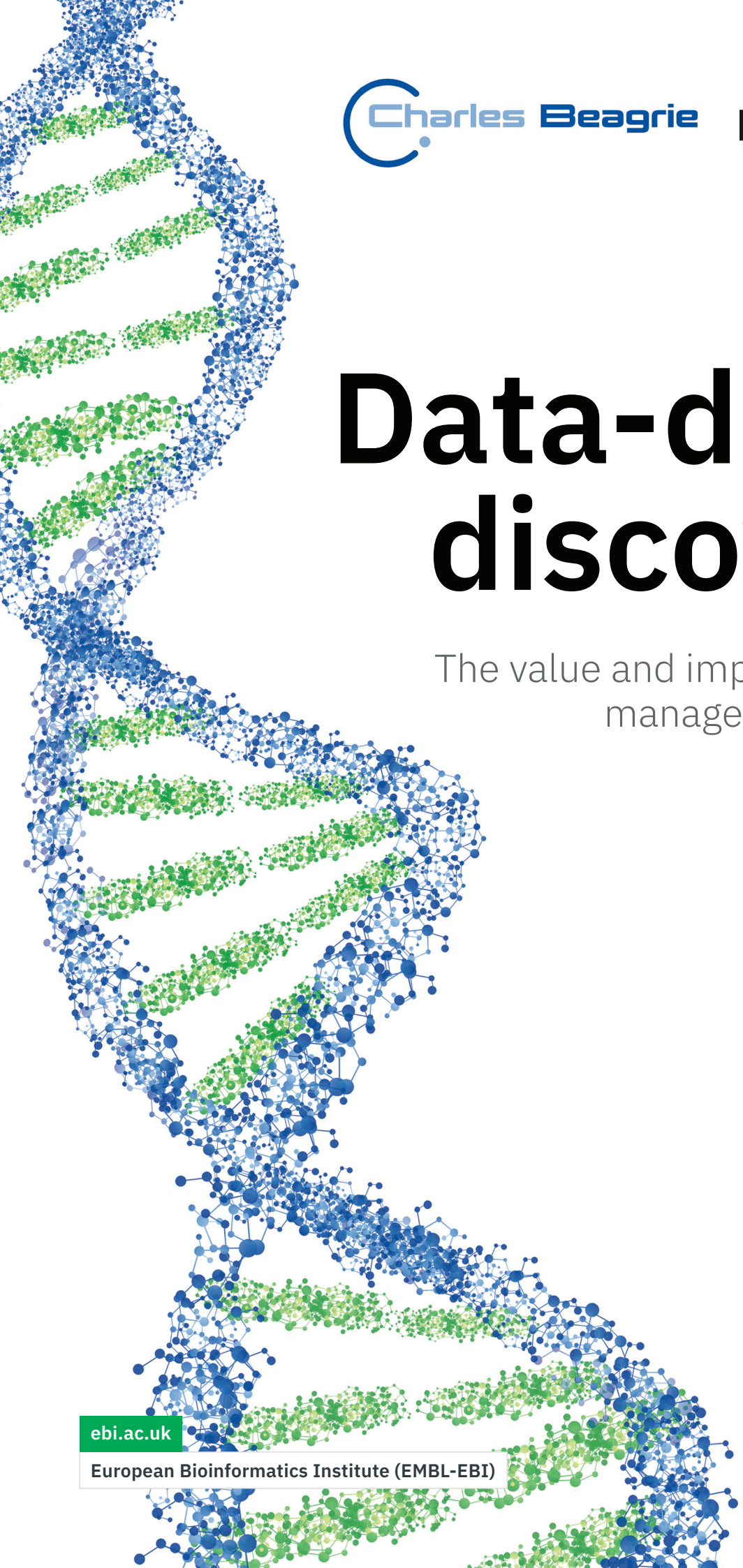


Data-driven discovery:

The value and impact of EMBL-EBI managed data resources

Neil Beagrie & John Houghton

Charles Beagrie Ltd

www.beagrie.com

ebi.ac.uk

European Bioinformatics Institute (EMBL-EBI)

About this publication

Cover image credit: AdobeStock image edited by Karen Arnott, EMBL-EBI.

This report and a separate version of the executive summary are available online in printable format, at www.embl.org/documents/document/embl-ebi-impact-report-2021 and www.embl.org/documents/document/embl-ebi-impact-report-summary-2021 respectively.

Contents

Acknowledgements

We would like to thank: EMBL-EBI staff who contributed to the study particularly Mary Barlow for her support and input throughout, Amonida Zadissa, Gaia Cantelli, Ewan Johnstone, and Rodrigo Lopez Serrano for their work on EMBL-EBI web statistics, and previous EMBL-EBI surveys; Karen Arnott of EMBL-EBI for her work on the layout and cover design for the report; Oana Stroe and colleagues in the EMBL-EBI Communications team for the comms effort to make the survey a success; the EMBL-EBI users who kindly participated in user pilot testing of the user survey for the report; and finally, all the survey respondents, who all gave valuable time and input to the study.

Neil Beagrie and John Houghton

July 2021

Executive summary

The findings in this report are the outcome of an independent evaluation undertaken by Charles Beagrie Ltd in 2020/2021.

Introduction

The European Molecular Biology Laboratory (EMBL) is Europe's only intergovernmental laboratory for life science research. Established to advance the study and understanding of molecular biology, nurture young talent, new ideas and technologies, it now performs its activities across six sites in five host nations.

EMBL's European Bioinformatics Institute (EMBL-EBI), located on the Wellcome Genome Campus near Cambridge, is Europe's hub for biomolecular data and an acknowledged world leader in the management and analysis of big data in biology. It is a highly collaborative organisation, with data resources run in partnership with organisations throughout the world.

Investment by EMBL member states and other funders alongside the contributions of collaborators, enable EMBL-EBI to host the world's most comprehensive and integrated collection of data resources in the life sciences.

Charles Beagrie Ltd undertook the first EMBL-EBI Value and Impact study in 2015-2016 and the current study after a gap of five years. Both studies have been conducted as part of an on-going programme, led by EMBL-EBI, to develop a framework and evidence base for demonstrating the economic value and impact of the open data resources.

The user survey for the current study launched in March 2021 and received 4 920 usable responses, providing an excellent foundation for analysis. The response to the user survey is one of the world's largest from surveys on open research and open data in recent years.

The current study used multiple approaches to assess the economic value and impact of 44 open data resources managed by EMBL-EBI. The quantitative economic approaches used in the study include: estimates of access and use value; contingent valuation; estimating the efficiency impacts of EMBL-EBI data resources; and a macro-economic approach that seeks to explore the wider impacts of EMBL-EBI data resources on returns to investment in research. These approaches allow 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. Methods rely on triangulated conservative estimates of EMBL-EBI user populations, levels of use, and investment value (expenditure) required by EMBL-EBI and its many collaborators in developing and delivering the data resources. Both these methods and their application, and so the estimates used by us in this study, are conservative.

In order to isolate attributable impacts, we also use a counter-factual approach, focussing on what users could and would have done if the EMBL-EBI data resources did not exist. For both research efficiency and returns to R&D, this enables us to distinguish between the impacts arising from activities *facilitated* by EMBL-EBI data resources and the impacts that *depend* on EMBL-EBI data resources.

Economic impact studies often focus on a single or very limited approach(es) and metrics, and counter-factuals are rarely used. Although this is simpler to present and easier to undertake, all individual approaches have weaknesses. A strength of this study is the breadth and depth of the multiple approaches employed and the confidence this can give in the overall picture of its findings.

COLLABORATIVE BY DESIGN

EMBL-EBI develops many of its data resources through collaborations with other organisations around the world. This is done in many ways: through joint grants, consortium agreements (e.g. Protein Data Bank in Europe), direct partnerships (e.g. Reactome), and federated models (e.g. European Genome-phenome Archive). Some data resources are jointly managed (e.g. UniProt).

The scale and depth of these collaborations vary, creating a complex ecosystem.

EMBL-EBI does not just store the data, it brings value by curating, standardising, and making the data FAIR (Findable, Accessible, Interoperable and Reproducible). For the sake of brevity, this report refers to “EMBL-EBI managed data resources”, regardless of the type of collaboration agreement.

EMBL-EBI are grateful to its partners and collaborators for their continued support in developing and maintaining robust data infrastructure for the scientific community.

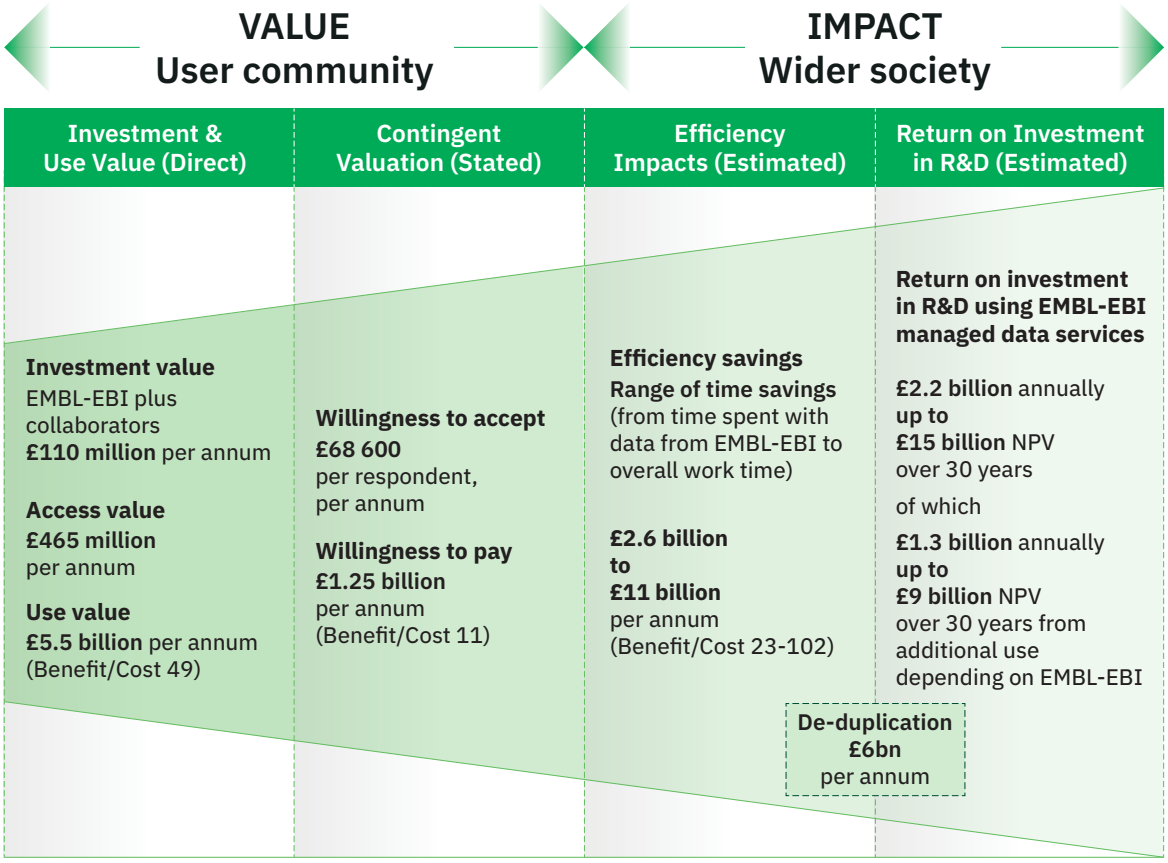
Key findings

The qualitative and quantitative analyses reveal that EMBL-EBI managed data resources are utilised widely and valued highly by their users.

We find that EMBL-EBI managed data resources present exceptional value for money in terms of the value returned and impact compared to the costs of running them.

FIGURE S1

The value and impact of EMBL-EBI managed data resources



Note: NPV is Net Present Value. All estimates rounded.

Source: Authors' analysis

The quantitative analysis summarised in Figure S1 (above) explores the value and impact of EMBL-EBI managed data resources and shows:

- **Use value:** The most direct measure of the value is the time users spend using EMBL-EBI managed data resources - **an estimated £5.5 billion per annum** (more than £3 billion higher than in our 2015-16 impact study, reflecting increased use). **This compares very favourably with the estimated £110 million total annual expenditure with the use value being 49 times the estimated total costs.**[1]
- **Contingent valuation:** Measures the value users place on a freely provided service by asking what they would be willing to pay for that service in a hypothetical market situation, which for EMBL-EBI managed data resources is **an estimated £1.25 billion per annum** (some £925 million higher than 2015-16). This contingent valuation estimate gives a sense of the minimum value of EMBL-EBI's managed data resources to users, **equivalent to 11 times the estimated total costs**.
- **Efficiency impacts:** Researchers reported that EMBL-EBI managed data resources made their research significantly more efficient. This benefit to users and their funders is estimated, at a minimum, to be worth £2.6 billion per annum worldwide, and at a possible maximum worth £11 billion per annum worldwide (almost 120 per cent higher than 2015-16). The estimated efficiency impacts give a sense of the **possible wider value of EMBL-EBI's managed data resources, equivalent to 23 to 102 times the estimated total costs**.
- **Return on Investment in R&D facilitated by EMBL-EBI managed data resources:** During the last year the use of EMBL-EBI managed data resources contributed to the wider realisation of research impacts conservatively estimated to be **worth some £2.2 billion annually** (almost 140 per cent higher than 2015-16), or up to £15 billion over 30 years in net present value.
- **Return on Investment in R&D depending on EMBL-EBI managed data resources:** Some 58 per cent 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 managed data resources underpinned **research impacts worth £1.3 billion annually** (£930 million higher than 2015-16), or up to £9 billion over 30 years in net present value, **that could not otherwise have been realised**.
- **De-duplication of effort:** A key value of open data lies in the de-duplication of research effort. A separate new calculation combining approaches in the study (therefore shown in Figure S1 as an insert) provides a supplementary way of estimating a major part of the value of research efficiencies. ***If*** the time saved by users from not having to (re)create the data enabled more research to be done

[1] Note that during the last year EMBL-EBI managed data resources underpinned more than 140 million hours of research.

(i.e., adding the value of the time saved (£4.3 bn) to the potential returns from the additional research facilitated (£1.5 bn)), it could be **worth almost £6 billion per annum** (more than 200 per cent higher than we would calculate for 2015-16).

The qualitative analysis reveals a similar picture of the value and impact of EMBL-EBI managed data resources:

- **More than two-thirds of all respondents (69 per cent) said that not having access to EMBL-EBI managed data resources would have a "major" or "severe" impact on their work or study**, and 90 per cent in total said not having EMBL-EBI would have a major, severe or moderate impact on their work. This represents a substantial increase compared to our 2015-16 study when the equivalent responses were 55 per cent and 84 per cent.
- The study shows that **EMBL-EBI's managed data resources have been making a major contribution to science throughout the COVID-19 Pandemic**. Some 32 per cent of respondents said they valued EMBL-EBI data resources more as a result of the pandemic and 25 per cent said they had used them more.

Value and impact statement have been based on estimated values of EMBL-EBI user populations and levels of use. This has focused on an estimated 450 000 to 500 000 unique primary users, those who access the data resources managed by EMBL-EBI directly and who could be invited to respond to our user survey, possibly 20 per cent of worldwide life science researchers. The study did find extensive indirect secondary use and a considerable number of secondary users of open data managed by EMBL-EBI, which are not included in our impact assessment.

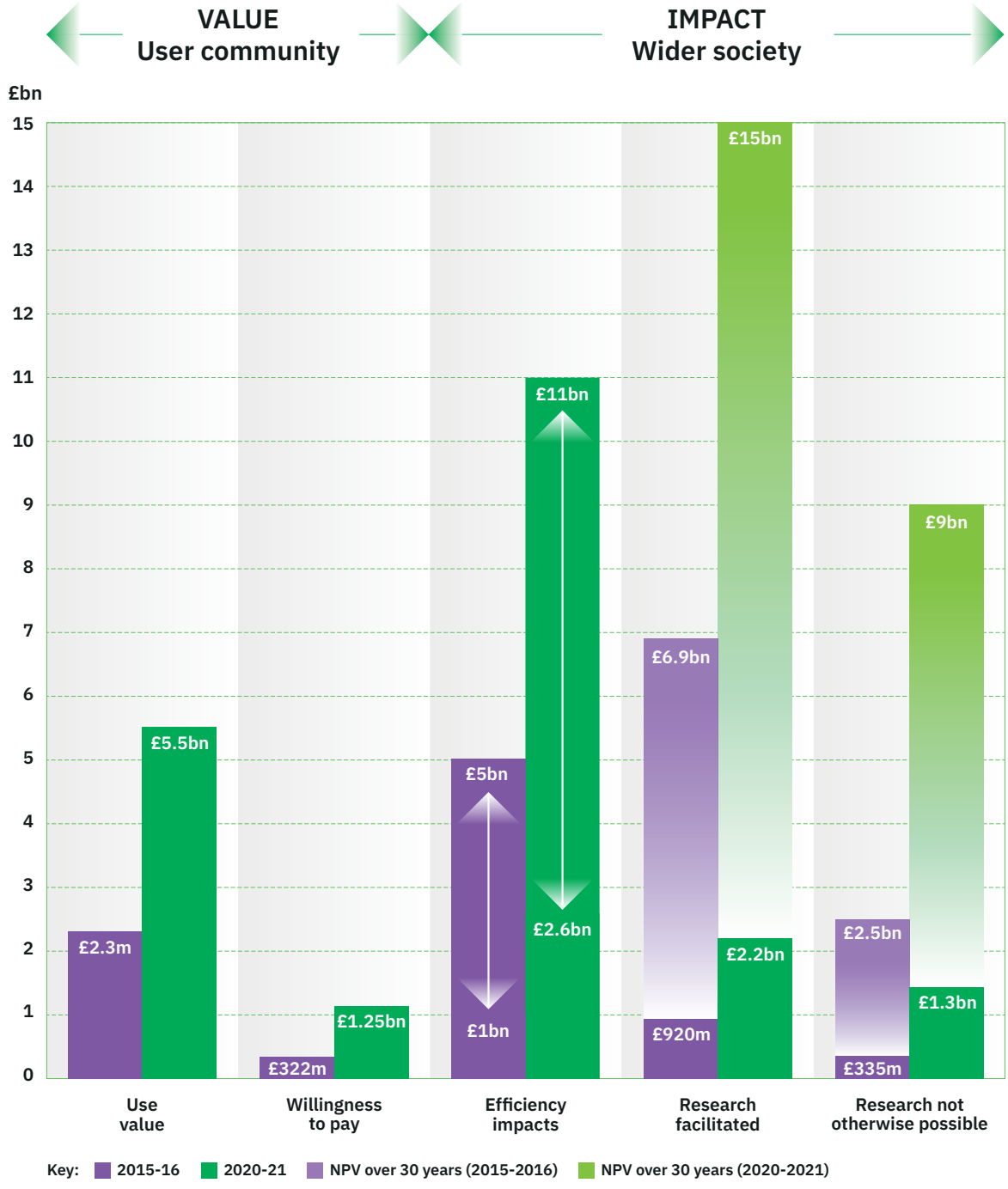
Putting these findings into context

The current study and the original value and impact study conducted in 2015-16 are snapshots in time and there are limits to how they can be compared. Nevertheless, the economic estimates can provide a sense of the scale of economic impacts at the time and changes over the last five years.

