Research Data Management: Benefits and Data reuse case studies

Part A—Long-form case studies

1: Neurodegeneration and Dementia Research
2: Biological Data Mining
3: Gene Expression Omnibus Data to Fight Cancer
4: The World vs E.Coli
5: Data-Powered Patents
6: Sharing Research Data and Infrastructure to Study Proteins
7: Using Flies to Understand The Human Brain
8: A Data-Based Approach to Preventing Curable Eye Diseases
9: Self Compacting Concrete
10: Preventing Drug Interactions On Your Mobile
11: Meeting Sustainability Objectives
12: Finding Offshore Hydrocarbons
13: Geological Data Made Easy
14: Protecting The Oceans By Coordinating Data Sharing Efforts
15: Corruption in Public Sector Procurement
16: Drones for Research
17: Data-Enhanced Archaeology
18: Supporting Science and Industry by Sharing Computer Code
19: Citizen Science at Zooniverse
20: History and Data
21: Saving the Earth from Mankind
22: Understanding Mobile Users and Evaluating Vulnerabilities
23: Understanding War
24: Surfing Gravitational Waves
25: Art From The Comfort Of Your Chair

Part B—Short-form case studies

1: A Performance Artwork Based on Datasets and Partnerships
2: Open Science Underpins Collaboration: The Structural Genomics Consortium (SGC)
3: Testing Doubts About the Reliability of Science: The Reproducibility Project
4: Large Volumes of Data Engage Large Communities: Sloan Digital Sky Survey (SDSS)
5: Scholarly Communication is About Combining Effort: Polymath Project
6: Is Citizen Generated Data Suitable for Academic Purposes? Volunteered Geographic Information (VGI)
7: A Specialist Research Data Archive: Crawdad
8: A Large-Scale Research Data Service: The European Bioinformatics Institute
9: Combining Data & Influencing Government Sustainability Policies
10: Objective Measures & Self-Reported Data Underpin Policy On Obesity: Health Survey for England (HSE)
11: Archival Acceptance as an "Indicator of Quality": Foot & Mouth Disease (FMD) Testimonial Data
12: Improving Policy by Providing Data: Live Music Exchange (LMX)
13: From Hard Copy Primary Sources to An Open Online Resource: Reading Experience Database, (RED)
14: Qualitative Data in Many Formats, Archived Online: Tate Encounters
15: Data Combination and Self Re-Use: Understanding Pauper Lives in Georgian London
16: Ireland-Bristol Trade in the Sixteenth Century
17: Film and an Ethnographic Approach: Buddhist Cosmology in Food

18: Data About Data Archiving: ICPSR's Data Sharing in the Social Sciences

19: Research Data Supports Restoration: Mackintosh Architecture

20: Evidencing Value of Artistic, Cultural or Sporting Activities: UK Subjective Well-Being Data (SWB)

Introduction

Research Data Management (RDM) is an overarching term encompassing the organisation, storage, and documentation of data generated during research projects. RDM deals with the organisation and curation of active research data, with its day-to-day management and use, and with its long-term preservation.

RDM is an important practice for both institutions and individual researchers. Data supporting results should be made available and preserved so as to allow its reuse and the verification of published research. Several other benefits can arise from the implementation of RDM, including increased citations, increased research collaborations, or increased visibility. Today, data has to be managed not only for preservation purposes, but also to fulfil the requirements of most research funders.

Although RDM has been around for a while, the above benefits are usually described qualitatively and the lack of a solid body of evidence makes advocacy difficult. We have sought to fill this gap by using case studies to present a rich and varied picture of the impacts underpinned by RDM.

Spread of disciplines

The case studies assembled here come from a wide range of research fields. Due to inherent differences between disciplines, the benefits of RDM become apparent in different ways. They are more tangible in certain fields, and more abstract in others. Nonetheless, the case studies demonstrate that RDM is a worthwhile activity for all institutions and researchers.

The examples below mostly involve large data management initiatives, as these are more likely to show the wide reach of the benefits of RDM. However, we would like to stress that even smaller data management efforts can have an impact. Unfortunately, this can be very difficult to track, as an individual researcher reusing data from other individual researchers is often lost in a sea of information. Similarly, impact sometimes cannot be traced to a specific source: in some studies, clear evidence of the impact of RDM is available, but they point to a whole repository rather than to a single study or dataset.

Enablers of impact from RDM

The effective implementation of RDM requires both cultural change and specific data skills. This makes its dissemination and practical realisation difficult and is the main obstacle to the above-mentioned benefits. It is, therefore, desirable to examine the RDM environment to investigate its enablers and what has worked historically to encourage future data curation and reuse.

Our research into the benefits or RDM led us to discovering some of the circumstances and situations that facilitate it, along with some of the reasons why this practice should be pursued. A summary of our findings is as follows:

- Open licensing (e.g., in the case of computer code and algorithms) is essential to allow crowd-sourced improvements.
- Data repositories and infrastructures are among the most significant enablers of impact: without them, very few of the impact case studies below would have been possible.
- Collaborations between international bodies or organisations strongly promote data re-use, especially in fields

where it was not possible for a single player to take charge. These collaborations create the right environment for sharing and re-use of research data: cultural change is encouraged along with the use of joint infrastructures at a national or international level.

- The impact of RDM is normally seen after a long time, when, i.e., after has been produced, curated, maintained, and reused. Thus, there is a need for sustained investment in this field, as benefits cannot be seen immediately.
- Aggregation of data and digitisation of documents are key to encouraging the development of digital humanities.

These initiatives often arise from the collaboration between museums, research libraries, and universities. When data that was spread between several sources (e.g., many different books/articles) or held in obsolete formats was organised and analysed through sound RDM, hidden findings could be uncovered.

Presentation format

The case studies are grouped into Long Form— more detailed with abstracts etc.— and Short Form. In all cases key information is presented in a table after the text.

Part A—Long-form case studies

1: Neurodegeneration and Dementia Research

Leveraging data from cohort studies to accelerate medical progress

About 850,000 people in the UK currently suffer from dementia. Dementia is a syndrome causing deterioration in memory, thinking, behaviour, and the ability to perform everyday tasks. While it usually affects older people, it is not a part of the normal aging process. Dementia is one of the main reasons for disability in older people, making them highly dependent upon caregivers, families, and social services. This leads to an estimated socioeconomic cost of over £26 billion a year.

A £53 million collaboration between academia and industry, Dementias Platform UK (DPUK) is working to turn the best dementia research into treatments as quickly and efficiently as possible. The project aims to find ways to delay the onset of dementia, to alleviate its symptoms, to improve the quality of life of the patients, and to slow down or halt the progression of the diseases.

The DPUK portal

The DPUK collaboration is a coordinated way to conduct research in the fields of neurodegenerative disease and dementia. Its aim is to leverage the data gathered in cohort studies, over 30 of which (totalling over 2 million participants) can be accessed through the DPUK portal. Data is shared by the cohort studies using a secure data analysis portal with high levels of ethics and data security. For an overview of the cohort studies involved and details on each of them, the DPUK includes a useful cohort directory, which provides information on principal investigators and an overview for each study.

Users include both academic researchers and industry stakeholders such as pharmaceutical companies, and the initiative also aims to develop new methods to enhance research in the field. The collaborative approach to data sharing pursued by DPUK is shown by its connections with important national and international stakeholders, including the UK Dementia Research Institute, the European Medical Information Framework (EMIF), the European Prevention of Alzheimer's Dementia Consortium (EPAD), and the Canadian Consortium on Neurodegeneration in Aging (CCNA).

Use of DPUK resources and reach of the program

DPUK is a relatively young initiative, however, its potential for impact is significant. For instance, the £6.9m NIHR-MRC Dementia Deep and Frequent Phenotyping study is currently using DPUK to try to improve the success of clinical trials for treatments in Alzheimer's disease. It involves 250 volunteers from study cohorts in the DPUK collaboration and will perform up to 50 tests that have never been used before to detect dementia, over the course of 12 months. The project participants can be categorised thanks to the data in the DPUK, and include both people at risk of developing Alzheimer's disease and others who are not considered at risk.

The DPUK consortia includes six UK and international companies: Araclon Biotech, MedImmune, GlaxoSmithKline, Ixico, Janssen Pharmaceuticals, and SomaLogic. These companies are especially interested in the detailed health information on 10,000 participants with early-stage dementia, and aim to conduct novel research in the field of neurodegenerative diseases. This shows how research data management initiatives such as DPUK have the potential to enhance knowledge exchange, and to bring real benefits to the public through collaboration with the private sector.

Title	1: Neurodegeneration and Dementia Research
Subtitle	Leveraging data from cohort studies to accelerate medical progress
Abstract	Dementia and neurodegenerative diseases affect both individuals and society as a whole. However, neither cures nor treatments are available at the moment. The Dementias Platforms UK collaboration aims to turn dementia research into treatments as quickly as possible, to both improve people's lives and decrease the socioeconomic cost related to these types of diseases.
Keywords	dementia; Alzheimer; cohort studies
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Impact on health and wellbeing
Facts and figures	850,000 people with Dementia, £53 million project (DPUK), £6.9 million project financed using DPUK
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2014
Organisations involved	University of Cambridge; Cardiff University; University of Edinburgh; Imperial College London; King's College London; University of Manchester; University of Newcastle; University of Oxford (Academic Lead); Swansea University; University College London
Academic citations	
Links	https://www.mrc.ac.uk/research/facilities-and-resources-for-researchers/dementias-platform-uk/
	http://www.who.int/mediacentre/factsheets/fs362/en/
	http://www.dementiasplatform.uk/
	https://www.mrc.ac.uk/news/browse/world-s-most-in-depth-study-to-detect-early-signs-of-alzheimer-s-disease/
	https://www.alzheimers.org.uk/info/20025/policy_and_influencing/251/dementia_uk

2: Biological Data Mining

Difficulties in data handling in the field of biology

If you are a researcher or deal with researchers as a part of your job, you probably know how difficult it is to handle data. While this statement is true in general, things get even worse when large databases from a range of heterogenous sources need to be brought together to allow further work and safekeeping. In the field of biology, using diverse datasets is becoming increasingly common, but it is very difficult in practice. This type of work, called integrative analysis, is possibly only when biological data is properly organised and can be queried easily.

The InterMine platform

Conscious of the above issues, researchers from the Micklem Lab at the University of Cambridge created InterMine, an open source data warehouse system for the integration and analysis of complex biological data. The system, which is still in active development, is now used by a number of major model organism databases. InterMine provides parsers for integrating data from many common biological data sources and formats, and also lets users add their own data. One of the most interesting features of the InterMine system is that it allows querying and data mining, even in the case of very large databases (e.g., the modENCODE projects contains >300GB of data). Such a feature is essential for researchers, as it would be very difficult for them to achieve the same level of performance independently.

The InterMine system was built with the uses in mind and provides an easy-to-use web application with advanced analysis tool. This allows end-users to access and explore data without any programming knowledge, which means that InterMine bridges a very wide gap for researchers. In addition, InterMine takes full advantage of the most recent cloud technology, allowing users to start instances of the system on the Amazon Cloud.

Impact and reach

As of March 2017, Google Scholar reports a grand total of 574 citations for the main individual data warehouses powered by the InterMine system (InterMine, FlyMine, MitoMiner, modMine, TargetMine, and YeastMine). By powering these, the InterMine data warehouse system indirectly allows advancements in biology research, including, but not limited to, the study of model organisms, mammalian localisation evidence, phenotypes and diseases, and drug discovery. The information in the data warehouses powered by InterMine allows biologists to re-use data that has been previously captured by other studies and build on these research findings, saving time and resources. The true reach of the Intermine system is likely to be even greater than the citation figures suggest, as data is often used without citations (as is often the case for freely available web resources).

Title	2: Biological Data Mining
Subtitle	Difficulties in data handling in the field of biology
Abstract	Biologists often need to deal with heterogeneous data sources, which makes their work difficult and time-consuming. The InterMine system provides an easy-to-use data warehouse solution that biologists can exploit for their studies with little programming knowledge.
Keywords	biology; data warehouse; data mining
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	574 citations
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2012
Organisations involved	University of Cambridge
Academic citations	Smith, R.N. et al. (2012). [InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics.](https://doi.org/10.1093/bioinformatics/bts577)
Links	http://intermine.readthedocs.io/en/latest/about/
	http://www.modencode.org/publications/about/index.shtml
	http://www.flymine.org/
	http://mitominer.mrc-mbu.cam.ac.uk/
	http://intermine.modencode.org/
	http://targetmine.mizuguchilab.org/
	http://yeastmine.yeastgenome.org/

3: Gene Expression Omnibus Data to Fight Cancer

Systems biology and enhanced cancer diagnosis

With 4,348 datasets and more than two million samples, the Gene Expression Omnibus (GEO) shares and hosts microarray, next-generation sequencing, and other forms of high-throughput functional genomic data. Run by the US National Center for Biotechnology Information (NCBI), the repository allows biologists to find out how several kinds of cells look from a molecular perspective. This is very useful, as knowing whether a gene is active in a cell is essential to know how to shape research.

The GEO platform

The platform is a priceless tool for researchers in the life sciences, as it allows them to upload their data for easy sharing with their peers. In addition, the GEO is an effective data management tool, as authors can upload their research data for safekeeping even before releasing them to the public along with their publications. The GEO also acts as a collaborative platform for authors and reviewers, as the data submitted can be accessed by various stakeholders for peer-review before manuscripts referencing them are accepted for publication in journals.

GEO data re-used

The biggest strength of the GEO lies in the fact that it allows data re-use and the development of new research. A Google Scholar query of "Gene Expression Omnibus" currently returns about 65,000 results, which shows the reach of the platform.

The research performed thanks to GEO data has an impact in various fields. One of the most promising is the application of systems biology to cancer research, allowing scientists to understand how the networks of molecular interactions in cancer cells behave. While the mathematical and computational tools to understand these complex mechanisms could, potentially, have been developed in the past, the availability of data sources such as the GEO was essential to advance our knowledge on cancer. Understanding genetic networks allows a prediction of the future evolution of cancerous cells, which, in turn, helps researchers in the pharmaceutical field understand how to design better anticancer drugs. This level of detailed analysis is enabling scientists to develop new treatment options that seek to modify individual proteins to halt cancer progression, or to modify the entire network of cells and reset it to a normal state, depending on the stage of the disease.

Other important cancer research exploiting the GEO platform relates to the way cancer progression is identified. As reported by Nature, cancerous cells have been found to infect other healthy cells in the immediate surrounding tissue by shedding exosomes, which contain proteins, DNA and RNA. These exosomes then combine with healthy cells and cause the replication of cancerous cells, effectively infecting the healthy cell and spreading to new tissues. By tracking exosomes, disease can be tracked more efficiently than by isolating tumour cells floating

in blood, which are far less abundant. While this was just hypothesised in 2014, the approach has reached the US market in 2016.

Title	3: Gene Expression Omnibus Data to Fight Cancer
Subtitle	Systems biology and enhanced cancer diagnosis
Abstract	Sharing genomic data on the Gene Expression Omnibus (GEO) repository has a significant impact on medical research and improves the efficiency of the publishing environment. New approaches to cancer research were developed, including a novel method of diagnosis based on the study of exosomes.
Keywords	medical research; cancer; repository; publishing
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Methodological impact (e.g., new approaches developed); Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	65,000 results from Google Scholar
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2001
Organisations involved	National Institutes of Health
Academic citations	
Links	https://www.ncbi.nlm.nih.gov/geo/
	http://www.nature.com/news/cancer-cells-can-infect-normal-neighbours-1.16212
	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3354560/
	http://www.nature.com/nbt/journal/v34/n4/full/nbt0416-359.html
	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S9-S1
	http://www.nature.com/news/don-t-let-useful-data-go-to-waste-1.21555

4: The World vs E.Coli

Researchers joining forces to sequence bacterial genomes

Affecting 16 countries and almost 4,000 people, the 2011 E.Coli outbreak took a terrible toll on Europe and beyond. The strain caused illness and deaths between May and June 2011, along with massive financial losses due to the danger of exporting contaminated produce.

The bacterial strain responsible for the outbreak was identified as O104:H4 by the Robert Koch Institute. However, scientists were looking for further information to understand where it originated. When the Beijing Genomics Institute openly released the full genome sequence of the outbreak strain, researchers from all over the world started to investigate it further. They managed to discover ten isolates of the strain from the outbreak thanks to a collaboration between universities, public, and private laboratories.

The investigation of the outbreak strain

Initially, researchers and public officials thought that the vehicle of the infection was the consumption of raw vegetables and salads. These were thought to be infected with a typical Shiga toxin-producing E. coli (STEC), thus, public authorities recommended that people avoid eating certain types of food. Further collaborative analyses proved that the strain was, instead, very rare and managed to explain the symptoms of the diseases and complications reported by the nearly overwhelmed hospitals. The Robert Koch institute confirmed that German sprouts were the source of the outbreak and a later assessment implicated Egyptian seeds that had been shipped to Germany in 2009. Interestingly, this conclusion was possible thanks to the combination of traditional and more recent methods of investigation, i.e., epidemiological methods and genome sequencing.

A literal example of “sharing is caring”

Understanding how the outbreak arose and where it originated was a very difficult task, but it was performed in record time. This is mostly because the datasets from sequencing initiatives were publicly released in a timely manner, so as to allow effective data re-use. The results from the analyses were covered via blogs, Twitter, and other websites, outside the traditionally slower scientific publication route. The combination between the analyses allowed the characterisation of the outbreak, which would have been extremely burdensome for a single organisation to undertake. On GitHub, 17 researchers gathered the results of their crowdsourcing effort to better understand the E.coli strain and compiled a list of 12 academic publications that arose from their work. Their crowdsourcing efforts were covered by 8 media outlets and showed that collaborations can increase a researcher’s profile and reputation while also serving a higher purpose such as analysing a bacterial outbreak. In addition, this collaboration showed that, when data is managed properly, newly-found highly virulent food-borne pathogens can be promptly analysed and tracked to their origin, thus, improving the safety of all citizens.

