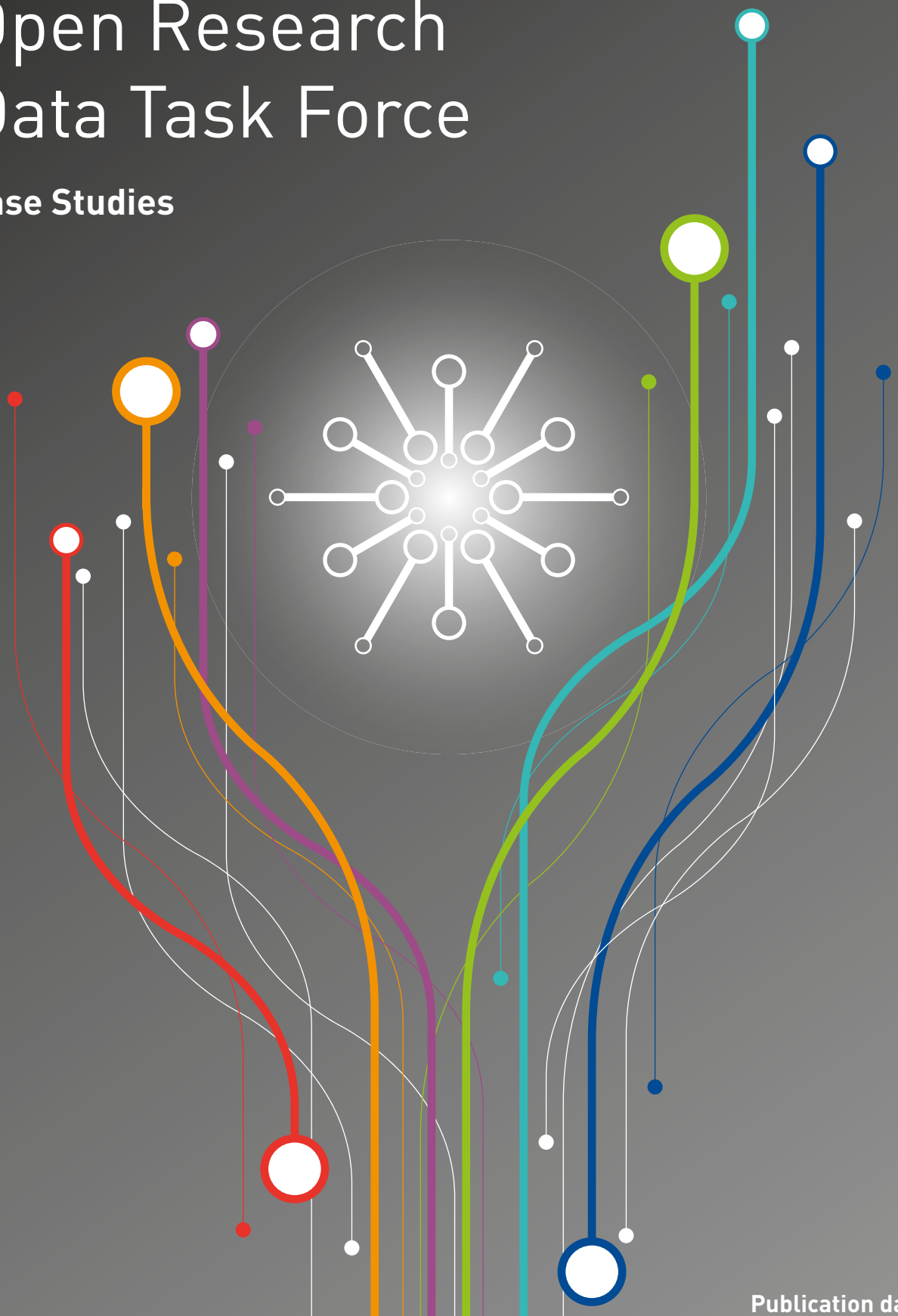
