



Measuring the impact of data institutions

Contents

Contents	1
About	2
Introduction	3
Summary of final quantified benefits	5
Marine Environmental Data and Information Network	6
About	6
Impact	7
Costs	9
An array of use cases	9
Farmbench	10
About	10
Impact	11
Costs	12
Sustaining the tool	12
360Giving	13
About	13
Impact	14
Costs	15
Expanding standards	15
Clinical Practice Research Datalink	16
About	16
Impact	17
Costs	18
Supporting public health research in the UK	19
OpenActive	20
About	20
Impact	21
Costs	23
Future prospects	23
Bottom-up data institutions	24
Synthesis and reflections	26

About

This report has been researched and produced by the Open Data Institute, and published in March 2022. Its lead authors are Aditya Singh and Jack Hardinges. To share feedback by email or to get in touch, contact the Data Institutions project lead, Jack Hardinges, at datainstitutions@theodi.org.

Introduction

At the Open Data Institute (ODI), we're interested in data institutions – organisations that steward data on behalf of others, often towards public, educational or charitable aims.¹ We think data institutions have a vital role to play in ensuring that data is used to create new technologies, products and services, as well as limiting the harm that misuses of data can cause people and communities.

We've done a lot of work on the different ways data institutions can steward data responsibly, such as by facilitating safe access to sensitive data² and empowering people to play a more active role in deciding how data is used³. This study is part of our ongoing effort to increase awareness and understanding of data institutions among policymakers, funders and others. The objective is to support them in making interventions that create an enabling environment for data institutions to thrive.

This study aims to improve understanding of data institutions by quantifying the impact they may have on the ecosystems within which they operate. There is already a body of evidence for the impact of open data and open standards; some of which has also been conducted by the ODI. While these provide valuable insight into the impact and benefits of data flows, and how they may be measured, there is limited evidence for the impact of data institutions and the stewardship they perform.

Data institutions vary in size, domain and maturity. We've previously identified the following roles they may perform:⁴

- Facilitating safe access – protecting sensitive data and facilitating safe access under restricted conditions.
- Empowering people – empowering people to take a more active role in stewarding data about themselves and their communities.
- Independent gatekeeping – acting as a gatekeeper for data held by other organisations.
- Publishing open data – creating open datasets that anyone can access, use and share to further a particular mission or cause.
- Developing infrastructure – developing and maintaining identifiers, standards and other infrastructure for a sector, or field, such as open standards.
- Generating insights – combining or linking data from multiple sources and generating insights and other services back to those that have contributed data.

¹ Open Data Institute (2021), [What are data institutions and why are they important?](#)

² Open Data Institute (2021), [How do data institutions facilitate safe access to sensitive data?](#)

³ Open Data Institute (2021), [Could 'bottom-up data trusts' help to tackle the climate crisis?](#)

⁴ Open Data Institute (2021), [What are data institutions and why are they important?](#)

Considering this diversity, this study focuses on illustrative case studies of five UK-based data institutions, corresponding with the different stewardship roles:⁵

- Marine Environmental Data and Information Network (MEDIN) (Facilitating safe access & Developing infrastructure) – a web portal which aggregates UK marine data, providing a centralised access point for users.
- Farmbench (Independent gatekeeping & Generating insights) – an online benchmarking tool which allows farmers to compare their performance to other farmers across the UK.
- 360Giving (Publishing open data & Developing infrastructure) – a charity which provides an open search platform for charitable grants data.
- Clinical Practice Research Datalink (CPRD) (Facilitating safe access & Independent gatekeeping) – a platform which collects anonymised patient data from UK GPs, and links this data to a range of other sources to provide a UK representative health dataset.
- OpenActive (Developing infrastructure) – a sector-wide initiative providing a set of open data standards, and support with using and adopting these standards, to facilitate the sharing of sports activity data.

These case studies demonstrate the ways in which different types of data institutions create value within their ecosystems. Our aim is for the findings to serve as evidence for policymakers when designing interventions – such as in the case of UK Government’s work on data intermediaries and the EU’s interest in data altruism organisations – and for data institutions themselves to be able to point to this evidence base in making the case for their work. We also hope that this study indicates how the value generated by these institutions, and approaches to documenting it, may be extrapolated to other data institutions.

We commissioned London Economics to support us with this research. Their approach for each case study involved three steps:⁶

- **Data ecosystem mapping** – Using ODI’s data ecosystem mapping methodology and creating a visual representation of the stakeholders in the ecosystem (for example, data institution, data provider, end user) and the value chains connecting these stakeholders (for example, data, financial benefit, reputation gain).
- **Stakeholder consultations** – Conducting semi-structured interviews with multiple stakeholders in each ecosystem, to substantiate the ecosystem map’s existing value chains and expand with additional chains.
- **Quantification** – Developing a framework to quantify the benefits from each data institution. This included producing estimates that could be obtained given the project’s scale, timeframe and data availability, while also scoping how further estimates could be obtained in future work.

⁵ The initial aim was to study six data institutions corresponding with each role. However, as we describe on page [xx], there remain unique challenges with identifying and measuring the impact of bottom-up data institutions.

⁶ See [Annex](#) for a detailed description of the methodology.

In the following section, there is a table summarising the quantified benefits from the data institutions we studied. Subsequent sections set out our findings from the individual case studies, including data ecosystem maps, a description of each data institution and their data stewardship role, and estimates of the value they generate for their stakeholders. The final section summarises and reflects on the findings across the case studies.