The current study shows higher levels of value and impact across the board compared to 2015-16, reflecting the increased level and intensity of use of EMBL-EBI managed data resources, the increasing maturity of the data resources, and evolving research practices. Figure S2 (below) shows the 2015-16 results alongside those from 2020-21.

FIGURE S2

The increasing value and impacts of EMBL-EBI managed data resources (2015-16 & 2020-21)



Note: All values are per annum apart from the Net Present Values (NPV) for returns on research which are over 30 years.

Source: Authors' analysis

It is difficult to make direct comparisons with other studies. To be truly comparable studies must use the same methods and processes of application (e.g., in activity costing, estimation of user population, etc.), they must focus on similar types of resources and be done in a similar timeframe and locale. However, the current study compares favourably against a range of studies of library and information services, national statistics and other information and data services (See Box S1 below).

BOX S1

Putting the value and impact of EMBL-EBI managed data resources into context

While individual studies focus on different information services and content, and use a variety of methods and measures, it is possible to explore their findings to give a sense of how the value and impact of EMBL-EBI data resources compare:

- Houghton (2011) estimated the benefit/cost ratio of the Australian Bureau of Statistics making data and publications freely available online and using Creative Commons licensing at 5.3 to 1.
- Tennison (2015) reported that a report by Nesta and the ODI adds to the evidence of the impact of open data. The report's analysis, undertaken by PwC, examines the effects of the Open Data Challenge Series (ODCS) and predicts the programme will result in a potential 10x return (£10 for every £1 invested over three years), generating up to £10.8m for the UK economy.
- Like other meta reviews, *Measuring the Value of California's Public Libraries, 2017-2020* found that investment in public libraries is a sound use of public funds: for every dollar invested in libraries, about $2-$10 are returned, with an average of between $3 and $6.
- King (2010) summarized findings relating to library services and concluded that: special libraries exhibit a return of 2.9 to 1, academic libraries 3.4 to 1 (for staff), and public libraries 5.8 to 1.

Source: Authors' analysis

1. Introduction

1.1 A brief description of EMBL-EBI

Founded in 1974, the European Molecular Biology Laboratory (EMBL) now operates across six locations: Barcelona, Grenoble, Hamburg, Heidelberg, EMBL-EBI, and Rome.

Established in 1994, EMBL's European Bioinformatics Institute (EMBL-EBI) based in Hinxton near Cambridge in the UK, is a global leader in the storage, analysis and dissemination of large biological datasets. It helps scientists realise the potential of big data by enhancing their ability to exploit complex information to make discoveries that benefit humankind.

While it is best known for its bioinformatics resources, the Institute is also at the forefront of computational biology research, and offers an extensive training programme that helps researchers turn data into knowledge. EMBL-EBI's Industry Partnerships office promotes pre-competitive collaboration and ensures that the institute's public offerings align with the needs of the private sector. As of January 2021, EMBL-EBI directly employed 697 staff (full-time equivalent) from 78 countries.

EMBL-EBI is a highly collaborative organisation, with data resources run in partnership with organisations throughout the world. It serves as the European node for globally coordinated data collection and dissemination projects, for example the worldwide Protein Data Bank (wwPDB) and the International Nucleotide Sequence Database Collaboration (INSDC). Its core databases are produced together with other world leaders, including the NCBI (USA), the National Institute of Genetics (Japan), the SIB Swiss Institute of Bioinformatics (Switzerland), the Wellcome Sanger Institute (UK) and the Cold Spring Harbor Laboratory (USA). There were 136 collaborative grants with researchers from 674 institutes in 62 countries in 2020.

The largest part of EMBL-EBI funding comes from the governments of over 20 EMBL member states, associate, and prospect member states. Other major funders include the UK Research and Innovation (UKRI) via the UK's Biotechnology and Biological Sciences Research Council (BBSRC) and Medical Research Council, as well as the European Commission, the US National Institutes of Health, the Wellcome Trust, and members of the EMBL-EBI Industry Programme. The UK Government sustained capital investment has helped EMBL-EBI develop a robust compute infrastructure to efficiently manage the growth and diversity of biological data held in its public data resources.

Investment by funders enables EMBL-EBI to offer the world's most comprehensive and integrated collection of open access molecular databases for the life sciences. This creates substantial benefits and efficiencies for users and funders alike.

1.2 Background to this study

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

EMBL-EBI has been applying a range of methods for exploring the value, benefits, and impacts of research data resources, including those developed by Charles Beagrie Limited and Prof John Houghton and applied in previous studies of the economic impact of research data centres and EMBL-EBI (Beagrie and Houghton 2014, Beagrie and Houghton 2016).

Over the last nine years Neil Beagrie and John Houghton have completed studies assessing the economic value and impact of the Economic and Social Data Service (Beagrie, Houghton et al 2012), the Archaeology Data Service (Beagrie and Houghton 2013a), the British Atmospheric Data Centre (Beagrie and Houghton 2013b), and the European Bioinformatics Institute (Beagrie and Houghton 2016). These studies combine qualitative and quantitative methodologies to measure the value and impact of open research data and associated services and tools.

These methods have broken ground in measuring the value and impact of major research data resources and open research data. In an international state of the art review, the Economic and Social Data Service impact study was one of three studies highlighted to the UK Department of Business, Innovation and Skills as being particularly good examples of 'good practice' in the measurement of the economic impacts of large-scale science facilities (Technopolis 2013, p31-2).

EMBL-EBI and the authors have worked together, applying these methods not only to evaluate the institute's data resources but also to develop potential synergies with components of EMBL-EBI's own evaluation programme.

1.3 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. This includes use of counter-factual models. The programme is based on statistics/user data gathered both internally (annual report) and through surveys (general user and training).

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 (Research Information Network 2011).

EMBL-EBI was also 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 (Guittard et al 2013).

Two EMBL-EBI managed resources, the European Nucleotide Archive (ENA) and the Protein Data Bank in Europe (PDBe), are among the repositories examined in a research article exploring extensive reuse of data from bioinformatics resources in research articles and patents. This was used to demonstrate long-term value of biological data in life science research and biotechnology industry (Bousfield et al 2016).

In 2015-2016, Charles Beagrie Ltd undertook the first EMBL-EBI Value and Impact study. The study employed a range of economic approaches to explore the costs and cost savings involved in using EMBL-EBI managed data resources, their value to users and their impacts on the wider commercial, healthcare, and research communities (Beagrie and Houghton 2016).

More recently, ELIXIR (in which EMBL-EBI is a partner) has used approaches exploring impact pathways and indicators as a means of demonstrating the public value to funders and other stakeholders of a virtual and distributed research infrastructure for life science data (Martin et al 2021).

1.4 The discovery phase for this study

We divided this study into two stages. The initial discovery phase aimed to assess the potential impact of changes over the last five years and any major risks and mitigation measures needed for a new project.

Key findings from the discovery phase were:

- Developments in methodological approaches since 2016 suggest that there is, if anything, an increasing awareness and use of the economic approaches that we adopted in the previous study.
- Changes to the EMBL-EBI user population included substantial increases in the number of users and the intensity of their use.
- EMBL-EBI user surveys and web stats showed that the reach and use of resources was now even more international in scope than it was in 2015. Use from China and India in particular has increased significantly.
- There had been major changes to other survey response and completion rates. Over the period between 2015 and 2019 there has been around a 75 per cent reduction in the response rate from similar online surveys (see Appendix 1).

1.5 Layout of this report

Section 2 describes the techniques used to collect the data necessary for analysis and the approaches and methods used for measuring value and assessing economic impact.

Section 3 presents a brief description of the findings from the 2021 user survey, while the quantitative analysis of the value and impact of EMBL-EBI data resources is presented in Section 4. These are followed by some concluding remarks (Section 5).

Appendix 1 presents a description of the development of the survey questionnaire and results of the user survey conducted for this study, and presents the data underpinning the economic estimates.

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

2. Approaches and methods

2.1 What we are measuring

This study focuses primarily on the economic value and impact of the data resources managed by EMBL-EBI for its user communities. Those communities largely consist of researchers, with the managed data resources supporting academic research, as well as academic teaching and study, public and commercial research.

The study explores the value and impact of EMBL-EBI managed data resources as a way of better understanding the value and impact of EMBL-EBI (the organisation) - as was the case in the 2016 study. Hosting and curating these data resources represent approximately 80 per cent of EMBL-EBI frontline activities with the remaining 20 per cent that are not in scope being computational biological research.

The individual data resources managed by EMBL-EBI are effectively value-added services providing more than the data alone and drawing on a variety of central and group facilities. “Data resources” is a shorthand for services that can include: making available and maintaining a data resource; related ontologies and authority lists; user training; IT infrastructure and tools; and collaborations and partnerships.

EMBL-EBI has an extensive and diverse range of collaborators and partners who also contribute directly and/or indirectly to many of the data resources it hosts. This leverages the investment made by EMBL-EBI’s funders. There are also mutual benefits that flow in either direction from those collaborations. In section 2.4 we explore these collaborations and the treatment of investment and costs faced by collaborators.

Our approaches to estimates of economic impact are based on primary users, those who access directly the data resources managed by EMBL-EBI and who could be invited to respond to our user survey. Primary users of EMBL-EBI data resources include not just direct end-users but curators of other external data resources who incorporate those open data in their own services. As a result, there is extensive indirect secondary use and a considerable number of secondary users of EMBL-EBI managed open data, which are not included in our impact assessment.

In this study, we are measuring and estimating:

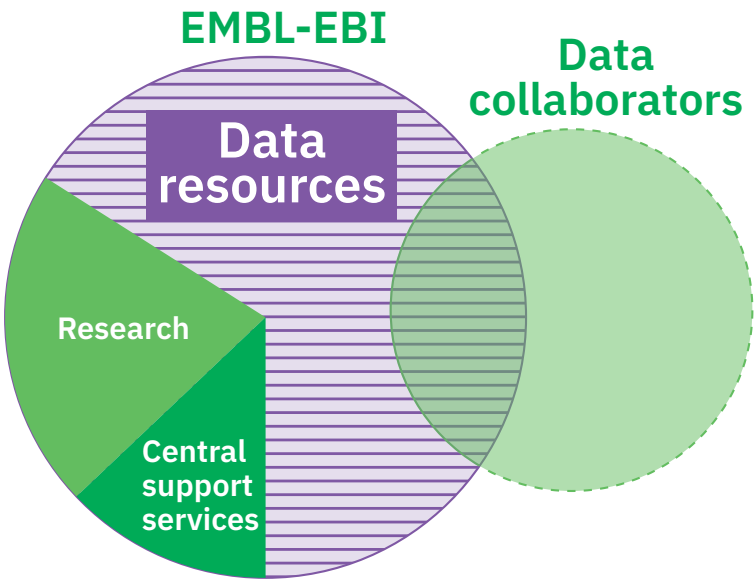
1. **Benefits:** The value and economic impact from use of data resources managed by EMBL-EBI.
2. **Costs:** The cost to EMBL-EBI funders for data resources managed by EMBL-EBI (i.e. operational budgets for the data resources plus an appropriate share of EMBL-EBI central services and capital investment) and the relevant costs incurred by collaborators in the data resources managed by EMBL-EBI.

3. **Benefit/Cost ratio:** The benefit/cost ratio of the data resources managed by EMBL-EBI.

Both the value and impact estimates and benefit/cost ratios are included throughout, as any choice between alternatives should consider both the benefit/cost ratio and net benefits.

FIGURE 2.1

Focus of study is the data resources managed by EMBL-EBI



We are estimating

The **value** and **impact** of data resources managed by EMBL-EBI

and

the **benefit/cost** of data resources managed by EMBL-EBI

Value and impact = from the use of the data resources

Expenditure (cost) = EMBL-EBI's expenditure on data resources (including share of central support services and capital investment) + relevant external contributions to data resources (from collaborators/partners, etc)

Source: Authors' analysis

2.2 A review of methods

Estimating the economic value and impact of open research data

Assessing the value and impacts of open research data and related data resources is still 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 realised.

We provided a general review of these methods in our 2016 report (Beagrie and Houghton 2016 section 3.1). Developments in approaches since 2016 suggest that there is an increasing awareness and use of the basic approaches that we adopted in the previous study.

For example, Sullivan, et al. (2017) adopted the methods and many of the parameters used in our 2016 study to explore the *Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank*, suggesting that a full user survey be conducted as the basis for a full impact assessment.

In *Investing in Science: Social Cost-Benefit Analysis of Research Infrastructures*, Florio (2019) puts forward a proposal for using cost-benefit analysis to evaluate the socioeconomic impact of public investment in large scientific projects (such as genomics platforms). Florio also shows how these costs and benefits can be expressed in the form of Net Present Value and other summary indicators.

Similarly, an increasing number of studies are exploring the time efficiency gains from the sharing and reuse of research data (e.g., Pronk 2019), and comparing those gains against the time costs involved in making data available for reuse, and the necessary infrastructure costs.

Fell (2019) undertook a Rapid Evidence Assessment which provides an excellent summary of work focussing on the economic impacts of open science (i.e., open access publishing and open research data).

2.3 Methods used to collect data

Desk research

Desk research included analysis of: recent evaluation literature; surveys of open research and open data communities; web publications and EMBL-EBI service publications; direct interviews with EMBL-EBI staff; 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 calendar years (January 2019 – December 2020) for the EMBL-EBI managed data resources included in the user survey. These are discussed further in section 4.1.

A user survey

An online survey was conducted aimed at measuring the value and impact of EMBL-EBI managed data resources for their users. We used our 2015-16 survey questionnaire with a number of updates and additions (See Appendix 1). The onset of the COVID-19 pandemic delayed the user survey for the 2020-21 study from Autumn 2020 until March 2021.The survey opened on 1st March 2021 and closed on 31 March 2021.

The questionnaire used a range of standard survey approaches, including the use of "critical instances", such as the last data accessed/downloaded (for users) and Likert scaling in selected questions. 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.

Users provided a wide range of examples of the benefits to them personally, and/or to research and the public in free text comments in the survey. These have been quoted throughout the report.

Our discovery phase for the study identified that there had been a significant decrease in response rates to similar surveys over the five years since our previous survey. Considerable effort was therefore put into promoting the survey to users. As a result, the survey enjoyed a good response rate and reasonable completion rates, especially given the topics and number of non-mandatory questions.

The EMBL-EBI impact survey of 2021, with 4 920 usable responses, is one of the largest recent datasets for an international survey covering open data or research.

2.4 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 managed data resources 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 in this study 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 managed data resources,

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 2.3).

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 estimated from user interviews and surveys, through questions about the time and costs involved in the discovery, access, and use of EMBL-EBI managed data resources.

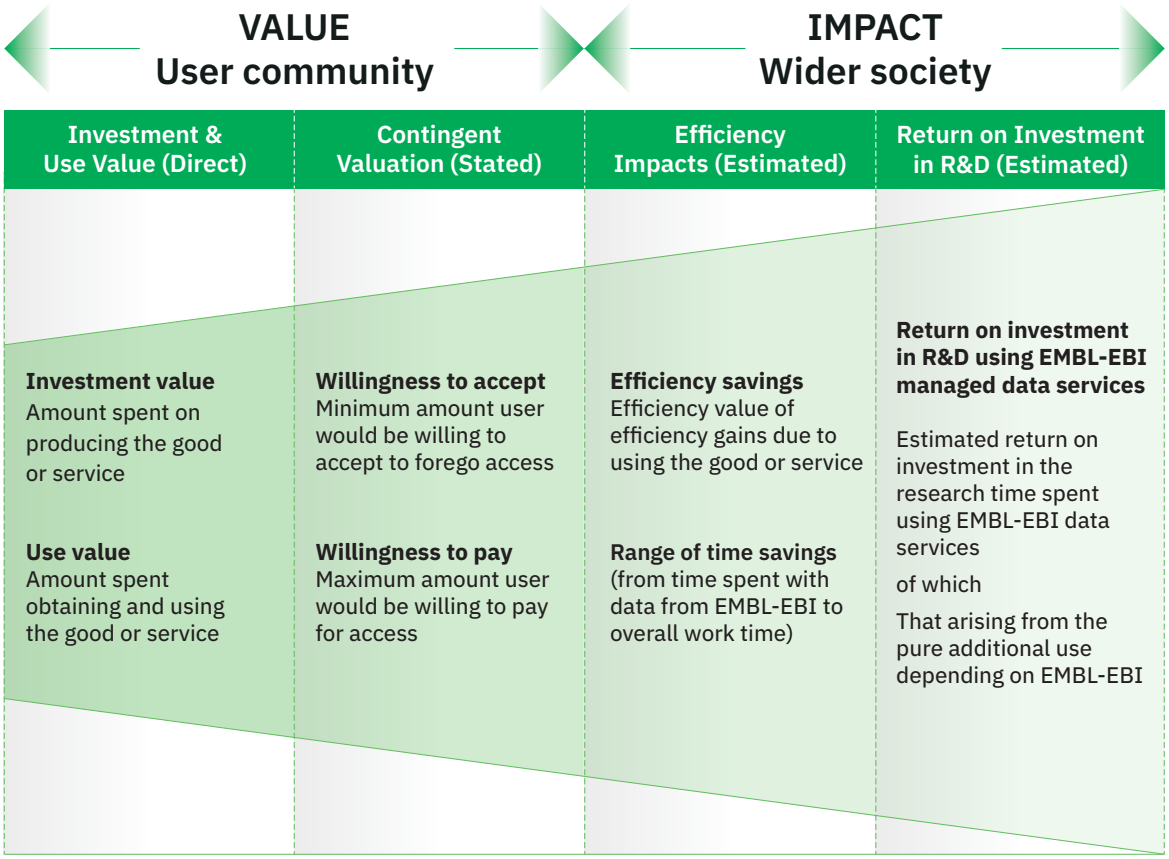
Activity costing follows the guidelines of the HM Treasury (i.e., Green Book, Magenta Book, etc.). Costings for job categories are taken from Times Higher Education surveys and related UK sources. Working times and loadings follow standard practices. As users are spread throughout the world, costings for survey respondents from each country are adjusted according to average income per capita using World Bank data (See Section 4).