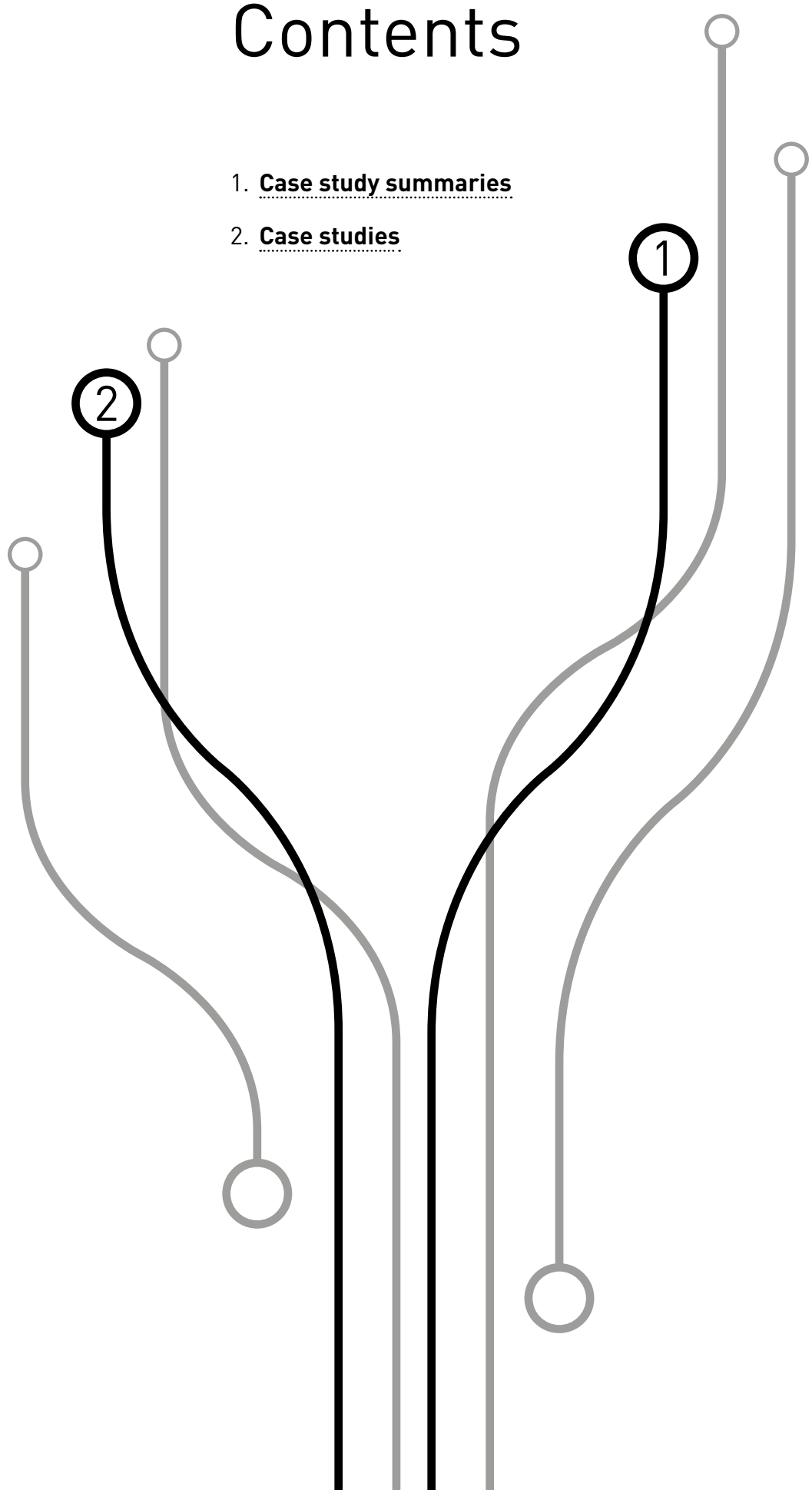
Open Research Data Task Force