Title	4: The World vs E.Coli
Subtitle	Researchers joining forces to sequence bacterial genomes
Abstract	During the 2011 E.coli outbreak, the release of the bacterial genome in the public domain allowed researchers to reach conclusions quickly and effectively. The strain was tracked to a seed shipment from 2009 and the crow-sourced analysis received high media attention.
Keywords	infection; e.coli; outbreak
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Impact on health and wellbeing
Facts and figures	17 authors working in a crowdsourcing effort 12 academic publications 8 media outlets covering the initiative
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2011
Organisations involved	BGI; Life Technologies; University Muenster; Health Protection Agency
Academic citations	
Links	http://blogs.nature.com/news/2011/06/the_german_e_coli_outbreak_40.html https://en.wikipedia.org/wiki/2011_Germany_E._coli_O104:H4_outbreak http://www.sciencemag.org/content/332/6035/1249 https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki https://www.cdc.gov/ecoli/general/ http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19890 http://www.dw.com/en/british-scientist-identifies-genetic-sequences-in-new-e-coli-strain/a-15133914

5: Data-Powered Patents

Citation patterns in granted patents in the field of biology

In 2014, open access data resources were cited by more than 8,000 patents in the field of biology. British researchers from Cambridge (affiliated with the Wellcome Genome Campus, ELIXIR, the EMBL, and Ganesha Associates) combined information from different biomolecular data repositories and looked for re-use in industry, uncovering this important result. Their work demonstrates both how academic and industrial innovators can leverage data to advance applied science and the extent of secondary data re-use in the field of biology.

The analysis of data citations in patents

The researchers linked open access scientific literature from Europe PubMedCentral to the metadata in biological repositories, such as the Protein Data Bank, the European Nucleotide Archive, and SureChEMBL. Their efforts were supported by text mining, a practice that uses algorithms to look for patterns, trends, and other information. Text mining is possible only when data is properly and consistently managed, which means that the information described in this case study primarily exists as a result of good research data management.

The main aim of the researchers' work was to investigate the secondary use of bioinformatics resources to understand the flow and impact of data in the field of biology. No previous studies addressed this and there was room for highly improvingof the indicators used to evaluate the impact and reach of academic work.

The use of biological data sources in official documents such as patents shows that they don't always concern basic research and that they can be a "fundamental component of the digital knowledge management framework needed in a modern society".

Other implications of the study

This investigation of the impact of biological data on patents highlights the importance of repositories, but it also sheds light on the thorny topic of citations and researcher evaluation. One of the ways researchers' performance is assessed is by considering the impact of their work. This is most often approached by analysing indicators such as the number of published articles and citations by other researchers, usually excluding citations to data. On the other hand, the above-mentioned work on citations from patents clearly shows how citations of academic articles only show a partial picture. This leads to the conclusion that a more comprehensive assessment of researchers' performance should include robust indicators of data citation in both articles and other types of documents, including scientific and technical documents such as patents, guidelines, reports, or grant applications.

In addition, research funders have been increasing their demands in terms of open data policies, thus, it is essential to develop metrics that evaluate how open research data is used. These can offer a range of benefits, including improving reward mechanisms for scientists, supporting the management and sustainability of data repositories, and understanding the societal value of data policies.

Title	5: Data-Powered Patents
Subtitle	Citation patterns in granted patents in the field of biology
Abstract	The analysis of data citation patterns in the field of biology showed how over 8,000 patents were based on publicly available data. While this proves the usefulness of repositories as a whole, it also shows how the evaluation of researchers should consider data citations and alternative sources, too, as these are often key to uncover the broader industrial and societal value of academic research.
Keywords	citations; patents; text mining
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed); Socio/Economic impact
Summary Impact Type	Impact on economy and business
Facts and figures	8,000 patents thanks to open data
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2014
Organisations involved	University of Cambridge
Academic citations	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](http://dx.doi.org/10.12688/f1000research.7911.1) . F1000Research.
Links	https://www.jisc.ac.uk/guides/text-and-data-mining-copyright-exception
	http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies
	http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/

6: Sharing Research Data and Infrastructure to Study Proteins

A research group's impact on the study of circular dichroism

With over 375,000 analyses run, the DICHROWEB online analysis tool is a clear example of how data interoperability can enhance research. Before the tool was released, researchers in the field of biology needing to study protein structures had to deal with a range of different programs using a variety of inputs and outputs that were not compatible with one another. Consequently, even though those programs were publicly available, it was very difficult to use them interchangeably. Bonnie Ann Wallace (Professor of Molecular Biophysics at the Birbeck College, University of London) and her team developed DICHROWEB as a web-based tool allowing the use of many algorithms to analyse circular dichroism (CD) and synchrotron radiation (SRCD) spectroscopy data.

The DICHROWEB server and the Protein Circular Dichroism Data Bank

The DICHROWEB tool accepts any recognised data format in the field and allows access to the Protein Circular Dichroism Data Bank (PCDDB), a repository on the topic. The PCDDB was also created by Prof Wallace, in collaboration with Dr Bob Janes from Queen Mary University of London. The PCDDB is a publicly available repository for CD and SRCD data to enable sharing, dissemination, and re-use in the field. DICHROWEB and the PCDDB are linked, so that users of the PCDDB can run their own computational analyses on the data deposited. This improves the research workflow, as a direct connection between the repository and the analysis tools means that users are not required to download or import/export data.

Prof. Wallace and her team implemented a large number of functions in the DICHROWEB portal, both novel and from the literature. These take into account advances in X-ray crystallography, NMR spectroscopy, and bioinformatics, thus, exploiting existing knowledge and building on it to uncover new methods and findings.

How DICHROWEB and the PCDDB are helping researchers

DICHROWEB and the PCDDB are being used by both academic researchers (for scientific investigations and teaching) and industrial stakeholders. DICHROWEB has over 3,600 registered uses and the tool gathered over 1,000 academic citations, while the PCDDB received 175,000 unique hits from 41 different countries.

Among the industrial users of the platforms, there are companies in the pharmaceutical, biotechnology, and food sectors, ranging from multinational firms to SMEs. In addition, biomedical product and clinical diagnostics suppliers are among the users of the PCDDB. These commercial players successfully exploited DICHROWEB and the PCDDB, developing analyses of the stability and solution properties of proteins and filing 11 patent applications citing the research group's papers (patent holders include Roche Holding AG, Pfizer Inc. and Collplant Ltd). The importance of Prof Wallace and her team's efforts have been widely

recognised and their work has been presented at several international conferences.

The resources are widely used in teaching both in the UK and internationally, including workshops for researchers (summer school) and University courses.

Title	6: Sharing Research Data and Infrastructure to Study Proteins
Subtitle	A research group's impact on the study of circular dichroism
Abstract	The DICHROWEB and PCDDB platforms are widely used for the study of proteins. Since their release, hundreds of thousands of users accessed them, from both academia and the private sector. In academia, the platforms fuel research and teaching, while they led to several advances in industry, including the development of 11 patents.
Keywords	protein; repository; infrastructure
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Impact on health and wellbeing
Facts and figures	375,000 analyses 3,600 registered users over 1,000 citations
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2001
Organisations involved	Birkbeck College
Academic citations	Lobley A., Whitmore L. & Wallace B.A. (2002). [DICHROWEB: an interactive website for the analysis of protein secondary structure from circular dichroism spectra](http://dx.doi.org/10.1093/bioinformatics/18.1.211). Bioinformatics. Whitmore L. & Wallace B.A. (2004). [DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data](http://dx.doi.org/10.1093/nar/gkh371). Nucleic Acids Res Lees J.G., Miles A.J., Wien F. & Wallace B.A. (2006). [A reference database for circular dichroism spectroscopy covering fold and secondary structure space](http://people.cryst.bbk.ac.uk/~ubcg25a/BI_reprint.pdf). Bioinformatics. Whitmore L. & Wallace B.A. (2008). [Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases](http://dx.doi.org/10.1002/bip.20853). Biopolymers. Wallace B.A. & Janes R.W. (2010). [Synchrotron radiation circular dichroism (SRCD) spectroscopy: an enhanced method for examining protein conformations and protein interactions](http://dx.doi.org/10.1042/BST0380861). Biochem Soc Trans. Whitmore L., Woollett B., Miles A.J., Janes R.W. & Wallace B.A. (2010). [The protein circular dichroism data bank, a Web-based site for access to circular dichroism spectroscopic data](http://dx.doi.org/10.1016/j.str.2010.08.008). Structure.

	Whitmore L., Woollett B., Miles A.J., Klose D.P., Janes R.W. & Wallace B.A. (2011). [PCDDB: the Protein Circular Dichroism Data Bank, a repository for circular dichroism spectral and metadata](http://dx.doi.org/10.1093/nar/gkq1026). Nucleic Acids Res.
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=36651
	http://dichroweb.cryst.bbk.ac.uk/
	http://pcddb.cryst.bbk.ac.uk/home.php
	https://www2.warwick.ac.uk/fac/sci/wcas/events/analytical_sciences_summer_school_winter_workshop/cdld/

7: Using Flies to Understand The Human Brain

Harvesting published data to power neurobiology research

How would you feel if you heard that our brain has many features in common with that of a fly? It may sound surprising, but the brain of a fruit fly replicates some aspects of our brain, which can be very helpful when studying human genetics, sleep, and even disease. Fruit flies are considered as model organisms, i.e. organisms that have simpler systems but still approximate to our own. The advantages of model organisms are that they can be manipulated more easily and, in the case of flies, they have a short lifespan. This means that several generations of fruit flies can be studied in as little as a few months.

The Virtual Fly Brain repository

Studying the brain of a fruit fly is no easy task, mostly due to the insect's very small size. Scientists have developed different approaches to study these organisms, including dynamically labelling neural connections, imaging brain activity in moving flies, and exploiting super resolution microscopy. However, these and many other efforts risk going unnoticed if their results are not gathered and organised for the sake of re-usability. This is what projects such as Virtual Fly Brain are doing, by integrating neuroanatomical and expression data from the published literature as well and image dataset onto a single brain template. It is a demanding endeavour due to the need to code existing data and make them suitable for sharing, but the researchers' goal is to make it easier for their peers to find relevant anatomical information and reagents in the fruit fly's brain. The Virtual Fly Brain portal works as an open source 3D image viewer that can be queried to find further information and images on a number of features of the insect's brain. The project is a compelling example of universities and funders collaborating to improve the effectiveness of data sharing and re-use. Partners include the University of Edinburgh, the University of Cambridge, the European Bioinformatics Institute (EMBL-EBI), the Medical Research Council, and Northern American institutions in the FlyBase Consortium (Harvard University, Indiana University, and the University of New Mexico).

A 2012 academic article referencing the Virtual Fly Brain project has already garnered more than 20 citations, indicating that the portal is increasingly recognised as a useful resource in the field of fruit fly neurobiology.

What are the benefits of studying fly brains?

Fruit fly neurobiology is a diverse research field and there are multiple uses for information generated on the topic. For example, mapping patterns of individual neural connections in fruit flies may lead to greater understanding of the communication that occurs between individual synapses in the fruit fly brain. This research can be applied to humans and will assist researchers in developing increased awareness of how the human brain processes information. Other interesting research is focused on the circadian regulation of sleep behaviour. It is thanks to fly genetics that we now know the molecular logic of circadian clocks, and further research on the

topic will allow us to explore the links between sleep and learning at the molecular level.

Title	7: Using Flies to Understand The Human Brain
Subtitle	Harvesting published data to power neurobiology research
Abstract	Studying the brain of fruit flies is helping researchers uncover how our brains work at the molecular level. Data repositories such as the Virtual Fly Brain help them curate, share, and re-use data in a structured way.
Keywords	neurobiology; medicine; fly; brain;
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2012
Organisations involved	University of Cambridge; University of Edinburgh; Medical Research Council; EMBL-EBI
Academic citations	Milyaev, N., Osumi-Sutherland, D., Reeve, S., Burton, N., Baldock, R. A. & Armstrong, J. D. (2012). [The Virtual Fly Brain browser and query interface](http://dx.doi.org/10.1093/bioinformatics/btr677). Bioinformatics 28, 411-5.
Links	https://helix.northwestern.edu/article/why-fruit-fly-research-no-joke
	http://www.nature.com/articles/ncomms10024
	http://www.nature.com/nmeth/journal/v13/n7/abs/nmeth.3866.html
	https://doi.org/10.3389/fncel.2016.00142
	http://www.virtualflybrain.org/site/vfb_site/home.htm
	http://flybase.org/
	http://www.neurobiology.northwestern.edu/people/core-faculty/ravi-allada.html

8: A Data-Based Approach to Preventing Curable Eye Diseases

Leveraging data to inform health policies

Worldwide, 285 million people are estimated to be visually impaired, with 39 million being completely blind. A collaboration of 79 leading ophthalmic epidemiologists from all over the world, the Vision Loss Experts Group (VLEG) worked to lower these figures by releasing location-relevant open data on eye diseases. The group's work has been supported by several key grants, including the Bill and Melinda Gates Foundation, Fight for Sight, and the Brien Holden Vision Institute Foundation (Australia). VLEG data is now incorporated in the IAPB Vision Atlas, where maps on vision impairment and blindness are available for all to browse. Interestingly, the maps show snapshots from 1990 to 2010, showing the evolution of these conditions all over the world.

VLEG's data-gathering work

The group's work was performed in the context of WHO's Global Burden of Disease (GBD) study, which built the disability-adjusted-life-year (DALY) indicator. This considers premature mortality and years of life lost due to time lived in states of less than full health. The DALY indicator is used to assess the burden of disease consistently across diseases, risk factors and regions.

VLEG's work started with a very complex data-gathering exercise, which consisted of assembling epidemiological data from 187 countries and 15,000 items, including published articles and grey literature. While the former were normally available as published research outputs, the latter had to be secured from hospitals and practitioners, as grey literature is most often not in the public domain. The largest study of its kind ever performed, VLEG's work considered geographic factors and trends globally. Such a comprehensive dataset was obviously ideal for modelling and performing meta-analyses, which led to unique metrics on eye disease and its causes.

Impact of the VLEG

VLEG findings had far-reaching effects, both on economic and regional development and on policymaking. These can be summarised as follows:

- PwC used the data in its economic reports Price of Sight and Investing in Vision to calculate the costs and benefits of eliminating avoidable blindness. This was quantified as US$2.20 per capita in developed countries for the ten years starting from 2013.
- The World Bank used GBD and VLEG data to define funding strategies and priorities for developing countries. The President of the World Bank stated that the data will "set the terms of health policy, planning, and funding discussions for years to come".
- The World Economic Forum used VLEG data in the context of the Human Capital Initiative, aiming to demonstrate the detrimental effect of visual impairment on a nation's economic potential and productivity.

- Geographically-linked VLEG data led the Health Minister of Trinidad & Tobago to pledge to collect detailed population eye-health statistics to address this unmet need.
- Policy debates were driven by GBD and VLEG data in many countries, including the UK (Public Health England) and the USA.
- VLEG's work has been used to develop a clinical education and awareness programme in India.

Title	8: A Data-Based Approach to Preventing Curable Eye Diseases
Subtitle	Leveraging data to inform health policies
Abstract	The Vision Loss Expert Group (VLEG) gathered and released data on vision loss all around the world. The data is localised, which means that every country can tackle local issues to reduce the burden of eye loss. The findings of the VLEG are far-reaching and were picked up by large organisations such as PwC, the World Bank, and the World Economic Forum, shaping policies, debates, and educations programmes worldwide.
Keywords	ophthalmology; eye health; VLEG; GBD
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Impact on health and wellbeing; Impact on public services; Impact on politics and governance
Facts and figures	15,000 articles harvested $2.2 per capita to eliminate avoidable blindness in developing countries by 2020 collaboration between 79 scientists
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1990
Organisations involved	Vision Loss Expert Group
Academic citations	Kim, J.Y. (2012). [Data for better health - and to help end poverty](http://dx.doi.org/10.1016/S0140-6736\(12\)62162-X). The Lancet.
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=4989
	http://www.who.int/mediacentre/factsheets/fs282/en/
	http://atlas.iapb.org/about-vision-atlas/vision-loss-expert-group/
	http://atlas.iapb.org/about-vision-atlas/
	https://www.idf.org/sites/default/files/PwC 1. Price of Sight Final Report 2013.pdf
	https://www.hollows.org/getattachment/au/What-We-Do/Ending-Avoidable-blindness/Research/PwC-Investing-in-Vision.pdf.aspx
	http://www.iapb.org/9th-general-assembly
	http://reports.weforum.org/human-capital-report-2016/

9: Self Compacting Concrete

Advancements in engineering led by data aggregation

When authors release data along with their articles, they open the doors to all sorts of new discoveries. Sometimes, these are derived from the data in a single article, but, other times, new advancements in research are possible only when published data are aggregated and analysed as a whole. This is because the aggregation of datasets allows researchers to visualise trends and to better understand the phenomena that they are studying.