Contingent valuation

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

FIGURE 2.3

Methods used for exploring the value and impacts of EMBL-EBI managed data resources



Source: Authors' analysis

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. Expressed values are converted to UK pounds using spot exchange rates on 5 April 2021 based on the country of each survey respondent.

Efficiency impacts

Wider benefits and impacts can be explored by looking at the efficiency gains enjoyed by research 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 (implemented as described above).

Return on investment in R&D

A sense of the scale of the value and impacts of EMBL-EBI managed data resources can be derived from an exploration of the potential return on investment in the research time spent using EMBL-EBI managed data resources, using a modified Solow-Swan model (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 managed data resources (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.

BOX 2.1

What value is and is not being captured?

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 in R&D 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

A number of modelling parameters are required. Those relating to returns to R&D over time are derived from US and UK R&D Satellite Accounts, while the discount rate used is that specified by the European Commission for use in public policy analysis (See Box 4.1). Throughout this report all the estimates are rounded.[2]

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).

[2] Default rounding is middle up for positive numbers.

3. Qualitative analysis

3.1 User demographics

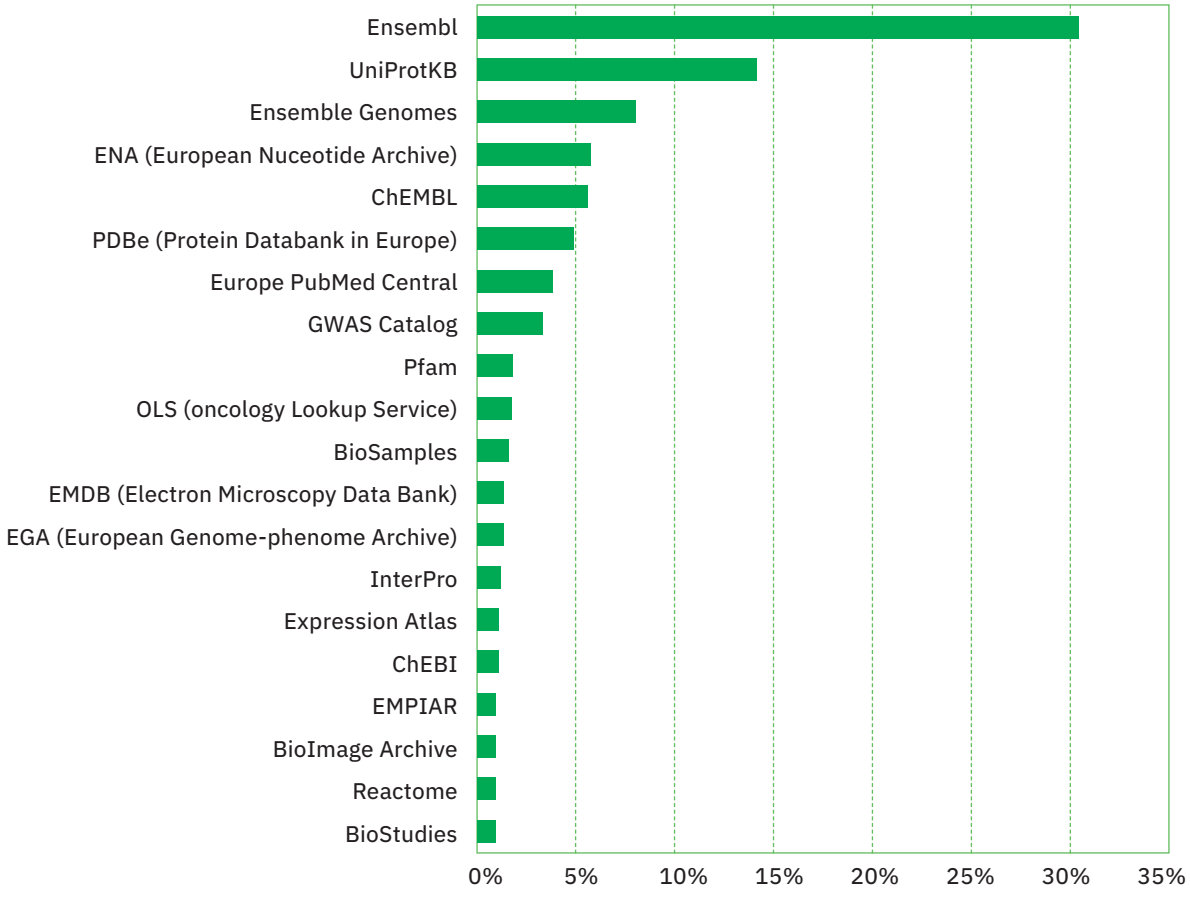
There were 5 662 responses to the user survey. Due to the number and variety of EMBL-EBI data resources, the survey was demanding, and a number of respondents did not complete enough of the questions to be included. After data cleaning, 4 920 responses were included (See Appendix 1 for details).

Responses were received from more than 100 countries all around the world, from Afghanistan to Vietnam. The largest number of respondents were from the United States (746), China (571), the United Kingdom (563), Germany (467), and India (274). More than 100 responses were also received from France (198), Spain (178), Italy (145), the Netherlands (112), and Japan (101).

The majority of respondents were in the academic sector (75 per cent), with 13 per cent coming from the corporate sector, and the remainder from other locations such as hospitals, government laboratories, and not-for-profits/charities (with research not being the main focus). This distribution reflects a higher proportion of non-academic and corporate users than did our 2015-16 survey, where 83 per cent were in the academic sector.

FIGURE 3.1

The Top 20 services most recently used (share of respondents, per cent) N = 4195



Source: Authors' analysis

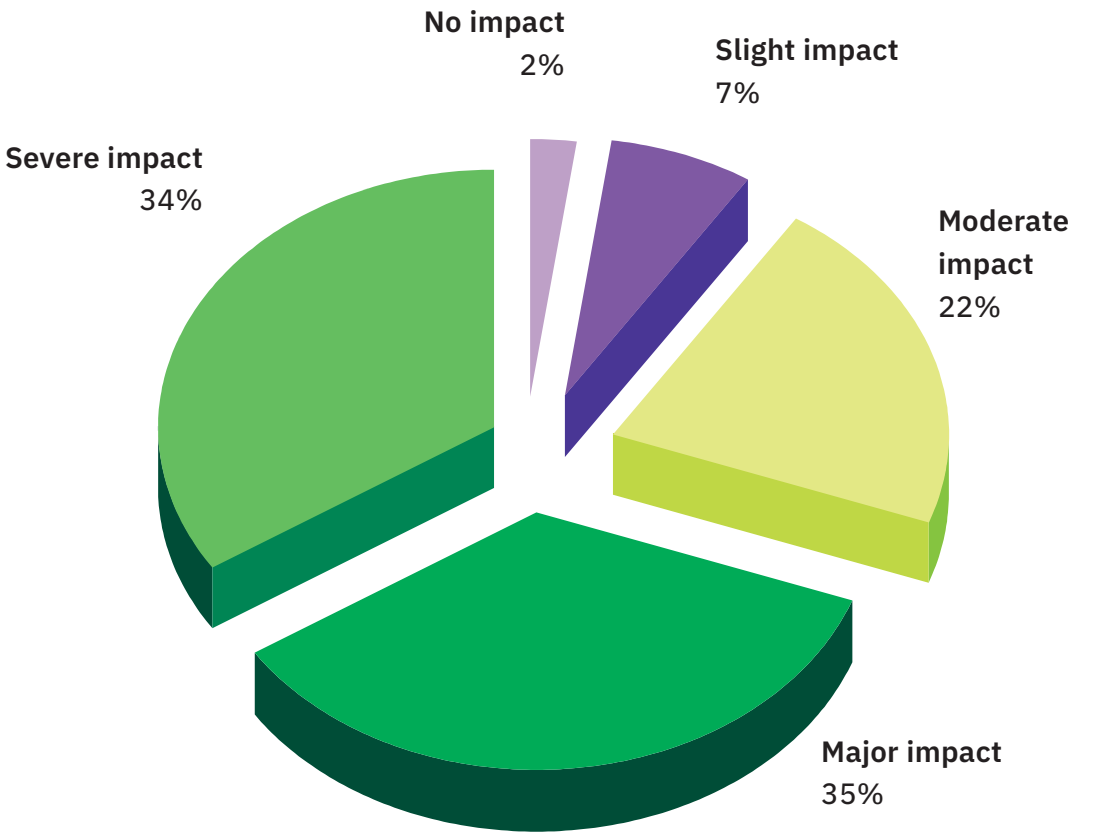
Forty-eight per cent of respondents described their work as mostly 'dry', working with computers (e.g., a bioinformatician), with 37 per cent saying mostly 'wet' laboratory (e.g., a scientist performing experiments in a lab or a manager overseeing 'wet' research). The remainder reported being in related science/research roles such as, hospital and clinical practice, or other work environments.

More than 30 per cent of respondents reported most recently using Ensembl, 14 per cent most recently using UniProtKB, 8 per cent Ensembl Genomes, 5.7 per cent ENA (European Nucleotide Archive), 5.5 per cent ChEMBL, 4.9 per cent PDBe (Protein Databank in Europe), and just less than 4 per cent Europe PMC (Figure 3.1).

The vast majority (83 per cent) of respondents use a web browser to access EMBL-EBI data resources, with almost 10 per cent downloading using FTP and a further 7.5 per cent using programmatic access.

FIGURE 3.2

The impact of not having access to EMBL-EBI (share of respondents, per cent) N = 4883



Source: Authors' analysis

3.2 The impact of access to EMBL-EBI

Almost 70 per cent of all respondents said that not having access to EMBL-EBI data resources would have a major or severe impact on their work or study (with a further 21 per cent saying it would have a moderate impact). This represents a substantial increase compared to our 2015-16 study when the equivalent response was 55 per cent.

> *"Losing access to EMBL-EBI services and resources would constitute a major setback for infection control and research."*
> [2021 User Survey respondent, USA]

> *"My work would be severely affected. The resources of the EMBL and EMBL-EBI allowed me to work even during the lockdown."*
> [2021 User Survey respondent, Italy]

> *"We use ChEMBL all the time to access public domain chemical structures & published data. Having to recreate that would not be possible for a single company."*
> [2021 User Survey respondent, UK]

> *"We are a genomics start-up company and without access to open data sources like EMBL-EBI services, the cost to build a market presence in a highly competitive environment would not be possible."*
> [2021 User Survey respondent, USA]

> *"EBI resources are the lifeblood of my research group. Without these services, a full two-thirds of my group's research efforts would suffer dramatically."*
> [2021 User Survey respondent, Belgium]

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

The majority of respondents (65 per cent) said they could not have obtained the last data resource they used from another source, with 35 per cent saying they could have (N = 4 141). More than 85 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 15 per cent said they could have done so (N = 3 962).

Combining Q20 and Q22, of the total 3 914 answering both questions, 2 252 (58 per cent) could neither have obtained elsewhere nor created/collected the last data they used themselves - up from 45 per cent in our 2016 survey. This gives an indication of the proportion of use of EMBL-EBI data resources that is additional use, which could not otherwise have occurred. To explore the impact of EMBL-EBI managed data resources on their user community, research respondents were asked to estimate any resulting change in their research efficiency. Among those reporting a zero or positive impact, the mean was 53 per cent (N = 2 965), with a median of 50 per cent. In similar surveys that we have run in different disciplinary fields, the reported efficiency impacts also have been considerable (Beagrie and Houghton 2014).

> *"My research efficiency on looking up and using genome references is 100 fold improved - it's like comparing cell phones with tin cans and bits of string."*
> [2021 User Survey respondent, USA]

> *"No data no science for me, you store, curate and offer what feeds my work and without that I would need to change completely my research."*
> [2021 User Survey respondent, Chile]

> *"My work will collapse without EMBL-EBI resources."*
> [2021 User Survey respondent, Israel]

> *“EMBL-EBI has transformed what I and my team and students are able to conceptualise and achieve. Wouldn’t be able to do 1/10 of what I can or have done over last 20 years.”*
> [2021 User Survey respondent, UK]

Just 27 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 some specific problems in what they were trying to do.

> *“All the isoforms of my gene changed with the last update and there is some incomplete data.”*
> [2021 User Survey respondent, Australia]

4. Quantative analysis

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 and the data underpinning this analysis can be found in Appendix 1.

4.1 Data limitations and estimates

Key operational data include financial and technical elements, including the annual expenditure on the operational provision of EMBL-EBI data resources, together with financial and in-kind contributions from collaborators in the ongoing development of knowledge-based (value-added) data resources; 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 resources.

There are number of important limitations and caveats:

- **Expenditure:** It is difficult to identify and apportion all the relevant elements of expenditure across the range of archival and value-added data resources. A number of estimates must be made when apportioning capital expenditures, in relation to collaborator contributions, and the value of in-kind contributions.
- **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 not all requests are successfully completed (e.g., when an FTP fails due to network interruption and is restarted), and sometimes a 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 EMBL-EBI's data resources 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 site), 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 allocation). The unique hosts counts can also include automated and robot/spider activities.
- **Users and data requests:** Our approaches to estimates of economic impact are based on primary users, those who directly access the data resources managed by EMBL-EBI. However, it is quite common for some primary users to curate a local copy of EMBL-EBI data or to integrate it with other sources for use by local or external users. In our 2021 user survey, 21 per cent did so. From our survey analysis, secondary user numbers could be double that of primary users. There may be some overlap

between primary and secondary users but a major part of this secondary use would fall outside EMBL-EBI (and our) primary user counts. As such, the host and request counts recorded by EMBL-EBI logs are necessarily understating the full extent of use.

Expenditure

In most fields, open research data resources are archival in nature (i.e., a static output of the research process that is archived and curated for sharing in a data repository). In such cases, the expenditure on various activities can be relatively easily attributed to the researchers, other depositors (i.e., government agencies, etc.), or to the data repository's operations. However, in biosciences in particular, some data are more actively and continuously developed as part of knowledge bases, and are value-added in nature.

- The relative scale of activity and open data in the biosciences also drives collaboration and the collaborations that have evolved can be very diverse. In these cases, the attribution of expenditure is more difficult still as there are more players and activities involved – not only the original researchers/depositors and repository operator, but a number of collaborators who may actively exchange and extract data from other sources and add value to the original data by making further links between the data, developing ontologies, sharing tools or interfaces, etc.
- These collaborations can involve a continuum from relative organisational independence of action (e.g., deposition or exchange), through to increasing inter-dependency between the organisations (e.g., some shared service components such as ontologies or shared interfaces), or ultimately a "jointly produced" service confederating their impact together (e.g., a jointly produced resource with delivery via a single website or with combined site web statistics). Figure 4.1 shows these main types of collaboration and their suggested cost implications for this study.

FIGURE 4.1

Main types of collaboration and their data resources cost implications for this impact study

Deposition*	Costs = cost of repository ops + deposition costs where appropriate
Example: European Nucleotide Archive (ENA)	
Exchange*	Costs = cost of repository ops + exchange costs for collaborators where appropriate
Example: Europe PMC (and PMC)	
Shared Service Components*	Costs = cost of repository ops
Example: Protein Data Bank in Europe (PDBe) and wwPDB partners	
Jointly Produced (i.e., confederated impact)*	Costs = cost of repository ops + collaborator costs
Example: UniProt	

Notes: * All data resources costs should make some allowance for the wide range of in-kind contributions, which may include data deposition costs. A data service may have multiple types of collaboration, multiple collaborators, and need to combine respective costs for each of them.

Source: Authors' analysis

When conducting impact studies, the implications for costs should be considered together with the implications for impact. We need to be able to match data resources' web statistics, user survey data and user estimates for services in scope, with respective data resources' investment expenditure. The focus of impact in this study is primary use. When partners have separate primary users who are not combined for impact reporting, a large part of the data creation and collaboration should be treated as a "sunk cost" and a "sunk impact (secondary use)" between the partners.

A sunk cost (also known as retrospective cost) is a cost that has already been incurred and cannot be recovered.[3] As we are expressing all costs and impacts in annual terms (circa 2020), all historical expenditure is treated as a sunk cost. Current expenditure that would have been incurred regardless of the data hosting and sharing can also be treated as sunk. Hence, be they historical or current annual, the costs of producing the original research data can be treated as a sunk cost, as it is a part of the research process that would have taken place whether the data were to be shared or not. So too can the direct costs of data deposition by researchers where data sharing is normally a requirement of the research grant or of publication.

[3] See https://en.wikipedia.org/wiki/Sunk_cost

Exchange, similarly, would normally involve sunk costs and sunk impact between the partners when they maintain different websites and have separate primary users. Curated data flow in both directions and they are respectively sunk costs for the investment expenditure or secondary use excluded from the impact. Exchange costs are incurred by all partners.

Partners may share some service components (e.g., ontologies, registries, infrastructure for data exchange or data collection) and incur their relative share of costs. All partners will benefit from the efficiencies and reduced duplication of effort arising from this.

Where collaboration goes further, to jointly produced services, and involves a shared group of primary users (e.g., a single website), the financial and in-kind contributions from all collaborators and partners in the ongoing development of the data resources will be in scope. “Jointly produced” currently involves a small number of data resources within this study but includes UniProt, which is one of the larger services and is therefore significant.

In all cases the data resources costs should include repository operational costs (whether funded directly to the resources or indirectly through the repository via central services). In some research fields it may still be necessary to account for data deposition costs (i.e., fields in which open access / open data are not the norm). All data resources costs estimates should make some allowance for the wide range of in-kind contributions (e.g., researcher/student time, university infrastructure and estates, etc.).

Taking EMBL-EBI annual expenditure in 2020 and subtracting expenditures relating to research activities, gives an estimated annual expenditure on data resources of around £78 million. This includes shares of annualised capital expenditure, central services and other overheads. While we have not obtained detailed data on collaborator expenditures (and recognise this may be very difficult, perhaps impossible, to do in the absence of activity-based costing), we have made an allowance based on available information and suggest that collaborators and partners may contribute around 30 per cent of total expenditure in scope. Combining these figures, we estimate total annual expenditure on EMBL-EBI managed data resources to be around £110 million.

User population

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

- EMBL-EBI log data reported data requests from around 41 million unique hosts during 2020. 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 website 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 ~7 times unique visitor overestimation in a month and around ~80 times overestimation in a year, 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 have been analysed across the data resources in scope *and* across the year (i.e., effectively treating the collective EMBL-EBI managed data resources as a single website), a first adjustment is to divide the reported 41 million unique hosts by 80 to estimate the number of unique users.