Summary of final quantified benefits

MEDIN	Farmbench	360Giving	CPRD	OpenActive
Time savings from making data more findable and accessible £1.84 million per year	Efficiency gains from reduced labour costs £212,577 per year	Duplication of funding avoided £10.4 billion	Number of academic papers/publications supported by CPRD 2,900	Additional bookings enabled 529,576 per month
Cost savings from making data reusable £1.32 million per year	Efficiency gains from reduced machinery costs £283,437 per year	Time savings for funders and grant recipients £1.5 million	Compensation for GPs recruiting patients into clinical trials enabled by CPRD £50,000 per year	Additional revenue enabled in the ecosystem £15.6 million per month
Cost savings from improved data management and storage £2.64 million per year	Efficiency gains from reduced expenditure on inputs (for example, fertilisers, pesticides) £2.6 million per year		Cost savings for pharmaceutical companies on phase 2 and 3 clinical trials £5 million	Time savings for individuals to find activities £4.7 million per month
Additional datasets shared 497	Number of academic papers/publications supported by Farmbench 4 to 5 per year		Cost savings for pharmaceutical companies on phase 4 (“pragmatic”) clinical trials £800 million	Time savings for activity providers on promotional activities £267,200 per month
				Societal benefits of increased exercise: Avoided death 79 per month Productivity loss £5.2 million per month Healthcare costs £4.1 million per month

Source: London Economics

Marine Environmental Data and Information Network

Improving access to UK marine data

About

The Marine Environmental Data and Information Network (MEDIN) was established in 2008 to make the collation, analysis and access to marine data easier and cheaper.

Marine data is crucial for many scientific disciplines, as well as for a variety of commercial activities such as port and fisheries management, environmental planning, marine conservation and weather forecasting. Marine data is also being increasingly used in public policy and planning. Marine data collection, however, is expensive due to the complex nature of the marine environment and the fact that marine data is always unique relative to time and geographical position.⁷ Another challenge is that data is often held in silos, which makes it difficult for organisations and researchers to find and access the various types of datasets they may need.⁸

MEDIN is an **open partnership representing government departments, research institutions and private companies**.⁹ Its purpose is to promote the sharing and reuse of marine data, as well as to improve access to marine data through:

- Data Archive Centres (DACs)
- a web portal
- a metadata standard to provide information about datasets
- workshops to improve the uptake, knowledge and use of MEDIN data guidelines and metadata standards.¹⁰

The MEDIN DACs have committed to making datasets openly available wherever possible and to removing barriers to data access such as having to register or pay for the services provided.¹¹

⁷ Eftic and ABPmer (2019), [MEDIN Cost Benefit Analysis 2019](#)

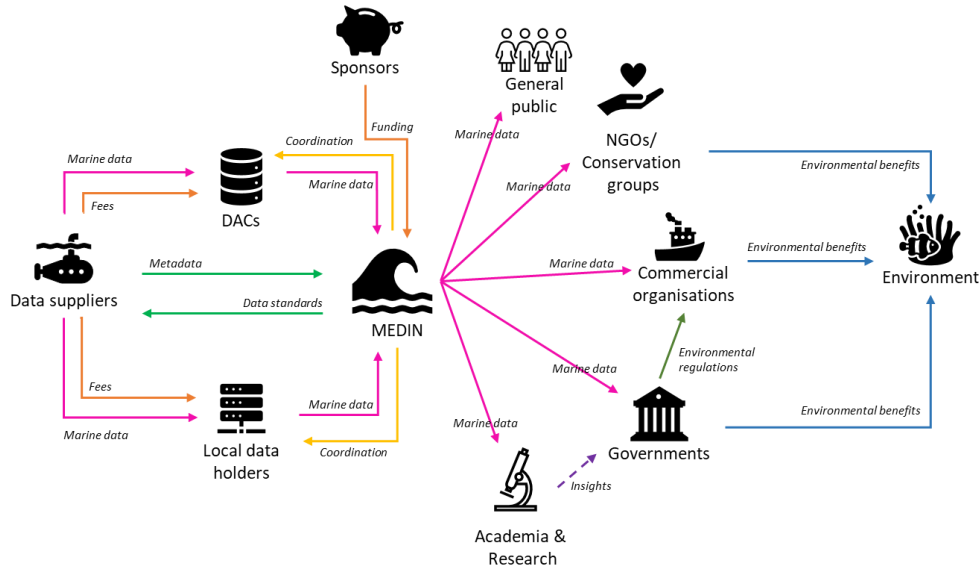
⁸ MEDIN (2010), [Managing UK marine data. Marine Scientist No.32](#)

⁹ Eftic and ABPmer (2019), [MEDIN Cost Benefit Analysis 2019](#)

¹⁰ MEDIN (n.d.), [About MEDIN](#)

¹¹ Organisation for Economic Co-operation and Development (2021), [Value chains in public marine data: A UK case study. OECD Science, Technology and Industry Working Papers 2021/11](#)

Figure 1: MEDIN data ecosystem map



Source: London Economics analysis

Impact

MEDIN conducted a cost-benefit analysis in 2019 with *eftec*, an environmental economics consultancy, and *ABPmer*, a group of marine consultants. Our research builds on their work.

Data end users save a considerable amount of time searching for marine datasets because of the availability and accessibility of data through the MEDIN data portal. These time savings may relate to academic and commercial research, marine licence applications and environmental impact assessments. It is estimated that the total value of the time saved each year due to MEDIN is **£1.84 million**.¹²

By being able to find and access secondary data through the MEDIN portal, organisations may be spared having to collect primary data themselves and avoid duplicating the work that other organisations have already done.

MEDIN provides guidelines on how data must be structured, as well as metadata which describes relevant information about each dataset. Organisations are therefore able to reuse datasets that have been uploaded by others. This saves approximately **£1.32 million per year** on primary research efforts.¹³

¹² *eftec* and *ABPmer* (2019), [MEDIN Cost Benefit Analysis 2019](#).

¹³ *Ibid*

This figure is likely to **underestimate the true savings related to research duplication** as it is a time cost measure and therefore does not include the cost of specialised equipment that is generally required to collect marine data.