A literature review on self-compacting concrete

In the field of civil engineering, a team of researchers from Belgium and the UK gathered data related to self-compacting concrete from over 250 published studies and conference proceedings and organised it in the form of a database. Their purpose was to develop a full assessment of the mechanical properties and behaviour of self-compacting concrete, based on a thorough interpretation of the available scientific information. Research on self-compacting concrete is very important because it is a high-flowing material that spreads easily and sets under its own weight, rather than relying on vibration to set like conventional concrete. It can be used in instances where there is limited space due to tight physical reinforcements. This material is advantageous over traditional concrete as it can be poured faster and requires less labour to operate equipment necessary for the conventional product to cure. As a result, contractors and builders can complete jobs faster and the finished surface is high-quality, captures fine detail and is well suited for precast models.

The scientists gathered insights on more than 1500 self-compacting concrete mixtures. The database built included information directly extracted from the articles considered, but also properties derived from the existing data that were not originally available. Rather than relying on small-scale studies that are subject to local conditions, materials and approaches, this research produced a range of possible material behaviours that could then be compared with models that have been previously developed for conventional concrete. Specific research findings and technical details from the above-mentioned study were published in the State-of-the-Art Report (STAR) by RILEM along with a journal article.

How the interpretation of data can inform construction practices

The main impact of the study is related to the fact that, in the past, it wasn't clear to what extent existing design codes and best practices applicable to "traditional" vibrated concrete would apply to self-compacting concrete. Thanks to this large-scale review and to the insights allowed by data aggregation, it has been clearly shown that there exist measurable differences between the two materials. However, it was concluded that, in most cases, design codes and material mechanical properties normally used for vibrated concrete also apply when using self-compacting concrete. In addition, the researchers showed that, when self-compacting concrete is poured in-situ, it is either equivalent or preferable to vibrated concrete from a mechanical standpoint.

Title	9: Self Compacting Concrete
Subtitle	Advancements in engineering led by data aggregation
Abstract	Data from over 250 academic sources was aggregated in the form of a database and used to inform future design of self-compacting concrete. This large-scale study allowed researchers to precisely describe the differences between self-compacting concrete and "traditional" vibrated concrete.
Keywords	concrete; database; self-compacting
Research subject area	ENGINEERING
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	250 articles 1500 concrete mixtures analysed
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2014
Organisations involved	University of Antwerp; HUB-KAHO University College; Belgian Building Research Institute; University of Cambridge; Ghent University
Academic citations	Khayat, K. & De Schutter, G. (2014). [Mechanical Properties of Self-Compacting Concrete. State-of-the-Art Report of the RILEM Technical Committee 228-MPS on Mechanical Properties of Self-Compacting Concrete](http://www.springer.com/gp/book/9783319032443). Springer. Craeye, B., Van Itterbeeck, P., Desnerck, P., Boel, V. & De Schutter, G. (2014). [Modulus of elasticity and tensile strength of self-compacting concrete: Survey of experimental data and structural design codes](http://dx.doi.org/10.1016/j.cemconcomp.2014.03.011). Cement and Concrete Composites 54.
Links	http://www.sciencedirect.com/science/article/pii/S0958946514000572 https://www.concretenetwork.com/self-consolidating-concrete/advantages.html

10: Preventing Drug Interactions On Your Mobile

Novel tools for patients and healthcare professionals

In the world, there are 36.7 million people with HIV/AIDS and 130-150 million people suffering from chronic hepatitis C. These people need to take medications such as antiretrovirals or direct antiviral agents which are prone to major interactions with other medications that they may be receiving. Clinicians need to know how to manage drug interactions, otherwise, the effectiveness of treatments may be highly reduced, or the patient could develop drug resistance or other complications. Thus, it is essential to optimise therapies to maintain drug levels in the body that attack the infections effectively without causing undesired side effects. This implies that some combinations of drugs should not be used, while others should be administered carefully, sometimes needing close monitoring and adjustments. Such a scenario is exactly where the HIV and Hep iChart apps can help.

The HIV and HEP iChart apps

Based on data generated at the University of Liverpool, two websites on drug interactions were developed, i.e., HIV Drug Interactions and HEP Drug Interactions. These inspired the iChart apps, holding the same information but available free-of-charge on both Android (HIV, HEP) and iOS (HIV, HEP) smartphones. The apps provide healthcare professionals with essential information on potential drug interactions between the medications they are likely to prescribe and other drugs such as prescribed, over -the-counter, recreational, or herbal medications. The databases are updated on a daily basis using the latest drug information, publications, and meeting data, which are then reflected in the apps. Before the apps were created, no such system existed.

Impact and reach of the iChart apps

The Independent called the HIV iChart app "a small medical miracle" and reported it was one of the most popular in the medical field just days after its launch in 2010. The apps have been downloaded in 128 countries by patients and doctors, with a total of over 17,000 unique downloads.

These figures show the usefulness of the technology, however, the do not prove impact by themselves. The websites and iChart apps developed by the University of Liverpool were endorsed by several national and international organisations. For instance, the National HIV Nurses Association recommends the use of the HIV iChart app on their website and the system was also endorsed by Dr Ian Williams, Chair of the British HIV association.

The main impact of the iChart apps is methodological. Before their creations, clinicians couldn't easily determine drug interactions when starting a new treatment regime, while these can now be promptly checked on a smartphone or tablet. This led to improved patient response and adherence to treatment and, obviously, reduced side effects. In addition, the efficiency of clinicians highly improved. The iChart apps should also be

lauded because they are free. Due to their usefulness, they could be made available for a charge, but instead they are released as a free service to improve dissemination to the widest possible audience. Thus, the work being done at the University of Liverpool is in line not only with the principles of research data management, but also with the higher-level idea of open science.

Title	10: Preventing Drug Interactions On Your Mobile
Subtitle	Novel tools for patients and healthcare professionals
Abstract	The iChart apps developed by the University of Liverpool help clinicians and patients with HIV or hepatitis C better deal with drug interactions. The apps allow clear and ubiquitous access to research data that has been arranged for maximum effectiveness and dissemination, thus, improving patient response and reducing the side effects experienced. In addition, clinicians can save time, as all the information they need is now available directly on their smartphones.
Keywords	hiv; hepatitis; drug interactions
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Impact on health and wellbeing
Facts and figures	128 countries 17,000 downloads
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2010
Organisations involved	University of Liverpool
Academic citations	
Links	http://www.hiv-druginteractions.org/
	http://www.hep-druginteractions.org/
	https://www.aids.gov/federal-resources/around-the-world/global-aids-overview/
	http://www.who.int/mediacentre/factsheets/fs164/en/
	https://aidsinfo.nih.gov/guidelines/html/1/adult-and-adolescent-arv-guidelines/367/overview
	http://www.independent.co.uk/life-style/health-and-families/features/hiv-app-a-small-medical-miracle-2147820.html
	http://www.nhivna.org/apps.aspx
	http://www.positivenation.co.uk/article.php?section_id=2&category_id=1&article_id=13

11: Meeting Sustainability Objectives

Reducing carbon emissions from agricultural production

It has been estimated that as much as a third of our greenhouse gas emissions are caused by agriculture. These need to be drastically reduced if we want to prevent the Earth from warming by more than 2ºC over the next century. Unfortunately, it is very difficult to quantify emissions from agricultural production, because these change widely based on geographic factors and differences in practice between land users. While many organisations are committed to sustainable practices, in reality it is often hard to achieve the desired targets. In the field of agriculture, a lack of tools to calculate on-farm emissions limits our understanding of how existing circumstances could be improved. The Cool Farm Tool was created to address this issue, by gathering the best available scientific knowledge into an open source greenhouse gas calculator. A collaboration between the University of Aberdeen, Unilever, and the Sustainable Food Lab, the tool can be accessed by both farmers and the general public.

The Cool Farm Tool

The calculation of gas emissions is based on empirical research from a broad range of published data sets and Intergovernmental Panel on Climate Change (IPCC) methods. The estimates of carbon emissions are based on:

- an empirical model built from an analysis of over 800 global datasets for N2O
- 100 global datasets for soil carbon sequestrations
- Peer-reviewed industry data for embedded fertiliser production.

This approach, which considers region and technology-specific emission factors, is possible thanks to the inputs from a variety of academic and industrial collaborators.

When using the Cool Farm Tool, users can see what are the effects of changes in their practices on the greenhouse emissions of their farms. The tool is suitable for all sizes of farms, from an individual producer to multinational. The Cool Farm Tool has been widely accepted by its users, mostly thanks to the endorsement received from commercial players.

One of the most important features of the greenhouse gas calculator is that it doesn't require sophisticated inputs such as those typically used by academic researchers. Instead, it was designed to work with data that is typically available to farmers, so as to promote its use at all levels.

Towards the reduction of the carbon footprint of farms

The Cool Farm Tool has been piloted in the Cool Farming Options project, where sponsoring companies used the tool to improve one or more of the farms in their supply chains. This allowed a collaboration with farmers, which led to the implementation of new and better features. The impact and reach of the tool are proven by its widespread use. Examples of use cases of the tool by major commercial players are as follows:

- PepsiCo decided to use the Cool Farm Tool in their 50-in-5 initiative (reducing carbon emissions and water use 50% in 5 years).
- Marks and Spencer and WWF-India used the tool to study the greenhouse gas

production related to Indian cotton within and outside the Better Cotton initiative.

- Costco assessed greenhouse gas emission from egg production in the USA, studying the relationship between geography and management practice. This led to suggestions being formulated to reduce emissions in Costco's farms.

Title	11: Meeting Sustainability Objectives
Subtitle	Reducing carbon emissions from agricultural production
Abstract	Meeting sustainability objectives is becoming increasingly important to reduce the impact of global warming. The field of agriculture has been deemed responsible for a third of our greenhouse emissions, thus, resources such as the Cool Farm Tool are essential to help people in the sector understand how they can reduce their environmental impact.
Keywords	agriculture; carbon footprint;
Research subject area	ENVIRONMENTAL SCIENCES
Type of RDM impact/benefit	Methodological impact (e.g., new approaches developed); Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Environmental impact
Facts and figures	800 global datasets for N_2O 100 global datasets for soil carbon sequestrations
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2011
Organisations involved	University of Aberdeen; Unilever; Sustainable Food Lab
Academic citations	
Links	https://coolfarmtool.org
	http://www.bbc.com/news/science-environment-36315952
	http://www.ncub.co.uk/sor14/aberdeen-efarming.html
	http://www.adas.uk/News/the-pepsico-50-in-5-initiative-halving-greenhouse-gas-emissions-from-potato-production
	http://bettercotton.org/
	http://www.abdn.ac.uk/news/4359/

12: Finding Offshore Hydrocarbons

Leveraging satellite data to improve the efficiency of exploration

The improved gravity data elaborated by researchers at the University of Leeds and Getech has a far-reaching impact. It has been used for over 50 exploration projects per year and has been valued $2.5 million per project by Shell. In the field of offshore hydrocarbon exploration, datasets are used extensively to minimise risks for both people and the environment. In particular, gravity data can help companies map geological structures under the seabed and can guide decision-making prior to the use of more expensive physical approaches to exploration such as controlled source electro-magnetic methods or 3D seismic surveys. Even though the use of marine gravity datasets derived from satellite observations (e.g., GEOSAT and ERS-1) is an established practice, the resolution of the gravity grids calculated from these was too low to be useful in practice.

The research at the University of Leeds and the Getech spin-out company

To solve the problem of low-resolution gravity grids, researchers developed a whole new dataset based on satellite observations (here, in its most updated version). Their work required processing and the creation of new techniques and approaches, including tests of the robustness of the calculations. Academic staff and a spin-out company of the University of Leeds called Getech carried out the analysis and quantified the improvement in the original satellite data as more than 10%. An accurate analysis of the benefits of the newly-created dataset has been performed for India and the researchers' work is described as a major improvement compared with existing satellite-derived datasets.

Impact of the improved gravity data

The use of better gravity data has a direct impact on the effectiveness and efficiency of marine exploration. Oil companies used the improved information to inform their work, including high-profile players such as Eni, Shell, Statoil, and Total. Getech also licensed the data to most of the world's leading oil companies and further developed their analysis to build a new dataset called Trident and released in 2008. Trident data is another example of successful research data management, as it builds on the work of three independent authors (Getech, the Danish National Space Centre, and Sandwell & Smith).

Trident data reportedly led to a 10% improvement in the reliability of marine gravity data for prospecting and Getech customers stated that a total of more than 50 exploration projects per year use the dataset. The research also decreases the risks for employees, as the first part of exploration projects can be performed digitally and does not involve going on-site.

Finally, the dataset and the research performed brought significant advantages to Getech, which received more than £1.2 million between 2002 and 2004 from oil companies wishing to use the marine gravity dataset. This contributed to enabling the company to be floated in 2005.

Title	12: Finding Offshore Hydrocarbons
Subtitle	Leveraging satellite data to improve the efficiency of exploration
Abstract	Satellite data has been re-used to produced a more up-to-date and precise dataset helping with offshore exploration. The improved gravity data prepared by the researchers has shown a very high potential and is estimated to be 10% more accurate than previously-available work. The dataset has been used by major oil companies to drive decision-making and improve the safety of their exploration efforts.
Keywords	oil; exploration; gravity
Research subject area	EARTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Impact on economy and business; Environmental impact
Facts and figures	$2.5 million per project £1.2 million received 10% improvement compared to previous data
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1998
Organisations involved	University of Leeds; Getech
Academic citations	Maus, S., Green, C.M. & Fairhead, J.D. (1998). [Improved ocean-geoid resolution from retracked ERS-1 satellite altimeter waveforms](http://dx.doi.org/10.1046/j.1365-246x.1998.00552.x). _Geophysical Journal International,_ 134, 243-253. Fairhead, J.D., Green, C.M. & Odegard, M.E. (2001). [Satellite-derived gravity having an impact on marine exploration](http://dx.doi.org/10.1190/1.1487298). _The Leading Edge,_ 20, 873-876. Bansal, A.R., Fairhead, J.D., Green, C.M. & Fletcher, K.M.U. (2005). [Revised gravity for offshore India and the isostatic compensation of submarine features](http://dx.doi.org/10.1016/j.tecto.2005.03.017). _Tectonophysics,_ 404,1- 22.
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=6325
	http://www.getech.com/
	https://directory.eoportal.org/web/eoportal/satellite-missions/g/geosat
	https://directory.eoportal.org/web/eoportal/satellite-missions/e/ers-1
	http://www.ipgroupplc.com/media/ip-group-news/2005/2005-09-23

13: Geological Data Made Easy

The British Geological Survey and the OpenGeoscience portal

More than 20 different ways to re-use data have been found by the users of the British Geological Survey (BGS). Sponsored by the Natural Environment Research Council, the BGS is a world-leading geological survey and provides geoscientific data. The information held in the BGS database helps society use resources responsibly, deal with environmental change, and protect themselves from environmental hazards. The team of researchers behind the BGS cover a wide range of topics and the data they produce is available either for free or subject to the payment of a license fee. OpenGeoscience is BGS's open science initiative to increase the value of their work by allowing data re-use for the general public.

The OpenGeoscience initiative

OpenGeoscience can be described as an open access web service built by the BGS to provide users with access to a large number of geological datasets. These can be combined with other information from different sources to better understand the world. OpenGeoscience allows access to:

- the Geology of Britain map
- over a million borehole scans
- photos from the GeoScenic geological photo archive
- published paper maps and sections from 1832 to 2014 and publications from 1835 to the present
- software
- web services
- educational resources.

This wealth of information is available at no cost under an Open Government License, so as to maximise its impact and dissemination.

OpenGeoscience mash-ups

The impact of OpenGeoscience is difficult to quantify, as the data can be accessed freely with no registration. However, the website received coverage on the BBC website and the project was widely acclaimed by research institutions thanks to the data it offers. OpenGeoscience has been acknowledged by Michael Jones, CTO of Google Earth and Jack Dangermond, President of ESRI.

The BGS highly encourage the re-use of OpenGeoscience data and host a collection of mash-ups, i.e., examples of how BGS datasets were combined with others to create new interfaces or functionalities. These are produced by developers, research partners, and the BSG itself. Among the list of mash-ups using OpenGeoscience data, we highlight the following:

- The UK Onshore Geophysical Library: The platform releases 2D and 3D seismic data recorded in the UK.
- The UK Soil Observatory: The website is a gateway to find large-scale soil property datasets from NERC research centres.
- The EarthObserver app for iOS: The app was developed by Columbia University and allows users to "visit frozen icecaps, study geological maps, scout mountains to climb and trips on coastal waters and exploit a rich atlas of other earth and environmental imagery."
- The Earthquake Lite app for iOS and Android: The app gives live information on earthquakes, drawing from, among

others, the US Geological Survey, Geonet, and the BGS.

- Teaching resources: These include material on the Earth Learning Idea and ESRI ArcWatch websites.