Case Studies



Publication date:
July 2018

Contents



Case study summaries

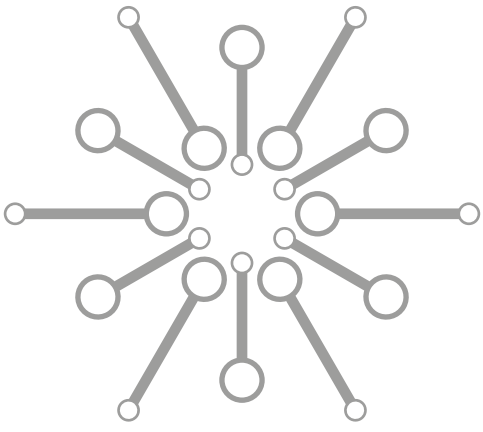
The Open Research Data Task Force was set up following Adam Tickell's advice to the Minister Jo Johnson to provide advice on open research data infrastructure and to deliver a roadmap for the UK. To achieve this, the Task Force undertook some preparatory work. The current landscape was reviewed and progress in the area charted in a **landscape study**.

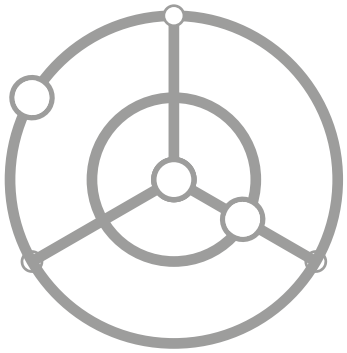
The investigation of eight institutional and disciplinary examples was commissioned to understand the roles and responsibilities of different organisations and communities in this space. The case studies covered use cases from astronomy, biosciences, digital humanities, crystallography, the Universities of Bristol and Salford, the Natural History Museum and Germany. The full case studies, together with this summary, are published alongside the final report of the Task Force, which makes recommendations to accelerate the UK's move to open research data (ORD).

The main research for these case studies was carried out in mid-2017 in order to support the Task Force in developing its recommendations and final report. The biosciences, crystallography and digital humanities cases have since been updated by members of the Task Force, but in the others there may have been more recent developments that are not fully reflected in the text.

Case study summaries:

1. **Astronomy**
2. **Biosciences**
3. **Crystallography**
4. **Digital Humanities**
5. **University of Bristol**
6. **University of Salford**
7. **Natural History Museum**
8. **Germany**





View the full case study on **page 12**

1. Astronomy

In astronomy, data is generated within large facilities that host at least one telescope. Each facility is set up to survey the sky every day, and store the images or any other data collected.

Cleaning, standardising, publishing and archiving is complicated, however, the workflow is incorporated into each physical facility from the start. There is no facility, as far as we could tell, that does not have an online presence: an archive, or some sort of searchable database for its data. Moreover, virtual facilities (like the **Virtual Observatory**) provide the environment for researchers to look through and theorise across many astronomical data centres internationally.

The large volumes of data collected by each facility every day, the transition from analogue to digital in 1970s and the physical inability of any one group of people to process and analyse this data has driven observatories to make their data open both for other researchers and the public. However, it's worth noting that in many circumstances there will be an initial period of exclusive access for particular research groups.

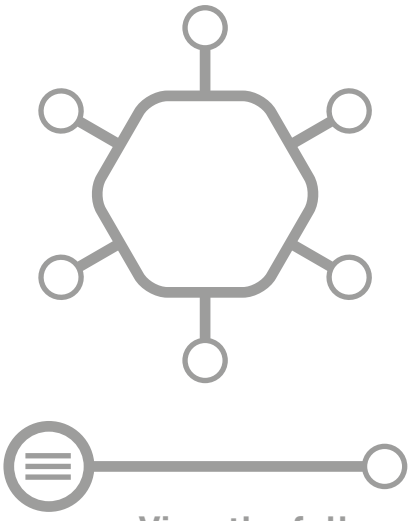
Although the research data in astronomy is almost inherently open, with facilities like the **Sloan Digital Sky Survey** promoting openness since 1989, the discipline is facing several challenges around research data infrastructure:

- Unusually large volumes of data that keep increasing every year (e.g. data that can fill 120 average laptops per night from a single facility)
- Adherence to standards for metadata and file formats is widespread, but they need to be continuously updated
- The infrastructure is mostly based on facility-supported or self-funded data centres, which are numerous and distributed, making data discovery more difficult
- Journal policies around depositing the underlying data are still divergent

The key learning point of this case study is that, for certain disciplines dealing in large datasets, the establishment of large facilities facilitates publication of new data. Researchers need to be able to access data to be able to use and reuse it.

2. Biosciences

The UK has played a leading role in the development of data resources which underpin the global bioscience research enterprise.