By following the guidelines on how to format data and sharing it with DACs to store, **organisations save time and resources when managing their data.**

Organisations can be sure that the data they collect will be stored safely, and that it is formatted in a usable way when they need to access it again. This saves organisations approximately **£2.64 million per year** on management and storage costs.¹⁴

There are also reputational benefits for data suppliers when they provide data to MEDIN, as it is seen as the socially responsible thing to do.

Researchers account for the largest proportion of MEDIN data users and make up half of all data downloads.¹⁵ This includes people working in academia, as well as in research councils. Marine data underpins many scientific research studies and is often required to place research within geographic and temporal contexts.¹⁶

Assuming that access to the MEDIN portal enables academics to publish a paper a year, an **additional 700 research papers** are published as a result of the MEDIN portal.

Government departments and agencies rely on MEDIN data to monitor reporting and status assessment of the seas, inform marine planning and assist in human activity (such as fishing or navigation), and in the creation of oceanographic models for forecasting by the Navy and the Met office.¹⁷ Just under 20% of MEDIN data downloads come from the government departments and maritime authorities.¹⁸

Non-governmental organisations (NGOs) and conservation groups are also interested in marine data, both as a general interest and to allow stakeholders to participate in marine planning and management processes.¹⁹ Approximately 8% of MEDIN users are NGOs and conservation groups that access data through the MEDIN portal.

¹⁴ Ibid

¹⁵ Organisation for Economic Co-operation and Development (2021), [Value chains in public marine data: A UK case study](#)

¹⁶ MEDIN (n.d.), [About MEDIN](#)

¹⁷ Ibid

¹⁸ Organisation for Economic Co-operation and Development (2021), [Value chains in public marine data: A UK case study](#)

¹⁹ MEDIN (n.d.), [About MEDIN](#)

Assuming that each of these users represents one project:

- around **250 government projects** are enabled through MEDIN each year
- around **84 conservation projects** are enabled through MEDIN each year.

Costs

Uploading data and metadata to DACs involves spending time becoming familiar with the MEDIN standards and preparing data, as well as ensuring that each dataset and the related metadata are MEDIN-compliant prior to upload. The cost that data suppliers incur uploading data is estimated to be approximately **£180,000 per year**.²⁰ The cost of running MEDIN, including employment of the core team, associated overheads, and external contracts related to maintenance and operation of the MEDIN network, are estimated to be **£720,000 per year**.²¹

An array of use cases

MEDIN demonstrates the wide ecosystem of impact that mature data institutions can have. **Commercial organisations** make up 21% of MEDIN users.²² Marine data is a key input for companies at many stages of projects, including when assessing an optimal site or route, carrying out environmental impact assessments, satisfying licensing requirements, and informing engineering designs on operational conditions.²³ Benefits for commercial parties include the potential to commercialise public marine data into highly complex and customised information products, as well as having access to data which can inform planning and operations.²⁴ These benefits, though significant, would be challenging to quantify in the present study.

There is also demand for marine data from the **general public**, although this is the smallest group of MEDIN users (5%). This demand can be due to both a general interest in the seas and a desire to use the coastline safely for recreational and tourism purposes.²⁵

The **environment** benefits through evidence-based policy actions from the government, which regulate how commercial organisations use the sea and coastline, as well as from evidence-based conservation efforts by NGOs and conservation groups.

²⁰ Efttec and ABPmer (2019), [MEDIN Cost Benefit Analysis 2019](#)

²¹ Ibid

²² Organisation for Economic Co-operation and Development (2021), [Value chains in public marine data: A UK case study](#)

²³ MEDIN (n.d.), [About MEDIN](#)

²⁴ Ibid

²⁵ Ibid

Farmbench

Enabling valuable industry insights by combining performance data across the UK farming industry

About

Farmbench is an online benchmarking tool operated by the **Agriculture and Horticulture Development Board (AHDB)**, which represents farmers, growers and others in the supply chain.

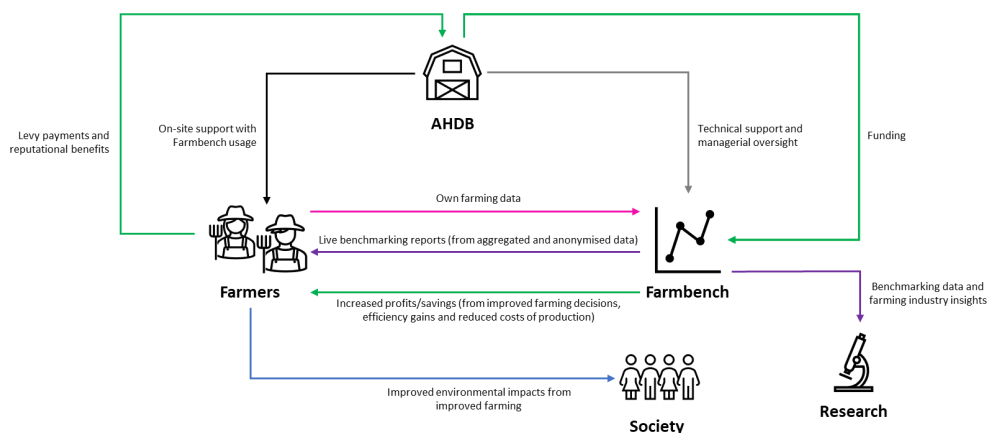
In exchange for inputting their own farming data into Farmbench, farmers can access **live benchmarking reports to compare their farm's performance to similar businesses in the UK**, which can help them **improve their farming decisions**.

Farmbench generates the benchmarking data by aggregating and anonymising the data entered by each individual farmer. The data entered by farmers predominantly includes **financial data** (variable and fixed costs) and **physical data** (for example, livestock) about their farming enterprise. **Farmers** are the bulk of Farmbench users. There are over 2,000 registered users, of which **1,200 farmers actively contribute their data to the tool**.²⁶

The AHDB created Farmbench in around 2017 as a cross-sector tool to replace the sector-specific tools it had previously operated, that is, benchmarking tools that each specialise in one type of farming (for example, dairy, beef and sheep, combined crops). Covering multiple enterprises creates **efficiency gains** as non-specialist farmers with multiple enterprises only have to use one benchmarking tool. This has **increased uptake of benchmarking** among farmers. Before Farmbench, only an average of 2% of users of each sector-specific benchmarking tool operated by the AHDB were non-specialist farmers. Now, about 40% of Farmbench users are non-specialist farmers.