Title	13: Geological Data Made Easy
Subtitle	The British Geological Survey and the OpenGeoscience portal
Abstract	Geological data was released by the British Geological Survey to align their resources to the principles of open science. Their efforts took the form of the OpenGeoscience portal, where data is shared through an Open Government License. Users of OpenGeoscience resources are encouraged to reshape the data to develop new products, called mash-ups, and more than 20 of these are available on the BGS website.
Keywords	geology; survey; NERC
Research subject area	EARTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Impact on economy and business; Environmental impact
Facts and figures	Over 20 mash-ups (projects derived from OpenGeoscience data)
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2010
Organisations involved	British Geological Survey
Academic citations	
Links	http://www.bgs.ac.uk/opengeoscience/
	http://www.bgs.ac.uk/data/services/mash-ups/
	http://mapapps.bgs.ac.uk/geologyofbritain/home.html
	http://www.bgs.ac.uk/data/boreholescans/home.html
	http://geoscenic.bgs.ac.uk/asset-bank/action/viewHome
	http://www.bgs.ac.uk/data/maps
	http://www.bgs.ac.uk/data/publications
	http://inspire.ec.europa.eu/events/conferences/inspire_2010/presentations/265_pdf_presentation.pdf
	https://ukogl.org.uk/
	http://www.ukso.org/
	https://itunes.apple.com/us/app/earthobserver/id405514799?mt=8
	https://play.google.com/store/apps/details?id=com.mobeezio.android.earthquakelite
	https://itunes.apple.com/gb/app/earthquake-lite/id372888894?mt=8
	http://www.earthlearningidea.com/PDF/129_Opengeoscience_1.pdf

14: Protecting The Oceans By Coordinating Data Sharing Efforts

UNESCO's efforts to preserve marine environments

Over 245,000 km^2 of our oceans are classified as dead zones, where excessive nutrients resulting from human activities have made marine life impossible. To put things in perspective, this equates to more than the surface of the United Kingdom. It is, thus, unsurprising that UNESCO established the Intergovernmental Oceanographic Commission (IOC) to promote international cooperation, coordinate new developments on the nature and resources of our oceans and coastal areas, and ensure both sustainable development and the protection of the marine environment. In 1961, the IOC started the International Oceanographic Data and Information Exchange (IODE) to collect, curate, and archive millions of ocean observations and make them available to member states.

The IODE programme and the UK's involvement

The IODE programme aims to facilitate and promote the discovery, sharing, and access to marine data. The programme encourages archival, preservation, and documentation of data, considering best practices and developing new ones where necessary. In addition, capacity building initiatives are promoted to help all member states get up to speed with the required knowledge. The IODE network of data and information centres (called National Oceanographic Data Centres, or NODCs) sponsor a number of data-focussed services in all member states and, just in the UK, 13 of these are available. Some examples include:

- GEBCO, the General Bathymetric Chart of the Oceans, which provides the most authoritative publicly-available bathymetry data for the world's oceans
- UK Argo, which shows data gathered by free-drifting robotic floats
- The Cruise Inventory, which hosts data on more than 15,000 UK research vessel activities.
- The NERC vocabulary server, which provides lists of controlled terms to describe marine data

The UK contributes to the IOC IODE programme via the British Oceanographic Data Centre (BODC) and the United Kingdom Hydrographic Office (UKHO), which help with its implementation and strategic development. UK representation at the IOC is devolved to the Natural Environmental Research Council (NERC).

Impact of sharing data on marine environments

Being coordination bodies, the IOC and IODE can have a far-reaching impact. The first observation that can be made is that no single country in the world could ever observe and analyse oceans alone. This means that the main impact of IODE is methodological, as their approach to data management allows a group of stakeholders to tackle a complex series of problems they could not deal with separately. Such a way to coordinate data sharing efforts can offer insights to researchers, private companies, policymakers, and the general public. Examples of scenarios

where sharing the data held by IODE NODCs helps are:

- Studying meteorology and coastal defence: Having real-time or near-real-time data on the weather conditions above the oceans can avoid damages due to natural disasters and mitigate the risks for people living along the coastlines.
- Scheduling shipping: The safety of distributing goods via ship is highly affected by tides, storms, and currents. Leveraging data to perform simulations greatly helps companies avoid shipping disasters due to, e.g., rogue waves.
- Managing living and non-living resources: Data collected by researchers in member states is used to advise the EU Commissioner for Fisheries about how much of each species can be captured in a year, which leads to the negotiation of the so-called Total Allowable Catches (TACs). In addition, data sources on the availability of materials such as sand, gravel, oil, gas, etc., are used to manage their future use.

Title	14: Protecting The Oceans By Coordinating Data Sharing Efforts
Subtitle	UNESCO's efforts to preserve marine environments
Abstract	UNESCO's efforts to protect marine environments materialised with the creation of IODE in 1961. The programme aims to improve data management in the field and to guide and coordinate the data gathering work by a large number of countries. Such a high-level initiative allows data to be shared very effectively, as no country could possibly gather so much information on its own. In addition,
Keywords	marine; ocean; coordination; unesco
Research subject area	EARTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Environmental impact; Impact on politics and governance; Impact on economy and business
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1960
Organisations involved	UNESCO
Academic citations	
Links	http://www.unesco.org/new/en/natural-sciences/ioc-oceans/focus-areas/rio-20-ocean/blueprint-for-the-future-we-want/marine-pollution/facts-and-figures-on-marine-pollution/
	http://www.ioc-unesco.org/
	http://www.iode.org/
	http://www.oceandataportal.net/portal/portal/odp2/home
	http://www.iode.org/index.php?option=com_content&view=article&id=465&Itemid=100201#United Kingdom

	http://www.uk-ioc.org/IODE
	https://darchive.mblwhoilibrary.org/bitstream/handle/1912/3927/23-3_glover.pdf?sequence=1&isAllowed=y

15: Corruption in Public Sector Procurement

Digital whistleblowing to quantify the cost of corruption

Corruption is one of the biggest challenges facing European society. Even though corruption takes different forms in each country, it harms the EU as a whole, as it lowers investments levels, affects public finances, and perturbates the fair operations of the market. Corruption can be political, criminal, private, or in the form of abuse of power, and four out of five EU citizens consider it as a major problem in their state.

The DIGIWHIST project

DIGIWHIST, an EU Horizon 2020-funded project, is a collaboration between six European research institutes, with the University of Cambridge representing the UK. The project aims to "increase trust in governments and improve the efficiency of public spending across Europe". While this is an extremely complex endeavour, data collection and management make it feasible. DIGIWHIST collects and analyses information on public procurement and accountability of public officials within the EU and in neighbouring countries. This information consists, e.g., of public procurement transactions and the ownership structures of successful firms, along with aggregated asset and income declaration data. These can be elaborated with the purpose of identifying potential conflicts of interest and systemic vulnerabilities in the 35 countries in scope for the project.

The DIGIWHIST project aimed to help governments be more transparent by releasing curated datasets using the above information. These are found on different websites, including the Index of Public Integrity and the EuroPAM portal. The former offers a graphical representation of a society's capacity to control corruption and ensure that public resources are spent without corrupt practices. On the other hand, the latter contains Public Accountability Mechanisms indicators for financial disclosure, conflict of interest restrictions, and freedom of information, along with public procurement data. Being based on a variety of international sources, the DIGIWHIST project needs data collection algorithms to update the available information. This shows the importance of data compatibility and interoperability between the different sources of information, which will affect the maintenance of the project beyond its lifetime.

The consequences of blowing a digital whistle

DIGIWHIST data was used by the European Parliament to conduct a study of the price impacts of corruption, with findings that suggest corruption costs Europe between €179 billion and €990 billion per year (in GDP terms). The study was presented via an EC press release and broadly covered by media outlets, including Politico and the BBC.

The study indicates that corruption in the procurement process hurts the economy, as it deliberately removes companies from competition and rigs procurement timelines in favour of preferred vendors. Companies are aware of these practices, and, as such, do not spend time bidding on public sector contracts, which results in automatic contract awards to select companies.

Nations where corruption is most prevalent include Poland, Romania, Lithuania, Cyprus and Croatia. Notably, these countries receive considerable sums of funding from the European Union. For example, Poland received €1.4 billion in 2013 and carried one of the highest risks of procurement-related corruption in the European Union. With this data, more comprehensive strategies can be developed to address corruption within the European Union, potentially leading to significant cost savings and reinvestment of funding into other goods and services.

Title	0 Corruption in Public Sector Procurement
Subtitle	Digital whistleblowing to quantify the cost of corruption
Abstract	The DIGIWHIST project gathered and elaborated information on public procurement and accountability of public officials within the EU and in neighbouring countries. This was picked up by the European Commission, which released a study showing how corruption may cost Europe up to €990 billion per year.
Keywords	corruption; european commission; transparency; public sector
Research subject area	STUDIES IN HUMAN SOCIETY
Type of RDM impact/benefit	Socio/Economic impact; Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	Impact on politics and governance; Impact on public services; Impact on economy and business
Facts and figures	up to €990 billion lost to corruption yearly
Original dataset from which the impact arose	http://digiwhist.eu/resources/data/
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2015
Organisations involved	University of Cambridge; Hertie School of Governance Ggmbh; Akki Atlathato Kormanyzas Kutatointezet Kft; Datlab Sro; Open Knowledge Foundation Deutschland; Universita' Cattolica del Sacro Cuore
Academic citations	
Links	http://cordis.europa.eu/project/rcn/194398_en.html
	http://digiwhist.eu/about-digiwhist/
	http://integrity-index.org/
	http://europam.eu/?module=about
	http://www.rand.org/pubs/research_reports/RR1483.html
	http://europa.eu/rapid/press-release_IP-14-86_en.htm
	http://www.politico.eu/article/corruption-costs-eu-990-billion-year-rand-study-fraud-funding/
	http://www.bbc.co.uk/news/world-europe-26014387
	http://ec.europa.eu/home-affairs/what-we-do/policies/organized-crime-and-human-trafficking/corruption_en

16: Drones for Research

Using unmanned aerial vehicles to power new approaches to scientific investigations

Almost 3 million drones are expected to be produced in 2017. These include commercial and personal unmanned aerial vehicles (UAVs or, more commonly, drones), i.e., devices that are capable of flying and performing various tasks, including taking photos and videos or transporting goods. As such, drones are potentially dangerous and there exists a set of rules on how to safely operate them. Nonetheless, considering their widespread presence in the market, it is unsurprising that they caught the interest of academic researchers, especially thanks to their unique photographing features. Being able to fly, drones can take accurate images of virtually anything on Earth, which opens the doors to applications in many research fields.

Managing drone data

Thanks to the ease of capturing images at a relatively low cost, drones have allowed a novel approach to remote sensing. However, the data captured is not useful unless it is properly managed, as having a cheap way to gather images can easily lead to an unstructured mass of information with little use. Some research fields where drone data is helpful are, e.g., agriculture, geography, and archaeology. In the case of these applications, the single images captured by drones are usually stitched to create mosaics describing areas and structures in either two or three dimensions. These larger images are normally georeferenced and can be overlaid onto GIS maps for an improved analysis and sharing. This all means that, before even starting to use information captured by drones, this needs to go through a complex data management and curation process, which involves advanced software, metadata standards, and cross-disciplinary expertise.

The benefits of drones for agricultural research and the environment

Images captured by drones have allowed critical advancements in the monitoring of crops. This is usually done by walking between plants and by assessing their conditions visually, however, some farms are so large that this is not practical or ineffective. Researchers at Utah State University take photos of fields using drones and combine them with data collected from the ground to assess the health of plants. The drones used include visual, near-infrared, and thermal cameras, to allow the researchers to build accurate snapshots of the fields. The images are also used to spot the exact location of dead or diseased plants, to evaluate irrigation needs, and to monitor canopy volume for pruning. The data gathered by drones is also being evaluated as a tool to detect plant diseases before outbreaks in a $1.74 million research programme at Kansas State University.

In the UK, curated drone image datasets have been used to study floods, a common occurrence in the country. The data gathered has been enriched with dates, so that the mosaics built by combining the drone images can be observed over time. This allowed researchers from the University of Salford to formulate conclusions and suggestions on how to best manage and control the overflow of water on farmed land and human structures such as railways.

Title	16: Drones for Research
Subtitle	Using unmanned aerial vehicles to power new approaches to scientific investigations
Abstract	Drones are becoming a constant presence in technology news and media, thus, it is not surprising that they also caught the attention of the research community. Drones allow researchers to capture aerial images easily and at a low cost, however, the data they gather needs to be properly curated to allow any applications. In the field of agriculture, drone-captured datasets are being used to spot plant diseases and help farmers better protect their yields. In addition, drone time-stamped drone image sets have been used to study how to best protect crops and land from floods.
Keywords	agriculture; drones
Research subject area	AGRICULTURAL AND VETERINARY SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Impact on economy and business; Economic; Environmental
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2016
Organisations involved	Utah State University; Kansas State University; University of Salford
Academic citations	
Links	http://www.gartner.com/newsroom/id/3602317
	http://dronesafe.uk/wp-content/uploads/2016/11/Dronecode.pdf
	https://www.exeter.ac.uk/media/universityofexeter/esi/pdfs/ESI_Concertina_12pp_version_6-FINAL.pdf
	http://proceedings.esri.com/library/userconf/proc15/tech-workshops/tw_395-290.pdf
	http://www.giscloud.com/blog/the-challenges-of-drone-mapping/
	https://www.rd-alliance.org/blogs/drones-emerging-scientific-tools-trade.html
	http://utahpests.usu.edu/files/uploads/UtahPests-Newsletter-fall16.pdf
	http://www.k-state.edu/media/newsreleases/mar15/uasinsect31815.html
	http://www.bbc.com/news/science-environment-35353869

17: Data-Enhanced Archaeology

Using modern tools to study ancient times

Archaeology is probably one of the fields where research data management can yield the most benefits. Why, you ask? Simply because creating archaeological data implies the destruction of primary evidence. Archaeologists have long been working with no strict research data management rules, however, it has become clear that a more structured approach needs to be pursued to uncover and share the wealth of fieldwork information that is still unpublished. Furthermore, even published archaeological data is often difficult to access, which makes the need for a centralised approach even more urgent. The Archaeology Data Service (ADS) aims to fill the gap, by liaising with national and local authorities and research funders to negotiate the deposition of data from both fieldwork and desk-based work. This includes a wide range of data types, such as reports, databases, images, maps, plans, numerical datasets, surveys, and drawings.

The Archaeology Data Service

The Council for British Archaeology and eight UK universities founded the ADS to collect, describe, catalogue, preserve, and provide user support on digital resources arising from archaeological research. Based at the University of York, the ADS also promote standards and guidelines for best practices related to the use of archaeological data. Data deposit on the ADS is mandated by the Arts and Humanities Research Council (AHRC) and the National Environment Research Council (NERC) and recommended by Nature Scientific Data and the British Academy. To date, the service includes over 1.3 million metadata records on all parts of the UK, covering historical periods from Early Prehistory to Modern times. Not all users of the ADS are academics and they include, among others, central and local government, consultants, private research organisations, businesses, and charities.

The ADS are involved in a number of archaeological data preservation initiatives, including very recent ones such as E-RIHS and ArchAIDE, which keep the organisation at the forefront of digital preservation and data sharing in the field.

Impact of the ADS

The impact of the ADS was estimated in 2013, as the organisation sought to measure the value of the services it provides. It was found that the economic benefits of the ADS exceed its operational costs and users reported a significant increase in research efficiency as a result of using deposited data. The increase in efficiency was valued as at least £13 million per annum and, today, may now be even higher due to the larger number of deposited datasets. In addition, based on the study, the ADS exhibits a 2-fold to 8-fold return on investment and 44% of respondents stated that they could not have obtained the data they downloaded in any other way. This strongly suggests that, in this specific field, research data management is essential due to the difficulty of retrieving information (or reproducing it, if possible).

The ADS has also had methodological impact on data management in the field and helped the Netherlands, Sweden, Germany, the US, Canada and Australia to establish similar facilities (e.g., DANS, in the Netherlands, Digital Antiquity, in the USA, and FAIMS, in Australia).

Title	17: Data-Enhanced Archaeology
Subtitle	Using modern tools to study ancient times
Abstract	In the field of archaeology, data is scarce and difficult to find. This is simply because it usually comes from excavations or physical operations on artefacts, which are normally expensive and can be performed only in certain conditions. The Archaeology Data Service (ADS) aims to fill the gap by freely providing more than 1.3 million metadata records on archaeological data and by driving developments in research data management in the field. The service enables increased efficiency thanks to data reuse and this is valued at at least £13 million per annum.
Keywords	archaeological data; repository
Research subject area	HISTORY AND ARCHAEOLOGY
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	£13 million per annum savings due to increased efficiency 2-fold to 8-fold return on investment 44% of interviewed stakeholders could not have carried out their work without the ADS
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1996
Organisations involved	University of Birmingham; University of Bradford; University of Glasgow; University of Kent at Canterbury; University of Leicester; University of Newcastle; University of Oxford; University of York
Academic citations	
Links	http://archaeologydataservice.ac.uk/about/background http://archaeologydataservice.ac.uk/about/funders http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=43451 http://archaeologydataservice.ac.uk/research http://repository.jisc.ac.uk/5509/1/ADSReport_final.pdf

18: Supporting Science and Industry by Sharing Computer Code

Sharing software as a form of research data

The most well-known operating systems, Windows and Mac OS X, are made of 50 million and 86 million lines of code, respectively. The algorithms a researcher may produce during his or her career usually pale in comparison, however, they are still very valuable. This is because, in many fields, researchers need to analyse their digital data to turn it into something that can be published and this most often requires some programming. Sharing code developed during one's research efforts is, unfortunately, not a widespread practice, even though there are a wealth of data repositories and journals where it can be deposited. This is due to reasons such as perfectionism or the fear that users will want support and bug fixes, for which researchers don't have time. However, recent development are showing how these objections can be overcome and, even better, how code sharing can lead to impact in both academia and industry.

The SPM software at University College London

The Statistical Parametric Mapping (SPM) software package is a standard in the field of imaging neuroscience. It is used to analyse brain imaging data and it originated in 1991. The original version of the SPM software was shared with the then nascent brain imaging community and really bloomed in 1994 when the Wellcome Trust Centre for Neuroimaging was opened at University College London. The computer code was developed and extended over the years to allow the analysis of data from new imaging equipment. In its current form, the SPM software can be used to analyse data from Magnetic resonance imaging (MRI), Magnetoencephalography (MEG), or Electroencephalography (EEG) scans and includes the application of dynamic causal modelling, enabling researchers to study how brain regions interact. The software is released under a General Public License (GPL), which allows users to freely run it, share it, and modify it in the spirit of open science.

