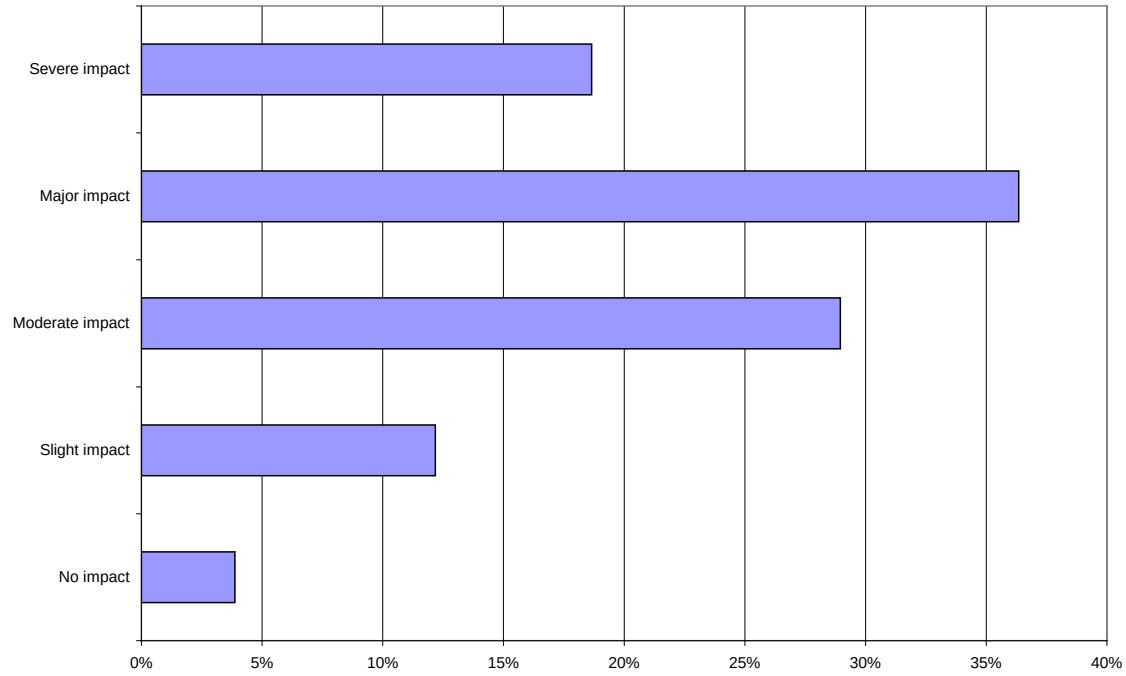


Source: Authors' analysis.

Q6. What impact would it have on your work or study if you could not access EMBL-EBI services and resources?

More than half of all respondents (55 per cent) said that not having access to EMBL-EBI services and resources would have a major or severe impact on their work or study, with a further 29 per cent saying it would have a moderate impact. The weighted average score was 2.54 out of 5, reflecting the balance of responses towards the importance and impact of EMBL-EBI services and resources.

Figure A1.6 The impact of not having access to EMBL-EBI (share of respondents, per cent) N = 4155



Source: Authors' analysis.

Focus and frequency of access

Q7 through Q19. Approximately how frequently did you access/download resources from (…) in the last 12 months?

The questionnaire offered the frequency categories: > once a day, daily, weekly, fortnightly, monthly, quarterly, 1-2 times a year, and not used. These were converted to approximate annual use frequencies as follows: > once a day = 440, daily = 220, weekly = 45, fortnight = 22, monthly = 12, quarterly = 4, 1-2 times a year = 1.5, and not used = 0 (deleted).

The respondents' average frequency of use of EMBL-EBI data services was 547 per year, with a total of almost 2.2 million accesses/downloads reported. There was considerable variation between the services (Table A1.1). The most heavily used services included European PubMed Central, UniProtKB, Ensembl, Ensembl Genomes, and EMBL-EBI Search.

Table A1.1 Estimated frequency of use of EMBL-EBI data services (per annum)

Service	*Mean use frequency among respondents*	*Sum of uses by respondents*
Europe PubMed Central (Europe PMC)	89	247,900
BioJS component registry	32	34,467
EFO (Experimental Factor Ontology)	26	26,503
GO (Gene Ontology)	39	98,032
OLS (Ontology Lookup Service)	28	32,530
1000 Genomes	26	37,145
ArrayExpress	20	26,314
DGVa (Database of Genomic Variants Archive)	22	18,281
EGA (European Genome-phenome Archive)	23	16,932
ENA (European Nucleotide Archive)	35	45,346
Ensembl	66	170,613
Ensembl Genomes	55	120,030
Expression Atlas	27	41,098
HGNC (HUGO Gene Nomenclature Committee)	44	57,132
Metagenomics	25	21,876
Enzyme Portal	32	32,467
IntEnz	34	23,496
InterPro	42	80,987
PDBe (Protein Databank in Europe)	52	104,142
Pfam	46	100,701
PRIDE (PRoteomics IDEntifications)	39	38,760
UniProtKB	83	216,247
ChEBI	40	27,564
ChEMBL	39	29,849
IntAct	36	24,861
Metabolights	36	16,830
Reactome	33	32,424
UniChem	31	15,042
BioModels	33	12,711
BioSamples	32	10,711
Mouse services (EMMA, etc.)	31	8,589
PomBase	82	22,427
VectorBase	33	13,760
WormBase	34	12,279
EBI Search (searching our data)	60	111,980
EBI sequence analysis tools	53	99,368
Rfam	41	32,648
NCBI taxonomy	52	93,745
TreeFam	34	23,435
Across all services	*547*	*2,179,210*

Source: Authors' analysis.

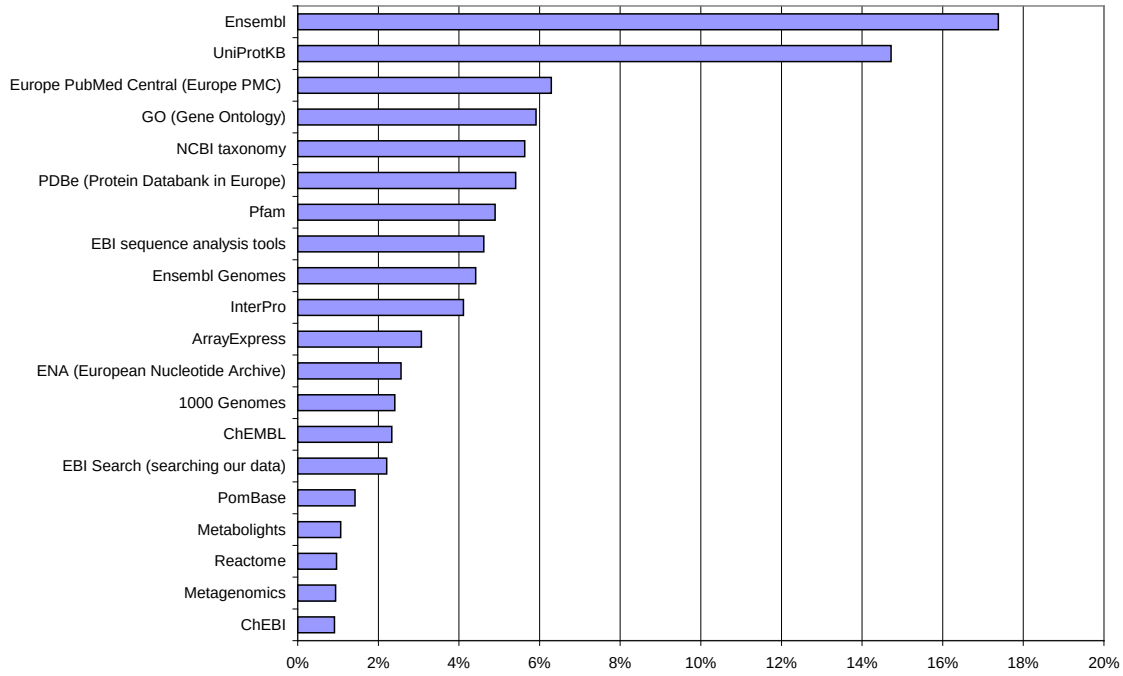
Services accessed, access time and mode

In order to randomise subsequent answers, we asked a critical incident question about the last use.

Q20. Which EMBL-EBI service resource did you last use?

More than 17 per cent of respondents reported most recently using Ensembl, 15 per cent most recently using UniProtKB, 6.3 per cent Europe PubMed Central, 5.9 per cent GO (Gene Ontology), 5.6 per cent NCBI taxonomy, 5.4 per cent PDBe (Protein Databank in Europe), and just less than 5 per cent Pfam (Figure A1.7).

Figure A1.7 The Top 20 services most recently used (share of respondents, per cent) N = 3941



Source: Authors' analysis.

Q21. How long did it take you to find and access the last resource you used from EMBL-EBI?

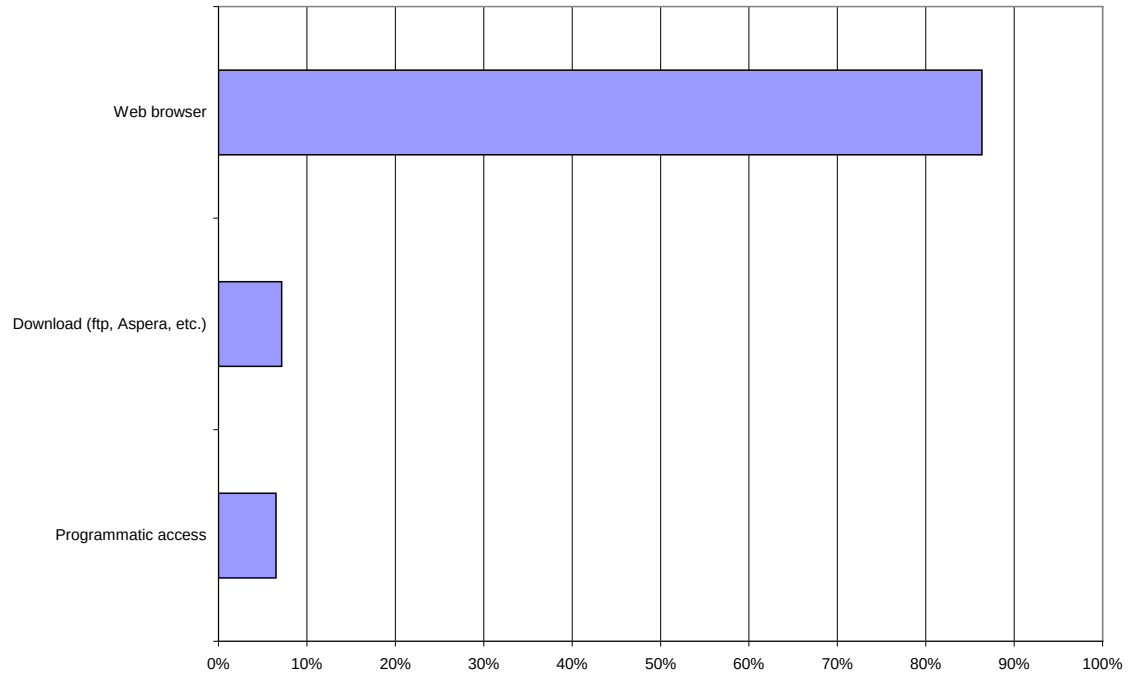
Respondents reported a mean access time of 51 minutes, although with wide variation.[17] The median reported access time was 5 minutes. Many found access easy, with 506 saying it took 1 minute or less, with comments such as: "it's bookmarked", "I know where to go", "it takes no time at all", and so on. At the other end of the scale, 51 respondents reported access times of 24 hours or more. Such cases were accompanied by comments about download speeds using FTP, difficulties setting up a programmatic access, and so on.

[17] During data cleaning we changed the zero time responses to half-a-minute, as there were a number of comments about access taking little or no time at all, and yet of course some time must be involved.

Q22. What mode of access did you use to obtain data from the last resource you used from EMBL-EBI?

The vast majority (86 per cent) of respondents use a web browser to access EMBL-EBI resources, with just 7 per cent downloading using FTP, and a further 6.5 per cent using programmatic access (Figure A1.8). These shares inform the previous question about access time, with mean reported access times of 83 minutes using FTP, 65 minutes using programmatic access, and 47 minutes using a web browser.

Figure A1.8 The mode of access used to obtain the data last used (share of respondents, per cent) N = 3919



Source: Authors' analysis.

Measuring value and impact

When thinking about the value and impact of EMBL-EBI data and services it is important to explore the counterfactual (i.e., what would users have done if EMBL-EBI did not exist).

Q23. If EMBL-EBI had not existed, would you have been able to obtain the last resource you used from another source?

The majority of respondents (55 per cent) said they could not have obtained the last resource they used from another source, with 45 per cent saying they could have (N = 3887).

Q24. If Yes, approximately how much time would it have taken to obtain the resource you last used from another source?