- The web logs do not include unique hosts that may use other access methods, such as FTP, Aspera, and APIs, if they do not use web access as a part of the process. Consultation suggests that the number of users that may use the alternative access methods exclusively (of web access) will be small, perhaps 2.5 per cent of the overall estimated users. Hence a second adjustment is to add 2.5 per cent to the estimated users.
- 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 site) 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. The user survey received 5 662 responses from 5 320 unique IP addresses (i.e., 1.06 people per unique host), suggesting a third adjustment to account for the balance of multiple users behind unique hosts at a given time.
- We would normally treat this as very much a minimum estimate of users, but 2020 has been different in so many ways. One impact of the COVID-19 pandemic has been a much greater tendency for working from home, which will tend to inflate the number of unique IPs recorded in logs as users work in at least two places and home use will involve more dynamic address allocation. The extent of this impact can be seen in log data. During 2020 there was an increase of 62 per cent in unique IPs visiting the data resources included in this study, compared to an annual average of 22 per cent over the four years 2015 to 2019. Moreover, there has been a rapid increase in the number of visits coming from fixed and mobile ISPs (Internet Services Providers), which increased by 19 per cent during 2019, and by 90 per cent during 2020. Assuming that COVID-19 may have led to somewhat higher growth in visits, it would seem reasonable to attribute half of the 2020 growth (31 per cent) to the working from home phenomenon. This suggests an adjustment (applied first), allowing for half of the annual IP growth during 2020 to be due to working from home.

These four adjustments suggest that the 41 million unique hosts reported in EMBL-EBI log data might represent some 450 000 to 500 000 unique direct users across the data resources during 2020. This is more than double that estimated for 2015 in our previous impact study.

This is likely to be a conservative estimate of the direct user population. Eighty-nine per cent of survey respondents said that research was part of their role. Hence, 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.

UNESCO (2021) reported 8.8 million full-time equivalent researchers worldwide circa 2018. Few countries report detailed statistics by field of research, with Australia being among them. The Australian Bureau of Statistics (ABS 2010) reported gross expenditure on R&D of almost AUD 28 billion, of which life-sciences (i.e., biological sciences, medical & health sciences, and agricultural & veterinary sciences) accounted for some 25 per cent. If these proportions were approximated worldwide and the ratios of staffing approximately matched funding ratios, then there might be around 2 million full-time equivalent life-sciences researchers worldwide who, at a maximum, would be *potential* users of EMBL-EBI data services. This estimate relates to full-time equivalent researcher counts, which based on our survey responses on hours spent on research might be 2.5 million head count.

This suggests that our estimate of around 450 000 to 500 000 direct users could account for around 20 per cent of the upper estimate of possible potential worldwide life-sciences researchers - a plausible, but conservative 'market share'.

4.2 Estimating the value and impact of EMBL-EBI managed data resources

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, then scaling to include non-wage labour costs using a 30 per cent uplift, based on the HM Treasury Green Book method. 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. These are adjusted for country of activity using World Bank income per capita data.[4] Across the respondents, this resulted in an average costing of around £38 per hour, including both staff at all levels and students.[5]

Investment and use value

The most direct indicators of value are investment value (i.e. the amount spent on the production and delivery of the good or service) and use value (i.e. the amount spent by users in obtaining the good or service). Measures of the investment made by users accessing the data resources suggest the minimum amount they are worth to them. For

[4] http://data.worldbank.org/data-catalog/world-development-indicators

[5] Standard guidelines as to working days and hours are used throughout (e.g. OECD, European Commission)

simplicity, both investment 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 (2020).

Investment value of EMBL-EBI managed data resources

As noted in section 4.1 (Expenditure) above, investment value includes EMBL-EBI's annual operational spending on data resources, their share of EMBL-EBI central services and annualised capital expenditure, as well as the relevant costs faced by collaborators in co-curating and adding value to the data. As the original creation of the data and its deposition (where a requirement of the research grant) are a part of the research process they are treated as 'sunk costs' (i.e., excluded). Similarly exchanged data are treated as a sunk cost. As we have been unable to source information on collaborative contributions directly, we have estimated them to account for 30 per cent of overall expenditure on EMBL-EBI managed data resources. During 2020 EMBL-EBI expenditure on data resources was around £78 million. Hence, we estimate total expenditure to have been around £110 million.

Use value of EMBL-EBI managed data resources

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

Access time/cost

Responses to the question about time to access and obtain the last data used varied widely, with a mean of 60 minutes reported.[6] However, there was a large number of low times reported and relatively few very long times reported, with the median reported being five minutes. This reflects the different methods of access.

> *"The services and resources available in EMBL-EBI enable in depth data analytics and meaningful interpretation of data. Ease of access to multiple, integrated sources of data is one of the key features of the various services."* [2021 User Survey respondent, USA]

Many found access easy, with 523 respondents saying it took one minute or less, with comments such as: *"It is a bookmark since many years"* [2021 User Survey respondent, Sweden], *"I knew what I was looking for"* [2021 User Survey respondent, Germany], *"Took less than a minute to find what I need"* [2021 User Survey respondent, Netherlands], and so on.

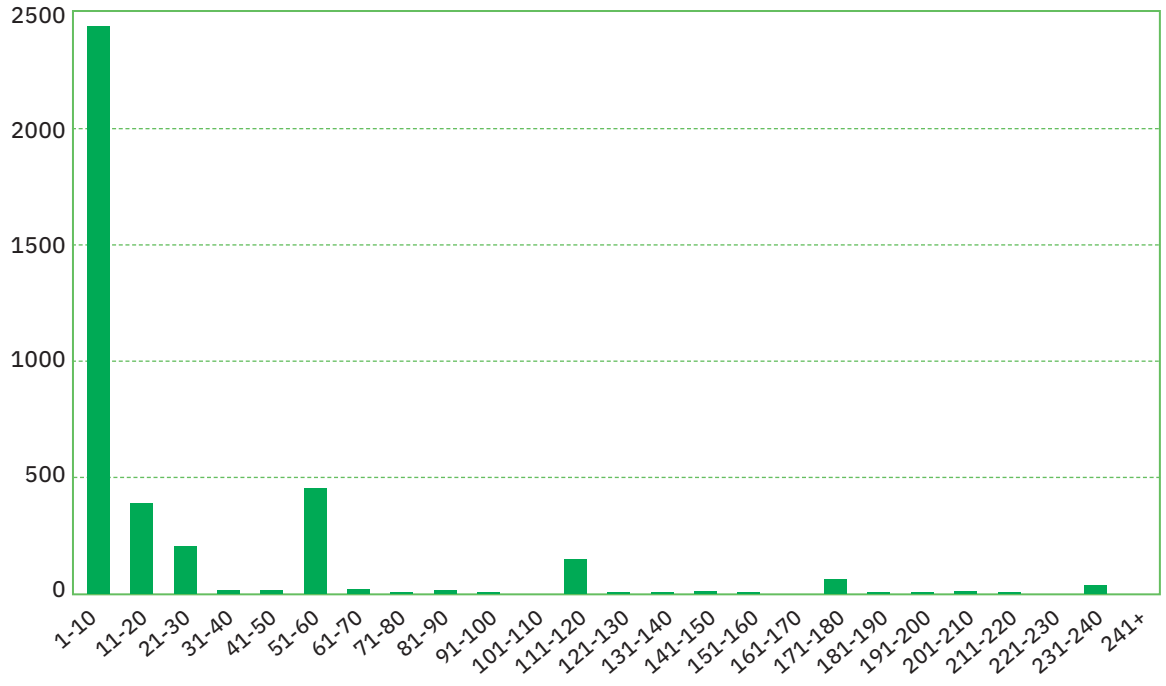
[6] It is important to note that there are limitations to use value as a metric. It is an expression of value, which does not change in direct proportion with the ease or difficulty of use. If it were easier to use EMBL-EBI data resources, there would be more use; and if it cost more time to use EMBL-EBI, it would be used less, with the impact on (use) value unlikely to be large - depending on whether expenditure on EBI data resources was realising increasing or decreasing marginal returns. It will also be driven by changes in expenditure on life-sciences research.

At the other end of the scale, 64 respondents reported access times of greater than 12 hours. Such cases were accompanied by comments about download speeds using FTP: *"The FTP is incredibly useful, but its structure is a tad byzantine"* [2021 User Survey respondent, France], difficulties setting up a programmatic access: *"My task required API access so I had to learn the API structure"* [2021 User Survey respondent, Germany], and so on.

Examining the distribution of reported access times suggests that the median would be a better indicator of central tendency (Figure 4.2). Hence, subsequent analysis is based on the median time to access and obtain the last data used.[7]

FIGURE 4.2

The distribution of reported access times – far from a normal distribution (N = 3 982)



Source: Authors' analysis

Simply multiplying the reported median time of last access by the mean hourly cost and the estimated number of users and data accesses during the last year suggests a total worldwide user access cost of around £465 million per annum (an average of around £1 158 per person across the survey respondents).

[7] In other cases the means and medians are similar, and the means are used for estimation.

*(median time of last access * mean hourly cost)*
** (estimated users * median accesses)*
=
£465 million pa

Use time/cost of EMBL-EBI managed data resources
Use value is expressed by the amount of money and time people are willing to spend using a product or services. An estimate of the wider use value of EMBL-EBI managed data resources 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.

The reported mean time spent on research per week was 31 hours per week and the median 30 hours per week (N = 3 136). Respondents reported spending a mean of 55 per cent of their research working time with data (N = 2 967) and 23 per cent of their time with data obtained from EMBL-EBI (N = 2 748). The medians were 50 per cent and 20 per cent, respectively.

Converting the reported times spent with data from EMBL-EBI to pounds at the average hourly cost suggests an approximate worldwide use value of £5.5 billion per annum (an average of £12 148 per person across survey respondents). This is equivalent to 49 times the estimated total costs (71 times EMBL-EBI's annual operational cost).

*(((mean time with EMBL-EBI data per week * mean hourly cost) * weeks pa) * estimated users)*
=
£5.5 billion pa
(benefit/cost = 49)

Note: During the last year EMBL-EBI managed data resources underpinned more than 140 million hours of research.

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 managed data resources was £2 757 (median £200), equivalent to £1.25 billion per annum across the worldwide user population (considerably higher than our 2015-16 study). This is equivalent to 11 times the estimated total costs (16 times EMBL-EBI's annual operational cost). [8]

*(mean willingness to pay * estimated users)*
=
£1.25 billion pa
(benefit/cost = 11)

Efficiency impacts

The value of the efficiency impacts of EMBL-EBI managed data resources among its research 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.

> *"I use EMBL-EBI every week if not every day to track down which protein or which gene I might be looking at - for antibiotic resistance, phenotype, virulence, etc. These databases are critical and central to my research."*
> [2021 User Survey respondent, USA]

Excluding the 27 responses reporting negative but un-quantified impacts, among those reporting a zero or positive impact, the mean was 53 per cent, with a median of 50 per cent (N = 2 965). Hence, the estimated efficiency impacts could be worth as much as £28 000 on average per person per annum, or more than £11 billion per annum across the estimated 20 per cent share of the worldwide research user population.

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 research/working time. If that were the case for all respondents,

[8] Because contingent valuation is based on preference theory, it is necessary to use the mean willingness to pay, as it is the only measure of central tendency that includes all stated preferences.

then the efficiency impacts would still, at a minimum, be worth an average £6 420 per person per annum, or £2.6 billion per annum across the worldwide research user population. This would, nonetheless, represent a benefit to users, and their funders, that at a minimum is 23 times estimated total costs (33 times EMBL-EBI's annual operational cost).

*(((estimated users * mean hourly cost) * share of time with data) * efficiency impact)*
=
£2.6 billion and possibly up to £11 billion pa
(benefit/cost = 23 to 102)

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 managed data resources (i.e., pure additional use that could not otherwise have occurred).

> *"UniProt is the best resource on the Internet for learning anything about proteins. Without it, our research would be dead in the water."*
> [2021 User Survey respondent, USA]

> *"Could not identify disease variants in our patients."*
> [2021 User Survey respondent, UK]

> *"Lab wouldn't exist."*
> [2021 User Survey respondent, Germany]

Some 58 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 58 per cent of use is effectively additional use. A further 8 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).

BOX 4.1

Estimating returns to research activities

Robert Solow won the Nobel Prize in Economic Sciences in 1987 for his work on a macro-economic model that subsequently formed the basis for Growth Accounting.[9] 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 2020), we then model the recurring returns.

Source: Authors' analysis

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 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).

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 managed 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 almost £1.3 billion annually and possibly as much as £9 billion over 30 years in net present value. Hence, the estimated value of additional use depending on the EMBL-EBI managed data resources is at a minimum 11 times estimated total costs.

[9] 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

*((((time with EMBL-EBI managed data pa / frequency of use) * hourly cost) * average returns to R&D) * additional share of total use)*
=
£1.3 billion annually
(£9 billion over 30 years NPV)
(benefit/cost = 11)

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

Net of access costs incurred using EMBL-EBI managed data resources, the potential savings from not having to obtain the data elsewhere or create/collect it themselves could be as much as £4.5 billion per annum. These savings are a part of the overall efficiency savings reported by survey respondents (accounting for around 39 per cent of total estimated efficiency savings).

*(hours to obtain * hourly cost) * (estimated users * share who could)*
+
*(hours to create * hourly cost) * (estimated users * share who could)*
=
£4.5 billion pa
(benefit/cost = 40)

Open data reduces the duplication of research effort. If the time/money saved by those users who could not have obtained the data elsewhere but could have (re)created it themselves were reinvested in research (i.e., the time saved enabled more research to be done), then the openly available data resources may have led to a total impact worth almost £6 billion per annum from reduced duplication of effort.

> *"A lot of my work would become more difficult if this service was not available - I would not be able to deposit structures so that others can use my work, and I would not be able to use deposited maps which are a valuable resource."*
> [2021 User Survey respondent, UK]

> *"Biomart and the other bundled resources are a god-send for me. Also the fact that the archives are very well maintained which allows me to version my pipelines really helps to reproduce the studies."*
> [2021 User Survey respondent, India]

Potential wider and longer-term impacts of EMBL-EBI managed data resources

One indicator of the potential wider and longer-term impacts of EMBL-EBI managed data resources is the impact of the research to which they contribute.

> *"Services and resources (e.g. DNA and genomes databases) are essential for the analysis, identification and publication of the DNA sequences we produce. Accessing published data for comparison purpose is also essential for our work."*
> [2021 User Survey respondent, Belgium]

> *"Centralized, validated publications, NGS data and sequences are crucial for the biotech sector"*
> [2021 User Survey respondent, Austria]

> *"It is a core component of our daily work, we/I visit EMBL/EBI at least once a week and when teaching it is one of our first Open access platforms we use. It is an important educational example for future generations. Simply put."*
> [2021 User Survey respondent, Spain]

Assuming an average 40 per cent return on expenditure on R&D and basing analysis on the estimated use value of EMBL-EBI managed data, modelling suggests that the wider value of the research facilitated by EMBL-EBI managed data resources during 2020 might be worth as much as £2.2 billion annually, or perhaps £15 billion over 30 years in net present value.

*(((mean time with EMBL-EBI managed data per week * mean hourly cost) * weeks pa) * estimated users) * average return to R&D)*
=
£2.2 billion annually
(£15 billion over 30 years NPV)

From the analysis presented above we can see that, of this, some £1.3 billion annually, or £9 billion over 30 years in net present value, 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 managed data resources.

> *"European Genome-Phenome Archive as part of EMBL-EBI is a valuable repository of cancer genomics data. Deposition in EGA is recognized by most scientific publications and serves as excellent platform of sharing scientific dataset."*
> [2021 User Survey respondent, China]

Of course, these values arise from the further research the data supports and its subsequent use, and cannot be directly attributed to EMBL-EBI managed data resources.

The latter are a contributing element. Nevertheless, these estimates give a sense of the scale and importance of the activities to which EMBL-EBI managed data resources make an important and widely appreciated contribution.

4.3 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 to the values below in an approximately linear fashion and, therefore, twice the users or double the intensity of use would double the value (and vice versa) and the reported values should be treated as indicative.

Value of EMBL-EBI managed data resources

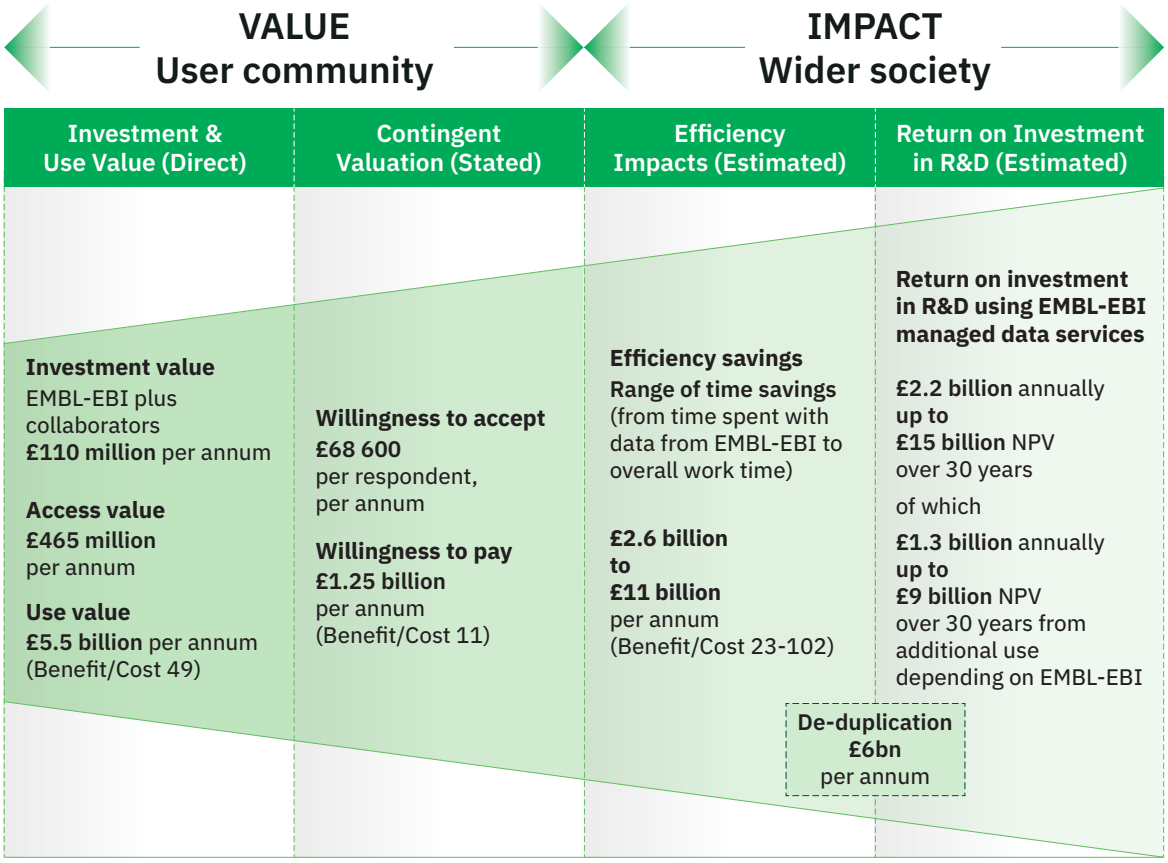
The most direct measure of the value of something is the amount of time and/or money people are willing to spend obtaining it (i.e., the activity-cost). Hence, the time users spend accessing EMBL-EBI managed data resources reflects what it is worth to them: an estimated £465 million during 2020 (72 per cent higher than in 2015).