²⁶ There has been some fluctuation in the number of farmers contributing to the tool over the years. For the Autumn 2020 cycle, the number of participating farmers was relatively lower than the previous year as a significant number of farmers were unwilling to use Farmbench because the entire process had to be conducted online during the pandemic.

Figure 2: Farmbench data ecosystem map



Source: London Economics analysis

Impact

Farmbench provides farmers with benchmarking data, industry insights and knowledge-sharing opportunities that can help them make better farming decisions that lead to cost savings or efficiency gains for their farming enterprise. This includes savings on inputs such as fertilisers, seeds and herbicides, as well as savings on machinery and labour costs.

An evaluation study of Farmbench (2017 to 2020) estimated annual savings per individual farmer from on-farm changes due to Farmbench to be about **£6,600**.²⁷ The study assumes that 20% of the 2,677 registered Farmbench users have made on-farm changes due to Farmbench and estimates the total annual efficiency gains across all farmers to be about **£3.5 million**.

Improved farming performance and cost reduction typically lead to more sustainable farming and improved environmental impacts that benefit both farmers and the wider society. These environmental benefits can include:²⁸

- reduced greenhouse gas emissions
- improved soil condition (and carbon storage) by adding matter and reducing pesticide usage
- reduced water usage
- improved water quality by protecting watercourses from sediment and pesticides in run-off
- improved landscape diversity and increased pollination.

²⁷ Teevan, P (2021), Final Evaluation Report: Farmbench Period 2017-2020

²⁸ Department for Environment Food & Rural Affairs (2021), [Sustainable Farming Incentive pilot: environmental outcomes and benefits](#)

In addition to farmers, the AHDB and the Farmbench team itself, **third-party researchers** and farm business consultancies also use Farmbench data. Farmbench receives four to five data requests from external researchers a year. Farmbench facilitates or enables this research by providing access to aggregated industry data. Without Farmbench, it would be possible (but costly) to consolidate cross-sector mixed farm data at a local level, but it would probably not be feasible to obtain the data at a national level. However, there is limited evidence available to quantify the value of the research enabled by Farmbench data at this stage.

Costs

The **main costs within the Farmbench ecosystem are those associated with developing, maintaining and operating Farmbench**, including providing user support to farmers. These costs amount to about **£1.2 million a year**, which mainly includes the wages of the AHDB employees operating Farmbench.

The time and effort that farmers spend entering data into Farmbench, as well as understanding the live benchmarking reports, should also be accounted for. However, these are likely to be small relative to the efficiency gains for farmers from using the benchmarking information.

Sustaining the tool

The sustainability of Farmbench depends on positive feedback and appreciation from the farmers. AHDB supports farmers in using the tool through a team of people dotted around the UK who interact directly with farmers and help them input their data. Farmbench data and benchmarking reports are also used to support **discussion groups and workshops between farmers**, where knowledge and best practice on how to improve farming performance are shared.

In turn, the AHDB benefits from an improved reputation and positive feedback within the industry as farmers make efficiency gains and value using Farmbench. Given that the AHDB is entirely funded through farmers' levy, a good reputation is key to ensuring its continuity and success.

Farmbench is currently working with other countries (for example, Australia and New Zealand) on similar initiatives, to further expand the geographical coverage of their benchmarking data.

360Giving

Creating open grants data that anyone can access, use and share.

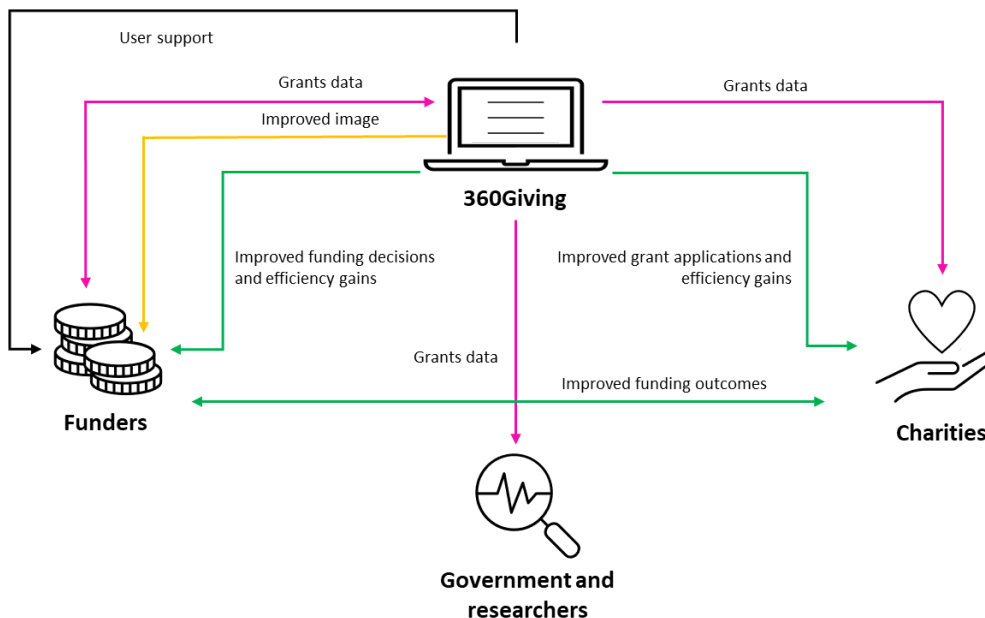
About

360Giving is a charity that aims to improve grant making in the UK. 360Giving operates an **open database** called GrantNav, which consists of UK grants data such as the amount, award date, recipient and corresponding grant maker's information.