For the reasons noted in footnote 3 above, zero time responses were changed to half-a-minute. Greater than 12 months responses were coded to 13 months, and the whole converted to hours with months as 20 days, and days 7.5 hours.

Respondents suggested that it would take them a mean of 32 hours to obtain the data they last used from elsewhere (N = 1529).

Q25. If you could not have obtained the last resource you used elsewhere, would you have been able to collect/recreate it yourself?

Almost 80 per cent of those who could not have obtained the data they last used elsewhere said that they could not have created/collected it themselves. Just 21 per cent said they could have done so (N = 3691).

Combining Q25 and Q23, of the 3 636 answering both questions, 1 643 (45 per cent) could neither have obtained elsewhere nor created/collected the last data they used themselves. This gives an indication of the proportion of use of EMBL-EBI data that is additional use, which could not otherwise have occurred.

Q26. If Yes, approximately how much time would it have taken to collect/recreate the resource you last used?

Again, zero time responses were changed to half-a-minute. Greater than 10 years was converted to 11, and the whole converted to hours with years as 220 days, months 20 days, and days 7.5 hours.

The mean time required to create/collect the data last used for themselves was 999 hours (133 days) (N = 688).

Qualitative benefits

Q27. To what extent do you benefit from using EMBL-EBI in any of the following ways?

Asked the extent to which they had benefited from using EMBL-EBI, responses suggested that most had benefited from data and tools, with training, user support and collaborations also widely beneficial (Table A1.2).

Table A1.2 Extent benefited from using EMBL-EBI (share of respondents, per cent)

	Data and tools	User support	Training	Collaborations	Other
haven't used	3.2%	38.1%	31.4%	46.0%	80.4%
no benefit	0.4%	4.2%	3.5%	5.5%	3.8%
low benefit	2.8%	12.1%	9.6%	11.0%	3.1%
medium benefit	14.7%	20.2%	19.2%	14.5%	4.7%
high benefit	43.7%	18.7%	22.8%	14.4%	4.7%
very high benefit	35.1%	6.7%	13.4%	8.6%	3.2%

Source: Authors' analysis.

Among users, data and tools received a weighted score of 3.14 from 5 (N = 3734), with training scoring 2.48 (N = 2533), user support 2.19 (N = 2228), and collaborations 2.18 (N = 1930).

Impacts on research

Q28. Is research a part of your role?

Research was a part of the role for 93 per cent of respondents (N = 4040).

Q29. Over the last twelve months, on average how many hours per working week did you spend on research?

The average time spent on research per week varied widely as some respondents were purely researchers while others were mainly teaching. Nevertheless after data cleaning the reported mean was 34 hours per week, with a number reporting long hours (N = 3370).

Before data cleaning there were 386 responses of greater than 50 hours per week, with a maximum of 100 hours per week. Surveys of academic working time often report long working hours. In the UK, Kinman and Wray (2013) found that more than three-quarters of their survey respondents employed on a full-time contract worked over 40 hours a week, and more than one-third in excess of 50 hours a week. Kinman and Jones (2004) noted that a considerable proportion of members of the Association of University Teachers were working in excess of 48 hours per week, and 59 per cent of their respondents employed on a full-time basis were working more than 45 hours in a typical week, and 21 per cent in excess of 55 hours. Similarly, a recent survey of academics in Australia found that they were working an average of a little more than 50 hours per week (Trounson 2015).

We set a maximum of 65 hours per week, re-coding the 138 responses (4 per cent) reporting greater than 65 hours to 65 hours. As a result, the mean fell from 35 hours to 34 hours per week—making little difference.

Q30. Can you estimate the approximate share of your total research working time spent with data during the last twelve months (e.g., creating, manipulating, and analysing data)?

To maximise responses, we asked respondents to estimate the share of their time spent working with data, using 10 per cent intervals up to >90 per cent, which was converted to 91 per cent to be conservative. Some respondents appear to have interpreted the question to mean the overall share with all data (correctly) and the share of that with data from EMBL-EBI (incorrectly). Consequently, where the reported share of time spent with data obtained from EMBL-EBI was greater than the share spent with all data it was re-coded as a percentage share of the reported time with all data (138 cases).

On that basis, they reported spending a mean of 56 per cent of their research working time with data (N = 3244), and 20 per cent of their time with data obtained from EMBL-EBI (N = 2953).

Q31. What do you think might be typical for others in the same research field?

In order to generalise from survey responses, it is important to know how typical respondents are in respect to their data use. Following similar adjustment of 187 responses to that noted above, respondents suggested that others in their research field might typically spend a mean of 53 per cent of their research working time with data (N = 2669), and 24 per cent with data obtained from EMBL-EBI (N = 2426). Thus, as we have found in similar surveys in different fields of research, respondents see themselves as typical of those in their field, giving very similar times with data for themselves and for others.

Q32. To what extent, if any, has your use of EMBL-EBI services and resources changed your research efficiency (i.e., time saved compared to if no EMBL-EBI existed)?

To explore the impact of EMBL-EBI services on its user community, respondents were asked to estimate any resulting change in research efficiency using the following categories: negative change,

no change, 5 per cent time saving, thence 10 percentage point intervals to >90 per cent. Again, we converted >90 per cent to 91 per cent to be conservative.

As in similar surveys that we have run in different disciplinary fields, the reported efficiency impacts are considerable. Among those reporting a zero or positive impact, the mean was 46 per cent (N = 3239), with a median of 50 per cent.

Just 18 respondents reported a negative impact, which could not be included in analysis as it is not numerical. A few comments by these respondents suggested that they had spent time on training and familiarisation but had not yet used EMBL-EBI data and services enough to gain a benefit.

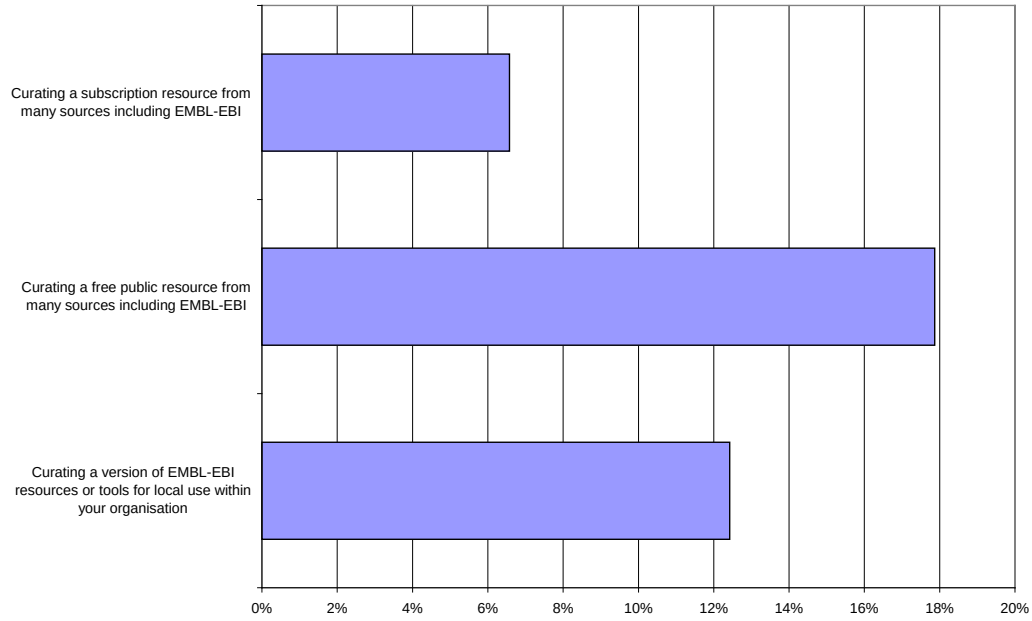
Curation and reuse

A number of users are known to obtain data from EMBL-EBI and curate it locally, making it available to other users in-house and/or incorporating it into data-based products and services used by others. As a result, the effective number of people using the data and services of EMBL-EBI is greater than the number of direct users. We included two questions in the survey to try to explore the extent of this indirect use.

Q33. Are any of the following part of your role: curating a version of EMBL-EBI resources or tools for local use within your organisation, curating a free public resource from many sources including EMBL-EBI, or curating a subscription resource from many sources including EMBL-EBI?

As expected, these forms of curation and sharing are relatively common, with 12 per cent reporting curating a version of EMBL-EBI resources or tools for local use within their organisation (N = 3534), 18 per cent reporting curating a free public resource from many sources including EMBL-EBI (N = 3548), and 7 per cent reporting curating a subscription resource from many sources including EMBL-EBI (N = 3469).

Figure A1.9 Curation and sharing of data obtained from EMBL-EBI (share of respondents, per cent)



Source: Authors' analysis.

Q34. If you use EMBL-EBI services and resources to curate a version for local use in your organisation, or to curate a public or subscription resource, can you estimate approximately how many users it has annually?

One-third of all respondents (1 396) answered this question and reported a number of indirect or secondary users, attesting to the widespread reuse of EMBL-EBI data not captured in its user statistics (e.g., page requests, download volumes, etc.).

However, the question was difficult to construct. Respondents were offered five curated user ranges: 1-10, 11-100, 101-500, 501-1000, and >1000. Translating these to mid-range counts and coding >1000 to 1001 unless a specific user count was given in comments (as was requested), makes the counts no more than indicative. Moreover, there were a small number of very high responses given in comments, with six responses of more than 10 000 and a maximum of more than 200 000. Indicatively, the reported mean count of secondary users was 678, and the median 55 (N = 1396).

Contingent valuation

The contingent value of a non-market good or service is the amount users are willing to pay for it and/or are willing to accept in return for foregoing it. Utilising this method requires specific wording of the questions as well as offering an opportunity for open ended comments to enable analysis of the thinking behind responses and the identification of protest answers (DTLR 2002). During analysis, 94 responses were identified as protest answers and not included in the analysis.

Q35. EMBL-EBI will never charge for services. Imagine you have the option to either carry on using EMBL-EBI or to sell your rights as an existing user to a third party. If you sold your rights to access EMBL-EBI, what is the *minimum* amount that you would be *willing to accept* as an annual payment in return for giving up all of your access to EMBL-EBI resources for a year?

The amount that users are willing to accept in return for giving up access is typically higher than the amount they would be willing to pay, primarily because the latter is constrained by what they can afford to pay.

Respondents' comments as to the rationale for their answers to these questions provide invaluable insights into their thinking about the value of such services. Among the reasons reported for being willing to accept only very high amounts in return for giving up access is the belief that the resource is invaluable, with respondents entering amounts in the millions of pounds. Another group of respondents thought through the implications of not having access, suggesting that they could not do their research without it, and putting in amounts equivalent to their annual or sometimes multi-year salary or research grants, with amounts ranging from around £100 000 to £1 million. Others did a range of "back of the envelope" calculations, such as the amount it would cost to obtain the data elsewhere or to create/collect it themselves.

The principal (and, perhaps, principle) reasons given for saying they would not be willing to accept anything in return for giving up access include such comments as "it's priceless" and that they believe that science and research data should be open and free and would not accept anything in return for it. There were 46 responses explicitly referencing Open Access/Open Data principles.

As a result, the range of amounts that users would be willing to accept is wide. The mean of values respondents would be willing to accept in return for giving up their access for a year was more than

£125 000 per annum, with a maximum of £70 million. The median value was £300 (N = 2079), and the difference between these mean and median values illustrates the wide range of responses.

Q36. EMBL-EBI will never charge for services. For this question, however, please imagine that access were no longer free. In this hypothetical case, what is the *maximum* amount you would be *willing to pay* as an annual subscription for your individual access to resources from the EMBL-EBI (or would ask your employer/funder to pay for a single user licence on your behalf)?

Similar reasons inform what respondents would be willing to pay as an annual subscription for access to EMBL-EBI data and services, with what they or their organisation could afford to pay being an oft-cited limitation. A number of respondents expressed a willingness to pay nothing because they believe that science should be open and data free.

The mean of the value respondents reported being willing to pay for access for a year was £1 628, with a median of £126 (N = 2191).

Appendix 2: A Summary of Survey Comments

The following responses have been selected from the comments in the survey. They illustrate a range of views, including testimonials, and comparisons with other services. All suggestions for improvements have been included. Selected comments are provided in Box A2.1.

Q6: What impact would it have on your work or study if you could not access EMBL-EBI services and resources?

Respondents were asked to estimate what impact (on a scale of no impact to severe impact) it would have on their work or study if they could not access EMBL-EBI services and resources, and 463 respondents added free text comments on this question.