Contingent valuation is an alternative approach to estimating what something is worth, measuring what users say rather than what they do. We estimate that EMBL-EBI users would have been willing to pay £1.25 billion for their access during the last year (285 per cent up on 2015).

These estimates give a sense of the (minimum) direct value of EMBL-EBI managed data resources to its user community and compare very favourably with the estimated £110 million annual expenditure on the data resources (i.e. a minimum direct value to users that is equivalent to around 11 times the estimated total costs).

FIGURE 4.3

The value and impact of EMBL-EBI managed data resources



Source: Authors' analysis

Impact of EMBL-EBI managed data resources

Users report that their use of EMBL-EBI managed data resources 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. These were worth an estimated £2.6 billion, and possibly up to £11 billion, during the last year across the worldwide user community—a benefit to users and their funders that, at a minimum, is equivalent to 23 to 102 times the estimated total 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 would typically 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 £5.5 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 £2.2 billion annually, or £15 billion over 30 years at net present value. So, during the last year, EMBL-EBI managed data resources facilitated the realisation of research impacts conservatively worth £15 billion in net present value, and possibly more because biomedical research is often characterised by higher-than-average returns to R&D - not least in a year characterised by a rapid response development of COVID-19 vaccines.

In the context of rapid response, a key value of open data lies in de-duplication of research effort. A separate new calculation combining approaches in the study (therefore shown in Figure 4.3 above as an insert) provides a supplementary way of estimating a major part of the value of research efficiencies. If the time saved from not having to (re) create the data themselves enabled more research to be done, it could be worth a further £1.5 billion per annum. Hence, the openly available data resources led to a total impact estimated to be worth almost £6 billion per annum from reduced duplication of effort.

Of course, it is important to consider the counter-factual and ask: how much of this impact might have been realised without EMBL-EBI, through alternative means? Some 58 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 managed data resources would have underpinned future research impacts worth an estimated £1.3 billion annually, or £9 billion over 30 years in net present value, that could not otherwise have been done (i.e., depends upon and is directly attributable to EMBL-EBI managed data resources).

It is important to note that these estimates of value and impact focus on direct users. A number of direct users are known to obtain data from EMBL-EBI and curate it locally, making it available to other secondary 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 managed data resources directly and indirectly is greater than the number of direct users alone, and our estimates do not capture the value for these secondary users of EMBL-EBI managed data resources.

4.4 How do the economic impacts compare with the previous study?

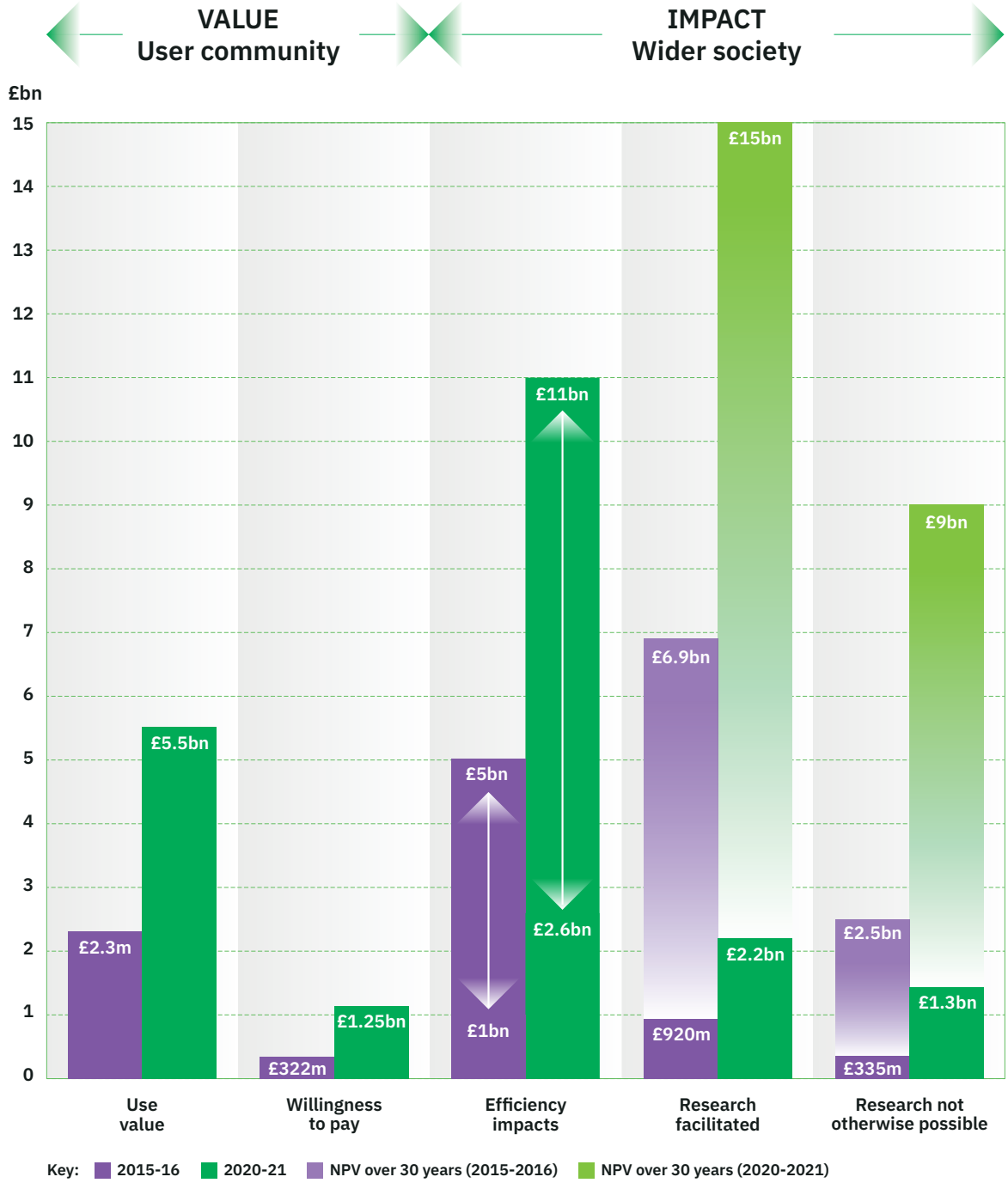
The current study and the original value and impact study conducted in 2015-16 are snapshots in time, and there are limits to how they can be compared. Nevertheless, the economic estimates provide a sense of the scale of economic impacts at the time.

presents the summary results on the value and impact of EMBL-EBI managed data resources from both the 2015-16 and 2020-21 impact studies - with the 2016 results shown alongside those from 2021. Reflecting the increased level and intensity of use of the managed data resources, the increasing maturity of the services and evolving research practices, the current study shows higher levels of value and impact across the board. [10]

[10] Mean access costs have fallen over the period as data resources improved their accessibility and/or users became more familiar with them.

FIGURE 4.4

The increasing value and impacts of EMBL-EBI managed data resources (2015-16 & 2020-21)



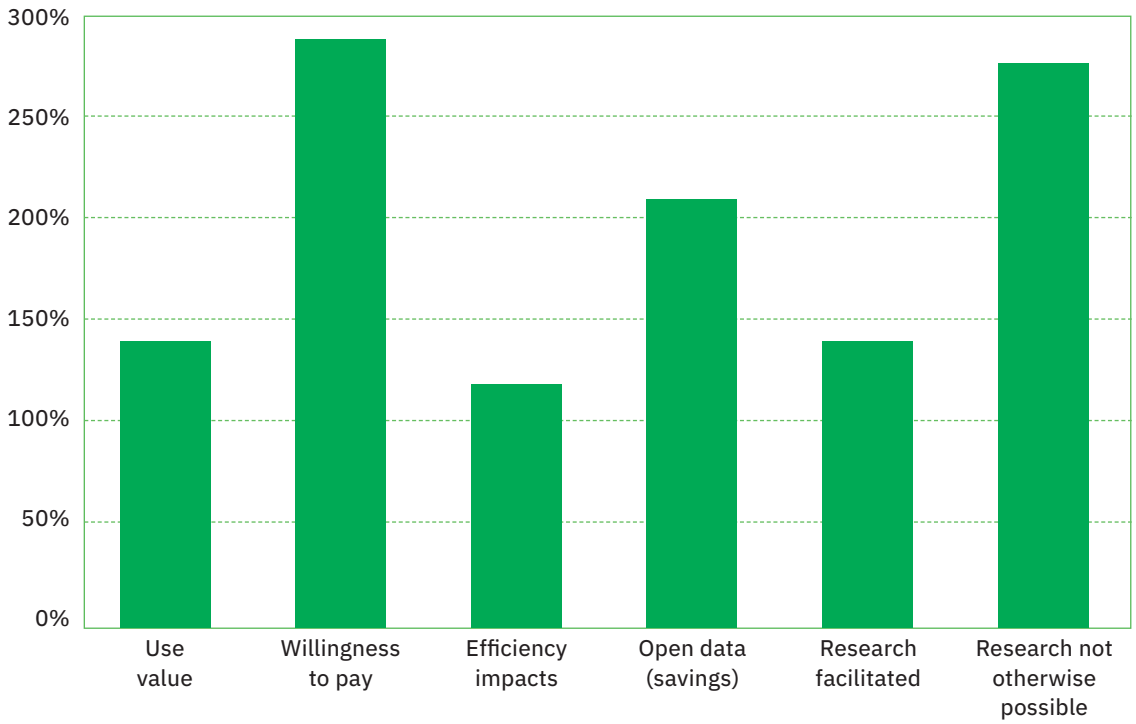
Note: All values are per annum apart from the Net Present Values (NPV) for returns on research which are over 30 years.

Source: Authors' analysis

Another way to look at these numbers is to compare the percentage change in each of the key value and impact indicators across the 2015-16 and 2020-21 studies. While many caveats apply, it is clear that key values and impacts have increased over the five years - by between 140 per cent and almost 290 per cent (Figure 4.5). Again, this reflects the increasing scale of use of EMBL-EBI managed data resources, and increasing reliance upon them as it becomes ever more difficult for users to operate without such resources.

FIGURE 4.5

The value and impacts of EMBL-EBI managed data resources (percentage increase 2015-16 to 2020-21)



Note: Open Data savings arise from the de-duplication of research effort in data creation.

Source: Authors' analysis

BOX 4.2

Putting the value and impact into context

While individual studies focus on different information services and content and use a variety of methods and measures, it is possible to explore their findings to give a sense of how the value and impact of EMBL-EBI managed data resources compare:

- Houghton (2011) estimated the benefit/cost ratio of the Australian Bureau of Statistics making data and publications freely available online and using Creative Commons licensing at 5.3 to 1.
- Tennison (2015) reported that a report by Nesta and the ODI adds to the evidence of the impact of open data. The report's analysis, undertaken by PwC, examines the effects of the Open Data Challenge Series (ODCS) and predicts the programme will result in a potential 10x return (£10 for every £1 invested over three years), generating up to £10.8m for the UK economy.
- Like other meta reviews, *Measuring the Value of California's Public Libraries, 2017-2020* found that investment in public libraries is a sound use of public funds: for every dollar invested in libraries, about $2-$10 are returned, with an average of between $3 and $6.
- King (2010) summarized findings relating to library services and concluded that: special libraries exhibit a return of 2.9 to 1, academic libraries 3.4 to 1 (for staff), and public libraries 5.8 to 1.

Source: Authors' analysis

5. Conclusions and observations

5.1 Conclusions

> *"Repositories like this are like temples (You can call it as Temples of Knowledge)."*
> [2021 User Survey respondent, India]

This study explores the value and impact of EMBL-EBI managed data resources in 2020 as a way to better understand the value and impact of EMBL-EBI (the organisation) - as was the case in the 2016 study. It includes 44 different data resources managed by EMBL-EBI supporting a diverse range of research and other use in the life sciences. The assessment approaches apply a combination of quantitative and qualitative methods and a counter-factual to provide a full picture of the nature and dimensions of value, and explore the range of impacts.

Both qualitative and quantitative analyses show that EMBL-EBI managed data resources are widely used, appreciated, and valued by their users:

- A very significant increase in research efficiency (when compared to our 2015-16 study) was reported by users as a result of their use of EMBL-EBI managed data resources, which we estimate to be worth at least at least £2.6 billion per annum worldwide (and potentially up to £11 billion per annum worldwide).
- Showing the extent to which EMBL-EBI managed data resources facilitate research that could not otherwise have been done, 58 per cent of respondents said they could neither have created/collected the last data they used nor obtained it elsewhere. It is estimated that the returns to research that could not have been done without access to EMBL-EBI managed data resources is worth some £1.3 billion annually, and up to £9 billion over 30 years (net present value).
- Sixty-nine per cent of respondents reported that it would have a major or severe impact on their work if they could not access EMBL-EBI managed data resources.

The quantitative and qualitative analyses independently show a similar picture of the value and impact of EMBL-EBI data resources: they are complementary, reinforce each other, and lend credence to the findings. 2020 was an extraordinary year because of the outbreak of the COVID-19 pandemic. The study shows that EMBL-EBI's managed data resources make a major contribution to science throughout the pandemic - 32 per cent of the survey respondents said they valued EMBL-EBI data resources more as a result of the pandemic, and 25 per cent said they had used them more.

The current study and the original value and impact study conducted in 2015-16 are snapshots in time, and there are limits to how they can be compared. Nevertheless, the

economic estimates can provide a sense of the scale of economic impacts at the time. It is clear that values and impacts have increased substantially over the five years between our 2015-16 and 2020-21 studies. Estimated user numbers have also more than doubled over this period. An increase in EMBL-EBI operational funding for the managed data resources from £47 million in 2015 to £78 million in 2020 alongside sustained UK host nation capital funding over this period has been crucial to enabling and supporting this growth.

Overall, we find EMBL-EBI managed data resources present exceptional value for money in terms of the value returned and impact compared to the costs of running them.

5.2 Observations

1. **Audience:** The findings of this study will be relevant to EMBL-EBI and its funders, its collaborators and partners, the wider life sciences community and others interested in open science policy and funding.

2. **The breadth and depth of this study:** The study is unique in terms of its broad coverage of data resources and multiple approaches to assessing value and impact. Such a study is data and time intensive, but the results potentially have wide application and longevity (even if the study is focussed on impact and value in a single year).

3. **Survey fatigue:** Survey techniques face an increasing challenge from survey fatigue. There has been a sharp decline over the last five years in survey response rates (as a percentage of the user population) and some decline in the time respondents devote to filling out questionnaires (which can negatively impact data quality). Although we had a record response to the 2021 survey, a significant communications effort was needed to achieve this. If current trends continue, assessing user populations or impact for open data via user surveys could become increasingly difficult and, ultimately, may no longer be viable.

4. **Activity-based costing:** Assessing total costs for data resources requires activity-based costing by a wide range of partners. Costs are not a simple topic and in practice can be very complex. Costs for a specific data service in any organisation may be distributed across many departments, activities and budget headings. Costs and staff may not be exclusively allocated to a specific data service (e.g., staff may also be engaged in research). In addition, issues such as multi-year capital investments, relevant share of capital and central services, and the value of in-kind contributions, all affect total costs. Guidance on activity-based costing and lifecycle costing for data resources has been the focus of a range of projects in recent years (e.g., National Academies of Sciences, Engineering, and Medicine. 2020, Beagrie et al 2008 & 2010). Costing collaboration, particularly for shared service components or jointly produced services as they continue to grow in importance, could be a future focus of further work in the sector.

5. **Counter-factuals weakening:** The continuing growth of open data in the life-sciences may potentially weaken the applicability of a counter-factual approach. As open data becomes the norm for research in the life sciences, researchers will increasingly have fading memory of, or no experience of, what to compare it with. It will be harder for survey respondents to answer questions relating to changes in research efficiency or reliably benchmark open data resources against previous alternatives.

6. **User estimates:** As noted in section 4.1, there are many difficulties in estimating user numbers (i.e., the number of *people* using the services) from unique IPs visiting or cookie-based tracking. However, as accurate an estimate as possible is necessary for value and impact estimates. Too often, unique IPs counts are taken at face-value without adjustments as equating with user counts, which can significantly exaggerate impact. EMBL-EBI has appointed a member of staff to compile web stats specifically to meet impact requirements and supported approaches used in this study that seek to achieve a more accurate estimation of user population. We hope the estimates of value and impact in the findings will be welcomed and widely recognised as more reliable as a result.

7. **Increasing automation:** A further complication is arising from increasing automation of data access, retrieval, and analysis (e.g., programmatic interfaces, data mining, etc.), which distances the human user from data access events. Ultimately this may affect the accuracy of answers given in user survey questionnaires, and make it increasingly difficult to reconcile the data resources' log data with data gathered through user surveys.

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., 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/

Beagrie, N. and Houghton, J. (2016) *The Value and Impact of the European Bioinformatics Institute Full Report* www.embl.org/documents/document/embl-ebi-2016-impact-report/

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.

Bousfield, D., McEntyre, J., Velankar, S., et al. (2016) Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources [version 1; peer review: 3 approved]. F1000Research 2016, 5(ELIXIR):160 https://doi.org/10.12688/f1000research.7911.1

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

Digital Science (2020) *The State of Open Data 2020*, December 2020, https://doi.org/10.6084/m9.figshare.13227875

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.

Fell, M. J. (2019) The Economic Impacts of Open Science: A Rapid Evidence Assessment, *Publications* 7, no. 3: 46. https://doi.org/10.3390/publications7030046

Florio, M. (2019) *Investing in Science: Social Cost-Benefit Analysis of Research Infrastructures*, MIT Press. https://mitpress.mit.edu/books/investing-science

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

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.