Benefits of sharing the SPM code

The first impact of the SPM software is methodological, as nothing else could perform the same tasks before it was released to the public. The code supported the birth of an entirely new scientific field, imaging neuroscience, in which academic neuroscientists, healthcare professionals, companies, and neuroimaging consultants are also involved. Thanks to its novelty, the SPM software is currently the most widespread package for brain imaging analysis, used in 64% of studies (compared to about 14% in the case of its closest competitors).

Companies such as Imagilys, Siemes, and Brain Innovation BV were able to develop their own software products thanks to the SPM code. In particular, the BrainVoyager QX software by Brain Innovation BV is being used by about 2,000 scientists and clinicians and a single license retails for €5,000, showing the value that the SMP software added to the private sector.

In addition, SPM has been widely used in global drug research by large companies such as GSK

and AstraZeneca thanks to its GPL licensing conditions. Other companies benefiting from the code and using it on a regular basis include Imanova Ltd, Eli Lilly, and Mango Solutions.

Title	18: Supporting Science and Industry by Sharing Computer Code
Subtitle	Sharing software as a form of research data
Abstract	Sharing software is not as common as sharing other types of research data, however, it is sometimes very impactful. The Statistical Parametric Mapping (SPM) software in the field of neuroimaging is an example of how sharing computer code can lead to far-reaching effects. In this case, making the code public allowed the creation of a new field of study and led the software to become the leader in the sector. In addition, thanks to the licensing chosen, some companies were able to exploit the code to create derivative products, which are generating substantial income, while major pharmaceutical companies are using it in the field of drug research.
Keywords	software; code; algorithms
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Economic; Technological; Health
Facts and figures	64% of users in the field use the software €5,000 for each license (software derived from sharing)
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1991
Organisations involved	University College London
Academic citations	Carp, J. (2012). [The secret lives of experiments: Methods reporting in the fMRI literature] (http://doi.org/10.1016/j.neuroimage.2012.07.004). NeuroImage.
Links	https://en.wikipedia.org/wiki/Source_lines_of_code
	https://www.nature.com/news/2010/101013/full/467753a.html
	https://www.software.ac.uk/which-journals-should-i-publish-my-software
	http://www.fil.ion.ucl.ac.uk/spm/
	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=44229
	http://www.fil.ion.ucl.ac.uk/

19: Citizen Science at Zooniverse

Sharing data and efforts via an online platform

It is estimated that, every day, we create 2.5 quintillion bytes of data. That corresponds to about 2.5 million 1TB hard drives filled daily! It is very difficult to store all this data, let alone analyse it. In some cases, it is possible to deal with large amounts of information using computer algorithms, e.g., artificial intelligence and machine learning. However, this is not always the best option. Sometimes, data needs to be looked at and analysed by humans, as the use of algorithms is not always suitable when considering some types of nuanced or complex datasets.

The Zooniverse initiative

The Zooniverse website hosts a collection of web-based citizen science projects that use the efforts of volunteers to help researchers deal with the flood of data that confronts them. In some fields, researchers recognise that the "human ability for pattern recognition—and our ability to be surprised - makes us superior" to computerised approaches to data analysis and prefer to recruit human volunteers rather than writing algorithms. Researchers are able to create projects through a Project Builder and then share their data with the general public. Over 1.5 million registered volunteers collaborate to power the largest and most successful crowd-sourced research effort, enabling excellent science that wouldn't otherwise be possible.

To date, the Zooniverse platform hosts as many as 58 projects, dealing with arts, biology, climate, history, language, literature, medicine, nature, physics, social science, and space. Zooniverse projects constitute an excellent example of research data management and data sharing and rely on the concept of wisdom of crowds to produce reliable and accurate data. Contributors can discuss their findings on Zooniverse discussion boards, which allow in-depth collaborative analysis.

Why crowd-source research?

Involving the general public in academic research is, obviously, no easy feat. People who have never done research won't be familiar with methods, rules, or specific features of each field. However, in some cases, their participation delivers impressive advantages, such as:

- The ability to deal with extremely large datasets: In its first six months, Galaxy Zoo (one of the space Zooniverse projects) provided the same number of classifications as would a graduate student working round the clock for 3.5 years
- The opportunity to calculate quantitative estimates of errors: This is a necessary step in the scientific process, as the accuracy and reliability of research findings have to be quantified in most field.
- People, as opposed to algorithms, can spot details and features that a computerised approach would likely miss.
- Crowd-sourced research allows people to get in touch with science, thus, making it a powerful educational tool where the user not only learns but also contributes to discoveries.

If these rather qualitative advantages were not enough, let us quantify the impact of the Zooniverse project: Overall, more than 130 articles have been published starting from 2008

and dealing with a number of topics. None of this would have been possible without good research data management!

Title	19: Citizen Science at Zooniverse
Subtitle	Sharing data and efforts via an online platform
Abstract	It is not always easy to deal with large amounts of data. At times, algorithms can help researchers make sense of their large datasets, however, sometimes the human mind cannot be replaced. In these cases, platforms like Zooniverse come to the researchers' help, allowing them to have citizen volunteers analyse scientific data and enable new scientific discoveries. More than 130 articles were published thanks to citizen science, showing how wise research data management can allow the crowd-sourcing of scientific research.
Keywords	crowd-sourced research; citizen science
Research subject area	INFORMATION AND COMPUTING SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	More than 130 articles published Over 1.5 million registered volunteers At least 58 web-based citizen science projects
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2009
Organisations involved	Zooniverse
Academic citations	
Links	https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf https://www.zooniverse.org/about https://en.wikipedia.org/wiki/The_Wisdom_of_Crowds https://www.citizensciencealliance.org/philosophy.html http://galaxyzoo.org/?_ga=1.20284209.840913750.1493028984

20:History and Data

The link between disciplines that seem diametrically opposed

Our planet is more than 4.5 billion years old and the human species evolved to its current form about 200,000 years ago. Civilisation has been around for a tiny fraction of this time, about 6,000 years. Unfortunately, very little is known about ancient times and all we can say is based on remains and artefacts dug up by archaeologists, and examined by experts and researchers.

For more recent times, a greater number of documents, books, and artefacts is available, however, studying them is not always easy. These items are often spread between several libraries, archives, and offices, making the lives of scholars and historians rather difficult. At times, access to some material may even be restricted, due to an item's fragility, value, or importance.

This is where ancient and modern times meet: Thanks to transcriptions and research data management, digitised historical information can finally surface and be accessed by all.

The British History Online digital library

The Institute of Historical Research (University of London) and the History of Parliament Trust founded British History Online (BHO) in 2003. The portal hosts over 1,270 volumes on the history of Britain and Ireland with a focus on the period between 1300 and 1800. These include datasets, guides and calendars, maps, primary sources, and secondary texts from all sorts of subjects, from religion to historical geography.

The material held by the BHO can be accessed by researchers from all over the world and has mostly been digitised through a technique called double re-keying. This approach allows the accurate reproduction of hard texts into a digital form, as transcriptions are performed by two different typists and then compared for maximum accuracy (estimated as greater than 99.995%). In addition, the maps available on the website were scanned at a high resolution and offer an impressive level of detail.

The approach pursued by the BHO portal aims to enable the use of physical collections that are difficult to access, thus, empowering researchers, teachers, and the general public. This solves two problems:

- There is no concern over the fragility and safety of ancient artefacts and printed materials
- Information can be accessed from anywhere, no matter where the originals are held.

Fortunately, the BHO is not alone in its efforts to digitise ancient texts. Other major players in the field include Oxford's Bodleian Libraries, the Vatican Libraries, and the British Library, but also projects such as Look-Here!, which produced a series of case studies on the digitisation of fine arts.

The impact of digitised heritage information

The main impact of the BHO is methodological, as researchers now have a new way to access the material they need. Academic users reportedly use the BHO for several activities, "from resource discovery activity at the beginning of a research project ('finding new works') to the consultation of known works and tasks associated with writing ('checking

references')". Users of the BHO also reported that the digital library changed their research practices, allowing them to discover information more easily and find "unexpected treasures". This initiative is powered by a sound research data management approach, which results in researchers being able to perform more accurate research and perform speculative enquiries "without a great investment of time".

In terms of wider impact, the BHO receives more than 330,000 visitors per month and a Google Scholar query returns 1,410 results. Press coverage of the digital library is extensive for a repository in the field of history, with 9 mentions in news (between January 2017 and January 2017).

Title	20: History and Data
Subtitle	The link between disciplines that seem diametrically opposed
Abstract	Access to ancient books, manuscripts, and artefacts is often limited due to their fragility and importance. In addition, they are spread between several locations, which makes the work of historians difficult. The digitisation of heritage data by the British History Online digital library bridges the gap by making material available to researchers from all over the world. The application of research data management in this field led to changes in the researchers' workflow and earned the library a large number of citations in the academic literature and mentions in the news.
Keywords	British history; digitisation; library
Research subject area	HISTORY AND ARCHAEOLOGY
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	330,000 unique visitors a month 1,410 Google Scholar results 9 mentions in news
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2003
Organisations involved	University of London; History of Parliament Trust
Academic citations	
Links	https://en.wikipedia.org/wiki/Age_of_the_Earth https://www.universetoday.com/38125/how-long-have-humans-been-on-earth/ http://www.bl.uk/reshelp/inrrooms/stp/register/appointment/mssapprove/consultmanuscripts.html http://www.history.ac.uk/

	http://www.british-history.ac.uk/
	http://bav.bodleian.ox.ac.uk/
	http://digi.vatlib.it/
	http://www.openbookpublishers.com/product/283
	https://www.webarchive.org.uk/wayback/archive/20140615015752/http:/www.jisc.ac.uk/media/documents/programmes/digitisation/analysis_bho.pdf
	https://www.webarchive.org.uk/wayback/archive/20140614150536/http:/www.jisc.ac.uk/media/documents/programmes/digitisation/Impact_Synthesis report_FINAL.pdf
	https://scholar.google.co.uk/scholar?q=%22British+History+Online%22&btnG=&hl=en&as_sdt=0%2C5
	https://www.google.co.uk/search?q=%22British+History+Online%22&client=opera&biw=958&bih=913&source=lnt&tbs=cdr%3A1%2Ccd_min%3A1%2F1%2F2016%2Ccd_max%3A1%2F1%2F2017&tbm=nws
	https://vads.ac.uk/lookhere/about.html

21: Saving the Earth from Mankind

How can we preserve Earth and develop sustainably?

It is well-known that biodiversity plays an important role in the survival of our planet as we know it, but it also has critical effects on business and industry. In 2008, the United Nations estimated that between $2 trillion and $4.5 trillion were lost due to deforestation and land degradation alone. The loss of biodiversity has been linked to a range of other risks with estimated severities ranging (in $ terms) "from tens of billions for inland flooding and infectious disease, to many hundreds of billions for food price volatility and chronic disease".

In the past, several events showed the dangers of biodiversity loss and ecosystem degradation, especially when they are summed to the environmental changes induced by technological and societal changes. Clearly, we have to manage risks to the environment by creating progressive approaches to preserve biodiversity on Earth, and these can be effectively supported by data. In particular, managing and sharing information on biodiversity arising from research is key to protecting our planet, as many issues in the field require knowledge from a number of different domains.

The Global Biodiversity Information Facility (GBIF)

In 1999, the Biodiversity Informatics Subgroup of the Megascience Forum, set up by the Organization for Economic Cooperation and Development (OECD), recommended the creation of the GBIF. This decision arose from the need to "make biodiversity data and information accessible worldwide" and the panel argued that such an approach would promote sustainable development by allowing access to sound scientific information.

The GBIF provides access to hundreds of millions of records created by research institutions and is the largest biodiversity database on the internet. Data deposited in the repository features evidence on more than 1.6 million species collected over three centuries of natural history. The datasets consist of both data collected by researchers and automated monitoring programmes.

The GBIF works as a network of nodes, coordinating the biodiversity data facilities of the countries and organisations that take part in the initiative. The GBIF aims to promote collaborative research and the sharing of skills, experiences, and technical capabilities.

What has been achieved by sharing biodiversity data

To begin with, GBIF's data management efforts are pushing researchers to publish their data using common standards, which enables a higher compatibility between datasets. Research that once was impossible is now enabled by the GBIF, and is used to inform decisions on the conservation and sustainable use of resources on Earth.

More than 1,400 peer-reviewed publications by authors from 81 different countries cited the repository as a source of data, and dealt with "the impacts of climate change, the spread of pests and diseases, priority areas for conservation and food security". According to the GBIF's portal, a new paper based on the data is published every

day and the trends on data re-use are steadily increasing.

In addition, GBIF members have set up national websites that use GBIF material and approaches to inform their citizens and policymakers about biodiversity.

The variety of applications for GBIF data is almost endless and a wealth of examples of GBIF-mediated data re-use exist. For more information, please see GBIF's 2016 Science Review and the rich list of re-use case studies on GBIF's portal (ranging from 2013 to 2017). GBIF data is also currently used by journalists and featured in several news articles.

Title	21: Saving the Earth from Mankind
Subtitle	How can we preserve Earth and develop sustainably?
Abstract	Our planet needs to be protected, as human development tends to ignore sustainability and the effect business has on the environment and on biodiversity. Luckily, things are slowly changing, and better decisions and policies supported by research data can now be made. The Global Biodiversity Information Facility (GBIF) plays a critical role in enabling sustainable development by hosting evidence on more than 1.6 million species and its data is featured in more than 1,400 research papers. GBIF data is used by scientists, policymakers, and journalists alike.
Keywords	biodiversity; environment; sustainable development
Research subject area	ENVIRONMENTAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact; Methodological impact (e.g., new approaches developed)
Summary Impact Type	Political; Economic; Environmental
Facts and figures	Evidence on more than 1.6 million species 1,400 peer-reviewed articles citing GBIF data 81 countries
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1999
Organisations involved	Organization for Economic Cooperation and Development
Academic citations	
Links	http://online.wsj.com/ad/article/execdigest-biodiversity https://www.pwc.co.uk/assets/pdf/wef-biodiversity-and-business-risk.pdf http://www.gbif.org/what-is-gbif http://www.gbif.org/resource/82873 https://www.google.co.uk/search?q=%22Global+Biodiversity+Information+Facility%22&gws_rd=ssl#q=%22Global+Biodiversity+Information+Facility%22&tbm=nws

22: Understanding Mobile Users and Evaluating Vulnerabilities

User data as an investigation tool

Within the space of a few years, smartphones and mobile technology have revolutionised the way we live. These technologies keep evolving at a very high pace, which puts manufacturers under constant pressure to innovate. One of the issues they face is that they need information on what people are likely to want, and this is very difficult to estimate. New functions, form factors, or accessories can be effectively hypothesised only when designers can formulate an educated guess on what their prospective clients desire. A possible way to gather insights for new ideas is to build on the ways customers use the products they currently own.

The data-gathering efforts of the Device Analyzer project

Originating at the University of Cambridge, the Device Analyzer project, which includes an Android app, aims to understand what people do with their smartphones. The Device Analyzer project acts as an intermediary between users and industry to provide anonymised datasets on mobile phone usage. These include information such as the features of a phone that are used, frequency of missed calls, number of messages sent, etc. The researchers gather data from the users' phones, then anonymise it while preserving useful information. The curated dataset is periodically updated to the project's server at the University of Cambridge.

At the moment, there are 30,978 contributors to the project, which has led to several scientific publications. This corresponds to over 100 billion records of Android smartphone usage from countries all over the world. Access to the dataset is possible by sending a brief project description to the project team and by signing terms and conditions that protect the participants' privacy.

Impact and reach of the Device Analyzer dataset

To date, the dataset has been shared with over 100 companies, universities, and research institutes. These include high profile commercial stakeholders in the technology market, such as AT&T, Cisco, IBM, Intel, Microsoft, NEC Labs, and Spotify. On the other hand, those who installed the Device Analyzer app on their smartphones are provided with analytical data related to their historical usage. This allows the Device Analyzer team to formulate suggestions about the best data plans based on internet usage and recommendations for new app downloads based on user trends and preferences over time.

Results from the Device Analyzer project were also mentioned by several mainstream technology websites including, among others, Ars Technica, engadget, Gizmodo, and The Register. Among the most important research findings disseminated by mass media is the fact that "on average 87.7% of Android devices are exposed to at least one of 11 known critical vulnerabilities", which need to be patched through updates by smartphone manufacturers. Thanks to Device Analyzer data, the University of Cambridge also launched the Android Vulnerabilities website, which shows what

networks, devices, and manufacturers are safest through the so-called FUM score.