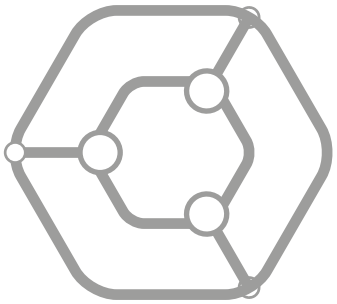
View the full case study on **page 18**

It is home to the **European Molecular Biology Laboratory European Bioinformatics Institute** (EMBL-EBI) and a key player in international initiatives such as **ELIXIR** and **Euro-Bioimaging**. There has been rapid growth in the volumes of data produced by researchers in the biosciences. This has been particularly evident in genomics, fuelled by the rapid fall in the costs and availability of DNA sequencing technologies. But across many other areas of the biosciences - from imaging to phenomics - new technologies are making it easier to generate increasingly large volumes of data. The increasing volume, heterogeneity and complexity of biological datasets is creating major challenges for data analysis, stewardship and re-use.

The life sciences are evolving rapidly and data requirements are diverse and complex. Even for a single experiment, data formats may vary widely, and require a range of open source and proprietary software. There are challenges in determining how to characterize reliably and reproducibly the details of specific environments where researchers collect data. Decisions on what data to keep and to share openly may too be problematic; with the relative value of different types of data not immediately apparent. Often, the data generated by a single project does not fit within the remit of a single database; submission to multiple databases, each with distinct metadata requirements, reporting standards and submission systems may present serious barriers to adoption.

In fields such as bioinformatics and 'omics research (genomics, proteomics etc.), data sharing has for some time been accepted as the norm. Hence, progress towards open data has been more rapid than in many other subject areas; with a high degree of community support for infrastructure needed to enable this. Over the past four decades, a huge variety of knowledge bases and deposition databases have been established to serve the needs of the community. Communities like the **COMBINE** consortium, organisations like the **NCBI**, and more recently European research infrastructures such as **ELIXIR**, have provided a key focus for the development of policies, practice and the underpinning infrastructure for data sharing and open data. However, the lack of a sustainable model for funding, even for well-established core resources, is the main challenge with keeping research data open within biosciences.

The main learning points for the Task Force are that aggregation through large infrastructure services have been a success and had a lot of impact on sharing outputs and data. Biosciences leads in open data due in part to the need for cross-national collaboration. Stringent requirements around the outputs have led to development of policies and standards across fields.



View the full case study on **page 28**

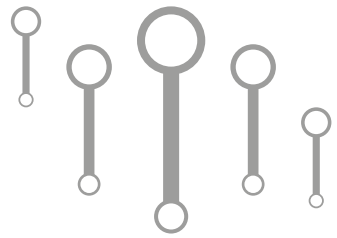
3. Crystallography

Within crystallography, research data is focussed around the chemical and molecular crystal structures. These are catalogued/deposited/archived within large searchable and mostly open databases, with some exceptions.

The crystallography databases are used increasingly, logging more than 7 million sessions in one year, demonstrating clear benefits to the discipline. Although some costs are underwritten by organisations like the **NIH** and the **Wellcome Trust**, and preservation is increasingly feasible from a technical point of view, a major challenge remains the sustainability of covering storage costs.

From the early inception of this discipline, it was common to include raw data in published papers. Since crystal structures and methodologies are prone to error, and it is important to have the correct images published, the key service that operates in this discipline is validation before deposit. The **International Union of Crystallography** (IUCr) leads on several such efforts, and has a relatively sustainable model for funding these.

The sustainability model for the crystallography databases is arguably one of the better models, proven by their continuous existence, some over 50 years, and with millions of users. The quality of the data is a big issue, and the Task Force understands that this has been addressed via validation services built into database and journal submission processes.



View the full case study on **page 34**

4. Digital Humanities

In digital humanities, research data is formed of texts, books, manuscripts, images, or other visual media. Thus, the heterogeneity, idiosyncrasy and complexity of the data are key features.

Whether data in the digital humanities is openly available is often a function of how it was collected; a publicly-funded digitisation project would usually result in openly available data; many scholarly editions are available with a licence. The inherent controversy in the meaning of "data" and the importance of personal interpretation on data for humanities researchers is not conducive to sharing. Skills in the handling of data are less widely and deeply distributed in the humanities than in most other disciplines. Institutions rarely provide facilities for depositing digital humanities data, in spite of the increasing amounts being generated.