Guittard C., Müller M., Wolff S., Matt, M. (2013) *EvaRIO Case Study: EMBL-EBI December 2013 Deliverable D5.1* http://evario.u-strasbg.fr/uploads/autres-docs-BETA/EvaRIO_CAse_Study_EMBL-EBI.pdf

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

H.M. Treasury. (2020) *The Green Book: Appraisal and Evaluation in Central Government*, HM Treasury, London. Available http://greenbook.treasury.gov.uk/index.htm. http://www.hm-treasury.gov.uk/d/green_book_complete.pdf

H.M. Treasury. (2021) The Magenta Book: HM Treasury guidance on what to consider when designing an evaluation, HM Treasury, London. https://www.gov.uk/government/publications/the-magenta-book

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

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

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

Hrynaszkiewicz, I., Harney, J., and Cadwallader, L. (2021) A survey of researchers' needs and priorities for data sharing, Version 2 February 2021. https://doi.org/10.31219/osf.io/njr5u

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

King, D.W. (2010) Measuring Value: 30-years of Experience with Valuation Studies. Presentation, Return on Investment (ROI) - Lib-Value Workshop, George Washington University, June 26, 2010. http://www.libqual.org/documents/LibQual/publications/King_LV.pdf

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.

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 https://www.sussex.ac.uk/spru/documents/sewp161

Martin, C.S., Repo, S., Arenas Márquez, J., et al. (2021) Demonstrating public value to funders and other stakeholders—the journey of ELIXIR, a virtual and distributed research infrastructure for life science data. *Ann Public Coop Econ*. 2021; 1– 14. https://doi.org/10.1111/apce.12328

Measuring the Value of California's Public Libraries. https://www.library.ca.gov/services/to-libraries/value-of-libraries. See also Cole, N. and Stenström, C. (2020) The Value of California's Public Libraries, *Public Library Quarterly*. DOI: 10.1080/01616846.2020.1816054

National Academies of Sciences, Engineering, and Medicine (2020) *Life-Cycle Decisions for Biomedical Data: The Challenge of Forecasting Costs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25639

Pronk, T.E. (2019) The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal*, 18(1), p.10. DOI: http://doi.org/10.5334/dsj-2019-010

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

Sullivan, P., Brennan-Tonetta, P., Marxen, L. J. (2017), *Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank*. https://cdn.rcsb.org/rcsb-pdb/general_information/about_pdb/Economic%20Impacts%20of%20the%20PDB.pdf

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

Taylor & Francis Group (2013) *Open Access Survey: Exploring the views of Taylor & Francis and Routledge authors March 2013*, https://www.tandf.co.uk//journals/explore/Open-Access-Survey-March2013.pdf

Taylor & Francis Group (2014) *Taylor & Francis Open Access Survey June 2014*, https://www.tandf.co.uk//journals/explore/open-access-survey-june2014.pdf

Taylor & Francis (2019) *The Taylor & Francis 2019 Researcher Survey*, October 2019, https://authorservices.taylorandfrancis.com/wp-content/uploads/2019/10/Taylor-and-Francis-researcher-survey-2019.pdf

Tennison (2015) *The economic impact of Open data: what do we already know?* https://www.huffingtonpost.co.uk/jeni-tennison/economic-impact-of-open-data_b_8434234.html

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

UNESCO (2021) *UNESCO Science Report: The Race Against Time for Smarter Development*, UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000377433

Appendix 1: A summary of the user survey

This section provides a summary of the survey development, process, and results from the user survey conducted as a part of this study.

Development and process

We reviewed with EMBL-EBI staff all 39 data resources included in the 2015 user survey, removing those which were no longer current/in scope (e.g., had moved to another location), updating name changes where appropriate, and adding new services that have become established since 2015. In total we included 44 data resources in the 2021 user survey, deleting 7 and adding 12.

TABLE A1.1

List of EMBL-EBI managed data resources included in the 2020 Impact Survey

BioImage Archive	IntAct
BioModels	IntEnz
BioSamples	InterPro
BioStudies	MetaboLights
ChEBI	MGnify
ChEMBL	Mouse Resources: EMMA (European Mouse Mutant Archive)/IMPC (International Mouse Phenotyping Consortium)
Complex Portal	OLS (Ontology Lookup Service)
EFO (Experimental Factor Ontology)	OmicsDI
EGA (European Genome-phenome Archive)	Open Targets
EMDB (Electron Microscopy Data Bank)	PDBe (Protein Databank in Europe)
EMPIAR (Electron Microscopy Public Image Archive Resource)	PDBe-KB
ENA (European Nucleotide Archive)	Pfam
Ensembl	PRIDE (PRoteomics IDEntifications)
Ensembl Genomes	QuickGO/GOA (Gene Ontology)
Enzyme Portal	Reactome
Europe PubMed Central (Europe PMC)	Rfam
EVA (European Variation Archive)	RNAcentral
Expression Atlas	SureChEMBL
GWAS (Genome-Wide Association Studies) Catalog	UniChem
HGNC (HUGO Gene Nomenclature Committee)	UniProtKB
Identifiers.org	VectorBase
IGSR/1000 Genomes	WormBase

Source: Authors' analysis

The user survey conducted for our previous study had proved very successful, both in terms of the level of response and the quality of the information obtained. Consequently, the questionnaire for the 2021 study was very similar. Nevertheless, there were a number of areas in which we developed the questionnaire further:

- We reviewed the survey questions relating to frequency of use, offering more refined categories than before;
- We extended the user survey questions on secondary use to ensure that we captured

as much information as possible;

- We included a small number of new questions (Q25-Q27) on the coronavirus pandemic and use of the European COVID-19 Data Portal;
- We added European countries individually to the drop-down list.

The survey questionnaire was developed iteratively by the project team, with external review and input from EMBL-EBI staff, and included pilot testing by external users.

Our discovery phase for the study identified that there had been a significant decrease in response rates to similar surveys over the five years since our previous survey. The EMBL-EBI annual user survey has been run in same format since it started, so that results can be compared over time. Over the period between 2015 and 2019, despite an almost doubling of unique IPs and higher usage, there has been a steady decline in both the number of survey respondents and in the survey completion rate. Taken together there has been around a 75 per cent reduction in the response rate from the presumed user population. Publishers Taylor and Francis conducted international researcher surveys on open access in 2013, 2014, and 2019. They sent targeted emails to researchers who have published in their journals. Their survey response rates have also been declining steeply: from 19 per cent (14 769 responses) in 2013 (Taylor & Francis 2013); to 9 per cent (7 936 responses) in 2014 (Taylor & Francis 2014); and to 3 per cent (2 755 responses) in 2019 (Taylor and Francis 2019), a trend which is very similar to the EMBL-EBI decline in responses 2015-2019. Considerable effort was therefore put into promoting the survey to users. We agreed a high-level survey communication strategy for launch of the user survey with EMBL-EBI.

In addition to direct email and social media invitations to participate in the survey (as used in 2015), the survey invitation was placed on the webpages of 17 EMBL-EBI data resources who were able to do so and on the news section of the EMBL-EBI website; in the newsletters of EMBL and partners such as ELIXIR; and a short video by the EMBL-EBI directors inviting participation in the survey was placed on YouTube. Given low historic participation rates by Chinese users, some targeted promotion also occurred in China with the invitation translated into Chinese and circulated by partners based there.

As a result, the survey enjoyed a good response rate and reasonable completion rates, especially given the topics and number of non-mandatory questions. The EMBL-EBI user survey of 2021, with 4920 usable responses, is one of the largest recent datasets for an international survey covering open data or research. For comparison the Springer Nature and Digital Science *State of Open Data 2020 Report* survey had 4 500 responses (Digital Science 2020); the *PLOS Data Sharing Survey* of 2020 had 1 395 responses (Hrynaszkiewicz et al 2021); the Taylor & Francis *Researcher Survey* of 2019 had 2 755 responses (Taylor & Francis 2019); and the Technopolis *Measuring the value and impact of the Europe PMC repository Survey* of 2018 had 1 560 responses (Technopolis 2019).

We received 5 662 responses to the survey. The completion rate (those reaching the end of the survey) was a very acceptable 76 per cent, and respondents spent an average

of around 10 minutes completing the survey. Due to the number and variety of EMBL-EBI data resources, 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, 4 920 responses were included in the analysis.

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 646 responses that did not get beyond questions on frequency of use of various data resources. A further 96 responses were deleted as they appeared to be protest and/or non-answers.

Survey findings

The findings from the 2021 user survey are presented below on a question-by-question basis.

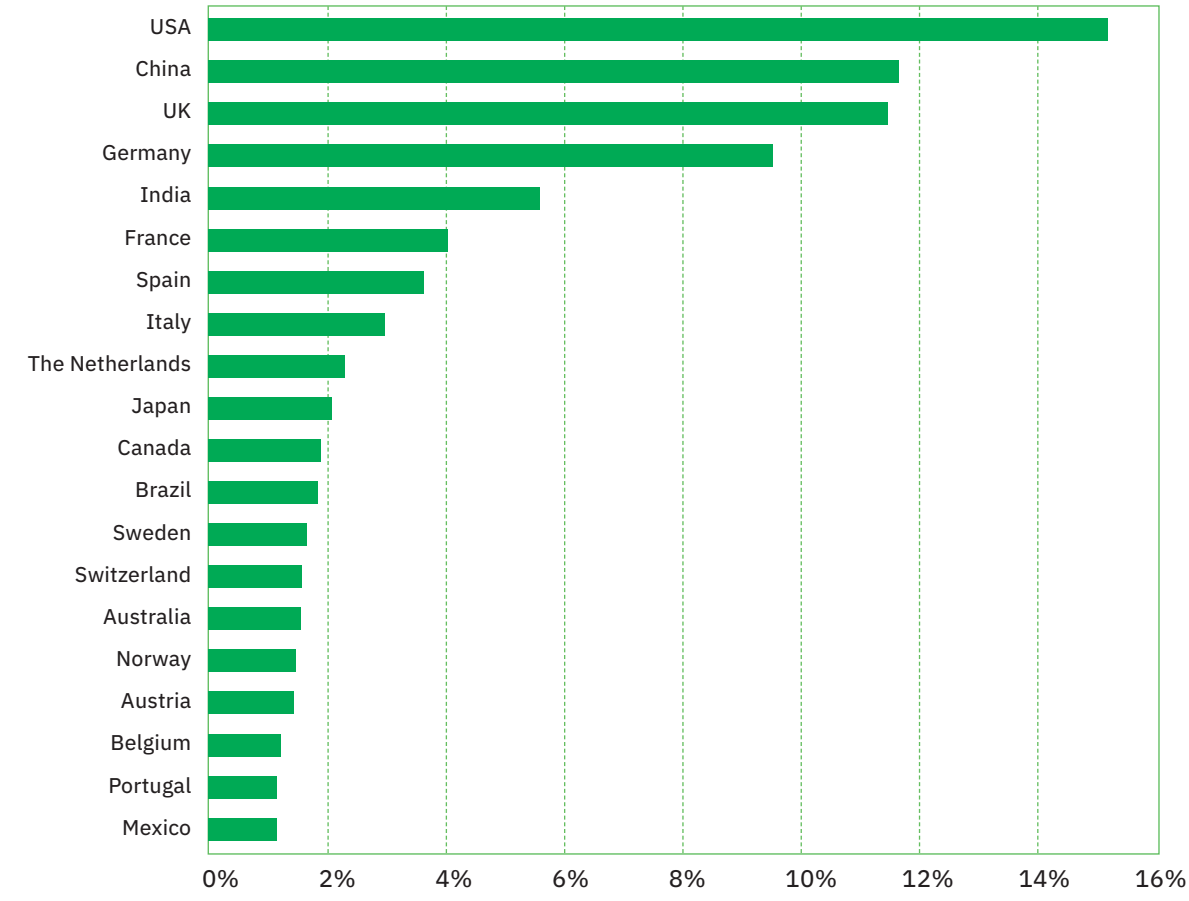
Demographics

The first few questions sought to explore 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 in which you work or study?
(Your answer to this question will be used to identify your local currency)
Responses were received from more than 100 countries all around the world, from Afghanistan to Vietnam. The largest number of respondents were from the United States (746), China (571), the United Kingdom (563), Germany (467), and India (274). More than 100 responses were also received from France (198), Spain (178), Italy (145), the Netherlands (112), and Japan (101) (Figure A1.1). The responses to this survey reflect much wider use of EMBL-EBI data resources than did our 2015 survey, with much greater use from non-European and developing countries.

The respondents' currency was derived from their answers to Q1, and their value responses converted to British Pounds (GBP) at the spot exchange rate reported by Google Finance on 5th April 2021 (i.e., as close to the time of the survey as possible).

FIGURE A1.1

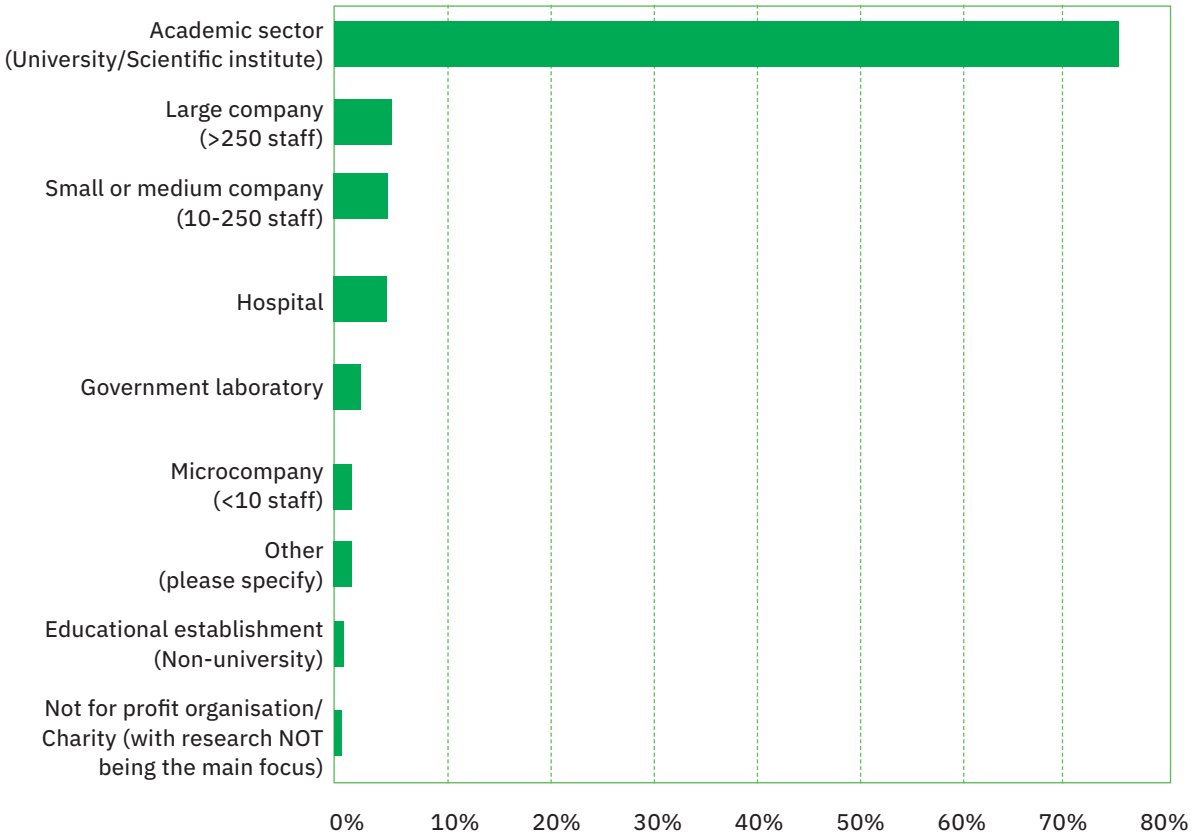


Source: Authors' analysis

Q2. Your Main Affiliation & Sector?

The majority of respondents were in the academic sector (75 per cent), with 13 per cent coming from the corporate sector (Figure A1.2). This distribution reflects a higher proportion of non-academic and corporate users than did our 2015 survey, where the respective distribution was 83 per cent for the academic sector and 9 per cent for the corporate sector.

FIGURE A1.2

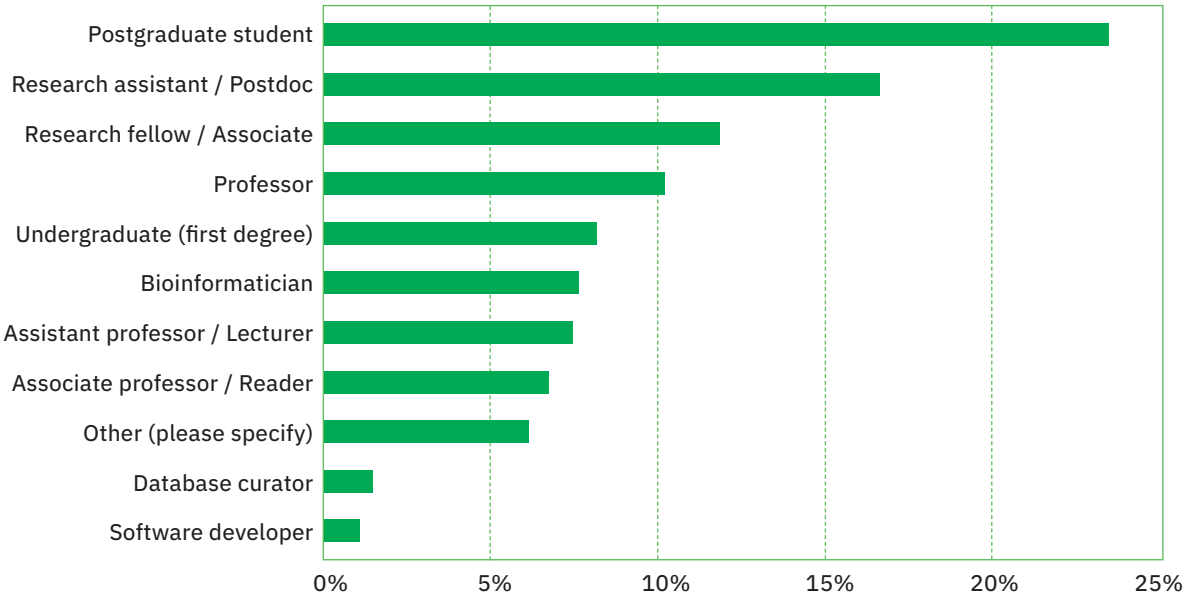


Q3 & Q4. What is your main role within this affiliation?