Major themes included:

- EMBL-EBI services are important, essential, fundamental, or difficult to replace. They rely/depend on them for their own work (N=125);
- EMBL-EBI services are useful, valuable, or helpful (N=18);
- Use them daily, regularly, frequently, or routinely (N=63);
- Use multiple EMBL-EBI tools and services (N=52);
- Loss of EMBL-EBI would impact time, efficiency, or integrated workflows (N=20);
- Could find alternatives for some things (e.g., NCBI, UCSC, Galaxy, Genbank) (N=50);
- EBL-EBI services are high quality or superior to other sources, have unique elements, or are important to the scientific community as a whole (N=52);
- Free, Open, Sharing (N=22);
- Using them in teaching (N=23);
- Value the workshops, training, and courses (N=34); and
- EMBL-EBI helps them keep skills up to date (N=14).

Selected comments are provided below in Box A2.1.

Figure A2.1: Relative frequency of comments including selected keywords for Q: What impact would it have on your work or study if you could not access EMBL-EBI services and resources?

Source: worditout.com

Box A2.1: Selected comments on the impact of EMBL-EBI services.

Q6: What impact would it have on your work or study if you could not access EMBL-EBI services and resources? Selected comments:
A lot of the databases and repositories are cloned at the NIH or JGI in the USA but the quality/organization of the EMBL-EBI datasets is in many cases superior to others.
Access to a composite centralised database of information is absolutely essential in the design of genomic experiments.
All members of my lab (~10 researchers) rely on these resources on a daily basis.
An access to structured, computer readable data in a way that is stable over time is critical to the activity of our microenterprise.
as many of the services are open source so it put a great impact on scientific community to use your free services. Because it is not possible for every institute to buy costly software. So using free source is a great deal and the best thing is that the result we receive is very authentic and very efficient.
as molecular medical parasitologist EMBL-EBI services and resources constitute an important part of day to day work.
At the moment we do all our sequencing at the EMBL GeneCore, via my EMBO YIP membership. This is very cost-effective at very high quality. Without the GeneCore we would have to spent much more, meaning less experiments.

Q6: What impact would it have on your work or study if you could not access EMBL-EBI services and resources? Selected comments:
Because, today I and other researchers also we contribute indirectly when we perform our work making the data available in particular on public health. Of course, the development of tools and services provided by EMBL-EBI and other centres are extremely important because they affect not only the scientific community, as it would be a setback in access and availability of methods and open access to science and improvement of research. Its unimaginable.
Could not do my job w/out these resources. Period.
Doing alignments on my pc would take 1 day instead of 1 hour on EBI.
EBI services are very important to life science community. The tools that are continuously developed/created in consultation with user community continue to be useful in furthering the research in computational biology.
EMBL-EBI is a FANTASTIC resource, particularly in the realm of gene/protein function annotation and provides THE BEST gateway to other resources. Aside from previous downloads of EMBL-EBI resources (UniProt, InterPro) I remotely access/download from several of these services...frequently, as in I am doing that right now and using data from EMBL-EBI resources in pretty much everything else I am running.
EMBL-EBI is the main and most significant provider of free high quality and valuable bioinformatic data resources, education and services in Europe in the field of molecular biology.
EMBL-EBI resources are essential to all modern life sciences research. It would be impossible to work with data of any kind relating to genomes, genes, proteins, small molecules, 3D structures or other related data without reference to EMBL-EBI resources. We simply could not function without the core, reliably maintained data collections and the world-leading expertise that is centred on EBI for the organisation and analysis of biological data.
EMBL-EMBI services, along with some others like NCBI are indispensable in day-to-day activates. From simple tasks like sequence alignment (BLAST or Multiple sequence alignment) to major tasks like whole genome annotation and data retrieval for high throughput analysis of genomic/transcriptomic data using annotation files or visualization of results on Ensembl genome browser, EMBL-EBI resources and services are integral tools for any bioinformaticians irrespective of their specific field of work. Regular and free access of these are important to the community, and in cases like KEGG we have realized that by making its access restricted, scientific progress has indeed been affected. I strongly urge you to kindly maintain the EMBL-EBI services the way they are, because the community cannot afford to forfeit it!
[…] As a community scientist, I believe Europe PMC is a major community resource.
I am a public high school teacher in a high poverty school in the US. One of the classes I teach is a dual enrolment (with a local community college) year-long survey of biology with lab. I use EMBL-EBI services as well as a tutorial prepared to work with those services as an extended, several class day lab in this course. The school would not pay for access to a database. My student group, generally the top students in the school, includes future engineers/biologists as well as future informed citizens who should know something about bioinformatics.
I am a structural biologist and all my bio-information is drawn from EMBL-EBI services. I rely heavily on it to understand my protein targets and all projects benefit from data obtained from EMBL-EBI. From construct design, ideas as to how and where my target function down to providing my research group with known data and in writing papers with details on my targets.

Q6: What impact would it have on your work or study if you could not access EMBL-EBI services and resources? Selected comments:
I am very grateful to the EMBL-EBI for providing free of charge HIGH QUALITY tools and resources that have been tested and are also accompanied by both instructions and references explaining how they work and the limits of safe interpretation and use. These tools enable me to conduct higher quality research with greater insight.
I could use Genbank/NCBI for nucleotide issues but I NEED EMBL-EBI for lots of questions regarding protein (and proteomic) studies / translations / modifications / and also 3D modelling.
I routinely use multiple service provided by EBI […]. I also routinely incorporate the use of these EBI resources in helping other researchers to analyse and understand their high-throughput data sets.
I use EMBL-EBI services on a daily basis. Most of these cannot be sourced elsewhere on the internet and my institution lacks the resources to provide them locally. EMBL-EBI training sessions have been by far the best NGS and RNASeq courses I have attended.
I use many bioinformatics services via EMBL-EBI. If I could not access these, I would rely more on NCBI and smaller species-specific databases. My work efficiency would drop, and I would lose the functionalities that are specific to EMBL-EBI.
[…] EBI offers the largest collection of open access databases, software tools and educational programmes, all of which are well-curated and reliable sources of information. I use these resources on daily bases, and I would feel a strong negative impact on my work, if the EMBL-EBI services and resources become unavailable.
It would be a considerable loss not to be able to call on the expertise on the leading edge of research, and would be a huge setback to innovation and data extrapolation to both experience and novice bioinformaticians and researchers.
It would have a major impact as many systems would need to be migrated to NCBI services.
No access to EMBL-EBI services and resources = No access to updated public data EMBL-EBI is the heart of European Research in term of Bioinformatics and public Data Access.
Oh! My work entirely depends on services offered by EMBL-EBI, without which I would be a painter with both hands tied behind!
One of the best resources available for protein information (curated or computationally annotated) is UniProt. I routinely use this site for our research. I dont know of any other sites that could fill this role. […]
Services from EMBL-EBI are a well-established and trusted resource for bioinformatic and proteomic research, as well as for structural biology and drug discovery/design. Even more, it gets better every year, making life easier for users, and ultimately speeding up the research process.
Some aspects of the work that my colleagues and I undertake could be done via e.g. NCBI GenBank applications or other open access tools but EMBL-EBI has a particularly useful suite of bioinformatics tools and with some very user-friendly interfaces. For example, although using BLAST to check a sequence at GenBank can be quicker than using FASTA via EBI, the latter output can be more useful. Also applications such as MAFFT and pairwise alignment tools such as the EMBOSS suite, are extremely useful and easy to use. Hence, although on one level some aspects could be replaced, there would be a major impact for us as EMBL-EBI has all the tools in one convenient location.

Q6: What impact would it have on your work or study if you could not access EMBL-EBI services and resources? Selected comments:
The availability of such high quality (trustworthy, reliable, comprehensive, up to date) information underpins much of what we do in the lab. […]
The EBI, along with NCBI, are crucial repositories for genomic data. Organizing and displaying these makes the public investment in generating them far more valuable than they'd be otherwise.
The EMBL-EBI services and resources are absolutely necessary to the scientific community, it would be impossible to make research without its resource and services.
The EMBL-EBI services are much less convenient to use the NCBI counterpart.
The EMBL-EBI services are the backbone of most if not all computational biology projects. It's an invaluable resource that many researchers don't often think about. Countless applications and software frameworks rely on databases and back-end web services provided by the EMBL-EBI.
the provision of open high quality data sources is vital for scientific research. Funders need to do more to recognise the value of data services, and should realise that the challenges of handling big data depend on high quality freely available data and data services in the public domain.
These services and resources allow a small lab like mine to have near instant access to a wide array of information, computer algorithms and analytical tools not normally available. They help us stay up-to-date and competitive in a very fast moving research environment. It is difficult to imagine doing the type of research that we do without these resources - they are highly valued and much appreciated!
[…] the entire proteomics community benefits from quality databases and mechanisms for dataset sharing. The EBI is an essential participant in the HUPO Human Proteome Project. EBI also stimulates cross-omics analyses.
We rely on many of the informatics services and databases that EMBL-EBI support, and also for the on-going development of new platforms and methodologies to support the ever changing data landscape. They lead the field in providing real-world solutions that are readily accessible to the unskilled and still useful to the skilled researcher Their activity in development of data standards is also crucial to our work.
We work in public sector R&D policy and project monitoring and EMBL-EBO is one way in which we gain insight into leading-edge thinking in an important aspect of informatics, which touches on our work in and Big Data for example.

Box A2.2: Q6: Impact of service access: Suggestions for improvements:

Q6: What impact would it have on your work or study if you could not access EMBL-EBI services and resources? Suggestions for improvements:
Make a search like NCBI where I can search in all resources of EMBL-EBI. Remove useless stuff from startpage (think Google, less is more).

Questions 7 to 19 covered the frequency of use of the individual services. While some of the comments (detailed below) were illuminating in terms of explaining what users liked (or did not like) about the services, they did not lend themselves to classification which could shed light on economic benefit.

Q8: Approximately how frequently did you access/download Literature or JavaScript services in the last 12 months?

Sixty-two comments were received overall, and selected comments are in Box A2.3.

Box A2.3: Literature and JavaScript services

Q8: Approximately how frequently did you access/download Literature or JavaScript services in the last 12 months? Selected comments:
For searching of an appropriate literature I am using other browsers as well (usually the goggle scholar)
I default to pubmed for literature searches, however google scholar does take me to EPMC.

Q9: Approximately how frequently did you access/download Ontology services in the last 12 months?

Forty-four comments were received overall, and selected comments are in Box A2.4 & A2.5 below.

Box A2.4: Ontology services

Q9: Approximately how frequently did you access/download Ontology services in the last 12 months? Selected comments:
GO is hardly an EBI service; it's an international collaboration, only a very small part of which is carried out at the EBI. I know EBI likes to take credit for as much as possible, but this is just insulting the community's intelligence.

Box A2.5: Ontology services: suggestions for improvements

Q9: Approximately how frequently did you access/download Ontology services in the last 12 months? Suggestions for improvements:
Needs to be crossed reference available in the Genbank file and the view of the flat file of Genbank files need some light tool like find by specific pattern which available in NCBI.

Q11: Approximately how frequently did you access/download resources from the following Genes, Genomes and Variation services in the last 12 months?

Seventy-two comments were received overall, and selected comments are in Box A2.6 & A2.7.

Box A2.6: Q11: Genes, Genomes and Variation services

Q11: Approximately how frequently did you access/download resources from the following Genes, Genomes and Variation services in the last 12 months? Selected comments:
Ensembl directly supports my research; while broadly comparable to UCSC its insights are much more accessible.
Have given up using EGA because managed access procedure outdated and not workable.
I accessed it many times until I could download the data.
I f****** love everything about ensembl. It is probably THE BEST tool out there. No one else has managed to put together that much information with such flexibility and capability for analysis, all while maintaining such a high degree of user friendliness. As someone who doesn't work with humans or mice, ensembl is invaluable. No other resource has taken the time to do such a great job for the other mammalian species out there. Seriously if it weren't for ensemble, many of the projects I have worked on and are working on would have been much more difficult and would have had lower quality results and performance. If there wasn't Ensembl/ensembl genomes and I had to wade through the sluggish, inconsistent and painful NCBI stuff instead, I'd probably find a different job.... Fund ensembl till the sun explodes!
I use the Gramene portal quite a lot in my research, it's much better than Phytozome for comparative work across species and actually uses useful annotations for its sequence files. Not quite sure how it relates to Ensembl Genomes, but it seems to be much the same thing.
I used GENCODE, whose gene model is based on Ensembl. I have not used Ensembl directly (i.e., fetching data from ensembl.org rather than other third-party websites).
While I don't use all of these tools directly, others do feed data into tools I use. For instance, I don't often access HGNC, but the standardized gene symbols are critical and accessed through Entrez Gene.

Box A2.7: Q11: Genes, Genomes and Variation services: Suggestions for improvements:

Q11: Approximately how frequently did you access/download resources from the following Genes, Genomes and Variation services in the last 12 months? Suggestions for improvements:
Genomes of cattle need to be added.
[…] the Perl API and Ensembl BioMart Perl API both need some serious work. The BioMart API is very inflexible and provides no error handling, so using it to retrieve larger and more complex datasets can require multiple runs. You can end up with incomplete data sets and not have any idea that a problem occurred. The documentation on both needs improvement. It would be great if you could support other languages (e.g. Python) or have a *nix command line tool like entreztools.
It is annoying not to be able to blast in a specific project assembly project in ENA

Q13: Approximately how frequently did you access/download resources from the following Proteins, Proteomics and Enzymes services in the last 12 months?