Most of the data is provided by the grant makers. 360Giving also develops **open data standards** for grant makers to publish their data on GrantNav, and it enriches the submitted grants data with information from other sources, such as company data from Companies House.

The 360Giving database currently covers **217 grant makers in the UK** and over 280,000 grants worth **£110 billion**. The data can be accessed and downloaded by anyone from the 360Giving website. 360Giving also helps both grant makers and recipients understand how to make full use of the open database and related tools.

Figure 3: 360Giving data ecosystem map



Source: London Economics analysis

Impact

Grant makers can use the 360Giving data to improve their grant-making decisions, including diversifying their funding to new and/or underfunded charities, identifying new potential grant recipients and coordinating with other grant makers to distribute grant money. The data was particularly useful during the **Covid-19 pandemic** and allowed grant makers to **distribute emergency grants as evenly as possible**.

Grant makers can make efficiency gains by identifying and leveraging previous due diligence and compliance checks conducted by other funders on potential grant recipients, as it **reduces duplication of efforts and frees up resources**. The information on 360Giving also provides grant makers with an idea of where they sit within the ecosystem of grant funding.

Assuming **10 hours of time** are saved by grant makers in funding decision making; and a **minimum hourly wage rate equal to the UK national living wage of £8.91**, there are **cost savings of £89.10 per funder and about £19,000 for all 217 funders** participating in the 360Giving database.

Improved funding decisions by grant makers can potentially help them avoid redundant funding and instead allocate funds where they are most needed and valuable, improving overall UK grant-making outcomes in the long term.

Assuming that a share of grants worth 10% of the total grant value covered by 360Giving avoids being funded in duplication as a result of grant makers using the database to make more informed funding decisions, the estimated value of this reduction in **duplication of funding is over £10 billion**. These funds are then likely to be redirected to organisations that haven't previously received funding. This figure also indicates how the measure of value from stewardship may be closely linked to the value and nature of the data stewarded (in this instance, high-value philanthropic grants).

Grant recipients and charities use the 360Giving data to **improve their grant-funding decisions** by looking at information on what makes a successful applicant, beyond the minimum requirements that are typically published by the funder on the grant application. This can help charities free up resources and make efficiency gains by strategically prioritising grant applications.

Assuming 10% of grant recipients relied on the 360Giving database, with 10 hours of time saved in grant applications; and assuming a minimum hourly wage rate equal to the UK national living wage of £8.91, there is an **estimated cost saving of £89.10 per grant recipient and close to £1.5 million for the estimated 16,832 grant recipients** that used the 360Giving database.

Government and researchers use the 360Giving data for their own research and planning purposes. For example, some government departments that are grant makers do not have access to a structured grant database internally, and use the 360Giving database instead. 360Giving benefits the government and researchers by allowing them to minimise the costs associated with accessing UK grants data as part of their work.

Costs

The **main costs are those related to developing, maintaining and operating 360Giving**. These include the wages of the 360Giving team, which amounts to about **£250,000 to £300,000 a year**. Additionally, the technical maintenance and development of the tools and database, which is outsourced, costs about **£150,000 a year**.

Expanding standards

While there have been many studies on the impact of open data, 360Giving demonstrates how even relatively specialised datasets can engage a wide network of stakeholders, and generate impact. Since the inception of the 360Giving grants database, sharing funding data has **increasingly become a standard** among grant makers. Only very few of the largest funders in the UK still do not publish their funding data. In part because of the Covid-19 pandemic, **some grant makers have created new capacity within their organisations for uploading data** as fast and as regularly as possible.

360Giving is currently working on expanding the set of fields within the database to include additional details on the grant maker and recipient, and a metadata repository. It is also working on a new equity standard, given strong appetite from grant makers to include information on the ethnic diversity of grant recipients' leadership teams, as improving diversity is a priority for grant makers.

Clinical Practice Research Datalink

Providing access to anonymised UK population health data to approved organisations and studies

About

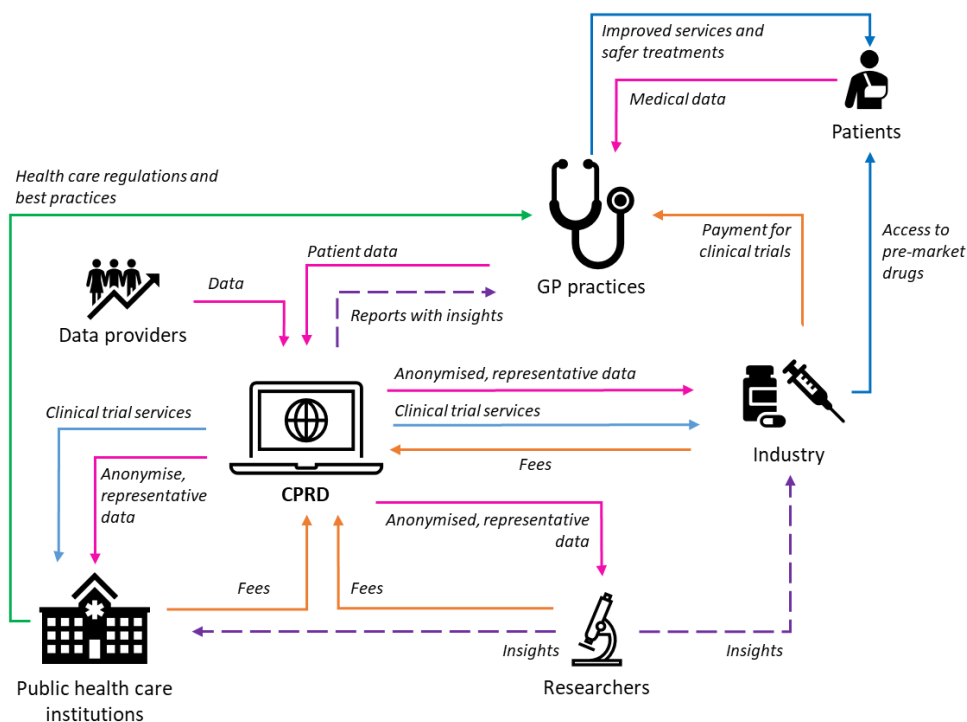
Clinical Practice Research Datalink (CPRD) is a not-for-profit research service that supports public health and clinical studies. It collects **anonymised patient data from a network of over 2,000 GP practices across the UK**, which can be linked to a range of other health-related data to provide a longitudinal, representative UK population health dataset. CPRD data is used by academia, industry and government organisations worldwide to investigate drug safety, the effectiveness of health policies, healthcare delivery and disease risk factors.