Respondents were taken to different versions of Q3/Q4 depending on their sector. Among the 3 707 academics, more than 23 per cent reported being a postgraduate student and a further 8 per cent were undergraduate students (Figure A1.3). Almost 30 per cent of the 1 220 non-academic respondents reported their role as researcher, 15 per cent bio-informatician, and 8 per cent principal investigator (Figure A1.4).

FIGURE A1.3

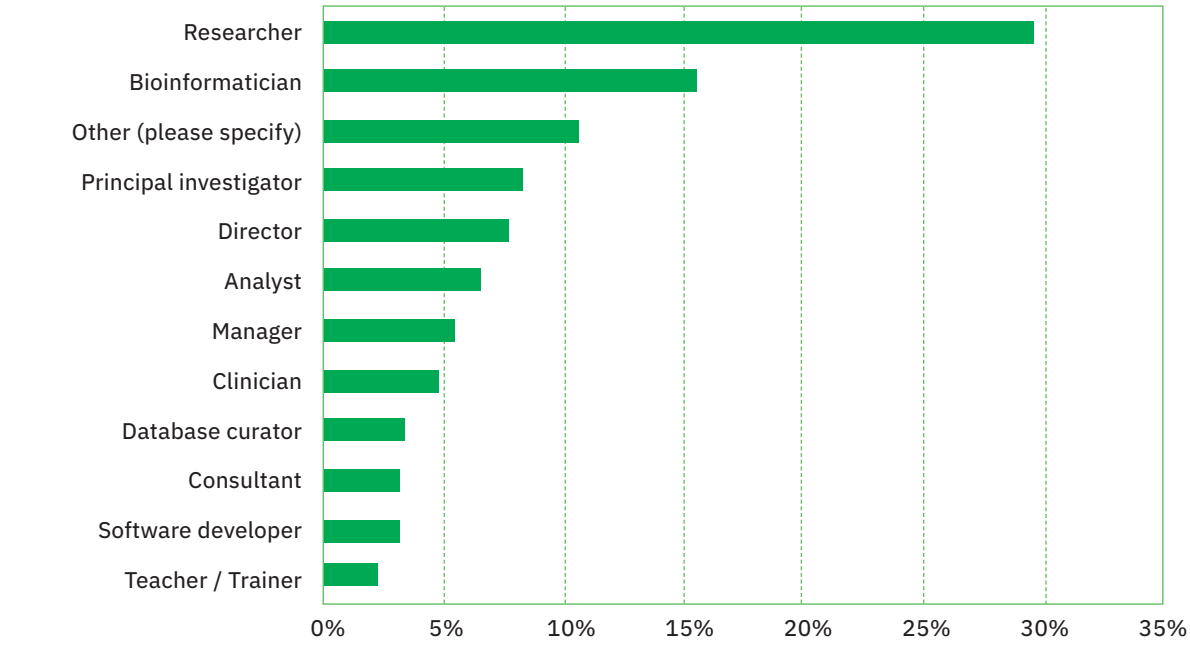
Main role in academic sector (share of respondents, per cent)
N = 3707



Source: Authors' analysis

FIGURE A1.4

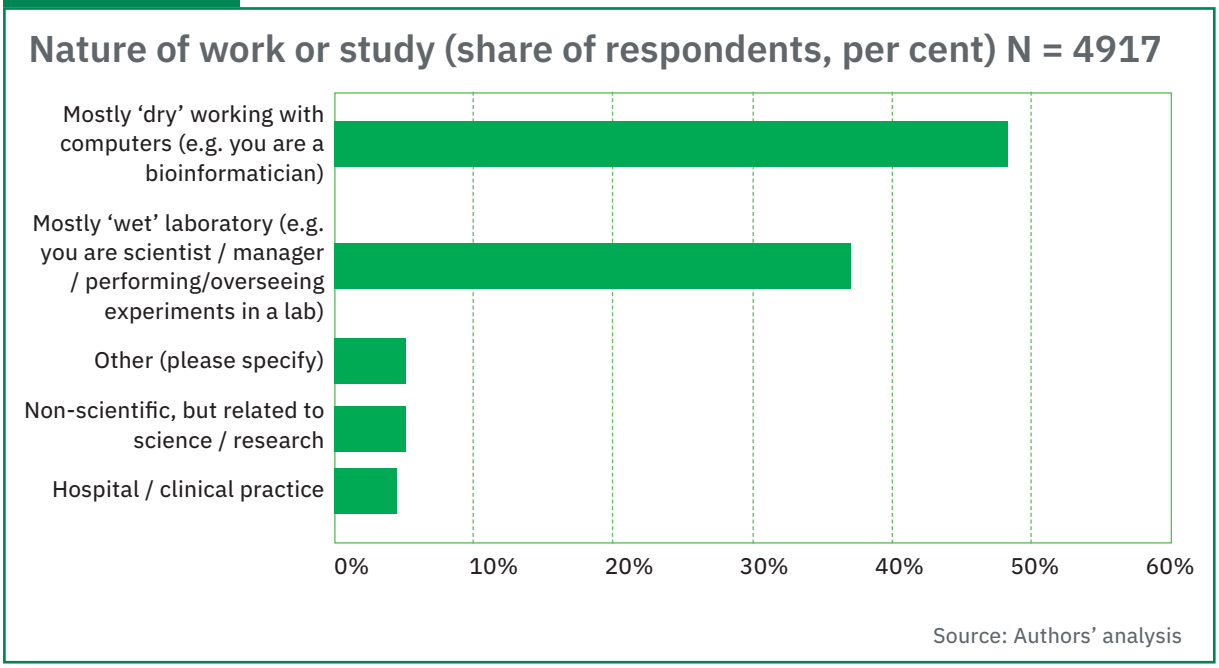
Main role in non-academic sector (share of respondents, per cent)
N = 1220



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 'dry', working with computers (e.g., a bioinformatician), with 37 per cent saying mostly 'wet' laboratory (e.g., a scientist performing experiments in a lab or a manager overseeing 'wet' research). A further 9 per cent reported being in non-scientific (but related to science/research) and hospital and clinical practice environments.

FIGURE A1.5



Q6. What would be the impact on your work or study if you could not access EMBL-EBI services and resources?
Sixty-nine per cent of all respondents 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 21 per cent saying it would have a moderate impact. These numbers reflect an increase in the importance of EMBL-EBI data resources to users compared with our 2015 survey, when the respective figures were 55 per cent and 29 per cent.

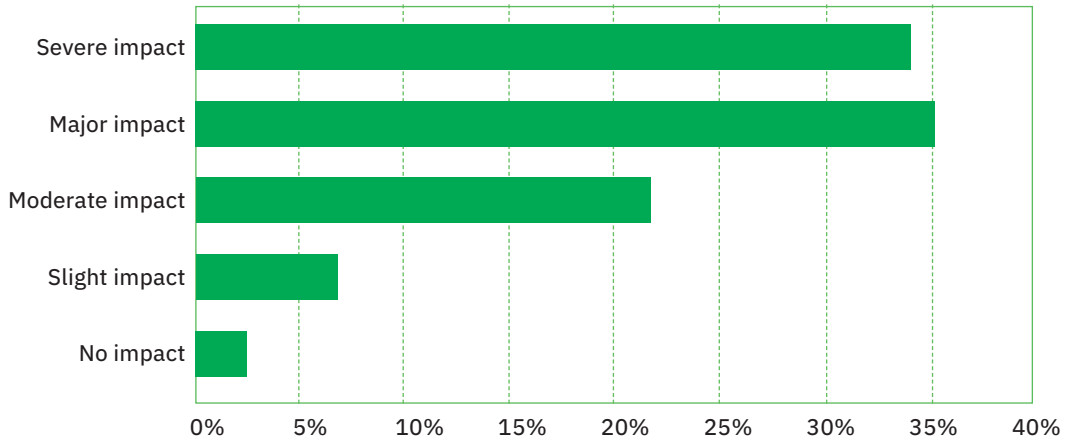
> *"Losing access to EMBL-EBI services and resources would constitute a major setback for infection control and research."*
> [2021 User Survey respondent, USA]

> *"We use ChEMBL all the time to access public domain chemical structures & published data. Having to recreate that would not be possible for a single company."*
> [2021 User Survey respondent, UK]

The weighted average score was 2.92 out of 5, reflecting the balance of responses towards the importance and impact of EMBL-EBI data resources. Again, reflecting an increasing importance since our 2015 survey.

FIGURE A1.6

The impact of not having access to EMBL-EBI (share of respondents, per cent) N = 4883



Source: Authors' analysis

"We are a genomics start-up company and without access to open data sources like EMBL-EBI services, the cost to build a market presence in a highly competitive environment would not be possible."
[2021 User Survey respondent, USA]

"EBI resources are the lifeblood of my research group. Without these services, a full two-thirds of my group's research efforts would suffer dramatically."
[2021 User Survey respondent, Belgium]

Focus and frequency of access

Q7 through Q16. Approximately how frequently did you access/download resources from (...) in the last 12 months?

The questionnaire offered the frequency categories: > 5 times a day, 2-5 times a day, daily, weekly, monthly, quarterly, 1-2 times a year, and not used. These were converted to approximate annual use frequencies as follows: > 5 times a day = 1320, 2-5 times a day = 770, daily = 220, weekly = 45, fortnight = 22, monthly = 12, quarterly = 4, 1-2 times a year = 1.5, and not used = 0 (subsequently deleted).

TABLE A1.2

Estimated frequency of use of EMBL-EBI data resources (per annum)

Europe PubMed Central (Europe PMC)	213	591 977
Ensembl	195	621 503
UniProtKB	195	582 158
Ensembl Genomes	154	418 812
PDBe (Protein Databank in Europe)	111	220 983
BioStudies	99	144 330
EMDB (Electron Microscopy Data Bank)	98	77 402
ChEMBL	96	96 642
OLS (Ontology Lookup Service)	93	119 839
Identifiers.org	93	95 696
HGNC (HUGO Gene Nomenclature Committee)	91	152 080
PDBe-KB	87	109 072
QuickGO/GOA (Gene Ontology)	87	179 957
EFO (Experimental Factor Ontology)	84	83 729
Pfam	83	173 023
BioSamples	83	119 695
BioImage Archive	75	22 383
GWAS Catalogue	74	113 187
ENA (European Nucleotide Archive)	73	123 152
Open Targets	73	58 569
WormBase	72	28 431
InterPro	70	124 824
IGSR/1000 Genomes	69	85 858
Mouse resources: EMMA/IMPC	68	31 188
BioModels	67	36 599
MGnify	66	34 311
EMPIAR	63	36 960
PRIDE (PRoteomics IDEntifications)	63	62 789
ChEBI	62	47 889
SureChEMBL	60	26 523
UniChem	59	31 815
Expression Atlas	59	109 529
EGA (European Genome-phenome Archive)	59	69 819

VectorBase	59	21 293
RNACentral	58	50 110
Enzyme Portal	58	52 254
IntAct	57	40 770
Reactome	57	68 360
EVA (European Variation Archive)	56	55 408
OmicsDI	56	23 283
Complex Portal	52	22 185
IntEnz	51	29 937
Rfam	49	47 461
MetaboLights	41	18 651
Frequency ALL	*1 158*	*5 203 997*

Source: Authors' analysis

> *"I use several databases on a daily basis almost several times every day. Without access to the pdb or EMDB and primary sequence data my work would be almost impossible."*
> [2021 User Survey respondent, Japan]

The respondents' mean frequency of use of EMBL-EBI data resources was 1 158 per year, with a total of 5.2 million accesses/downloads reported. There was considerable variation between the services (Table 1). The most heavily used services included Ensembl, European PubMed Central, UniProtKB, Ensembl Genomes, and PDBe.

Data resources accessed, access time and mode

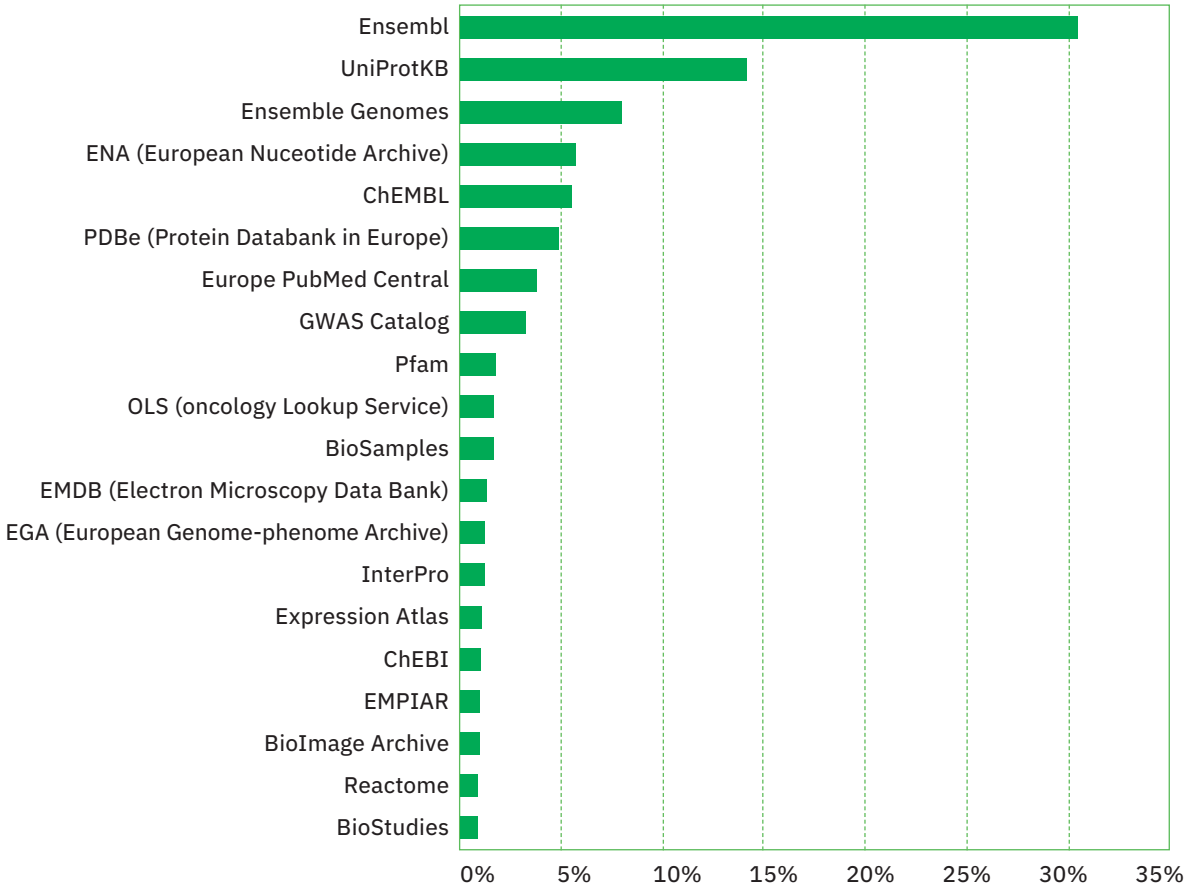
In order to randomise subsequent answers, we asked a critical incident question about the respondents' last use.

Q17. Which EMBL-EBI hosted data resource did you last use?

More than 30 per cent of respondents reported most recently using Ensembl, 14 per cent most recently using UniProtKB, 8 per cent Ensembl Genomes, 5.7 per cent ENA (European Nucleotide Archive), 5.5 per cent ChEMBL, 4.9 per cent PDBe (Protein Databank in Europe), and just less than 4 per cent Europe PMC (Figure A1.7). Of course, last use also reflects frequency of use rather than importance per se.

FIGURE A1.7

The Top 20 services most recently used (share of respondents, per cent) N = 4195



Source: Authors' analysis

"I use EMBL-EBI every week if not every day to track down which protein or which gene I might be looking at - for antibiotic resistance, phenotype, virulence, etc. These databases are critical and central to my research."
[2021 User Survey respondent, USA]

"The genome and proteome databases and tools made available by EML-EBI are absolutely essential to our work, both fundamental research and teaching."
[2021 User Survey respondent, UK]

"Uniprot is always the first stop whenever a new protein comes up! Always!"
[2021 User Survey respondent, USA]

Q18. How long did it take you to find and obtain the resource you last used from EMBL-EBI?

Respondents reported a mean access time of 60 minutes, although with wide variation (N = 3 982). The median reported access time was 5 minutes. Many found access easy, with

523 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, 64 respondents reported access times of 12 hours or more. Such cases were regularly accompanied by comments about download speeds using FTP "*The FTP is incredibly useful, but its structure is a tad byzantine*", difficulties setting up a programmatic access "*My task required API access so I had to learn the API structure*", and so on.

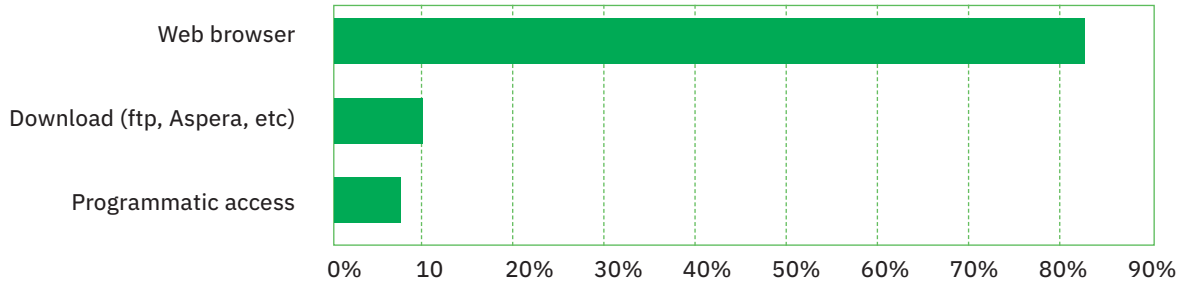
Q19. What mode of access did you use to obtain the last data/tool you used from EMBL-EBI?

The vast majority (83 per cent) of respondents use a web browser to access EMBL-EBI data resources, with almost 10 per cent downloading using FTP, and a further 7.5 per cent using programmatic access (Figure A1.8).

These shares inform the previous question about access time, with mean reported access times of 121 minutes using FTP, 86 minutes using programmatic access, and 50 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 = 4222



Source: Authors' analysis

Measuring value and impact

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

Q20. If EMBL-EBI did not exist, would you have been able to obtain the same data (and/or tools with equivalent functionality) you used from another source?

The majority of respondents (65 per cent) said they could not have obtained the last resource they used from another source, with 35 per cent saying they could have (N = 4 141).

Q21. If Yes, approximately how much time would it have taken to obtain the same data (and/or tools with equivalent functionality) from another source?