Title	22: Understanding Mobile Users and Evaluating Vulnerabilities
Subtitle	User data as an investigation tool
Abstract	The Device Analyzer project gathers data on smartphone usage and curates it for re-use by companies, universities, and research institutes. The project's data led to the development of important statistics on the vulnerability of smartphones using the Android operating system. The researchers found that this is related to the slow pace of system updates and only manufacturers have the tools to address the problem.
Keywords	Android; smartphone; mobile phones
Research subject area	INFORMATION AND COMPUTING SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	
Facts and figures	30,978 contributors; 87.7% of Android devices are exposed to at least one of 11 known critical vulnerabilities
Original dataset from which the impact arose	https://deviceanalyzer.cl.cam.ac.uk/
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2012
Organisations involved	University of Cambridge
Academic citations	Wagner, D., Beresford, A. & Rice, A. (2015). [Security metrics for the Android ecosystem](https://www.cl.cam.ac.uk/~drt24/papers/spsm-scoring.pdf) Thomas, D.R., Beresford, A., Coudray, T., Sutcliffe, T. & Taylor, A. (2015). [The lifetime of Android API vulnerabilities: case study on the JavaScript-to-Java interface](https://www.cl.cam.ac.uk/~drt24/papers/spw15-07-Thomas.pdf) Wagner, D., Rice, A. & Beresford, A. (2013). [Device Analyzer: Understanding smartphone usage](http://www.cl.cam.ac.uk/~acr31/pubs/wagner-understanding.pdf)Wagner, D., Rice, A. & Beresford, A. (2013). [Device Analyzer: Large-scale mobile data collection](http://www.cl.cam.ac.uk/~acr31/pubs/wagner-bigdata.pdf) Ding, N., Wagner, D., Chen,X., Pathak, A., Hu, Y.C. & Rice, A. (2013). [Characterizing and Modeling the Impact of Signal Strength on Smartphone Energy Consumption](http://www.cl.cam.ac.uk/~acr31/pubs/ding-signalstrength.pdf) Wagner, D., Rice, A. & Beresford, A. (2011). [Device Analyser](http://www.cl.cam.ac.uk/~acr31/pubs/wagner-daabstract.pdf)
Links	https://play.google.com/store/apps/details?id=uk.ac.cam.deviceanalyzer
	http://androidvulnerabilities.org/#vulnerabilities
	https://arstechnica.co.uk/security/2015/10/university-of-cambridge-study-finds-87-of-android-devices-are-insecure/
	https://www.engadget.com/2011/06/19/device-analyzer-android-study-wants-to-track-your-every-move-if/

	http://gizmodo.com/study-85-of-android-devices-exposed-to-at-least-one-c-1736677079
	https://www.theregister.co.uk/2015/10/12/android_patching_survey/
	http://androidvulnerabilities.org/

23: Understanding War

Data-powered insights into the motives and consequences of war

In 2016, the world saw four conflicts that, alone, totalled more than 109,000 fatalities. These include the wars in Afghanistan, Iraq, and Syria, plus the Mexican drug war, each causing at least 1,000 deaths each. Unfortunately, there are another 54 ongoing conflicts worldwide, causing between 4 and 5,701 fatalities each during 2016 alone.

It is extremely difficult to understand all the underlying issues and motives that lead to war in a given country. Examples include religion, revenge, ethnic cleansing, bargaining failures, commitment problems, and more, but more often it is a mix of these. For researchers, journalists, and politicians it is essential to have access to accurate and reliable quantitative data on wars and their reasons to try to uncover "the origins of conflict, conflict dynamics, and conflict resolution".

Researchers and political scientists study wars on an ongoing basis. They create variables and indicators that have to be responsibly managed and shared with the public so as to increase everyone's awareness of international conflicts and, potentially, avoid or prevent future ones.

War data repositories

Efforts to disseminate resources in the field of conflict research can be found in both Europe and the USA. Various initiatives such as the European Network for Conflict Research (ENCoRe) and the Correlates of War project aim to promote coding standards, datasets, and integration of resources in the field. They are committed to the principles of replication, documentation, review, and transparency of data collection, which are essential when such a delicate topic is considered.

Normally, data is categorised by year, so that information can be read and downloaded in the right context. Data is usually coded to fit scholarly requirements of comparability and to allow researchers to pick up the same phenomenon across time and space. Resources in this field cover different time periods, however, they all try to "better prepare researchers and policy makers for future conflicts".

Topics included in war data repositories include civil wars, violent protest, riots, state repression, terrorism, and more. By combining the data, scholars can analyse a wide range of risk factors and, hopefully, help reduce the likelihood of armed conflicts.

Sharing conflict information

One of the obvious places where data on wars can be re-used is mass media. Journalists can access conflict data repositories to complement their articles with factual information and add credibility to their writing.

Being based in academia, the above-mentioned platforms offer scope for war data to be picked up for the preparation of papers, conference presentations, and seminars. As of April 2017, the Correlates of War project was mentioned by over 4,000 scientific outputs on Google Scholar, while the Uppsala Conflict Data Program (UCDP at Uppsala University, participant in the ENCoRe project) was cited over 3,800 times.

The UCDP has a particularly interesting approach to sharing data, as the information it holds is

georeferenced and displayed on a map for maximum ease of use. The map shows detailed textual information on conflicts and their actors, along with useful metadata that allows filtering and the discovery of hidden links. This is only possible thanks to a sound approach to research data management.

Title	23: Understanding War
Subtitle	Data-powered insights into the motives and consequences of war
Abstract	In the world, there are currently 58 ongoing conflicts. These cause tens of thousands of fatalities each year and are related to reasons that are obscure, complex, and often difficult to understand. Researchers and political scientists have been trying to uncover the reasons for war for a long time. Today, they can leverage data to explain conflicts, and geotagged datasets can be organised to build visualisations that greatly facilitate the understanding of contexts and actors in a war. The use of research data management is instrumental in helping us fully grasp the reasons for conflicts and, hopefully, preventing future ones.
Keywords	War; conflicts
Research subject area	STUDIES IN HUMAN SOCIETY
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Political; Societal
Facts and figures	4,000 citations: Correlates of War project 3,800 citations: Uppsala Conflict Data Program
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2012
Organisations involved	Uppsala University; ENCoRe; University of California, Davis
Academic citations	
Links	https://en.wikipedia.org/wiki/List_of_ongoing_armed_conflicts
	https://web.stanford.edu/~jacksonm/war-overview.pdf
	http://www.pcr.uu.se/research/ucdp/program_overview/about_ucdp/
	https://www.encore.ethz.ch/stsm.html
	http://www.correlatesofwar.org/
	http://www.pcr.uu.se/digitalAssets/61/c_61976-l_1-k_2016-ucdp-annual-report.pdf
	https://scholar.google.co.uk/scholar?start=20&q=%22the+Correlates+of+War+project%22&hl=en&as_sdt=1,5
	https://scholar.google.co.uk/scholar?q=%22Uppsala+Conflict+Data+Program%22&btnG=&hl=en&as_sdt=1%2C5
	http://ucdp.uu.se/#/exploratory

24: Surfing Gravitational Waves

Understanding gravity and Einstein's General Theory of Relativity

Our solar system includes 8 known planets, 5 dwarf planets, 470 natural satellites, and over 707,000 minor planets. Our galaxy, the Milky Way, is estimated to host more than 100 billion planets. The universe is so immense that our understanding of it is, inevitably, limited. Einstein's General Theory of Relativity helps shed some light on the complex physics happening beyond Earth's atmosphere, however, some of his work is still regarded as unproven predictions.

One of the most complicated ideas in Einstein's theory is that of gravitational waves. These are defined as ripples in the fabric of space and time and can be noticed when very massive objects such as super-heavy neutron stars or black holes rotate around each other. Detecting gravitational waves is very difficult, because they can only be measured using extremely large experimental setups and complex calculations.

Data from the Laser Interferometer Gravitational Wave Observatory (LIGO)

LIGO is an experimental setup created to study gravitational-wave astrophysics. It consists of multi-kilometre-scale detectors that use laser interferometry to measure ripples in space time caused by passing gravitational waves. It was designed and built by LIGO Laboratory's team of scientists, engineers, and staff at the California Institute of Technology (Caltech) and the Massachusetts Institute of Technology (MIT), and collaborators from the over 80 scientific institutions world-wide that are members of the LIGO Scientific Collaboration.

The LIGO Scientific Collaboration (LSC) is a group of scientists studying the fundamental physics of gravity. They collaborate to develop and refine data characterization techniques, data quality vetoes, and extensive analysis pipelines that take the raw data and produce astrophysical results. The LCS is an open collaboration, meaning that all interested scientists can contribute to the project.

Such a large-scale initiative comes with a few obstacles, namely in terms of coordination and management of data. Results arising from LIGO experiments are shared through the LIGO Open Science Center. They need to be accurately managed and appropriate software for analysis maintained, so as to allow smooth access to physicists all around the world. In addition, the interpretation of LIGO data requires information from seismic activity and environmental factors, which must be curated and contextualised.

Impact of LIGO data

The management and use of data in the domain of physics can sound complex and abstract to most readers. Nonetheless, it is extremely important, as it shows how some significant discoveries are possible only when data is accurately managed. LIGO was built between 1994 and 2002, while the famous detection of gravitational waves only came in 2015, after the analysis of years of data (over 1 petabyte so far) and various technical improvements.

Why did the collection, curation, and sharing of such a massive amount of data matter, if it "only" confirmed the existence of something as imperceptible as gravitational waves? As the scientists responsible for the discovery stated, "gravitational waves carry information about their

dramatic origins and about the nature of gravity that cannot otherwise be obtained. Physicists have concluded that the detected gravitational waves were produced during the final fraction of a second of the merger of two black holes to produce a single, more massive spinning black hole. This collision of two black holes had been predicted but never observed".

LIGO findings have received impressive coverage by the media, with about 111,000 hits on Google News as of April 2017. Other impacts from LIGO data pale in comparison, however, experimental results are currently used in academic papers and for education and outreach activities. Furthermore, LIGO data is used by LSC members to produce additional information on past detections, including visuals, video, audio, and other media.

Title	24: Surfing Gravitational Waves
Subtitle	Understanding gravity and Einstein's General Theory of Relativity
Abstract	Since 2002, the LIGO experiment has been running and collecting data in the domain of astrophysics. This very large facility created over 1 petabyte of information, which required management for several years. This effort has allowed researchers to confirm the existence of gravitational waves, ripples in space time hypothesised by Einstein. In addition, data from the experiment is being used to produce peer-reviewed publications and additional information on past detections, such as visuals, audio, and other media.
Keywords	gravitational waves; relativity
Research subject area	PHYSICAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	111,000 news mentions Over 1 petabyte of data
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1994
Organisations involved	California Institute of Technology; Massachusetts Institute of Technology; consortium >80 institutions
Academic citations	
Links	https://en.wikipedia.org/wiki/Solar_System
	http://www.space.com/19103-milky-way-100-billion-planets.html
	https://en.wikipedia.org/wiki/Gravitational_wave
	http://www.independent.co.uk/news/science/gravitational-waves-simple-explanation-video-a6869761.html
	https://www.ligo.caltech.edu/page/about
	https://losc.ligo.org/

	https://dcc.ligo.org/M1000066/public
	https://en.wikipedia.org/wiki/First_observation_of_gravitational_waves
	https://royalsociety.org/topics-policy/projects/science-public-enterprise/case-studies/
	https://www.google.co.uk/search?hl=en&gl=uk&tbm=nws&authuser=0&q=%22LIGO%22&oq=%22LIGO%22

25: Art From The Comfort Of Your Chair

Aggregating arts metadata to power a digital museum

In 2016 alone, a total of more than 18 million people visited the British Museum, the National Gallery and Tate Modern in London. In many museums and countries, access to a wealth of paintings, sculpture, and artefacts is provided free of charge. However, researchers in this domain are often unable to visit museums whenever they need access or see a given work of art.

In this field, research data management has encountered a few obstacles as "data in the culture sector is perceived by and large as ticket sales, audience numbers and visitor footfall". This view is understandable yet inaccurate: educational institutions, museums, and other organisations own significant information on their collections and this can be openly used and shared to increase their reach.

The Europeana platform

More than 3,000 institutions contribute to Europeana.eu, an online repository including "54,364,816 artworks, artefacts, books, videos and sounds from across Europe". All these resources remain with their original organisation, however, the platform collects metadata and image miniatures to allow searching through a single interface. The role of Europeana is not that of digitising arts, but simply to accept and index metadata using the purpose-built Europeana Data Model.

The portal started "as a big political idea to unite Europe through culture by making our heritage available to all for work, learning or pleasure". They believe that culture can catalyse social and economic change, but this is only possible when it is "readily usable and easily accessible for people to build with, build on and share". Europeana allows access to an incredible wealth of material, however, they estimate that only 34% of the currently-digitised outputs (about 300 million objects) are available online for creative reuse. The portal is committed to making material online in open formats by developing standards, embracing new technologies, changing copyright, and look at new business models.

Among the participants to the project are UK contributors such as the Bodleian Libraries at the University of Oxford and the Wellcome Library.

The impact of sharing cultural heritage

Data shared through the Europeana portal is used to promote creative thinking in many disciplines, including Communication Design, Industrial Design, Fashion Design & Domestic Design. Through classes, students learn how to reuse digitised art and create unique designs inspired by digital heritage. Interestingly, the process of reusing digitised art has sometimes led artists to the creation of new analogue art in the form of installations, prints, or domestic designs.

Another project originating from Europeana involved art, technology, and storytelling. Storypix, a company based in Amsterdam,

created a web exhibition platform that allows institutions holding art to present their collections in public spaces such as digital advert panels or bus stops. This type of evidence of Europeana's impact is gathered through their Europeana Challenge, which looks for reuse of openly licensed content demonstrating social and/or economic impact.

Finally, Europeana recently started a new series of blogposts describing research arising from its resources. Named Europeana Treasures, the new series highlights how browsing and exploring the repository can lead users to discover new sources they weren't aware of.

Title	25: Art From The Comfort Of Your Chair
Subtitle	Aggregating arts metadata to power a digital museum
Abstract	Educational institutions, museums, and other organisations hold a wealth of information on paintings, sculpture, and artefacts. This has historically been kept private or indexed locally, however, the Europeana project aims to aggregate metadata in the field of arts and make it publicly accessible. Over 3,000 organisation contribute to the portal, which hosts more than 54 million records from across Europe. The Europeana project has been using its material to promote creative thinking in a number of disciplines and led to the creation of a web exhibition platform for institutions holding art. In addition, Europeana tracks research arising from its data through the Europeana Treasures blogpost series.
Keywords	cultural heritage; digitisation
Research subject area	LANGUAGE, COMMUNICATION AND CULTURE
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Technological; Cultural
Facts and figures	Collaboration of 3,000 institutions 54,364,816 artworks, artefacts, books, videos and sounds from across Europe
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2008
Organisations involved	>3000 organisations
Academic citations	
Links	http://www.alva.org.uk/details.cfm?p=423
	https://www.theguardian.com/culture-professionals-network/culture-professionals-blog/2014/mar/28/arts-culture-sector-data-impact
	https://en.wikipedia.org/wiki/Europeana
	http://www.europeana.eu/portal/en
	http://strategy2020.europeana.eu/
	http://pro.europeana.eu/page/copyright-reform
	http://labs.europeana.eu/blog/take-a-first-glimpse-at-re-media-s-showcase

	https://www.storypix.org/
	http://research.europeana.eu/blogpost/europeana-treasures-1-creating-a-narrative-of-children-s-literature-books-and-illustrations

Part B—Short-form case studies

1: A Performance Artwork Based on Datasets and Partnerships

Living Symphonies, a landscape sound installation portraying forest wildlife, plants, and weather, was funded by the Arts Council England, with delivery partners Sound and Music and the Forestry Commission England.

A sophisticated software model of the forest ecosystem was created. Datasets were produced in preparation for the work (221 Gb) and during the tour (213 Gb). They included ecological surveys of four forest sites, photographic surveys, forest sound recordings, weather datasets, music session recordings, documentation of the live installations, testimonials from visitors, and more.

Outputs and lessons learnt

Working in forests meant no reliable Internet connection. Working with partners meant sharing data. Consequently, data was first stored and duplicated onto hard drives before later being synchronised through the cloud. Dropbox provided ease of use, stability, and a simple sharing interface.

A free Toolkit on Planning and Producing Artworks in the Natural Environment was published, sharing good practice.

The artists retain all copyright in the work and datasets and they are willing to make data open access through Creative Commons licenses. However, as reported in the LEARN Toolkit of Best Practice for Research Data Management, there was never a requirement to make such data publicly available, nor funding provided to cover the costs of preparing, explaining and contextualising the datasets. There are also unresolved issues in hosting such large quantities of material.

Title	1: A Performance Artwork Based on Datasets and Partnerships
Keywords	music; sound; software
Research subject area	EARTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	>430GB of data created
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2014

Organisations involved	Forestry Commission England and Sound And Music; Arts Council England
Academic citations	
Links	http://www.livingsymphonies.com/
	http://www.artscouncil.org.uk/
	https://doi.org/10.14324/000.learn.11
	https://www.forestry.gov.uk/forestry/infd-96vmjg
	http://soundandmusic.org/create/planningandproducingartworksinthenaturalenvironmenttoolkit

2: Open Science Underpins Collaboration: The Structural Genomics Consortium (SGC)

SGC carries out precompetitive research in structural biology at universities in the UK, Canada, Brazil, Germany, Sweden, and the USA. Funders include pharmaceutical companies, public bodies, and non-profit organisations. Removing intellectual property restrictions and contributing newly solved protein structures to the Protein Data Bank is a key element of the approach, as well as sending samples to researchers upon request.

Benefits and impact

SGC demonstrates an open data, public-private partnership. By not claiming patents and making data available through open access, costs for individual collaboration and materials transfer agreements estimated at "hundreds of thousands of dollars" were eliminated. This was achieved by sharing "over 1,500 protein structures and 75 kinase structures".

The SGC was the fourth-largest contributor to the Protein Data Bank (PDB), providing around 10% of new structures deposited each year. The PDB itself has been estimated to be worth over US$12 billion, and to impact more than 80% of biomedical research grants, which further illustrates the importance of the data sharing efforts by the SGC.