Anonymised patient data is shared with the CPRD by GPs across the country, with one in five primary care practices uploading their patient data. CPRD data currently encompasses over 60 million patients, including 16 million currently registered patients of which a medical history from the last 20 years exists for a quarter of them. This data is then linked to secondary care and other health datasets, as well as to area-based datasets by the CPRD. Before patient data reaches the CPRD, it is **automatically pseudonymised**, pushed to NHS Digital, where it is linked to NHS secondary care data, and then passed from NHS Digital to CPRD **in anonymised form**.

CPRD provides access to UK patient datasets that are quality-assured, longitudinal, representative and collected in real time. Organisations from the private and public sectors can access CPRD data provided they receive approval through the CPRD's Research Data Governance Process.

CPRD also offers services such as the CPRD SPRINT, for commercial and non-commercial organisations to recruit patients for clinical trials. CPRD SPRINT (SPeedy Recruitment INto Trials) supports organisations that are rapidly recruiting patients living with chronic conditions into phase 2 and 3 trials. This service uses CPRD data to identify potentially eligible patients in the community. Similarly, CPRD can help organisations to run post-market clinical trials (also known as “pragmatic trials”), in which data is used to understand how medications work in the real world.

Figure 4: CPRD data ecosystem map



Source: London Economics analysis

Impact

CPRD data has been used in **2,900 peer-reviewed publications**, with approximately 60% of these publications (1,740) being from the last five years.

Half of CPRD users are commercial and all of the top 20 largest pharmaceutical companies are CPRD users. CPRD enables access to anonymised patient-level primary care and linked data, which supports a wide range of commercial applications. End users can also get access to expert advice on data, validation and study design from the CPRD.

CPRD SPRINT services are considerably less expensive and time-consuming than traditional methods of patient recruitment. Using CPRD SPRINT services to recruit individuals reduces the cost by half, which means that companies could save around **£250,000** on a clinical trial involving 100 patients.²⁹

Assuming that CPRD will run 10 SPRINT services for 10 clients per year (10 trials), we estimate that this would result in a **cost savings of at least £5 million** for the pharmaceutical industry per year.

²⁹ In practice, the cost per patient varies significantly. We have assumed an average cost of £5,000 per patient, based on information from stakeholder consultations.

In addition, depending on the size of the trial, companies could complete their trial six months faster if they opt for the SPRINT service. This time saving benefit may be incalculably large in cases where firms are competing to get a drug to market first.

CPRD “pragmatic trial” services are also considerably less expensive than traditional methods. CPRD can run a 3,000-person pragmatic trial for £10–20 million for a commercial organisation, which would cost around £100 million to run without CPRD; an **estimated saving of between 80 and 90%**.

In the future, if CPRD has the capacity to provide recruitment services for 10 pragmatic trials annually, this would result in a total cost saving of **£800 million** for the pharmaceutical industry.

GPs can earn additional income for their practices through clinical trials and surveys. These payments depend on the type of trials being conducted and the value would be agreed upon between the GPs and the company conducting the trial – the value of these payments is confidential. Based on a review of the literature, we assume that the minimum payment to GPs for recruiting patients into a trial is £50 per patient.³⁰

If 10 trials consisting of 100 patients each were conducted each year through CPRD, the total benefit to UK-based GPs would amount to at least **£50,000 per year**.

GPs also receive regular practice-level prescribing and patient safety quality improvement (QI) reports from CPRD.

Costs

CPRD is a not-for-profit organisation that aims to aid public health research and recoup the costs of doing so through services to the private sector. Because of this, it is assumed that the total revenue earned by CPRD is approximately equal to its total operating costs. Total annual revenue for CPRD is estimated to be just over **£10 million**, based on the number of multi-study licence agreements sold and the proportion of revenue which comes from these agreements.³¹ A multiple study licence agreement for commercial organisations costs **£330,000 per year**. A multiple study licence agreement for non-commercial organisations costs **£75,000 per year**.

³⁰ Raftery, Jet al (2008), Payment to healthcare professionals for patient recruitment to trials: systematic review and qualitative study. *Health Technology Assess*; 12(10)

³¹ These data points were provided to us through stakeholder consultations.

Supporting public health research in the UK

CPRD demonstrates how safe data access controls can support access to uniquely valuable datasets, which would otherwise be unavailable to stakeholders. Major pharmaceutical trials are increasingly being conducted in the UK because CPRD lowers the cost of recruiting to trials and provides high-quality data. This benefits patients as participating in trials means they can get access to treatments before they are available on the market. Research on Covid-19 heavily relied on CPRD datasets and services.

CPRD linked datasets provide a fuller picture to public health institutions, enabling them to make informed improvements in patient safety and the delivery of care. CPRD lowers the cost of collecting data and therefore the cost of conducting medical research. This may increase the already substantial returns to medical research.