Greater than 12 months responses were coded to 13 months, and the whole converted to hours with months as 20 working days, and days 7.5 hours.

Respondents suggested that it would take them a mean of 64 hours to obtain the data they last used from elsewhere, with a median of 1 hour (N = 1 259).

Q22. If you could not have obtained the last data/tool you used elsewhere, would it have been practical for you to collect/recreate it yourself?
More than 85 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 15 per cent said they could have done so (N = 3 962).

Combining Q20 and Q22, of the total 3 914 answering both questions, 2 252 (58 per cent) could neither have obtained elsewhere nor created/collected the last data they used themselves - up from 45 per cent in our 2015 survey. This gives an indication of the proportion of use of EMBL-EBI data resources that is additional use, which could not otherwise have occurred.

Q23. If Yes, approximately how much time would it have taken to collect/recreate the data/tool you last used?
Greater than 10 years was converted to 11, and the whole converted to hours with years as 220 working days, months 20 days, and days 7.5 hours.

The mean time required to create/collect the data last used for themselves was 1 675 hours (223 days), median 150 hours (20 days) (N = 496).

Qualitative benefits

Q24. 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, based on the share of responses of medium to very high benefit, responses suggested that most had benefited from training and user support, but teaching and collaborations were also reported to be widely beneficial (Table A1.3).

TABLE A1.3

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

	User support	Training	Teaching	Collaborations
Haven't used	42%	40%	47%	44%
No benefit	2%	2%	3%	3%
Low benefit	6%	7%	7%	6%
Medium benefit	15%	18%	15%	15%
High benefit	23%	22%	18%	19%
Very high benefit	12%	11%	10%	13%

Source: Authors' analysis

Among users, training received a weighted score of 2.64 from 5 (N = 2 256), with user support scoring 2.63 (N = 2 253), teaching 2.57 (N = 1 973), and collaborations 2.45 (N = 2 083).

Impacts of the COVID-19 pandemic

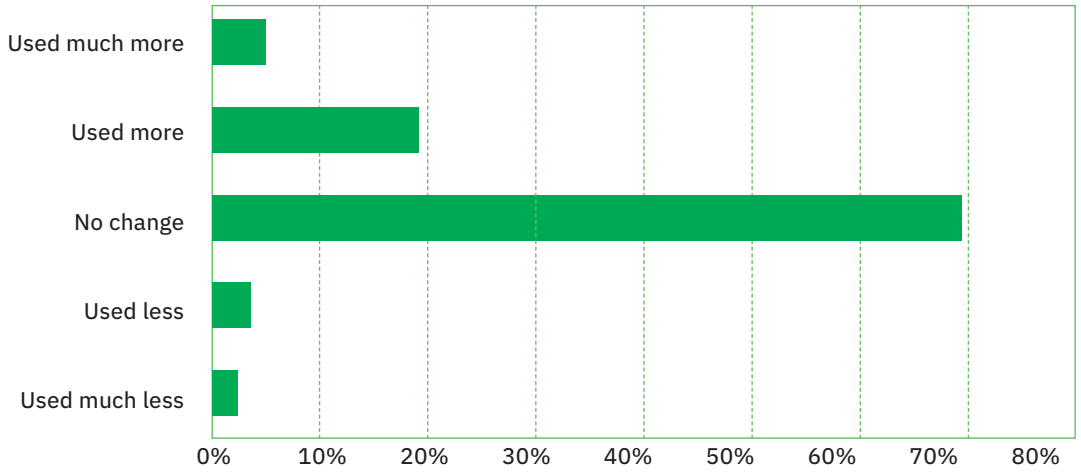
During 2020 the COVID-19 pandemic had a major impact and we sought to explore how it had affected users of EMBL-EBI data resources and whether they had used the European COVID-19 Data Portal.

Q25. COVID-19 had a major impact in 2020. How did it change your use of EMBL-EBI data resources?

The majority of respondents (69 per cent) reported that it had not changed their use of EMBL-EBI data resources, but almost 25 per cent reported using the resources more (19 per cent) or much more (5 per cent).

FIGURE A1.9

The impact of COVID-19 on data use (share of respondents, per cent) N = 3985



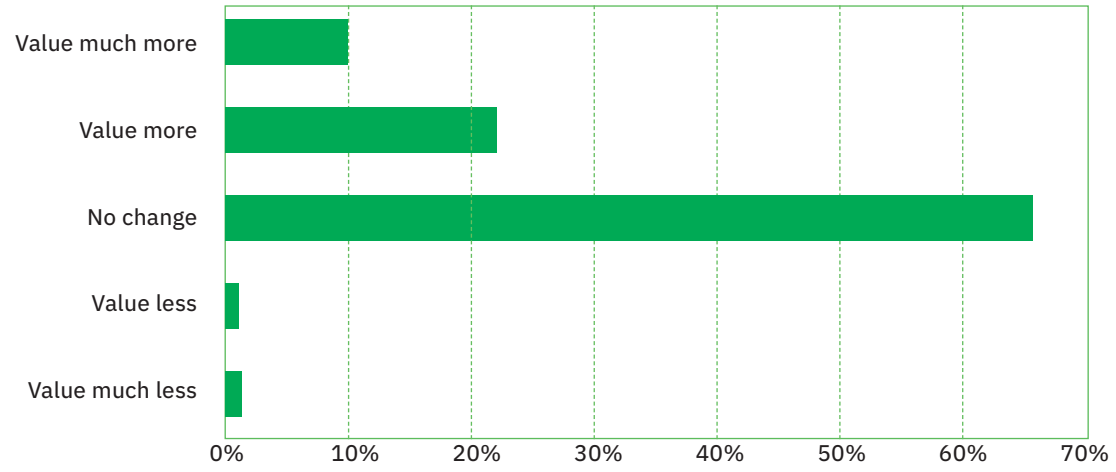
Source: Authors' analysis

Q26. How has the COVID pandemic changed your views on how much you value EMBL-EBI data resources?

A similar pattern emerged when asked how the COVID-19 pandemic had changed their views on the value of EMBL-EBI data resources, with 65 per cent reporting no change and 32 per cent reporting that they valued the resource more (22 per cent) or much more (10 per cent).

FIGURE A1.10

The impact of COVID-19 on view on the value of the data resources (share of respondents, per cent) N = 3976



Source: Authors' analysis

Q27. Have you used the COVID-19 Data Portal?

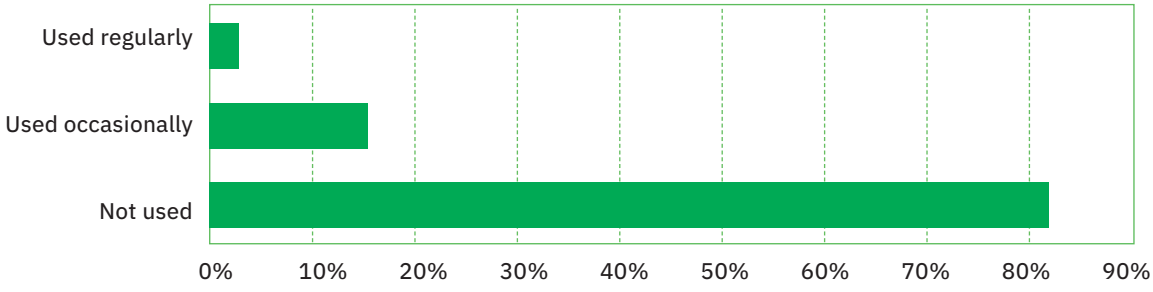
In response the COVID-19 pandemic and the needs of researchers and others working on solutions, EMBL-EBI quickly established a COVID-19 Data Portal (the subject of a separate case study).

Not surprisingly among respondents who are predominantly regular research users of EMBL-EBI data resources, and familiar with navigating them, the majority (82 per cent) had not used the COVID Data Portal. However, 15 per cent of respondents had used the COVID Data Portal occasionally, and a further 3 per cent had used it regularly.

> *"I used the COVID-19 portal extensively in mid 2020 while supporting COVID-19 research and clinical testing."* [2021 User Survey respondent, USA]

FIGURE A1.11

Use of the EMBL-EBI COVID-19 Data Portal (share of respondents, per cent) N = 4005



Source: Authors' analysis

Impacts on research

Q28. Is research a part of your role?
Research was a part of the role for 89 per cent of respondents (N = 4 138).

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 may have been mainly involved in teaching and other activities.

Before data cleaning there were 106 responses of 65 hours per week or more, 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 more than 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 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 100 responses (3.2 per cent) reporting greater than 65 hours to 65 hours. After data cleaning the reported mean was 31 hours per week, and the median 30 hours per week (N = 3 136).

Q30. Can you estimate the approximate share of your total research working time spent with data during the last twelve months (e.g. creating, collecting and analysing data as distinct from other tasks e.g. writing articles, management, etc)?
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 (206 cases).

On that basis, they reported spending a mean of 55 per cent of their research working time with data (N = 2 967), and 23 per cent of their time with data obtained from EMBL-EBI (N = 2 748). The medians were 50 per cent and 20 per cent, respectively.

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 317 responses to that noted above, respondents suggested that others in their research field might typically spend a mean of 51 per cent of their research working time with data (N

= 2 321), and 25 per cent with data obtained from EMBL-EBI (N = 2 131). The medians were again 50 per cent and 20 per cent, respectively.

Thus, as we have found in similar surveys in different fields of research, respondents see themselves as typical of those in their field, reporting very similar times with data for themselves and for others. This lends some confidence when extrapolating survey responses to all users.

Q32. To what extent, if any, has your use of EMBL-EBI hosted data resources changed your research efficiency (i.e. time saved compared to if no EMBL-EBI existed)?
To explore the impact of EMBL-EBI hosted data resources on their 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.

In similar surveys that we have run in different disciplinary fields, the reported efficiency impacts have been considerable. However, a limitation to this question is now emerging. Some 448 respondents (15 per cent) reported an impact of >90 per cent. Comments suggested that for many there was simply no alternative way to do their research. Hence, their response implied that it had a 100 per cent impact.

> *"No data no science for me, you store, curate and offer what feeds my work and without that I would need to change completely my research."*
> [2021 User Survey respondent, Chile]

> *"My work will collapse without EMBL-EBI resources."*
> [2021 User Survey respondent, Israel]

We converted >90 per cent responses to 91 per cent to be conservative. Among those reporting a zero or positive impact, the mean was 53 per cent (N = 2 965), with a median of 50 per cent.

> *"EMBL-EBI has transformed what I and my team and students are able to conceptualise and achieve. Wouldn't be able to do 1/10 of what I can or have done over last 20 years."*
> [2021 User Survey respondent, UK]

> *"My research efficiency on looking up and using genome references is 100 fold improved - it's like comparing cell phones with tin cans and bits of string."*
> [2021 User Survey respondent, USA]

Just 27 respondents reported a negative impact, which could not be included in analysis as it is not numerical. A few comments by these respondents (5 cases) suggested that they had some specific problems in what they were trying to do.

> *"All the isoforms of my gene changed with the last update and there is some incomplete data."*
> [2021 User Survey respondent, Australia]

"Uploading sequence data to ENA has been a complete nightmare and has set me back weeks of work."
[2021 User Survey respondent, Belgium]

Curation and secondary use

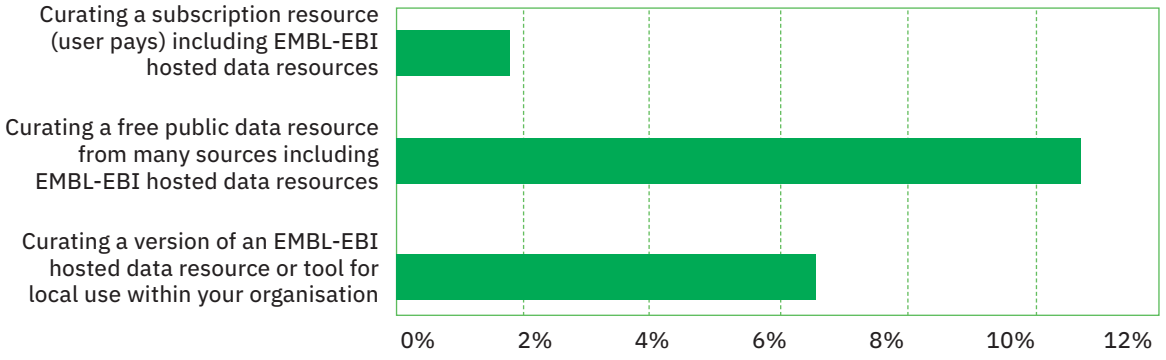
We know that there are many who curate a direct or value-added copy of EMBL-EBI hosted data resources and make them available to 'secondary' users. These secondary users will not be measured in EMBL-EBI statistics as they do not visit the data resources directly. Hence, we are interested in exploring how extensive curation and secondary use might be.

Q33. Are any of the following part of your role?

Some 20 per cent of respondents reported curating (N = 3 816). Of those curating 42 per cent reported curating a version of an EMBL-EBI hosted data resource or tool for local use within their organization, 68 per cent curating a free public data resource from many sources including EMBL-EBI hosted data resources, and 12 per cent curating a subscription resource (user pays) from many sources including EMBL-EBI hosted data resources.

FIGURE A1.12

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



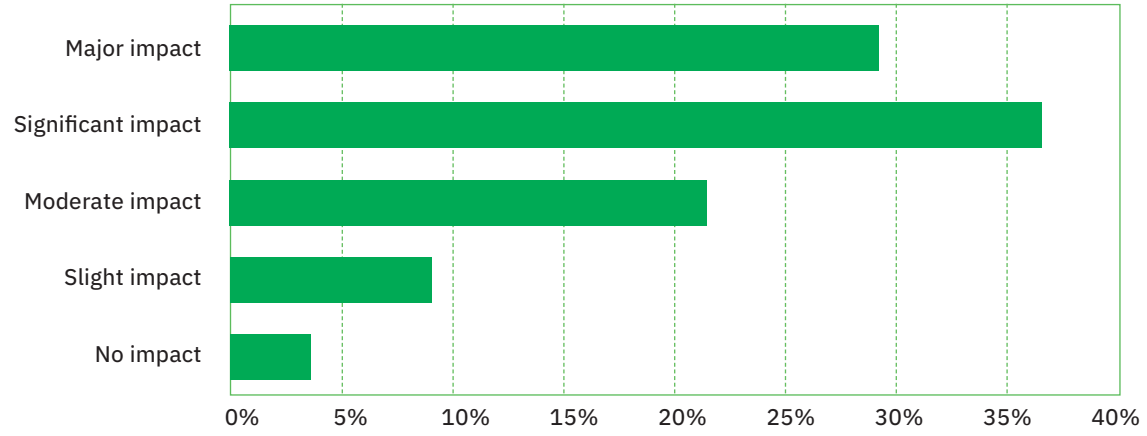
Source: Authors' analysis

Q35. How would you describe the impact of the EMBL-EBI hosted data resources' contribution to your version/resource?

Less than 14 per cent of respondents said that EMBL-EBI hosted data resources had a slight or no impact, with 30 per cent suggesting a major impact and a further 37 per cent suggesting a significant impact (N = 755).

FIGURE A1.13

Impact of EMBL-EBI data resources on their curated version (share of respondents, per cent) (N = 755)



Source: Authors' analysis

Q37. Approximately how many people used your version/resource during the last 12 months?
One-third of all respondents who reported a curating role (246) 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 and to answer, and there are obvious measurement and confidentiality issue involved. To try to understand the responses as best as possible, we asked a further question of clarification.

Q38. How has the number of people in the previous answer been estimated? (e.g. User registration, Unique IPs, Cookies etc., over a month or year)
Based upon responses to this question we tried to interpret the numbers reported and come up with an estimate of the number of secondary users. Due to the range of non- and incomplete answers this proved to be extremely difficult.

However, indicatively, after estimations and analysis the mean count of secondary users was 277, and the median 10 (N = 244). Our best guess would be that there may be close to 70 000 secondary users reported among respondents.

If the share reporting curating and their median users were typical of all users, then there may be as many as a million secondary indirect users worldwide who fall outside EMBL-EBI primary direct user counts.

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). It is essential that respondents reply as individuals. During analysis, 162 responses were identified as protest and/or collective answers and were excluded from the analysis.

Q39. Imagine you have the option to either carry on using EMBL-EBI or to give up your existing user access in exchange for a payment. What is the minimum amount that you would be willing to accept as an annual payment in return for giving up all of your individual access to EMBL-EBI hosted data 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. This gives a sense of what something is worth to respondents, independent of what they may be able to afford to pay for it.

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 some 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 project salary or research grants. 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 first estimate is that of being able to retire comfortably now (under 40 senior researcher and consultant), as I would not be able to do my job without these resources, as Proteomics relies on identification and quantification of thousands of proteins, therefore accurate databases, a user friendly interface with comparison and identifier mapping is central to all I do, this it is a very safe assumption that this field would be irreparably crippled without Uniprot."* [2021 User Survey respondent, UK]

The principal (and, perhaps, principle) reasons given for saying they would not be willing to accept anything in return for giving up their access included 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 (really their answer was infinity, not zero). There were 34 responses explicitly referencing Open Access/Open Data principles. On balance it seemed best to exclude the zero answers for this reason.

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 £68 000 per annum. The median value was £424 (N = 1 754), and the difference between these mean and median values illustrates the wide range of responses.

Q40. EMBLEBI will never charge for services. For this question, however, please imagine that access was 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 the data resources hosted by EMBLEBI (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 a single user annual subscription for access to EMBL-EBI data resources, 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. Again, it seemed best to exclude the zero answers.

The mean of the value respondents reported being willing to pay for access for a year was £2 757, with a median of £200 (N = 2 009).

A number of respondents commented on the rationale for their answers, with comparisons to the annual per user cost of related software licences, other database or journal subscriptions being common.

> *"Well, we pay quite a bit per researcher for Qiagen Ingenuity Pathway Analysis and Partek Flow. EMBL services are ten times more comprehensive (they are the basis for the information hosted on IPA anyway without attribution). Without these services a whole ecosystem of science collapses."*
> [2021 User Survey respondent, USA]

European Bioinformatics Institute (EMBL-EBI)
Wellcome Genome Campus
Hinxton, Cambridge, CB10 1SD
United Kingdom

www.ebi.ac.uk
+44 (0)1223 494 444
comms@ebi.ac.uk

@emblebi
/EMBLEBI
/EBImedia
/company/ebi/

EMBL-EBI is a part of the European Molecular Biology Laboratory.

EMBL member states and associate member states: Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, Switzerland, United Kingdom, Argentina, Australia

Prospect member states: Estonia, Latvia