Fifty comments were received overall and selected comments are in Box A2.8 & A2.9.

Box A2.8: Q13: Proteins, Proteomics and Enzymes services

Q13: Approximately how frequently did you access/download resources from the following Proteins, Proteomics and Enzymes services in the last 12 months? Selected comments:
Pfam is an amazing resource for someone working with novel species. Pfam allows me to make much more refined annotations and interpretations of de novo transcriptomes. Without it, I would be limited to taking crap shoots with BLAST. Pfam lets me more accurately assess how close two potential homologs are and resolve ties or other instances where homology is confounded. Uniprot is a wonderful resource. The UI is intuitive, searching is very easy, and so on. If I'm looking at/for proteins, I go straight to Uniprot.
UniProtKB is my favourite portal to use, because when it's good it combines every relevant link into one.
UniProtKB is the engine that makes work on less-studied organisms possible. Our research program would collapse without it.

Box A2.9: Q13: Proteins, Proteomics and Enzymes services: Suggestions for Improvements:

Q13: Approximately how frequently did you access/download resources from the following Proteins, Proteomics and Enzymes services in the last 12 months? Suggestions for Improvements:
As a crop researcher, I would like to see more detailed entries for non-Arabidopsis proteins, maybe with links to other useful portals and some more pre-run analyses (especially on the various phylogeny servers). Even so, it's been indispensable to my work.

Q15: Approximately how frequently did you access/download resources from the following Chemistry, Reactions/Pathways, Interactions and Metabolomics services in the last 12 months?

Twenty-six comments were received overall, and selected comments are in Box A2.10 below.

Box A2.10: Q15: Chemistry, Reactions/Pathways, Interactions and Metabolomics services

Q15: Approximately how frequently did you access/download resources from the following Chemistry, Reactions/Pathways, Interactions and Metabolomics services in the last 12 months? Selected comments:
As much as I like Reactome, the quality of the data available for plants is not exactly great.
I wish Metabolights to be established as a standard metabolomics data repository.

Q17: Approximately how frequently did you access/download resources from the following Model Organisms and Biological Systems services in the last 12 months?

Fourteen comments were received overall, and selected comments are in Box A2.11.

Box A2.11: Q17: Model Organisms and Biological Systems services

Q17: Approximately how frequently did you access/download resources from the following Model Organisms and Biological Systems services in the last 12 months? Selected comments:
Pombase is fantastic, and there is no similar service in the world.

Q19: Approximately how frequently did you access/download resources from the following Taxonomy and Sequence Analysis services in the last 12 months?

Twenty-nine comments were received overall, and selected comments are in Box A2.12.

Box A2.12: Q19: Taxonomy and Sequence Analysis services

Q19: Approximately how frequently did you access/download resources from the following Taxonomy and Sequence Analysis services in the last 12 months? Selected comments:
At what point did NCBI taxonomy become an EBI resource? I use the NCBI web pages. See my response above regarding the Gene Ontology.
The main search tool of www.ebi.ac.uk is excellent - the easiest way to get to which ever data resource one needs for a particular gene/protein.
The servers for searching and sequence analysis tools are wonderful. Normally one would have to rely on installing tools locally, which can be a huge pain, or hoping that some small server in an academic lab is still funded each time the tool is needed. With EBI this isn't the case, it provides great performance for a huge set of tools. For Rfam, see my comments about Pfam. NCBI Taxonomy is critical.
Typically get NCBI taxonomy from NCBI, not from EBI, unless the specific use is linked to an EBI service.
why replicate NCBI taxonomy?

Q21: How long did it take you to find and access the last resource you used from EMBL-EBI?

Two hundred and fifty-two comments were received overall, and a word cloud of the results can be seen in Figure A2.2 below with selected comments in box A2.13 & A2.14 below.

Figure A2.2: Relative frequency of comments including selected keywords for Q21

Source: worditout.com

Box A2.13: Q21: Time to find and access EMBL-EBI resource

Q21: How long did it take you to find and access the last resource you used from EMBL-EBI? Selected comments:
Access is generally straightforward, although once deeper into data it can become a little confusing.
Array express has been and is spot on! More so, after it started to mirror what's in GEO as well. I use ArrayExpress mainly for relevant short read sequencing experiments (RNA-seq and ChIP-seq) as finding these from ENA due to their ill formed meta-data handling is a headache.
As a reviewer or editor I sometimes try to evaluate the MS data of a manuscript. I would like to repeat a database search with mgf files. […]I am wondering that most of the reviewers are still not interested to see the MS data.
Googled Uniprot ---> Searched protein ----> retrieved sequence.
I am very pleased with the search functions in ArrayExpress because it's far superior than NCBI GEO.
I needed to compute LD for EUR panel of 1000 genomes between 42 SNPs we are interested in. I found no help how to do it on 1000 genomes homepage, but BioStar forum had some codes/suggestions. It took me a long time (about 2 days) to ftp 1000 genome data. And still I could analyze only 41 SNPs (now I am investigating if vcftools just silently skipped rs4795541). I heard "DataSlicer' mentioned, but could not find any instructions how to restrict my download only to haplotypes overlapping our 42 SNPs of interest.
I prefer the previous BioMart version. We need to access to historical records!!
I reside in a remote island where the internet connectivity is poor still it works very well thanks to your server.
I wanted to show some co workers that an antibody they were trying on mouse tissues wasn't working because mice do not have that gene. With ensembl I was able to pull up the human gene (CD66b), identify the region it was located, identify the syntenic mouse region and pull both into a great visualization showing an alignment of each region. While I was playing with the situation above, I used the ensembl MySQL server to obtain gene-

Q21: How long did it take you to find and access the last resource you used from EMBL-EBI? Selected comments:
transcript-protein annotations for all of my reference genes for an upcoming project. If it wasn't for the website access and visualization tools and access to the mysql database, these tasks could have easily taken a day or two.
It depend on the local internet connection where I am.
It's bookmarked, so this isn't a particularly useful survey question. If I were to search EBI pages I'd usually use Google […]
It's fast, but some genomes are not annotated well, example rat sequences.
Sometimes, it is difficult to access all the relevant annotations. For example, in Plant Ensembl Genomes to find out a function of a transcript, the page needs to be reloaded as it is on a separate page, then to see its annotated domains, again new page loads and that makes the process slow. […]
The services are quick and instant, with no long waiting hours to get the results, the servers are efficient and maintenance is excellent. Never ever had any problem in working with EBI's resources.
UniProtKB is very straightforward. Some improvements for heavy users would be very useful, for example a list of favorite proteins, that you can access from the first page. That would make it even easier to get around.
Usually can't download any articles, just read the abstracts, which is rather frustrating.
Very happy with all of the databases and downloadable resources from EMBL-EBI.
While the EBI home webpage often tries to hide all its resources, Ensembl itself is generally very accessible, especially with a little experience.

Box A2.14: Q21: Time to find and access EMBL-EBI resource: Suggestions for improvements

Q21: How long did it take you to find and access the last resource you used from EMBL-EBI? Suggestions for improvements:
[…] many pages contain large amounts of data, that can take time to sift through to find what you want - could do with better indexing and structure.
Finding the same experiments through ArrayExpress is quick and easy. The Samples (BioSamples) database should have this information as well but the link to the database on EBI main web page is never easy to find.
Access to the mgf files is sometimes impossible and sometimes they are difficult to find. This should be improved.
improve navigation architecture see: http://bioinformatica.ucr.ac.cr/variationviewer/ XXXX: XXXX
Please do not underrate the integration with Google, Wikipedia and other public resources.
Re ENA (short reads): It is not very intuitive to use, and difficult to find anything such as a studies on an organism. Unless you have the accession numbers.
ENA search: how can I download the results in FASTA or any standard data format from the web? My 'last time' was a 'give-up'.

Q22: What mode of access did you use to obtain data from the last resource you used from EMBL-EBI?

Ninety comments were received overall and selected comments are in Box A2.15 &A2.16.

Box A2.15: Q22: Mode of access to data from EMBL-EBI

Q22: What mode of access did you use to obtain data from the last resource you used from EMBL-EBI? Selected comments:
LD blocks ”around“ single SNP is available in Web browser, but download button did not work. API is soooo frustratingly complicated! so ftp, vcftools and plink is what I have to resort to.

Box A2.16: Q22: Mode of access to data from EMBL-EBI: Suggestions for improvements

Q22: What mode of access did you use to obtain data from the last resource you used from EMBL-EBI? Suggestions for improvements:
It would be good if EBI linked to programmes in the way that NCBI GenBank has with MEGA 6 where you can add sequences from GenBank directly into your alignment via a plug-in to the MEGA 6 program.

Q24: If Yes, approximately how much time would it have taken to obtain the resource you last used from another source?

Five hundred and fifty comments were received overall, and selected comments are in Box A2.17.

Box A2.17: Q24: Time to obtain data from another source

Q24: If Yes, approximately how much time would it have taken to obtain the resource you last used from another source? Selected comments:
Ensembl genome browser shares data with UCSC genome browser. Both are bookmarked in my web browser. However, the focus of the two tools is different. If I look for information about specific genes/transcripts/proteins together with sequence data and a visual presentation, I turn to Ensembl. If I would like to see epigenetic datasets or genomic regions outside of genes, I use UCSC genome browser. It takes several more clicks in the UCSC genome browser to obtain similar gene information as in Ensembl.
Ensembl Genomes (e.g., bacteria) are very easy to access in bulk due to excellent consistency in annotation and naming schemes. This is very precious.
I could have obtained this information through extremely slow manual parsing of genbank files that would have taken ages to do and would have ended up with inaccurate and incomplete data. Since NCBI only offers individual FTP downloads for a handful of model eukaryotes, I would have been stuck downloading a huge f******amount of data, most of which I’d subsequently throw out. Or, I could use entrez tools and try and download all of my transcripts and protein sequences for my dozen or so reference species using Entrez. This would have been limited to 3 downloads per second. Then I’d have to sit there and figure out which of the 30 files having the same sequence name was the right thing. I’ve tried it before, when trying to download transcripts for rhesus macaques, I ended up with over 500k genbank files, it seriously sucks. When I wanted to do this on ensembl it took me 10 minutes to download the data by FTP and then another hour to learn the SQL schema and pull my mappings. Even better was that I could use the same MySQL schema to download the NCBI accessions for those same sequences in an afternoon. This was for all 12 reference species.

Q24: If Yes, approximately how much time would it have taken to obtain the resource you last used from another source? Selected comments:
I could have used the NCBI RefSeq Protein database, which I find messy and poorly annotated. I don't remember exactly which resource I was looking for, so I can't say if I was looking for more information than just the sequence. The annotation and links to relevant databases are the main reasons why I prefer UniProt over RefSeq. No costs would have been involved, except time and effort.
I have been comparing various aspects of different genes that have come up in a screen. If I didn't have pombase, I would have to manually enter information on each gene on different websites (e.g. blast, uniprot, string etc.), some of which I am not familiar with, and wait for things to load. Palmas provides all this information for each gene on one page.
I last used UniProtKB to find taxonomic information for a list of protein identifiers and summarize the results. My previous methodology involved querying several databases including NCBI, and this has complications because other databases do not provide all of the cross-references and taxonomic information in one place. The cost involved for not having this resource is more time developing methods and running analyses, and this has a major impact on my ability to make discoveries in a timely manner.

Q26: If Yes, approximately how much time would it have taken to collect/recreate the resource you last used?

Two hundred and twenty-five comments were received overall, and selected comments are in Box A2.18.

Box A2.18: Q26: Time to collect/recreate resource

Q26: If Yes, approximately how much time would it have taken to collect/recreate the resource you last used? Selected comments:
2-3 days from raw data, but >90% of data, programs can be picked from other places w/o costs, if EBI did not exist (e.g. pfam db), than most likely authors would move most valuable resources somewhere else, currently this is very small cost to pay for hosting, but problem is to find reliable one where everything will work for years (only such services will get citations) and EBi is quite good at that.
even with EU funding like the EMBL-EBI this would take years and years (10+?) data collection on this scale would be impossible to cost out on a private network.
if I had not been able to retrieve a model from biomodels database it would have meant to identify the original publication (which may be obscure), encode the model manually, and replicate the curation process that was performed by the biomodels team in order to verify it. Costs would be mainly salary of the person wbo who would have to do it.
First I would get some humans, then I would invent sequencing technology, then I would sequence them, then I would hire a thousand bioinformaticians to figure out where the genes were and the transcripts and so on. Then I'd have re-created Ensembl.
Good lord, please don't make me recreate Pombase from scratch. There's just too much information.
Grant proposal, ethical application, experimental design, study subject enrollment, instrumental analysis.