OpenActive

Developing open data standards to publish and use data about activities

About

OpenActive is a community-led initiative for the sport and activity sector. OpenActive has developed **open standards for data** about where, when and what activities are taking place, to make it easier for people to find and book activities online.³²

OpenActive brings together organisations from across the sports and physical activity sector to publish and use open data, and has more than 120 member organisations in the UK that are helping to embed OpenActive standards into their platforms.³³ Although OpenActive provides some infrastructure, its main role is to provide data standards and encourage data publication.

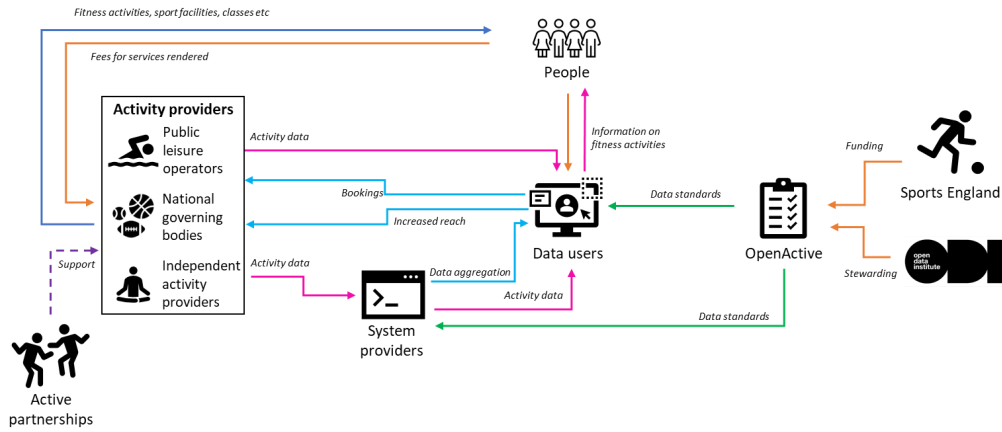
The standards developed by OpenActive are applied by system providers, data aggregators and data users so that the information about activity opportunities follows the same format and can be easily integrated into various services and tools such as OpenSessions (a platform funded by Sports England that allows members to promote activity sessions to wider audiences) or Playfinder (an online platform to book activities and sports facilities). Activity providers may be **individuals, clubs, grassroots organisations** such as charities, or larger institutions such as schools or gyms.

Innovators can use this open data to build products and services that help people find the information they need to get active. It improves the customer experience, helps activity providers reach new audiences, and helps more people get active.

³² OpenActive (n.d.), [OpenActive](#)

³³ OpenActive (n.d.), [Our members](#)

Figure 5: OpenActive data ecosystem map



Source: London Economics analysis

Impact

Activity providers using OpenActive standards are able to share and promote information about their activities more effectively with system providers and data users, and therefore customers are more likely to find, book and participate in activities.

It is estimated that OpenActive standards have led to an **additional 325,893 opportunities** published per month, resulting in **almost 530,000 additional bookings**.

OpenActive generates just over £15.5 million per month through additional activity bookings. Based on information gathered through stakeholder interviews, between 10% and 15% of this revenue is claimed by the booking system as commission, while the rest accrues to the activity providers. **Activity providers therefore generate an additional £13.6 million in revenue, and booking systems £1.9 million, per month.**

Customers spend less time trying to find activities they want to participate in, and are more likely to find activities in their price range, in convenient locations, and that fit into their schedules. Information about the amount of time saved through data sharing is not currently documented, and it could be that the only reliable way of recording this would be through a representative survey of consumers and activity providers.

For each of the additional 530,000 bookings, assuming that consumers save an average of one hour searching time per activity booked, and a UK national living wage of £8.91 per hour, we estimate **the total value of time saved by individuals to be £4.7 million per month.**

Activity providers that share their information with system providers and/or data users according to the OpenActive standards save time on marketing and promotion of their activities. For example, grassroots activity providers can upload information to OpenSessions, which can be accessed and shared by booking systems or activity finders. Without OpenActive standards to enable this, these activity providers would have to promote their classes themselves, either through social media or by leafleting.

Based on the conservative assumption that activity providers save £10 per month on promotion and marketing, we estimate that open data publication supported by OpenActive saves them **£300,000 per month** on trying to reach their audiences.

Increasing access to data leads to an increase in the number of hours of physical activity completed by people in the UK: around half of those participating in opportunities published through OpenActive would not have been able to do so had it not been for the open data standard.³⁴

Participating in physical activity promotes a healthier lifestyle, as well as having a positive impact on social connectedness. There are also wider societal benefits such as a reduction in premature deaths due to inactivity, fewer losses to the economy and lower healthcare costs that are borne by the state.

It is estimated that the increase in the number of hours spent exercising that open data enables can lead to the avoidance of almost 79 premature deaths per month, and a reduction of **£4.1 million** in healthcare costs.

³⁴ Frontier Economics (2019), Open standards for data: Evaluation the economic and social returns

Costs

OpenActive has **operating costs of around £200,000 per year**, which is used to pay for the initiatives' staff costs. A further **£70,000 is needed each year for stimulus funding**, such as running pilot programmes and raising awareness about the open data standards. The **maintenance and operational costs of running OpenSessions** (a systems provider working with OpenActive data) amount to £15,000 per year now that the platform has been developed.

Future prospects

At present, only activities relating directly to a sport or physical activity are covered under the data standards. However, stakeholders mentioned that OpenActive could broaden this to include other types of activities in the future, such as social prescribing. Social prescribing is a holistic approach to people's health and wellbeing in which people with long-term conditions can be linked to community groups and services for practical and emotional support.³⁵ For example, a local agency may prescribe activities such as gardening or a cooking class for an individual suffering from depression or loneliness. These types of activities are often prescribed alongside physical activities, and could therefore be a potential area of expansion for OpenActive.

³⁵ NHS (n.d.), [Social prescribing](#)

Bottom-up data institutions

Empowering people to take a more active role in stewarding data about themselves and their communities

In our work exploring data institutions, we've come across a number that have the objective of enabling individuals – typically those from whom data has been generated, or to whom the data relates – to play a more active role in data stewardship.