Standardisation of metadata is an issue; several projects within digital humanities have attempted to address this issue with limited success. In the UK, four of the original five strands of the **Arts and Humanities Data Service** established in 1996 are still in operation, but, apart from the **Archaeology Data Service**, they are not in high demand. Low traction from the sector has resulted in difficulty building a sustainable model and demonstrating the economic benefits.

In the Digital Humanities, the Task Force has observed that research data is heterogeneous, complex, and varies in size. There are issues surrounding incentivisation as well as acceptance of the 'research data' concept. Beyond archaeology, the provision of open data services is uneven, and depends largely on the enthusiasm of small groups of enthusiasts. The sustainability of such services is open to doubt.

5. University of Bristol

University of Bristol is a research-intensive institution that has been building its research data management infrastructure since 2009 with **CAiRO**, a Jisc-funded project.

The university runs its own data repository, **data.bris**, and encourages researchers across all disciplines to publish their data via its institutional policy. Since 2015, when the policy was approved, all research data services, including the repository, transitioned to being centrally funded and operated from within the Library. The services are supported by 5 posts (3.8 FTE in total).

The university is among the leaders in the provision of research data services, and is working hard to generate greater awareness and take-up of those services. Research have the ability to decide which data sets they want to share, and, as internal storage is limited, are encouraged to deposit data in subject-specific repositories.

Due to long investment, leading to internal demand for services, the University of Bristol is at the forefront of research data management and infrastructure development. The Task Force has observed the value of central funding and ongoing permanent resources employed in running and driving the services.

View the full case study on **page 42**

6. University of Salford

University of Salford is a teaching-focussed institution. Salford implemented a research data management policy in 2016; the University Open Access policy briefly mentions depositing data.

Local policies and operating codes of practice have been mostly driven by receipt of a significant percentage of research funding from **EPSRC**. Salford researchers are increasingly aware of open access and the open research agenda. The Library employs one dedicated research data manager, who supports researchers. However, many questions and requests around IP, data protection and other issues are not dealt with centrally. Salford is a relatively small institution, and development of its policies and services has depended on dialogue and partnership with a range of institutional stakeholders. The RDM service remains in its early stages, and more work is needed to enhance its visibility and to demonstrate its relevance across the university.

The Task Force recommendations will need to address a wide range of institution types. In spite of its teaching focus, University of Salford has invested in a basic research data infrastructure for its small body of researchers, including full-time staff to coordinate activity. It has seen an increasing number of researchers proactively contacting their central resource about opening their research data.

View the full case study on **page 48**

7. Natural History Museum

The **Natural History Museum** in London holds millions of physical and digital items in its collections, which are themselves part of even larger collections within the UK and worldwide, with estimates of several billion items in total.

The aim of NHM, and generally of the discipline of biodiversity, is to document the diversity of life on earth, which involves continuous and systematic classifications of organisms within taxa and by geography. This enables NHM to have a comprehensive metadata (taxonomy) for each item. Moreover, the actual names (taxa/codes of nomenclature) are published widely since this is a disciplinary norm. Several initiatives, involving the NHM as a partner with other museums, HEIs and funders, are currently dealing with the key challenge to consolidate the nomenclature.

The NHM's policy is to release all its data with a CC0 license and images under CCBY, immediately, with the usual exceptions for sensitive, commercial or confidential information. NHM also shares its collections via several citizen science projects: **Herbarium@home**,

View the full case study on **page 52**

zooniverse, **Notes from Nature**. The **NHM's Data Portal** is open to everyone both to encourage innovation and better research, as well as for contributing to and correcting errors in the 8 million records, which amount to less than 5% of all its collections. The museum is working on a number of mass digitisation projects, which will bring further challenges around the scale and volume of data, in terms of handling and making it open.

The Task Force has seen that the Natural History Museum both relied on and drove the open science agenda for the discipline of biodiversity. The inherent nature of this field enabled NHM and researchers in biodiversity to build databases and data infrastructures that open the data to other academics and the wider population.