Q26: If Yes, approximately how much time would it have taken to collect/recreate the resource you last used? Selected comments:
Immense costs.
I can't make a genome karyotype by myself.
I could have used a different software package - e.g. MEGA6 that uses a different alignment programme [i.e. Clustal and Muscle] but it would not be as authoritative as MAFF-T.
I could not recreate UniProt myself, and that's just one of the resources I use.
I don't think I can personally replace 15 years of Ensembl coding :)
I will need to download the data annotations and then use personal script to extract it in format which has only the required information in it and not DB specific. Finally I will need to run enrichment analysis myself and figure out ways to visualize it.
I work as a curator so have a good idea of the amount of work involved in this. Recreating the entire entry would require going back to the original papers and other resources to extract all relevant information. Cost of curator salary for 12 hours plus price of access to journals behind paywall: very roughly £200.
I would have grown the B73 plants, collected tissue in liquid nitrogen, and processed it into RNA. Then, I would amplify the sequence of the gene using custom primers and clone it into a pENTR vector, which I could sequence. All of these materials and reagents are ones my lab regularly stocks, so there wouldn't have been any large extra costs except for time. Now, if we were talking about the kind of thing I normally use the EMBL-EBI portals for, namely looking at families of genes across multiple species, the project would begin to take on a scale of multiple months to a year and become quite a bit more expensive.
I would have had to isolate bacterial strains that secrete the enzyme that i am interested it and then sequence that gene myself to create a database of that gene.
if I want to find tissue specificity data for a specific gene in a specific animal-model, i'd have to do up to a weeks worth of literature searching and reading. If I want to compare one species to another species, i'd have to all of that again - maybe another weeks work. Uniprots curation of genes & literature and all that is ln own about each gene - across all models - is absolutely fantastic - so within minutes - I have all the information I need. It is an essential resource for my daily curation duties.
If the dataset was not in NCBI GEO, I would probably have tried to get the data directly from the author, which could be really quick if they have it on an institutional website, or very slow (months) if they can't locate the data in their archives/collaborator has the data, or never if they don't want to share.
It is a ball park estimation. I used the text mined option for accessions for PMC articles. While I could download PMC from the US side, the TM module is not available there. I would have to develop some method to find the articles that have database accessions.
JGI and NIH offer similar repositories but a great automatic way of accessing EMBL-EBI data is through BioMart making it really convenient and faster.

Q26: If Yes, approximately how much time would it have taken to collect/recreate the resource you last used? Selected comments:
My work involves comparative genomics, the precomputed alignments, cohesive transcript naming system, and REST architecture makes Ensembl very easy to access programmatically. This could be pieced together using other services and/or built from scratch but it would have been harder or significantly harder depending on the route taken. The time required is for creating a custom resource for myself.
Probably it could be done, but it would take a considerable amount of time. First I would need to translate all DNA sequences to the protein sequences and then predict transmembrane domains and signal peptides.
protein x-ray crystallography - not by myself, but in collaboration with close colleagues. Costs would be at least a full person-year of labour (€50-100k) and possibly €50k or more of consumables.
Recreate Uniprot? You're kidding right? :) […]
Recreating the last used resource would have required going through multiple scientific papers, if available, and reproducing all the steps of the original authors. However in case the steps were not properly documented additional discussions with the original authors would have been required. Therefore the approximation provided above (i.e. 20 days) could be much larger in reality depending on the obstacles which could be encountered during the recreation process of the last used resource.
Sequencing of large pieces of mouse chromosomes at sufficient quality to rely on it for making genome editing decisions would take a significant body of time and money. This does not then take into account the extra info that ensembl provides in terms of annotation and the ability to examine homology and repeat sequences etc, etc.
The information I looked for is possible to obtain from the literature, but it is quite painful.

Q27: To what extent do you benefit from using EMBL-EBI in any of the following ways?

One hundred and twenty-nine comments were received overall, and selected comments are in Box A2.19 & A2.20 below.

Box A2.19: Q27: Benefits of using EMBL-EBI in specific ways

Q27: To what extent do you benefit from using EMBL-EBI in any of the following ways? Selected comments:
Access to training is difficult […] the course are oversubscribed. Why one student is chosen over another is unclear.
Acquisition of information on topics of cutting-edge research; motivation to maintain high standards and keep on ameliorating myself as a researcher; insight regarding future directions that help me take my research to the next level, especially regarding Regenerative & Translational Medicine issues in Orthopaedic Surgery & Traumatolgy.
as part of the industry consortium I am able to discuss industry relevant scientific approaches.
Data - very high benefit. Tools - low benefit without good instructions [...].
data sharing is essential to the function of our database, as we pool data from many sources, all about one model organism. The tools Data and support EMBL-EBI provide gives us the information we need to support 1000s of scientists hundreds of labs, not just in the west, USA and Europe, but worldwide, especially in developing countries where science & biomedical research doesn't get the support it needs to flourish.
EMBL-EBI is not very open to partnering with corporate players.
I didn't know many services that are shown in this survey. I would have used these more often if I knew them.
Negative benefit - in the past I have had to fend off the ebi tentacles trying to incorporate my own resources without giving appropriate credit.
The EBI is the ONE unique great place in Europe to do bioinformatics. Beyond all the above benefits, another extremely important one is that people who did their PhD, post-docs, or worked at EBI, are top bioinformaticians. EBI is essential for the European community of bioinformaticians.
The EMBL-EBI is a critical mass of bioinformatics expertise and as such one of the few ways to plan for the unplannable and remain competitive. It is also a phone number for any biocomputational problem at hand.
There are support members who post on the Biostars bioinformatics website, they're incredibly helpful.
trust integrity of databases and resources - level of confidence that the tools are appropriately validated and resourced (unlike with third party repositories)
User support is generally poor, navigation is very difficult, interface is not user friendly, will use UCSC resources first and only EMBL-EBI if not available at UCSC. Having said all of this collaboration with EMBL-EBI has been very valuable for the project, annotation of a major mammalian genome. But comunication is poor and no real involvement in the process, even though I am the leader of the genome assembly project within the consortium. It is very much a "this is how it is" proposition. Of course pretty much the same comments hold for the NCBI and UCSC, we get what we are given, and we are grateful for what we receive. Without them we would be way behind were we are with them.

Box A2.20: Q27: Benefits of using EMBL-EBI in specific ways: Suggestions for improvements

Q27: To what extent do you benefit from using EMBL-EBI in any of the following ways? Suggestions for improvements:
Could you please make public web-course on how to solve ”typical” data-mining problems?
I am looking for EBI come up with image retrieval service like KEGG pathway database
I find that your training modules, after you took down the 2CAN support portal are vastly inferior […]. Please bring it back!
Some data provided should be clearer about how they are generated and what they contain. e.g. the Humsavar Uniprot file is not clear how it was generated.
Training is always aimed at ‘young/early career researchers’. In this age of an uncertain job market these courses, which are always excellent, should be open to all - including those moving between closely related fields.

Q30: Can you estimate the approximate share of your total research working time spent with data during the last twelve months (e.g. creating, manipulating and analysing data)?

One hundred and twenty-six comments were received overall and selected comments are in Box A2.21.

Box A2.21: Q30: Estimation of research working time spent with data

Q30: Can you estimate the approximate share of your total research working time spent with data during the last twelve months (e.g. creating, manipulating and analysing data)? Selected comments:
Always use NCBI if possible as layout is better and easier to use. Use ENA to submit raw data before publication but not downloading.
EMBL/EBI tools used, all the data came from NCBI/Entrez. For me EMBL/EBI nucleotide searches are a disaster since EBI SRS stopped. Search criteria are either too specific or too general, simple use of keywords produces inapplicable data - the same keywords are successful with NCBI/Entrez!

Q31: What do you think might be typical for others in the same research field?

One hundred and twelve comments were received overall and selected comments are in Box A2.22

Box A2.22: Q31: Typical working time with data for others in the same research field

Q31: What do you think might be typical for others in the same research field? Selected comments:
accessing ESEMBL by command line is just way to hard compared to e.g. Genbank or Uniprot where you have direct download based on IDs. My view on EBI in general is that you like it if you are a 100% bioinformatician who programmed from before you could walk but the doorstep is just way to high for the rest of us. When your software for e.g. ENSEMBL access take to long to start using... i just go to another shop as e.g. Genbank who provide the same product but in a more customer friendly fashion.

Q31: What do you think might be typical for others in the same research field? Selected comments:
I have a feeling most people in US just do not use EBI resources directly.
I think some people I work with would benefit more from using EMBL-EBI data but they seem to be unaware of it.
I usually prefer other data sources such as UCSC, SRA, GEO, ENCODE. But I know that many colleagues like ENSEMBL or other services hosted at EBI.
I work with plants - generally speaking, the EMBL-EBI platform is quite limited for plants, especially non-Arabidopsis (which is a lot of research nowadays!).

Q32: To what extent, if any, has your use of EMBL-EBI services and resources changed your research efficiency?

One hundred and twenty-nine comments were received overall, and selected comments are in Box A2.23 & A2.24 below.

Box A2.23: Q32: Research efficiency

Q32: To what extent, if any, has your use of EMBL-EBI services and resources changed your research efficiency? Selected comments:
EBI promote strict standardization, which hugely improved my efficiency and everyone in this field's efficiency.
EMBL-EBI service is much faster than other tools.
For certain things it's absolutely critical - whereas other activities there are other sources - an average doesn't really make sense. In some ways it *changes* what is possible - not just allowing the same things to be done more efficiently.
Greatly increased the specificity of searches to specific gene/protein functions, this is of massive importance due to the large number of missed assigned sequences in general databases.
I can focus on data analysis and interpretation instead of data collection and curation. EMBL-EBI takes care of the fundamental and crucially important collection, curation, and distribution aspects.
If the resource was not there we could not look at targets that we would otherwise examine, therefore the impact is not on time but on the quality of our research.
In general all of the EBI tools I use are incredibly useful and are easy to use, so there are huge time savings there. If it weren't for ensembl I'd have to spend huge amounts of time manually creating my own ensembl-like resource. Which I've tried and gave up after encountering enormous headaches with entrez. With tools like Pfam and Rfam, I simply wouldn't have anything like them if EBI didn't.
It is an excellent resource with impressive interface […]
It is nice to have different sites in case one service is down.

Box A2.24: Q32: Research efficiency: Suggestions for improvement:

Q32: To what extent, if any, has your use of EMBL-EBI services and resources changed your research efficiency? Suggestions for improvement:
Can you look at speeding up ensembl web browser interface - it seems to have gotten slower recently. It makes me prefer UCSC genome browser.
I would like to have the possibility to extract flat files from BioMart for the domain *coordinates*, associated for each transcript. You have this option in BioMart for various parameters of genes, transcripts, proteins, but the option domain *coordinates* is absent. Perl scripting/SQL-query access is not optimal decision for me working with a small subset of (~5000) transcripts. Thanks.
[…] ArrayExpress and Expression atlas need proper visualization tactics.

Q34: If you use EMBL-EBI services and resources to curate a version for local use in your organisation, or to curate a public or subscription resource, can you estimate approximately how many users it has annually?

One hundred comments were received overall, and selected comments are in Box A2.25 below.

Box A2.25: Q34: Curation of data for local use

Q34: If you use EMBL-EBI services and resources to curate a version for local use in your organisation, or to curate a public or subscription resource, can you estimate approximately how many users it has annually? Selected comments:
I would like to explore synergies between the European Patent Office data and EMBL-EBI.
Our supercomputing facility has more than 5000 users in Norway and also serves important International collaborators in the life science computing field http://www.uio.no/english/services/it/research/hpc/abel/
The fasta.bioch.virginia.edu web site relies extensively on EBI resources for domain annotation, and has more than 3500 users per month.
we have 1600 registers users, and these are mainly principle investigators, but have about 900-1000 users per day, and about 3,500 regular hits from the same set of IP addresses. So most people who use our service are not ”registered“. Annually, that translates to be ~360,000+ use instances, but we estimated about 2,500 people use our service as an essential part of research -accessing our website daily to 3-5 times per week, and the remainder - about 1000 people - use it regularly (ie 1-5 times/month).

Q37: 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.)?

As noted in Annex I (above), 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 foregoing it. The method requires specific wording of the questions and an opportunity for open ended comments to enable analysis of the thinking behind responses and the identification of protest answers (DTLR 2002). During analysis, 94 responses were identified as protest answers and deleted.

The amount that users are willing to accept in return for giving up access is typically higher than the amount they would be willing to pay, primarily because the latter is constrained by what they can afford to pay.

Respondent comments as to the rationale for their answers to these questions provide invaluable insights into their thinking about the value of such services. Among the reasons reported for being willing to accept only very high amounts in return for giving up access is the belief that the resource is invaluable, with respondents entering amounts in the millions of pounds. Another group of respondents thought through the implications of not having access and suggested that they could not do their research without it. They then put in amounts equivalent to their annual or sometimes multi-year salary or research grants, with amounts ranging from around £100 000 to £1 million. Others do a range of ”back of the envelope“ calculations, such as the amount is would cost to obtain the data elsewhere or create/collect it themselves.