Title	2: Open Science Underpins Collaboration: The Structural Genomics Consortium (SGC)
Keywords	protein structure
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Economic
Facts and figures	over 1,500 protein structures and 75 kinase structures shared Protein Data Bank is worth $12 billion
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2003
Organisations involved	University of Oxford; University of Toronto; Universidade Estadual de Campinas; Karolinska Institutet; The University of North Carolina at Chapel Hill; Goethe University Frankfurt

Academic citations	
Link 1	http://www.rcsb.org/pdb/home/home.do
Link 2	http://ec.europa.eu/research/openscience/pdf/monitor/sgc_case_study.pdf
Link 3	http://ec.europa.eu/research/openscience/index.cfm?pg=researchdata&section=monitor
Link 4	http://www.rand.org/content/dam/rand/pubs/research_reports/RR500/RR512z1/RAND_RR512z1.pdf

3: Testing Doubts About the Reliability of Science: The Reproducibility Project

Funded by the Laura and John Arnold Foundation who seek evidence-based decisions, this project explored the reproducibility of scientific findings. It was a collaborative effort to replicate 100 psychology experiments involving:

- 270 researchers worldwide
- Interactions with original researchers
- Reuse of original materials (where possible)
- Open sharing of research designs and protocols
- Use of the Open Science Framework (OSF)
- Public availability of raw data and reports on the replications.

The project benefited from a fully open design, based on the idea that "a lack of transparency and accountability contributes to biases in the literature".

Results & Impact

Only about 40% of the original findings could be replicated, prompting debate about systemic problems in science and how to address them.

Two papers which are direct project outputs are among the most cited in their field for their age. References to the project and the involvement of its researchers in new initiatives evidence contributions towards:

- Driving changes among funders supporting reproducibility research
- Researchers beginning reproducibility projects in other fields
- Journals supporting reproducibility
- Research transparency being supported by policies.

The Reproducibility Project: Cancer Biology (RP:CB) follows and builds on the transparent, open approach and the Center for Open Science also has its foundations in the project, which demonstrated the importance of open data and protocols in producing robust scientific findings.

Title	3: Testing Doubts About the Reliability of Science: The Reproducibility Project
Keywords	reproducibility; open science
Research subject area	PSYCHOLOGY AND COGNITIVE SCIENCES
Type of RDM impact/benefit	Reproducibility
Summary Impact Type	N/A

Facts and figures	Only 40% of findings could be replicated
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2011
Organisations involved	University of Virginia
Academic citations	
Links	http://www.arnoldfoundation.org/
	http://ec.europa.eu/research/openscience/pdf/monitor/reproducibility_project_case_study.pdf
	https://cos.io/about/brief-history-cos-2013-2017/
	https://en.wikipedia.org/wiki/Reproducibility_Project

4: Large Volumes of Data Engage Large Communities: Sloan Digital Sky Survey (SDSS)

SDSS is conducted by the Astrophysical Research Consortium (ARC), a non-profit partnership among research universities and laboratories. Its objective is to reflect the large-scale structure of our universe in multi-coloured images and 3D maps. Accurate measurements are needed, and large datasets are produced via a 2.5-meter optical telescope in New Mexico.

Benefits and impact

SDSS enables mapping of the universe and has created the world's largest open access database about the universe. Data in the form of images, catalogues of objects and spectra are available through the SDSS website.

SDSS users are allowed to browse through images, measurements, data, and spectra. Their queries can relate to "the position, colour of a star or a galaxy or unique properties of interesting objects in the sky". Press releases and publications indicate the enormous value of the data to research.

SDSS is the largest survey contributor to Galaxy Zoo, a ground-breaking citizen science initiative that launched the multidisciplinary Zooniverse platform. Galaxy Zoo volunteers were instrumental to classifications of nearly 900,000 galaxies. Tools include lesson plans for using Galaxy Zoo in education.

The survey is ongoing, now in its fourth phase, and the team are committed to data releases until 2020.

Title	4: Large Volumes of Data Engage Large Communities: Sloan Digital Sky Survey (SDSS))
Keywords	space; citizen science
Research subject area	EARTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	Classification of nearly 900,000 galaxies
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1998

Organisations involved	Brazilian Participation Group; Carnegie Institution for Science; Carnegie Mellon University; Chilean Participation Group; French Participation Group; Harvard-Smithsonian Center for Astrophysics; Instituto de Astrofísica de Canarias; Johns Hopkins University; Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo; Lawrence Berkeley National Laboratory; Leibniz Institut für Astrophysik Potsdam (AIP); Max-Planck-Institut für Astronomie (MPIA Heidelberg); Max-Planck-Institut für Astrophysik (MPA Garching); Max-Planck-Institut für Extraterrestrische Physik (MPE); National Astronomical Observatories of China; New Mexico State University; New York University; University of Notre Dame; Observatório Nacional / MCTI; Ohio State University; Pennsylvania State University; Shanghai Astronomical Observatory; United Kingdom Participation Group; Universidad Nacional Autónoma de México; University of Arizona; University of Colorado Boulder; University of Oxford; University of Portsmouth; University of Utah; University of Virginia; University of Washington; University of Wisconsin; Vanderbilt University; Yale University
Academic citations	
Links	http://arc.apo.nmsu.edu/
	http://www.sdss.org/
	http://ec.europa.eu/research/openscience/pdf/monitor/sloan_digital_sky_case_study.pdf
	http://www.sdss.org/science/publications/
	https://www.galaxyzoo.org/
	https://www.zooniverse.org
	https://www.calacademy.org/educators/lesson-plans/galaxy-zoo-keeper-training

5: Scholarly Communication is About Combining Effort: Polymath Project

The Polymath Project blog invites all with a background in mathematics to address unsolved problems in combinatorial mathematics. A wiki page summarises all the knowledge developed for a specific problem. Research and discussion threads enable researchers to post their contributions and discuss solutions. The project set out to understand whether 'massive collaborative mathematics' was a possible path for research.

Benefits and impact

12 problems have been published (to date). 3 have now been solved, and Polymath 1, involving the efforts of 27 researchers, provided evidence that collaboration in mathematics can lead to faster results. This has stimulated debate within the combinatorics research community on incentives for collaborative research, and the characteristics that a collaborative approach should have.

The project developed a set of rules on how to publish problems and conduct related discussion, and for summarising and consolidating results: the Polymath general rules. The solution for crediting project contributors in publications (six, so far), has been an author pseudonym, D.H.J. Polymath.

The educational Crowdmath project is a spin-off, targeting high school and college students with the aim of educating young mathematicians to conduct research, "particularly through collaborative approaches".

Title	5: Scholarly Communication is About Combining Effort: Polymath Project
Keywords	collaboration; problem solving
Research subject area	MATHEMATICAL SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	3 out of 12 mathematical problems solved 6 published articles
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2009

Organisations involved	University of Cambridge
Academic citations	D.H.J. Polymath. (2014). [The "bounded gaps between primes" Polymath project - a retrospective](https://arxiv.org/abs/1409.8361)
	D.H.J. Polymath. (2014). [Variants of the Selberg sieve, and bounded intervals containing many primes](https://arxiv.org/abs/1407.4897)
	D.H.J. Polymath. (2014). [New equidistribution estimates of Zhang type](https://arxiv.org/abs/1402.0811)
	D.H.J. Polymath. (2010). [Deterministic methods to find primes](https://arxiv.org/abs/1009.3956)
	D.H.J. Polymath. (2010). [Density Hales-Jewett and Moser numbers](https://arxiv.org/abs/1002.0374)
	D.H.J. Polymath. (2009). [A new proof of the density Hales-Jewett theorem](https://arxiv.org/abs/0910.3926)
Links	https://polymathprojects.org/
	http://michaelnielsen.org/polymath1/index.php?title=Main_Page
	https://artofproblemsolving.com/polymath/mitprimes2016
	http://ec.europa.eu/research/openscience/pdf/monitor/polymath_case_study.pdf

6: Is Citizen Generated Data Suitable for Academic Purposes? Volunteered Geographic Information (VGI)

A popular VGI project, OpenStreetMap (OSM) is "a free, editable map of the whole world that is being built by volunteers largely from scratch and released with an open-content license". It generates academic interest in many topics, including data quality.

OSM's detailed data could be used in future to create up-to-date Land Use/Land Cover maps, which when otherwise produced by satellite imagery and field visits, have high production costs. Geo-tagged photos on social media are a rich source of VGI. A COST Action study (EU Horizon 2020) concluded that Geograph had the highest potential (Flickr and Panoramio also considered), "with a mean of only 12% of photographs considered as unusable", but automatic methods for excluding unusable photographs are an area for future work.

Addressing challenges

There is a need for data collection protocols: these might jeopardise contributors' motivation, but they would improve the quality and usability of the data. A proposed generic protocol for vector data suggests 3 principle ways to formalise data collection: "manual vectorisation; field survey; and reuse of existing data sources."

Quality assessment involves the comparison of OSM with authoritative road datasets: completeness and accuracy are important criteria. A flexible methodology for such comparison has been published.

Title	6: Is Citizen Generated Data Suitable for Academic Purposes? Volunteered Geographic Information (VGI)
Keywords	maps; volunteered data
Research subject area	EARTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	Only 12% of photographs considered unusable
Original dataset from which the impact arose	

Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2004
Organisations involved	OpenStreetMap Community
Academic citations	Brovelli, M.A., Minghini, M., Molinari, M. & Mooney, P. (2016). [Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets](https://dx.doi.org/10.1111/tgis.12182)
Links	https://www.openstreetmap.org/
	http://wiki.openstreetmap.org/wiki/About_OpenStreetMap
	http://www.mdpi.com/2220-9964/5/5/64
	http://www.geograph.org.uk/
	http://www.mdpi.com/2220-9964/5/11/217

7: A Specialist Research Data Archive: Crawdad

Research on wireless networks and mobile computing has had a community data repository in the form of CRAWDAD since 2005. The archive specialises in wireless trace data from many contributing locations, and has staff to "develop better tools for collecting, anonymizing, and analyzing the data." It has also been an early exponent of data citation, exploring problems and practices. ACM SIGMOBILE and Dartmouth College are current sponsors.

Tangible outputs

CRAWDAD's community build understanding of how real users, applications, and devices all use real networks, under real conditions. It boasts:

- 120 datasets
- 22 tools
- 10173 users from 116 countries
- 2378 papers using CRAWDAD datasets or mentioning CRAWDAD.

One recent dataset (CRAWDAD News) comprises experiments using open-source middleware NSense, which took four different pipelines from Samsung Galaxy S3 devices to compute aspects such as relative distance (Wi-Fi); social strength (based on Bluetooth contact duration); sound activity level; and motion.

Arguably the most popular dataset has been downloaded 1217 times, and cited 566 times: It offers "Four traces of Bluetooth sightings by groups of users carrying small devices (iMotes) for a number of days." There are 3 versions of this dataset available, plus another dataset in CRAWDAD is derived from it. Its six contributors come from 4 different countries, and include both academic and industry experts.

Title	7: A Specialist Research Data Archive: Crawdad
Keywords	wireless; networks
Research subject area	INFORMATION AND COMPUTING SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	2378 papers using CRAWDAD datasets or mentioning CRAWDAD 1217 downloads for the most popular datasets and 566 citations
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2004

Organisations involved	Dartmouth University
Academic citations	
Links	http://crawdad.org/index.html
	http://www.dlib.org/dlib/january15/henderson/01henderson.html
	https://www.sigmobile.org/
	http://citeulike.org/group/5303/tag/cambridge_haggle
	http://citeulike.org/group/5303/library

8: A Large-Scale Research Data Service: The European Bioinformatics Institute

The European Bioinformatics Institute (EMBL-EBI) "manages public life science data on a very large scale, making a rich resource of information freely available to the global life science community." It aims to exchange information, set standards, develop new methods, and curate complex genome information.

Value to the global community

EMBL-EBI data and services are almost entirely open resources: users are not required to register and are not directly identified or recorded. It has a £47 million annual operational expenditure, "with a minimum direct value to users that is equivalent to around 6 times the direct operational cost."

According to its users, the EMBL-EBI data and services are instrumental to making research significantly more efficient. This is estimated, at a minimum, "to be worth £1 billion per annum worldwide - equivalent to more than 20 times the direct operational cost".

In a survey which received 4,509 responses, researchers reported that they could neither have created/collected the last data they used themselves, nor obtained it elsewhere. EMBL-EBI data and services underpinned future research impacts "worth £335 million annually, or £2.5 billion over 30 years in net present value, that could not otherwise have been realised."

Title	8: A Large-Scale Research Data Service: The European Bioinformatics Institute
Keywords	genome information
Research subject area	BIOLOGICAL SCIENCES
Type of RDM impact/benefit	Reproducibility; Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Economic
Facts and figures	Efficiency gains worth £1 billion per annum Future research impact worth £335 million annually or £2.5 billion over 30 years
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)

Year (e.g., first data release, first output, year of impact)	1992
Organisations involved	European Bioinformatics Institute
Academic citations	
Links	http://www.ebi.ac.uk/
	https://beagrie.com/static/resource/EBI-impact-report.pdf

9: Combining Data & Influencing Government Sustainability Policies

The Smarter Travel research project at the University of Aberdeen examined and contributed to "a specific approach to sustainable transport planning. The approach, known as `Smarter Choices' or `soft measures'", consists of simultaneously informing people about their travel choices whilst seeking travel service improvements. Researchers "developed a specific quantitative methodology involving segmenting the population to give a flexible interpretation of behaviour, allowing different policies and messages to be targeted to different groups."

Impact of data-powered travel research

The study involved a contextual review, structured interviews, and analysis of a wide range of data sources. Data about intervention packages were gathered, so that inputs, outputs, outcomes and impacts of the interventions could be examined.

The research influenced transport and climate change agendas in both England and Scotland, being cited in policy guidance, evaluation frameworks, and new funding mechanisms. In addition, "in some instances use of time series data (correlated with specific events or interventions) and triangulation of different data sources has provided pointers as to the relative contributions of particular smart measures."

Census data hosted on InFuse were among the datasets used. InFuse is "a free service providing easy access to aggregate data from the UK 2011 and 2001 censuses", developed by the Economic and Social Research Council-funded UK Data Service Census Support team.

Title	09 COMBINING DATA & INFLUENCING GOVERNMENT SUSTAINABILITY POLICIES
Keywords	travelling; sustainability
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Political; Societal
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)

Year (e.g., first data release, first output, year of impact)	2008
Organisations involved	University of Aberdeen
Academic citations	
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=43341
	https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/4408/chap1.pdf
	http://infuse.ukdataservice.ac.uk/index.html

10: Objective Measures & Self-Reported Data Underpin Policy On Obesity: Health Survey for England (HSE)

The annual HSE is used to examine: ways of improving people's health; changes to the nation's health over time; and inequalities in health. It underpins government policy and planning of NHS services. Demographic, socio-economic, health, and lifestyle data are self-reported, whilst objective measures include measured height, weight, and waist circumference.

University College London (UCL) provides clinical and methodological expertise, produces annual reports, and leads secondary data analysis and policy evaluation. NatCen undertake operational- and field-work. HSE datasets are available through the UK Data Archive.

Expertise in methodology leads to rich, revealing data and research impact

HSE is "one of the few health surveys in Europe to obtain objectively measured anthropometric data, rather than relying on self-reporting which consistently under-reports obesity." Further, it "quantified the extent and escalation of obesity within the population as a whole as well as specific sub-groups, resulting in this issue being given significant attention in government."

In addition, researchers at UCL's Health and Social Surveys Research Group, multidisciplinary scholars with internationally renowned expertise in survey methodology, reused survey data to perform further analysis. They "compared the effect of mode and context of survey on response rates, non-response bias, and responses; and of demographic and socio-economic variation in survey participants by time and day interviewed."

Title	10: Objective Measures & Self-Reported Data Underpin Policy On Obesity: Health Survey for England (HSE)
Keywords	health survey; obesity
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A
Facts and figures	

Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1991
Organisations involved	University College London
Academic citations	
Links	http://www.natcen.ac.uk/
	http://www.data-archive.ac.uk/
	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=40405
	http://www.ucl.ac.uk/hssrg

11: Archival Acceptance as an "Indicator of Quality": Foot & Mouth Disease (FMD) Testimonial Data

Interdisciplinary, innovative research from the University of Cumbria showed that the FMD epidemic was a disaster for many rural people. The project focussed on human health and the social consequences of the FMD crisis.

The project involved "engagement with those affected by the crisis, via interviews and weekly diaries", and revealed perceptions of the crisis and its effects on life in Cumbria. Insights included the value of local knowledge in a disaster situation, and the ways people react to abnormal events. Researchers made policy recommendations based on longitudinal data and synthesised evidence.

High quality data

The Mass Observation social research approach influenced the project, which recruited a standing 'citizen' panel of 54 respondents who produced 3,200 weekly diaries of "enormous intensity and diversity over an 18-month period. The data were supplemented by in-depth interviews with each respondent, and focus group discussions, and in addition, 16 other interviews with stakeholders were conducted. All material was transcribed and digitised."

Research data "was selected to be archived by the Economic & Social Data Service (ESDS)" which was later integrated into the UK Data Service. The data collection archived there includes 42 semi-structured interview transcripts, 40 semi-structured diaries, 6 focus group transcripts, and 1 audiomontage transcript. Audio recordings of interviews and focus groups are available with the permission of the depositor.

Title	11: Archival Acceptance as an "Indicator of Quality": Foot & Mouth Disease (FMD) Testimonial Data
Keywords	crisis; diaries
Research subject area	MEDICAL AND HEALTH SCIENCES
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Societal
Facts and figures	3,200 weekly diaries leading to the understanding of how people react to abnormal events
Original dataset from which the impact arose	

Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2001
Organisations involved	University of Cumbria
Academic citations	
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=21479
	http://www.massobs.org.uk/
	https://discover.ukdataservice.ac.uk/catalogue/?sn=5407&type=Data%20catalogue

12: Improving Policy by Providing Data: Live Music Exchange (LMX)

LMX began in 2008 with a study of the UK's live music sector at the Universities of Edinburgh and Glasgow, funded by the Arts and Humanities Research Council (AHRC). It explored the "history, economics and sociology of the live music sector in the UK". Bibliographic data underpins a literature review on the impact of music festivals, with annotated bibliography.