We describe these as 'bottom-up' data institutions.³⁶ Some enable people to donate data about themselves to support further health research; some aim to harness data to support specific groups, such as gig economy workers. Many of them focus on addressing power asymmetries, and enabling increased participation from individuals and communities in decisions related to their data than they would otherwise have had. For example:

- Variant Bio works with historically marginalised populations to facilitate people-driven therapeutics. Communities are engaged prior to the beginning of research projects; their data is then collected and used within a framework that centres community concerns.
- Driver's Seat is an independent, driver-owned cooperative where members' data is used to derive insights that help them optimise their earnings and performance.
- Swash enables users to control what data is collected about their browsing habits, as well as to aggregate and sell access to this data to generate financial return.
- OpenHumans empowers individuals and communities to explore and share their personal data for the purposes of education, health and research.
- MIDATA enables users to contribute to medical research and clinical studies by granting selective access to their personal data.
- Gyeonggi Data Dividend ensures that any financial profits generated by selling access to data about transactions using the local currency are returned to citizens in the form of a dividend.

Concepts such as 'data cooperatives', 'data trusts', 'data unions' and 'personal data stores' fall within this 'bottom-up' category, and there has recently been a lot of interest, exploration and experimentation in this area, as evidenced by work from the Aapti Institute, the Data Trusts Initiative, the Ada Lovelace Institute, the Mozilla Data Futures Lab, and others.

³⁶Open Data Institute (2021), [Could 'bottom-up data trusts' help to tackle the climate crisis?](#)

For the purposes of this study, however, we were unable to find a suitable case study candidate. While there are initiatives based in the UK that align with the objective of bottom-up empowerment, it was difficult to determine whether they had attained sufficient institutional maturity and sustainability to demonstrate impact.

This is a nascent field, rife with experimentation, and bottom-up data institutions are likely to have a large impact in the future. We're currently undertaking research to explore the different interventions available to the UK government and other policymakers to support and enable bottom-up data institutions. Our experience with this study will enrich our insights on the state of such institutions and the nature of support they may need.

Synthesis and reflections

Data institutions come in many shapes and sizes, and operate in very different domains. This diversity means it is difficult to generate robust macroeconomic estimates of their impact. Our case study approach helped us map how an illustrative set of data institutions might create different forms of value within their different ecosystems. Some of the things we learnt along the way were:

- 1. It is difficult to collect the data necessary to generate robust estimates of the impact of data institutions.** Data institutions are at different stages of institutional maturity. For more established institutions, the quantification process was aided by the greater availability of data about them and how they interact with their stakeholders. However, in general, data on their impact was challenging to obtain. This challenge is most acute with bottom-up data institutions, arguably the most experimental approach to data stewardship, where we could not identify an appropriate case study to use to assess their impact. On the other hand, data on the costs associated with stewardship roles was relatively easier to obtain, and demonstrated that data institutions generate impact for a wide range of stakeholders at comparatively limited cost.
- 2. Despite their variety, there are some similarities in the nature of the impacts that data institutions have.** While estimates from different data institutions cannot be directly compared, our research brought to the fore important similarities in how data institutions create value. Some of the data institutions we studied enabled access to data for research (MEDIN, Farmbench, CPRD, 360Giving), saving time and resources by making data more findable, accessible, interoperable and reusable. Some data institutions also balance the need to broaden access to data with restricting it in ways that generate commercial revenue to sustain their activities (for instance, CPRD provides data access to the largest pharmaceutical companies which helps to sustain access for non-commercial users; 21% of MEDIN's users are commercial users, using marine data for use cases such as informing shipping routes and developing information products).
- 3. It is easy to undercount the broad, societal impacts of data institutions.** Reducing the resource intensity of farming (FarmBench) and the use of marine data in conservation projects (MEDIN) have clear environmental benefits. Improving access to clinical data for research (CPRD) and driving increased physical activity (OpenActive) have benefits for individual wellbeing and public health. Enabling more effective grant making (360Giving) supports a range of philanthropic outcomes. However, these benefits are disparate and downstream, which makes them difficult to quantify.

4. **Data institutions adapt and adjust their activities, which gives rise to evolving types of impact.** 360Giving is in the process of expanding to allow a more detailed set of data categories about grant making to be published; Farmbench is expanding its coverage to similar initiatives outside the UK; OpenActive is exploring expanding its work on data standards for social prescribing; and CPRD datasets have supported research into Covid-19.

5. **An underappreciated function of data institutions is that they build capacity among stakeholders in their ecosystems.** Some data institutions do this consciously to increase participation and uptake: Farmbench supports farmers in inputting data into the tool, and holds workshops to share insights from its analysis; CPRD provides access to expert advice on data validation and study design; 360Giving provides support and information to users on using their tools; and MEDIN conducts workshops to build knowledge and capacity on their tools and metadata standards.

Given the limited scope of this project, we're aware that we will not have captured all of the impacts of data institutions, and we have tried to be clear where we've made assumptions to generate estimates. We've also highlighted the types of impacts and potential methods that would aid such efforts and future work in this area.³⁷

Building on this study, avenues for future work could entail:

- employing additional data collection methods such as user surveys to generate richer and more accurate assessments of downstream impacts
- exploring ways to meaningfully compare quantified impacts across data institutions or begin to extrapolate to economy-wide estimates
- assessing the impact of bottom-up data institutions outside the UK that may have reached sufficient maturity, especially given the policy interest in the field.

Data institutions are characterised by their diversity. With this study, we've begun to describe and measure the ways in which they may typically engage stakeholders and create impact. We hope our findings will inform policy makers, funders and data institutions themselves as they advocate for continued support in their role as key enablers of trustworthy data stewardship.

³⁷ [See Annex.](#)