The principal (and, perhaps, principle) reasons given for saying they would not be willing to accept anything in return for giving up access include such comments as “it’s priceless”, and that they believe that science and research data should be open and free and would not accept anything in return for it. There were 46 responses explicitly referencing Open Access/Open Data principles.

Similar reasons inform what respondents would be willing to pay as an annual subscription for access to EMBL-EBI data and services, with what they or their organisation could afford to pay being an oft-cited limitation. There were also a number expressing a willingness to pay nothing, because they believe that science should be open and data free.

Figure A2.3: Relative frequency of comments including selected keywords and phrases for Q37: 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.)?

Source: worditout.com

Box A2.26: Q37: Basis for answers

Q37: 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.)? Selected comments:
[…] $90,000 would be enough to pay a bioinformatician for two years so they can rebuild as many of these resources as possible […]
#34 based on guestimating the cost per year of early retirement buyouts. #35 is based on EBI being at 2x more valuable to me than an annual subscription to Nature [...]
(33) These resources are so valuable to my work, that I can't function easily without them. My answer is the portion of my pay that goes to time I would spend collecting this data otherwise. […]
[...] What matters is cost (money) vs. benefit of having access to EMBL-EBI databases instead of having to gather data elsewhere (but note that when access to these databases would be restricted, I expect other institutes to take over functionality, so within months/few years I expect to have regained the lost functionality)
1. 200K per year for a substantial period of time, is based on the idea of giving up my freedom to use all the data and resources available at EMBL-EBI, and generate all the hypotheses and thinking based upon that resource. It is basically the amount I would accept to very reluctantly leave Omics based science and develop another career. 2. Nothing- this is fundamentally wrong. I understand the question, and the need to imagine this access fee. I would say that it is simply wrong for individual researcher's creativity to be blunted by the requirement for such charges. It is not about affordability- my wealthy European university may be able to afford it, but what about my collaborator whose university cannot? Also, it calls into question the ownership of data that in my opinion, belongs to all humanity.

Q37: 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.)? Selected comments:
[...] in the long run, other solutions would be sought i.e. other databases/services across the world. If that would fail because those would start charging fees as well, than likely biological research in our institute would badly suffer up to the point of abandoning biological research altogether. [...]
[…] seems comparable to TV licensing. In your thought experiment, I would expect you to charge according to revenue though, i.e. nothing for academic, lots for big corporations (I believe Decipher does this).
31. - I would not accept payment in return for this right - as an academic I want and need to do research, which I cannot do without this resource. If you need a number I would put it at roughly £1.5m - what I would earn until retirement, so that I could retire today :-)! 32. - the question is for an individual access license, which is not our preferred model. My group has currently ca. 20 members, and for a group license I would be willing to pay ca. £1000-1500 a year (which is the maximum we also pay for most software etc. under academic licenses)
31. Can't do my job without Ensembl so they would have to pay me my salary […]
32: Since I would not be able to do my job, I would need a minimum salary... […]
33 is lower than 34 because if I were individually offered a financial inducement to stop using EBI resources I could probably find alternatives [...] For a site license I think we would pay pretty substantially - e.g. I would value the EBI tools more highly than HGMD profession (which we pay for).
34 - Amount I would be willing to accept annually in order to find myself a job outside of academia […]
5000 is the cost of the time required to reconfigure research and teaching/training activity to operate without EBI access. 600 is the cost of my labs annual subscription to our central molecular biology lab and its core equipment and technical staff resources. Loss of either the Central lab or the EBI would have a similar impact on our ability to work efficiently and cost effectively.
50000 because without EBI services I cannot perform my daily job therefore I could just as well stop working and 100000 then seems a nice compensation. 2500 because that is 10% of my actual annual income ... I'm afraid there is no more to offer...
amount I would ask my funder or organisation to pay for me as a single user since the department (not the entire institution, i.e.: the university) would have to cover for single user licences for >100 staff members
[…] that moves the total amount needed to about US$200000, more that a US funding agency would part with. Sorry; I'm not playing the game by the rules but don't see how to.
[...] the amount would change dramatically if NCBI/DDBJ didn't exist (or started to charge!)... No access to data (neither at EBI, NCBI, DDBJ or elsewhere) simply means none of my current research would be possible. Full stop.
amount I would want to give it up is based upon first downloading everything i need first and should be alright for a year or two until gets too out of date. [...]
[...] It would cost a huge amount of money, time and resources to investigate alternative sources, re-write parsers, refactor pipelines, pay for some key curations etc. a complete guess of in the magnitude of £500,000-£2million one off and £100,000 a year to maintain alternatives/+ a little curation. Vendors with less data and tools than the EBI charge £500->£10,000 a seat.

Q37: 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.)? Selected comments:
Based on how much providers of specific software (e.g. Ingenuity Pathway Analysis) are asking per annual subscription.
basically I would need to hire a bioinformatic engineer. […]
Because I work for a micro biotech company and rely exclusively on being able to use freeware. [...] I really value EMBL-EBI as a source of data - the company would be crippled without free access to this information and expertise - the financial model of a micro company is based on free access to data and services!
Because these services are used by millions of people worldwide, I thought of a number that would reflect the amount of work and money that is required to keep all these services functioning and may help subsidize third would country users. [...]
[…] I would struggle to secure even this amount, but would consequently have to change many aspects of my research.
buying and selling are two different things in business. I could get around not using EBI and would be encouraged to do so but it would be at inconvenience and against my wishes.
can't see them paying >£1,000 for an INDIVIDUAL access - maybe £20,000 for a site license knowledge of user, dept and college-wide thoughts on what access to other resources should cost. e.g. KEGG, commercial software and difficulty of persuading PI's to work together to pay upfront for discounted annual shared licenses.
Charged services destroy science. Whenever I find one I look to change their model or create a free and open competitor
Cost of paying someone to perform precomputed data/analyses provided by EMBL and/or cost of software licensing to do analyses that EMBL provides free.
EMBL-EBI is a highly valuable resource for my work, I would have to give up working on certain projects if I had to sell my right to access EMBL-EBI. Hence I would demand a very high price that would give me the freedom to look for a new, interesting job for a while. [...]
EMBL-EBI services are essential for my PhD work. It is not possible for me to sell the rights without giving up on my PhD completely at this point. […]
[...] I use EMBL-EBI data for training and validation of analytic methods, without this data, the services I am working on WOULD NOT BE POSSIBLE or would require a similar initiative to EMBL-EBI.[…]
For #32, I think $2000 would cover the cost of all the free software and services through EMBL-EBI if I had to pay for these services through another company (database access, tools, etc.). This is very approximate, but I estimated a bit high because I believe the impact on my research would be greater than I may expect if those resources were taken away. [...]
[…] it would be the average salary for someone keep doing the kind a job I do, but frankly, I have no idea how she/he would do without UniProtKB. […]

Q37: 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.)? Selected comments:
For 32. If I didn't have access to EMBL-EBI services I would have to re-derive and maintain my own curated versions of all the databases I commonly use as well as those I only occasionally use or might use in the future as my research interests change. There is no sensible price you can put on access to the high quality curated data that EMBL-EBI manages and makes available for free so I wanted to put a figure in the billions of pounds since it would be impossible to reproduce anything close to what we have from EMBL-EBI or anything that would be remotely as useful. Unfortunately, the form would not allow me to do this! Here is the figure I wanted to put in: £100,000,000,000. I just had to keep deleting zeros until it worked! [...]
For answer 31, my research position and project wouldn't exist if there weren't EBI data and tools, so I would ask for all my wage, so to speak. […]
[…] my lab subscribes to a statistical package (GraphPad Prism) for about $2000 per year (a single user license that's non-expiring would be $600). I would say EMBL-EBI has similar usefulness, so that's why I chose this amount.
For question 1: amount I would need to pay 2-3 FTEs to handle data for my group. For question 2: assuming everyone is sharing the load, I would pay $10,000 annually. However, if this were to actually happen, other resources would be developed to replace EBI resources over time.
For the first question: it's one year's tuition fee for making for the wasted one year. […] the maximum amount (if determined by me myself) would be 12000 because that's what I heard about the pricing for HGMD.
[...] Pragmatically lack of the service would probably mean that we did different types of work as the value of the free databases for genomics and genetics researchers is very high, albeit spread across many different and overlapping resources. Loss of one would not be a critical issue, loss of the major ones would be a disaster. Without them genomics research would see much much less investment.
[…] I can't call myself bioinformatician with out access to PDB and UniproKB.
[…] the cost per user should be fairly low, to remain affordable. It's kind of like a Netflix subscription.
Giving away my EMBL-EBI access means that a large part of my research would be massively challenged. That places really high value for my rights. On the other hand, given that large number of users use EMBL-EBI resources, and something like Adobe Illustrator licence for vector-drawing costs about 350 GBP, the resources-diverse databases and tools offered by EMBL-EBI could be effectively placed at about 1200 GBP, for academic users.
[...] The use of WTCCC data, downloaded from EBI, is almost invaluable, but presumably had not EBI volunteered to host the data, would be available elsewhere. However, if EBI asked me to pay say 1000-2000 pounds per 5000 individuals genotypic data I downloaded (e.g. about 3000-5000 GBP for WTCCC1 data) I would consider this fair and very worth it for the amount of use I have got out these data for developing new methods.
Having no longer access would be disastrous in our research, we would have to buy additional journal subscriptions and recreate experiments to compensate. […]

Q37: 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.)? Selected comments:
[…] Without EMBL-EBI services I would have to rely on US-based sites that are not as intergrated into each other as they are at EMBL-EBI, and would therefore take more time to find the data and information I need.
[...] £2000 is the cost of getting a paper published in a journal with open access and all the grant bodies are insisting that papers are published in this way. However, there is no guarantee that use of these resources would lead to a publication in that year. [...]
I base this amount (for our whole department) on what I think our department would be willing to spend, and based on the pricing of TransFac BioBase. This is quite arbitrary, and would of course be heavily influenced by whether or not the UCSC and NCBI resources are available, and at what cost. If no other resources but those from EMBL-EBI were available, I imagine around 5000 euro would become acceptable.
[…]price per license may depend on the number of concurrent licenses and on the added-values to the basic service: personalized support, fast downloading speeds, special fees for presential or virtual training courses and meetings...
I would have to hire at least 4 full time equivalents to make up for the loss of EMBL-EBI resources.
I would need all the research money to fund a senior, excellent post doc to replace EMBL-EBI, so that is what I would ask for. As far as the amount to pay, a good product is worth some money, but if everybody starts to charge, there will be nothing left, so I would not be able to pay the same amount at all because I could then surmise that other parties that I also rely on would do the same, leaving me penniless. Hence the £50.
I would pay what was asked as this is an invaluable resource.
If I cannot do my job without the EMBL-EBI services, then I would like my wage as payment for not being able to use the services. No services = no job to do.
[…] For our work in genomics, bioinformatics, and even patient care in human genetics (NGS diagnostics), the EBI services are indispensable. It is hard to put a price tag on them, but we would be put at a big disadvantage without them.
It seems to me that the potential use for the data I use is of such high quality that my scientific institution could easily be persuaded to provide this amount of money.
[…] €1000 would be a pretty good price for all that EMBL has to offer, but don't get any big ideas: keep it free!
[...] If other databases such as NCBI continue to offer the same data for free then clearly none would buy my access rights to EMBL-EBI (and similarly I would not pay for them). If the entire free access model broke down internationally then I suppose the costs/value would be similar to those of a subscription to Nature.
Judging from the costs of paid services like Ingenuity Pathway Analysis which cost about $10000/year for access.
Most of the research is both embryonic and non-competitive/academic basic research. IMHO the computational power needed to make significant progress in my field is not available at any price which effectively means I might as well retire!