Research continued with the Live Music Census, a survey first carried out in Edinburgh in 2015, in other selected cities in 2016, and in 2017 across the UK. Stakeholders contributed to survey design and promotion, and LMX is "providing a method and framework we can all agree on for assessing the scope and value of live music in the UK."

Stakeholders value academic research and data

LMX contributed to the UK's policymaking process by providing "relevant data and data analysis and by improving communication between the sector's stakeholders." The academic, music industry, and government stakeholders involved in LMX are linked via the project website to cover all problematic areas in the sector, ranging from environmental sustainability to legal matters. Understanding of academic data is evident in a stakeholder quote:

"Data is really important to us because what we are doing is making a change from the sort of anecdotal evidence that we've used in the past... about the need for support for these venues."

Title	12: Improving Policy by Providing Data: Live Music Exchange (LMX)
Keywords	live music
Research subject area	LANGUAGE, COMMUNICATION AND CULTURE
Type of RDM impact/benefit	Socio/Economic impact
Summary Impact Type	Political
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2015
Organisations involved	University of Edinburgh; University of Glasgow
Academic citations	

Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=24040
	http://www.ahrc.ac.uk/documents/project-reports-and-reviews/connected-communities/impact-of-music-festivals/
	http://uklivemusiccensus.org/
	http://www.eca.ed.ac.uk/reid-school-of-music/research/projects/live-music-exchange
	http://livemusicexchange.org/links/
	https://emmawebster.org/category/live-music-exchange/

13: From Hard Copy Primary Sources to An Open Online Resource: Reading Experience Database, (RED)

Today, RED is the world's largest database about reading habits, with an open online contribution method. Developed by its hosts, the Open University, "RED is democratising scholarship about the history of reading by encouraging members of the public from any location to contribute and use information about readers through history."

In 1995, an Arts and Humanities Research Council (AHRC) funded project involved a hard-copy data collection form and the creation of an internal collection of primary materials to answer research questions. Two further AHRC projects saw first the digitisation of data (2006-2009) and then the involvement of international partners (2010-2011), along with development of a systematic, easy-to-use search function. Incrementally, RED has expanded into an online, open-access project with more than 30,000 entries.

Facts and figures

According to the RED portal, "evidence of reading presented in UK RED is drawn from published and unpublished sources as diverse as diaries, commonplace books, memoirs, sociological surveys, and criminal court and prison records."

The RED database:

- Includes documents the history of reading in Britain from 1450 to 1945
- Is the result of the contributions of 120+ volunteers, who created 6000 entries
- Has 1800+ users each month from more than 135 countries
- Provided expertise for partner projects in Australia, Canada, the Netherlands, and New Zealand
- Is cited as an "indispensable primary source" by scholars.

Title	13: From Hard Copy Primary Sources to An Open Online Resource: Reading Experience Database, (RED)
Keywords	reading
Research subject area	LANGUAGE, COMMUNICATION AND CULTURE
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Methodological impact (e.g., new approaches developed)
Summary Impact Type	N/A

Facts and figures	contributions of >120 volunteers, who created over 6000 entries >1800 users per month from >135 countries
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	1995
Organisations involved	Open University
Academic citations	
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=32227
	http://www.open.ac.uk/Arts/RED/
	http://www.open.ac.uk/Arts/reading/UK/about.php

14: Qualitative Data in Many Formats, Archived Online: Tate Encounters

Completed in 2010 and funded by the AHRC, Tate Encounters sought to identify barriers preventing access to publicly-funded culture, through knowledge of encounters with a leading international art museum. People who fit the demographic of non-attendees were of particular interest, and insights were sought into how museum professionals might respond to cultural policies advocating cultural diversity amongst audiences.

Data about participants' experience of encountering the art museum included: "300 student questionnaires; 200 student essays on Tate Modern and Tate Britain; 12 student workshops; 12 in depth student research projects; 5 extended participant family edited ethnographic films; 38 Tate staff interviews and interviews with 72 participants through the Research in Process events at Tate Britain."

The main impacts arising from the management from such a diverse set of data include:

- influencing Tate's online strategy
- influencing key decision takers
- modelling of categories of audience to allow for audience development.

Archiving heterogenous data

Although the project is closed, an edited selection of research data can be found on an archival website called Tate Encounters: Britishness and Visual culture. The site contains a large amount of audio-visual material including audio interviews and discussions, video films, photo-essays and research papers. These can be found through a menu of these types of content, through search box, or by clicking on a tag in a tag cloud. The menu page for each type of content explains more about how that type of content was created.

Title	14: Qualitative Data in Many Formats, Archived Online: Tate Encounters
Keywords	qualitative data; museum
Research subject area	STUDIES IN HUMAN SOCIETY
Type of RDM impact/benefit	Socio/Economic impact
Summary Impact Type	Societal; Cultural
Facts and figures	300 student questionnaires 200 student essays on Tate Modern and Tate Britain 12 student workshops 12 in depth student research projects 5 extended participant family edited ethnographic films

	38 Tate staff interviews and interviews with 72 participants through the Research in Process events at Tate Britain
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2010
Organisations involved	Tate Modern; Tate Britain
Academic citations	
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=43741
	http://process.tateencounters.org/
	http://www.tate.org.uk/about/projects/tate-encounters

15: Data Combination and Self Re-Use: Understanding Pauper Lives in Georgian London

Professor Jeremy Boulton at the University of Newcastle combined data from three projects relating to paupers' lives in Georgian London, and prepared a selection for public sharing which has been used by genealogists, academics, and others worldwide. Information on pauper's lives are now widely and easily available to users thanks to the conservation, contextualisation, and presentation of the above-mentioned cultural heritage.

The London Lives website hosts a selection of publicly available data.

Three phases of data creation and management

In the first research project, workhouse records and payments to parish pensioners were amongst sources used to create seven substantial datasets, reconstructing the life histories of over 50,000 poor individuals in the West End of London.

The next research project "involved the collection and analysis of very large datasets (containing over 300,000 records) which record information (workhouse careers, poor law payments, migration and employment histories, births, deaths, marriages) about named individuals." These datasets were linked to those created in the earlier research through a mapping exercise: this enabled an in-depth examination of mortality by social class.

Thirdly, baptism and burial fee records were transcribed into digital form for analysis. "Using GIS technology, the research mapped the spatial context of historical urban mortality."

Title	15: Data Combination and Self Re-Use: Understanding Pauper Lives in Georgian London
Keywords	pauper lives; London
Research subject area	HISTORY AND ARCHAEOLOGY
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	Over 300,000 records about over 50,000 poor individuals in the West End of London
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)

Year (e.g., first data release, first output, year of impact)	2004
Organisations involved	University of Newcastle
Academic citations	
Links	http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=21728
	https://www.londonlives.org/browse.jsp?archive=dataset
	http://research.ncl.ac.uk/pauperlives/

16: Ireland-Bristol Trade in the Sixteenth Century

Research from the University of Bristol examined Ireland's sixteenth century trade with Bristol, "to test the assumption that the Irish economy had a low growth potential prior to the English conquest of the late sixteenth and early seventeenth century." It was originally planned that the accounts for nine individual years would be included.

Researchers created a database of transcribed and translated, detailed customs accounts (originals in Latin). These records are invaluable "because poor record survival mean that no other economic records of equivalent value survive in Ireland, or elsewhere. The Bristol customs accounts thus constitute the best quantitative source that exists for examining Ireland's economic development over the course of the sixteenth century." Original documents are in The National Archives in London.

The research data management effort described above not only allowed scientific discoveries, but also enabled the preparation of peer-reviewed articles and multiple conference presentations. The structured approach followed by the researchers saw their research rated "Outstanding" by ESRC, the funder of the research.

Publishing and publicising the data

During the project, draft copies of the spreadsheets were made available on the project website. The data, which was prepared using a custom metadata schema, now has records in several key places:

1. University of Bristol's research Information system (EXCEL workbooks)
2. Funder website: data download links on the ESRC project pages
3. ROSE, the Bristol institutional repository
4. UK Data Service
5. A print, published book with the eleven customs accounts

Copyright of the data remains with the researcher depositor, and access through the UK Data Service requires registration so that the depositor is informed where re-use occurs.

Title	16: Ireland-Bristol Trade in the Sixteenth Century
Keywords	trade; Ireland; England
Research subject area	HISTORY AND ARCHAEOLOGY
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	
Original dataset from which the impact arose	http://www.bris.ac.uk/Depts/History/Ireland/datasets.htm

Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2006
Organisations involved	University of Bristol
Academic citations	Jones, E., & Flavin, S. (2009). [Bristol's trade with Ireland and the Continent, 1503–1601](http://www.fourcourtspress.ie/books/archives/bristols-trade-with-ireland-and-the-continent/). Four Courts Press Stone, R. (2011). [The overseas trade of Bristol before the Civil War](https://dx.doi.org/10.1177/084387141102300210) Flavin, S. (2011). [Consumption and material culture in sixteenth-century Ireland](http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0289.2010.00569.x/full) Smith, B. (2011). [Late Medieval Ireland and The English Connection: Waterford and Bristol c.1360-c.1460](http://www.jstor.org/stable/23265418)
Links	http://www.researchcatalogue.esrc.ac.uk/grants/RES-000-23-1461/outputs/read/b3d9fedd-c37b-4a75-8cfe-e59c887050b6
	http://doi.org/10.5255/UKDA-SN-6275-1
	http://www.bris.ac.uk/Depts/History/Ireland/
	http://hdl.handle.net/1983/1296
	https://data.bris.ac.uk/data/dataset/1e32a4005d1a5a2bb5a3ee5b4555bae9
	https://discover.ukdataservice.ac.uk/catalogue?sn=6275
	http://www.fourcourtspress.ie/books/archives/bristols-trade-with-ireland-and-the-continent/

17: Film and an Ethnographic Approach: Buddhist Cosmology in Food

Research outputs and data from the one-year project "Feeding humans and non-humans in Theravada Buddhism" are the basis of six documentary videos. This digital humanities initiative explored the relationship between humans and non-humans in South and Southeast Asia, through food offerings or feasts. Data produced included film from Sri Lanka, stills and sounds which were elaborated and then used to create documentary videos for sharing. Project funders included the Leverhulme Trust, the British Academy and the University of Bristol.

Video documentaries are shared

Videos are on a Vimeo channel entitled "Kitchen Cosmology" and include:

- Making food for the Buddha
- Three fruit trays for Pattini
- Milk rice for milk mothers
- Food for all
- Rice balls for the crows
- Breakfast for the Buddha, monks and hungry ghost

Documentary credits include composers of music scores, editors and the research Principal Investigator. Films are made available under a non-commercial public sector information licence. "The project favoured open formats and technologies over proprietary ones whenever possible" and a longer DVD combining the six documentaries has been released and deposited in the institutional repository. The HD film, A Buddhist Cosmology in Food, can be downloaded as a zip file and was picked up by a news outlet.

Title	17: Film and an Ethnographic Approach: Buddhist Cosmology in Food
Keywords	buddhism; food
Research subject area	LANGUAGE, COMMUNICATION AND CULTURE
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	6 videos created as a project output
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)

Year (e.g., first data release, first output, year of impact)	2014
Organisations involved	University of Bristol
Academic citations	
Links	http://www.bristol.ac.uk/religion/buddhist-centre/projects/cosmology-in-food/about/
	https://vimeo.com/channels/buddhistcosmologyinfood
	http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/non-commercial-government-licence.htm
	http://www.bristol.ac.uk/religion/buddhist-centre/projects/cosmology-in-food/about/
	https://data.bris.ac.uk/data/dataset/h10sdqnhz4a81tu03yj3qgk9s
	http://indianexpress.com/article/india/india-news-india/international-oral-history-day-2-highlights-2885293/

18: Data About Data Archiving: ICPSR's Data Sharing in the Social Sciences

This "meta" project surveyed 1,217 "principal investigators of social science awards made by the National Science Foundation (NSF) and the National Institutes of Health (NIH)" from 1985 to 2001, on research data issues. It elicited a 24.9% response rate and the survey's own data were made available on the ICPSR website in December 2016, after the publication of a paper in 2010.

Research findings included that "the majority of data collections produced by NIH and NSF awards have not been archived", and that informal sharing was more common: there were many reasons for this. The study highlighted the "benefits of developing and implementing a data archiving plan at early parts of the data life cycle."

What survey data is available?

The public dataset includes the questionnaire and documentation on coding and data are available for download in SAS, SPSS, Stata, R, ASCII, and Excel formats. Respondent confidentiality has been protected by masking of some variables, recoding, or collapsing of data. Researchers are encouraged to use the public use data first, and to apply for access to further data if needed. The entire dataset or its documentation has been downloaded 31 times by major institutions including the London School of Economics (as of 2 May 2017).

Title	18: Data About Data Archiving: ICPSR's Data Sharing in the Social Sciences
Keywords	meta-analysis
Research subject area	STUDIES IN HUMAN SOCIETY
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Long-standing (5+ years)
Year (e.g., first data release, first output, year of impact)	2010
Organisations involved	Inter-university Consortium for Political and Social Research

Academic citations	Pienta, A.M., Alter, G.C., Lyle, J.A. (2010). [The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data](https://deepblue.lib.umich.edu/handle/2027.42/78307)
Links	https://doi.org/10.3886/ICPSR29941.v1
	http://www.iassistdata.org/conferences/2006/presentation/2331
	https://deepblue.lib.umich.edu/handle/2027.42/78307
	http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/29941/utilization

19: Research Data Supports Restoration: Mackintosh Architecture

An Arts and Humanities Research Council (AHRC) funded project, 'Mackintosh Architecture – Context, Making and Meaning', delivered "the first authoritative survey" of Charles Rennie Mackintosh's architecture. A database of architectural projects was created, drawing on architecture office records and other sources.

Details such as project name and address, client name, start date, "whether the project was for a new building or for alterations to an existing one, and whether or not any of the job book entry was written in Mackintosh's hand" were entered into an Access database, which was later exported into a MySQL database by the University of Glasgow's Humanities Advanced Technology Information Institute (HATII). A website and online catalogue of over 350 architectural projects with biographies, architectural drawings and other images is available, now under the responsibility of the Mackintosh Curator, University of Glasgow Hunterian Museum and Art Gallery. The same gallery also hosted a successful Mackintosh Architecture exhibition, held from July 2014 to January 2015.

Restoration and re-use

In the aftermath of a devastating fire in May 2014 which seriously damaged the Mackintosh Building of the Glasgow School of Art (widely considered the architect's masterpiece), the research data was of immediate help. Based on project outputs, restoration works could be promptly planned and executed.

The project also supported research and conservation at: the Glasgow Art Club, The Hill House and The Willow Tea Rooms.

Title	19: Research Data Supports Restoration: Mackintosh Architecture
Keywords	restoration; building
Research subject area	BUILT ENVIRONMENT AND DESIGN
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort); Socio/Economic impact
Summary Impact Type	Societal; Cultural
Facts and figures	350 architectural projects shared
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)

Year (e.g., first data release, first output, year of impact)	2014
Organisations involved	University of Glasgow
Academic citations	
Links	http://www.ahrc.ac.uk/research/casestudies/preserving-and-showcasing-architectural-heritage/
	http://www.mackintosh-architecture.gla.ac.uk/catalogue/essay/?eid=about
	http://www.ahrc.ac.uk/research/casestudies/preserving-and-showcasing-architectural-heritage/

20:Evidencing Value of Artistic, Cultural or Sporting Activities: UK Subjective Well-Being Data (SWB)

Researchers from Nottingham Trent University used data from the Understanding Society study to investigate the effect of engaging with the arts, culture and sport on individuals' well-being.

Researchers used UK subjective well-being data (SWB), and looked at participants' reported participation in sport and art. This work is significant because it uncovered the non-economic impact of leisure activities, which is normally not studied. The researchers' investigation goes beyond the usual economic measures of the value of artistic, cultural or sporting activities and "showed that engaging in arts activities, visiting libraries, museums and historical sites and mild sport is associated with greater life... satisfaction." However, a published article states that "engagement in leisure activities is not found to spill over into job satisfaction (with the exception of certain sports)".

The research was informed by a consultancy report titled 'Arts, Cultural Activity, Sport and Wellbeing' supported by EM Media and the Arts Council East Midlands.

About the source data

The Understanding Society study, or the United Kingdom Household Longitudinal Study (UKHLS), began in 2009 and is conducted by the Institute for Social and Economic Research (ISER), at the University of Essex. A multi-topic household survey, its aim is "to understand social and economic change in Britain at the household and individual levels."

Title	20: Evidencing Value of Artistic, Cultural or Sporting Activities: UK Subjective Well-Being Data (SWB)
Keywords	understanding society; sport; culture
Research subject area	STUDIES IN HUMAN SOCIETY
Type of RDM impact/benefit	Efficiency in research and data re-use (e.g., reduce duplication of effort)
Summary Impact Type	N/A
Facts and figures	
Original dataset from which the impact arose	
Maturity of the initiative/data source	Recent (<5 years)
Year (e.g., first data release, first output, year of impact)	2016

Organisations involved	Nottingham Trent University
Academic citations	Wheatley, D. & Bickerton, C. (2016). [Subjective well-being and engagement in arts, culture and sport](https://dx.doi.org/10.1007/s10824-016-9270-0)
Links	https://discover.ukdataservice.ac.uk/series/?sn=2000053
	https://www.ukdataservice.ac.uk/use-data/data-in-use/case-study/?id=214
	https://discover.ukdataservice.ac.uk/series/?sn=2000053