Q37: 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.)? Selected comments:
My job would be impossible in its present form without access to resources such as Uniprot, quickGO, IntAct, Reactome.
My research will require at least two full time support staff to assemble and analyze the required data - this will cost ~£75,000 and hence this number. £7500 per research group per institute is a good amount to pay to have access to all resources - for instance HGMD charges ~£3500 to access their information. I would be willing to pay more but it will be difficult to convince my institute. [...]
My work is dependent on access to EMBL-EBI data, so I would never give up my rights to access EMBL-EBI data at any price. So I have set an amount sufficiently high to permit me to live without having to work. […]
My work would be impossible... I need to change jobs in this case, which I am very reluctant of doing since I just started and am very exciting about this research field of integrated biology/systems biology. therefore I would say my nett annual salary? [...]
Q1) what I'd have to pay to hire a bioinformatician to munge through a non-ensembl world and recreate the simplicity and internal consistency of access to genomic data across species I current get from ensembl. Q2) what I think I could get my company to foot for an annual single-user license. Eye-opening comparison...
Q30: I would not be able to do my job without EBI access so would require compensation for my salary. Q31: Not sure how to value this, but have estimated as equivalent to the cost of subscribing to 15 high-quality journals (Nature, Cell etc).
question 30: I consider the EMBL-EBI services to be priceless, I based my $$ for this question on this [...]. question 31: this is based on the high end of subscription services, as I consider EMBL-EBI to be a premium service, that given its great value to the community, could easily expect the community to be willing to pay a high fee for access.
Research funding bodies in Australia would never pay for subscriptions. Australian research intensive universities are cost cutting and would never support researchers access to a portal that was paid. A paid access portal to bioinformatics tools was started here 20 yrs ago, but the venture eventually collapsed since free alternatives became available on the web. And over time, these free alternatives became the standard. So they closed down.
Since my job relies on solved 3D enzymatic structures, the infrastructure provided my EMBL-EBI is invaluable, so £1000 is just a symbol to show that it offers a lot, but it is not depicting its value. What is depicting its value is the amount of £35000 that I would like if for any reason I could not use the services, which is a good to medium annual salary. When your job is based on something, then the annual payment should cover for your nonexistent job.
Slovenia is currently broke. If EMBL-EBI were to charge for its data, I'd switch to NCBI.
Some of the resources can be found freely in other database[s], but some specific services are exclusively available at EMBL-EBI. So for those exclusive series I might pay partial value of 1/5th of total value amount of total services.[…]
The 300,000 figure is my estimate of how much it would cost me yearly to re-create and maintain the data and tools essential to my work alone. Within the funding environment for non-medical research in Canada I cannot envision a granting agency or employer providing more than 10,000 for "personal" information technology. This

Q37: 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.)? Selected comments:
amount represents 20-40% of a typical research grant. Beyond this, no actual wetlab research would likely be able to occur.
The amount I can afford personally and by my organization is quite small, so how much I would be willing to pay is much smaller than what I would be willing to accept.
The amount we can reasonably pass on to our customers.
The amount which I can (maybe) justify diverting from cancer research budgets. I'd feel incredibly guilty about it, though.
[…] giving up my rights to the services would be given up performing research - I love doing reseach and it is what pay my mortgage - if i need to give up my job i would do for an amount that would allow me to live happily afterwards [...]
The figure assumes that GenBank and/or DDBJ remained available. If neither of them existed the figure would be considerably higher as the various tools would be more disparate and involve more time to locate. The figure I have selected is of the order of the institutional subscription to 2 essential journal resources [e.g. IJSEM and one other].
The gross cost of my salary is about €60000. I'm probably twice as efficient with EBI data than without, so this the level at which my employer (or grant funder) should value the data.
The prise of a good motorcycle.
[...] If access to sequence data incurs higher costs they would have to be included during research grant applications. If the costs are very high, grant structures would look differently, as access to the database is necessary for any -omics research. If database costs are prohibitive, some lines of research may not be possible in less well funded institutions, possibly would need to be concentrated in few institutions that can afford high subscription rates.
[…] I truly hope that we can preserve what's best for the generations to come and fight for what is right rather than what is affordable or convenient. I will definitely do my share on this quest.
[...] if these resources didn't exist I would become more adept at using other resources. However, as is, I consider them invaluable to my daily work. It would be very inconvenient to lose access, but if access was to be charged for, the individual rate would have to be lower - hence I need more money to give up, than I would pay to subscribe [...] Keep up the good work!
[...] in Cambridge we are in a good position to hire ex-EBI people (and we have done that already) who have worked on a particular database or a tool relevant for us. […]

Q37: 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.)? Selected comments:
This amount of money (in USD, EUR, or GBP) is the 'magical number' that most vendors use for licensing. This also roughly corresponds to the approximate resourcing of a team that would be dedicated to collecting, curating, and distributing part of the information similar to EMBL-EBI
We have developed software tools depending heavily on Ensembl. We have chosen EMBL-EBI as a) it is a European project, b) stands for continuity and c) is free of charges. We have seen KEGG become semi-commercial, and we no longer use it, Reactome is now our choice for pathway analysis. We analyze medical data and needed to purchase Biobase annual license (5000.--). I think it is almost an insult to only give the fivefold amount for an EMBL-EBI license, but financial resources are limited. The high price for selling our access reflects our long term software commitment to and invaluable appreciation for Ensembl, as we would need to reorganize our code to access an alternative.
[...] When an important service that we used for yeast gene functional annotation lost its "free" status, a few years ago, we just stopped using it and found open alternatives. The service was closed two years later, most likely for lack of paying customers. If EMBL-EBI services would not be available in an open and free manner, we would most likely go somewhere else.
Without free access I don't know if I would even have started my organisation
Your question will not accept a ridiculously high value I was trying to input so I said zero. [...]. If we start generating such comparators for public, open and collaborative community resources then where will it stop? X Factor for UniProt v Ensembl v PDB etc. I thought this battle was already fought over the human genome?! Stand up for yourselves and refuse to even countenance this silliness!
with the 50k difference I could hire a person to do all the annoying stuff to get programmatic access to sites like NCBI.

Appendix 3: A modified Solow-Swan model

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

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

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

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

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

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

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

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

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

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

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

gives:

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

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

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

Contributions to growth and total factor productivity

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

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

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

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

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

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

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

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

Estimating the rate of return to R&D

While there are recognized limitations to the traditional growth model approach, this basic framework has been widely used in estimating the rate of return to R&D. The standard approach to estimating returns to R&D is to divide the technology variable A in (1) into two components, a stock

of R&D knowledge variable R and a variable Z that represents a matrix of other factors affecting productivity growth. The production function then becomes:

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

and the counterpart of equation (5) becomes:

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

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

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

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

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

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

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

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

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

Incorporating the efficiency of research and accessibility of knowledge

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

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

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

We define an '*accessibility*' parameter ε as the proportion of the R&D knowledge stock that is accessible to those who could use it productively, and an '*efficiency*' of R&D parameter ϕ as the proportion of R&D spending that generates useful knowledge. Then starting with a given stock of

useful knowledge R^*_0 at the start of period zero, useful knowledge at the start of period 1 will be given by:

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

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

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

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

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

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

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

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

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

Assume now that there is a one-off increase in the value of ϕ and ε, from the constant values of ϕ_0 and ε_0 to new values of $(1 + \delta_\phi)\phi_0$ and $(1 + \delta_\varepsilon)\varepsilon_0$, respectively. Then the rate of return to R^*, that is:

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

is fixed, but the return to R will increase:

(17) $$\partial y/\partial R = \gamma . (Y/R) = \phi_1\varepsilon_1 \partial y/\partial R^* = \gamma . (Y/R). (\phi_1\varepsilon_1 /\phi_0\varepsilon_0)$$
$$= \gamma . (Y/R). (1 + \delta_\phi).(1 + \delta_\varepsilon)\varepsilon_0.$$

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

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

IMPACT CASE STUDY

EMBL-EBI

ChEMBL

ChEMBL

CHEMISTRY MEETS BIOLOGY

ChEMBL is a unique public knowledge base of chemical compounds and small molecules with their biological targets. It pulls together high-value information on compounds and their effects on biological systems from the available academic literature in a structured database.

EMBL-EBI dynamically links information from the academic literature in ChEMBL to chemical patent documents recorded in the SureChEMBL database. SureChEMBL provides live free access to chemical data extracted from the patent literature. Data in patents are important to drug discovery researchers because new discoveries often appear in patents 2-3 years ahead of the published scientific papers.

The ChEMBL database was originally developed as a commercial product, becoming an open and freely available EMBL-EBI service in 2008 with funding from the Wellcome Trust. In 2013, this was extended to include patent information held in the SureChEMBL database.

FUNDING

ChEMBL is primarily funded by

EMBL-EBI

wellcome trust Strategic Award

with supplementary funding from

CTTV Centre for Therapeutic Target Validation

NIH

imi innovative medicines initiative

Horizon 2020 European Union Funding for Research & Innovation

SEVENTH FRAMEWORK PROGRAMME

IMPACT

ChEMBL's focus on enabling all aspects of discovery, is utilised by academics and industries of all sizes, strengthening innovation from new research, and the discovery of new treatments and drugs benefiting human health and agriculture.

In the recent Strategic vision for UK e-infrastructure report[002], Professor Dominic Tildesley of Unilever identifies the ChEMBL database as key in their product development of anti-perspirants. They used the database, to identify active components for anti-perspirants and the ChEMBL data to build a model of their inhibition activity.

Below are quotes from four external users of the ChEMBL database which demonstrates how ChEMBL improves R&D, increases productivity and performance and underpins scientific investment.

USER IMPACTS

Sygenta (Industry: AgroChemicals)

syngenta

Syngenta is a leading agriculture company and employs 28,000 people in over 90 countries. 5,000 are in R&D and apply world class science to provide innovative crop solutions to transform how crops are grown.

"It has been estimated that without crop protection compounds (pesticides, fungicides, weed killers, etc.) 40% of the world's food would not exist. Our scientists use ChEMBL to support projects in our research towards innovative new products, and ChEMBL has links between both chemistry and biology data which makes it searchable in ways that the underlying literature would not be. People at the EMBL-EBI do a fantastic job in making a vast amount of data of different types openly available to researchers, and without the EMBL-EBI resources in general I'm sure life science research would be greatly hindered."
– *Mark Forster, Syngenta*

DATA STORAGE

13.5 MILLION RECORDS

ChEMBL contains information on more than 1.4 million compounds and 13.5 million records of their effects on biological systems.

SureChEMBL provides a live, updated daily, view of chemical patents, with approximately 50,000 new documents added per month.

UNIQUE RESOURCE

70%

If the ChEMBL resource did not exist 70% of its users stated that they wouldn't have been able to obtain the data from anywhere else.[001]

PUBLICATIONS

1,125

ChEMBL reference papers have been cited **1,125** times[003] with use of the database directly cited in **196** publications. Covering areas as diverse as colon cancer, drug design & development, virtual screening, experimental modelling & complimentary biology databases.

IMPACT CASE STUDY

EMBL-EBI

ChEMBL

ChEMBL

USERS IMPACTS Cont.

The University of Sheffield Information School

Placed 1st for the impact of research by the 2014 UK Research Excellence Framework.

"ChEMBL provides a unique resource in terms of publicly available data about compounds and their properties on a scale that is not available elsewhere. This facilitates the process of developing new methods for virtual screening and allows different research groups to compare methods more easily than would be possible without this data, which helps to advance the field more quickly. Without ChEMBL, the methods in general use for drug discovery might be poorer and have reduced success rates for virtual screening. This would lead to higher costs because of having to do more testing; fewer successful drug developments, and potentially an impact on human health as a result." - *Professor Val Gillet, Professor of Chemoinformatics, Head of School, University of Sheffield Information School*

MRC Technology (Charity)

An independent life science medical research charity working to bridge the gap between basic research and commercial application.

"ChEMBL database has proven to be invaluable to the computational group at MRCT. Whilst commercial sensitivity prevents the identification of specific targets and programmes active within MRCT, there are many cases of analyses covering a wide range of protein classes that have been moved forward with valuable data based on mining the ChEMBL data sources." - *Dr Andy Merritt*, MRCT

Cambridge MedChem Consulting (Drug Discovery)

Cambridge MedChem Consulting is a micro-company based in the UK providing a range of consultancy services in drug discovery and medicinal chemistry.

""ChEMBL acts as a standard source so everyone is working from the same base. There are an increasing number of independent consultants and small companies like MedChem. Without ChEMBL there would be some things I would just not be able to do. For example, going through the original literature would be prohibitively time-consuming (months/years). It is pretty unique as a data source, and the web services and API allow you to build custom tools and very efficient pipelines for access. It is an invaluable resource and the EMBL-EBI staff are helpful, enthusiastic and keen to promote it" - *Dr Chris Swain, Cambridge MedChem Consulting*

SUPPORTING PUBLICATIONS

001 **The value and impact of The EMBL-European Bioinformatics Institute**
Impact Survey conducted in 2015 with 4185 respondents of which 771 classified themselves as ChEMBL users.
Charles Beagrie Ltd

002 **A Strategic Vision for UK e-Infrastructureofessor Dominic Tildesley, Unilever PLC**
A independant report commissioned by UK Gov BIS Minister for Universities and Science. A roadmap for the development and use of advanced computing, data and networks
Professor Dominic Tildesley, Unilever PLC

003 **Google Scholar**

USER IMPACT

68%

Of surveyed ChEMBL users said it would have a major or severe impact on their work or study if they could not access EMBL-EBI services and resources.[001]

BRAIN GAIN

94%

94% of ChEMBL users were involved in research, spending an average of 34 hours a week on research activities, of which 60% was spent working with data.[001]

RESEARCH EFFICIENCY

46%

MORE EFFICIENT

The efficency value placed on our services represents a direct worth of between £5,382 to £26,000 per respondant with the overall average that EMBL-EBI services allowed users to be 46% more efficient in their work[1].

Charles Beagrie

All interviews and subsequent economic analysis were undertaken by Charles Beagrie Ltd. on behalf of EMBL-EBI.

IMPACT CASE STUDY

EMBL-EBI

Gene Ontology

THE IMPACT OF STANDARDISATION

The Gene Ontology (GO) project is a major initiative to develop a computational representation of biological function. GO provides a set of terms for describing gene product characteristics, and uses these standard terms to annotate genes and proteins. It also provides tools to discover, access and process these annotations.

The Gene Ontology Consortium (GOC) includes 34 groups from around the world that collaborate closely on developing and using the Gene Ontology. EMBL-EBI is a major contributor to this consortium across three key areas managed by distinct service teams; annotation, essential ontology development, tool development & provision. Tools such as Protein2GO, are used by a significant number of the consortium members for adding annotations. It is also used by an extensive range of external researchers in industry and academic institutions.

In this impact case study, we present short extracts from five interviews conducted with external users of the EMBL-EBI GO service. We use these extracts to illustrate how GO is underpinning scientific investment, improving R&D and increasing both productivity and performance.

www.ebi.ac.uk/goa

USER IMPACTS

Roche (Industry: Pharmaceutical and Healthcare)

Roche is a global, research-focused healthcare company headquartered in Switzerland.

"We rely on public databases for our analyses. Downloading data from EBI GO allows us to store the data locally and to run large or batch jobs on our server and GO updates are downloaded roughly every month [one unique visitor,12 downloads per year]. GO data is mostly used to query, for a given gene, the GO annotations in molecular function, cellular component and biological process. This information is part of our gene index tool which is used by around 1400 Roche scientists." *Isabelle Wells, Roche Pharma Research and Early Development.*

Counter-factual benefit: based on information supplied by EMBL-EBI and UK full economic costs, we estimate the cost of creating and maintaining EBI GO is around £498,500 per annum.

Cardiovascular Gene Ontology Annotation Initiative (Public Sector: Gene Annotation Project)

This collaborative project between UCL and EMBL-EBI provides more specific GO terms for, and annotations of, human genes that are implicated in heart development and cardiovascular processes.

"People are starting to use GO for Genome-Wide Association Study variants. In many cases a disease is caused by variants in a lot of different genes, but often these genes are all involved in a common pathway. There may, for example, be many genes required to keep the heart functioning, and the body can tolerate one gene variant within a single essential heart pathway but not (say) six. GO can help identify whether there are common variants causing the same disease by grouping genes together and improving the likelihood of the variants being detected by statistical analysis." *Ruth Lovering, Cardiovascular Annotation Coordinator, Centre for Cardiovascular Genetics, UCL*

Benefit: Enrichment analysis is a major application of the GO, and the addition of specific terms and annotations can increase the sensitivity of these analyses. In an experiment to determine the effect of new cardiovascular annotations it was shown that enrichment of metabolism for a key regulator of blood vessel constriction could be detected only with the additional annotations.

EFFICENCY SAVINGS

3 : 1

For every £1 invested in GO, an estimated £3 of efficiency savings are gained within academic and commercial settings.

HIDDEN REUSE

7000 GB

7000 GB of data was downloaded from GO in 2014. This data is fed into many different biological knowledge bases, allowing GO data to be used by thousands of additional users.

6  **18,525**

AgBase downloaded GO 6 times in 2013, alongside other relevant inputs. Agbase had 18,525 unique visitors in 2013.

12  **32,000**

dictyBase downloaded GO 12 times in 2013, alongside other relevant inputs. It had 32,000 users and 66,000 visits to the dictyBase gene page, which contained the relevant GO information.

IMPACT CASE STUDY

EMBL-EBI

Gene Ontology

USERS IMPACTS Cont.

Thomson-Reuters MetaCore (Industry: Bioinformatics database)

"Within MetaCore, we include both Thomson Reuters' and external ontologies including GO to enable our customers to choose their desired enrichment. Although our own ontologies are used by the significant majority of customers for enrichment and analytics, all entities within the databases are registered to differing GO processes. Some users will use different ontologies at different stages of the study workflow, and particularly academic users may apply the GO ontologies as a way to confer validation prior to external publication." *Gavin Coney and Mark Hughes, Thomson Reuters*

Benefit: Reproducibility or validation of scientific results add a vital assurance of quality for research. GO provides users with a good integration method between various ontologies and different datasets. When Thomson-Reuters supplement that with extra, highly curated information and advanced analytics it creates a lot of value to end users.

AgBase (Public sector database)

AgBase is a curated genomic database containing functional annotations of agriculturally important animals, plants, microbes and parasites.

"Using the EBI biocuration infrastructure means that we don't need to build our own data entry systems. Initial, one-off development of a biocuration interface I would estimate to take about 3 months of programmer time as an initial effort; about 1 week every 2 months afterward to do the quality control. Based upon my experience of data entry and number of papers curated over time, the Protein2GO curation interface increases [bio-curator] productivity by 10-15%..." *Fiona McCarthy, Associate Professor, School of Animal and Comparative Biomedical Sciences, University of Arizona, US*

Benefit: based on further information supplied in the interview and equivalent UK posts and full labour costs we estimate their use of the EMBL-EBI Protein2GO tool and infrastructure provides labour efficiencies to AgBase of around £50,700 per annum.

dictyBase (Public sector database)

dictyBase is an online bioinformatics database for the model organism *Dictyostelium discoideum*

"Protein2GO provides invaluable help to us, a small model organism database. It would have been possible to proceed without Protein2GO but at a high cost, severely delaying other projects that our users waited for, such as adding new genomes to the database." *Petra Fey, Senior Scientific Curator and Stock Center Manager, dictyBase, Northwestern University, Chicago, IL*

Benefit: based on further information supplied in the interview and equivalent UK posts and full labour costs we estimate their use of the EMBL-EBI Protein2GO tool and infrastructure provides labour efficiencies to dictyBase of around £14,300 per annum.

SCALE OF USAGE

2.8 MILLION WEB REQUESTS

There are over 50 different freely available functional analysis tools and 2 major commercial bioinformatics services that include GO annotations in their programs.

DIRECT IMPACT

Estimates of direct impact to interviewed users, using all the information supplied in interview and UK full labour costs, it was estimated that their use of the EMBL-EBI biocuration infrastructure provided labour efficiencies per annum as follows:

AgBase: £50,700

dictyBase: £14,300

Roche: £498,500*

*Estimated cost of developing and maintaining GO

Charles Beagrie

All interviews and subsequent economic analysis were undertaken by Charles Beagrie Ltd. on behalf of EMBL-EBI.

IMPACT CASE STUDY

EMBL-EBI

Variant Effect Predictor

Ensembl®

THE IMPACT OF STANDARDISATION

A genome is an organism's complete set of DNA, including all of its genes, and holds the key to greater understanding of an organism's development, how and why we differ and what makes us susceptible to diseases. Many organisms including humans have now been fully sequenced and reference data sets are held in Ensembl's databases.

We can start to understand what is happening in individual genomes by comparing them with reference genomes, and combining that information with knowledge from all other fields of molecular biology.

Ensembl's Variant Effect Predictor (VEP) is a powerful open software tool that can analyse most types of variation data. It uses the extensive annotation in Ensembl to provide detailed functional predictions and annotation on the effects of variants.

FUNDING

Ensembl receives funding from the Wellcome Trust with additional funding for project specific components from the BBSRC, EC, NIH, CTTV and EMBL.

wellcome trust

EMBL-EBI CTTV NIH BBSRC European Commission

IMPACT

VEP is deployed in many critical areas of research such as cancer and rare diseases, where strong links have been established between changes in the genome and disease development. VEP also supports conversion of variant data into the format most familiar to clinicians (HGVS codes) allowing the knowledge gained to be directly applied.

VEP is often adopted above other similar tools due to its high performance, the stability of funding, extensive user support, and it's open licence does not restrict users in distributing their results. Below are some findings from interviews with three diverse users of the VEP tool demonstrating the research it supports.

USER IMPACTS

The 100,000 Genomes project will sequence 100,000 whole genomes from NHS patients by 2017. The project is focussing on patients with cancer and patients with a rare disease and their families. It is hoped that the project's legacy will be a service ready for adoption by the NHS, new medicines, treatments and diagnostics, and a country which hosts the world's leading genomic companies. *(Genomics England 2015)*

illumina® Genomics england

Illumina is a leading developer, manufacturer, and marketer of life science tools and integrated systems for large-scale analysis of genetic variation and function. It is the sequence provider for the world-leading 100,000 Genomes Project. VEP is part of the annotation engine that is used to deliver annotated genomes for this project, and also for Illumina's VariantStudio software product.

"The Illumina VariantStudio data analysis software application enables researchers to quickly identify and classify disease-relevant variants, and then communicate significant findings in a structured report. VariantStudio talks to an annotation tool which has VEP at its core. VEP was selected by Illumina because it was more robust and more production-ready than other annotation tools. Because of VEP's superior quality and accuracy its users are able to catch some edge cases where annotation would be otherwise incorrectly handled."

Elliott Margulies, Illumina

GENE VARIANTS

88 MILLION

The human genome is made of 3.2 billion bases of DNA which code for approximately 20,000 protein coding genes. Scientists from around the world catalogued 88 million variants.

RARE VARIANTS

64 MILLION

The 100,00 Genomes Project considered that the majority of variants, 64 million were considered rare in frequency, occuring in only 1% or less of the population.

RARE DISEASE

6-7%

Rare diseases are individually very uncommon but in total they affect a surprisingly large number of people, between 6 and 7 percent of the UK population.

PROCESSING POWER

The VEP software tool can process more than **4 million** genetic variants per hour.

IMPACT CASE STUDY

EMBL-EBI

Variant Effect Predictor

USERS IMPACTS Cont.

As part of the **Deciphering Development Disorders study,** a collaboration between the Wellcome Trust Sanger Institute and NHS regional genetic services to understand a range of development disorders in children, just under 14,000 families had their DNA sequenced. The VEP tool was a central part of the analysis.

"we achieved a diagnostic yield of 27% among 1133 previously investigated yet undiagnosed children … In families with developmentally normal parents, whole exome sequencing of the child and both parents resulted in a 10-fold reduction in the number of potential causal variants that needed clinical evaluation … Most diagnostic variants identified in known genes were novel and not present in current databases of known disease variation." (Wright et all 2015)

"The benefit of VEP is in annotating and predicting the likely consequences of variants identified in the study, allowing us to identify disease-causing variants much more efficiently and effectively. This is important for the DDD team, clinicians and families. About a third of families will likely receive a diagnosis. Only a small number will be treatable, but the information is valuable for counselling and helping them to make informed choices about having further children, based on whether the variant is likely to be inherited or spontaneous".

Dr Caroline Wright, Wellcome Trust Sanger Institute

The Daniel MacArthur Lab and the **ExAC** project are prominent users of the VEP tool. The lab is jointly based at Massachusetts Hospital and the Broad Institute. VEP is central to three major projects in the lab including the Exome Aggregation Consortium (ExAC). ExAC is an international coalition of investigators with a focus on data from exome sequencing and variant discovery on the regions of the genome that encode proteins, known collectively as the exome. It is by far the largest single aggregation of coding variants in the world and a key comparison data set for childhood-onset Mendelian diseases.

"VEP is central to the ExAC project which is building a large reference database of human genetic variation. This is using exome sequencing to understand variation in human genes, and uses VEP to predict functional variation. By the end of 2015 we expect to have aggregated data from around 100,000 individuals and identified over 15 million genetic variants. Between launching in October 2014 and June 2015 EXAC has had over 1.5 million page views. This represents more than 80,000 unique users over 8 months."

We went with VEP for three main reasons:

- The quality and completeness of annotations: VEP leverages the Ensembl gene set. On manual inspection of results there were inaccuracies in the other tools and VEP was correct.
- The software is beautifully documented, which makes it easy to expand and build plugins.
- VEP integrates seamlessly with Ensembl. We often need to pull in other types of data and can do this smoothly with VEP".

Daniel MacArthur, Massachusetts Hospital/Broad Institute and lead analyst ExAC project.

PROCESS POWER

83%

From a survey of over 2,500 Ensembl users, 83% reported a high or very high benefit of the data and tools supplied.

DIRECT IMPACT

1 in 3

In the developed world, cancer will affect one in three people at some stage in their life. Without research we condeming tomorrow's generation to today's treatments.

RESEARCH EFFICIENCY

46%

MORE EFFICIENT

From a survey of over 4,000 EMBL-EBI users, the efficency value placed on our services represents a direct worth of between £5,382 to £26,000 per respondant with the overall average that EMBL-EBI services allowed users to be 46% more efficient in their work.

Charles Beagrie All interviews and subsequent economic analysis were undertaken by Charles Beagrie Ltd. on behalf of EMBL-EBI.

The LFCF Programme was managed by Biotechnology and Biological Sciences Research Council (BBSRC) on behalf of Research Councils UK (RCUK).

BBSRC
bioscience for the future