8. Germany

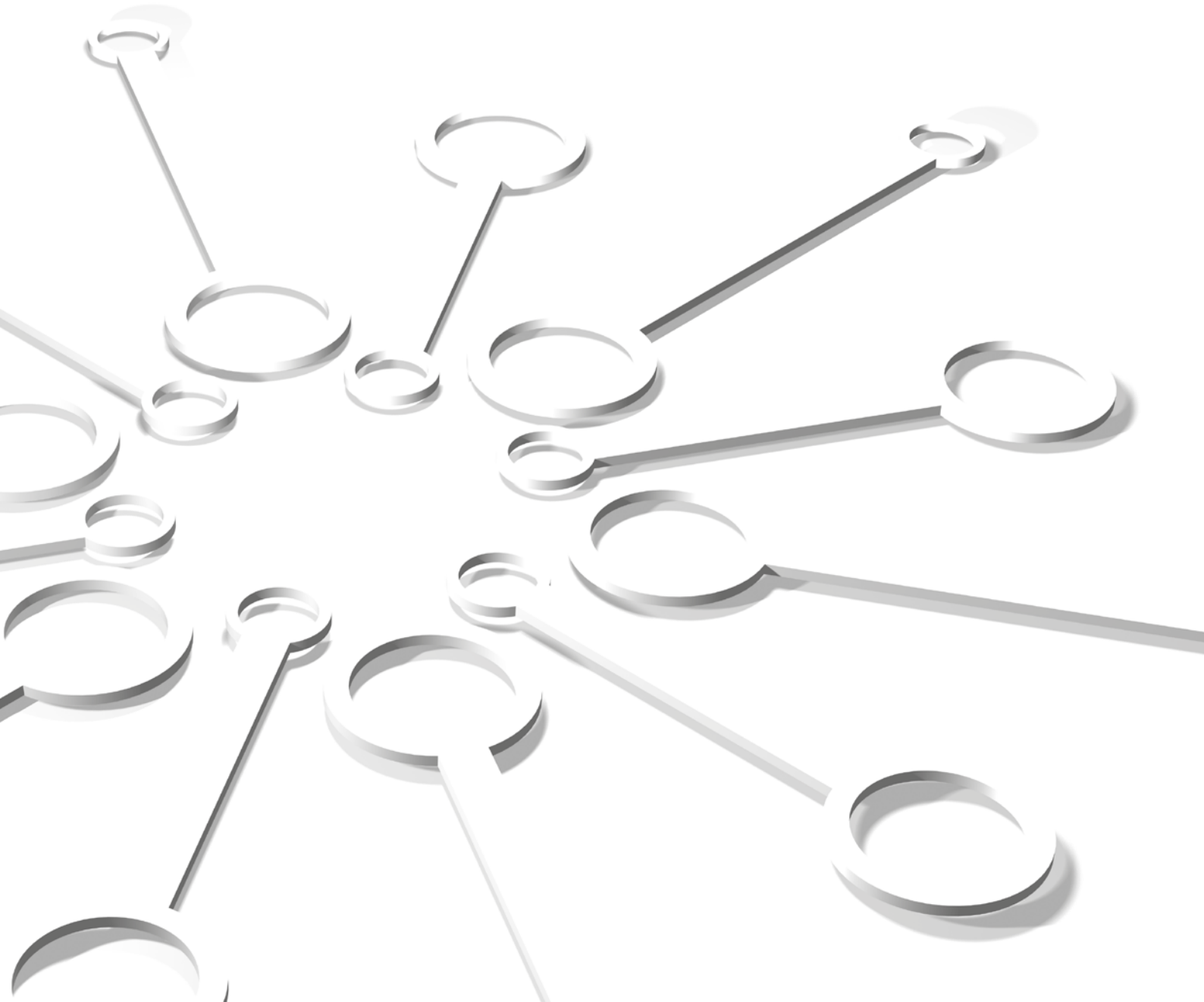
Germany has a complex range of organisations involved in funding and undertaking research.

View the full case study on **page 58**

Major funders include the **DFG**, the **Federal Ministry of Education and Research**, **DAAD**, and the **Alexander von Humboldt Foundation**. Around 100k of Germany's 360k researchers work in universities; but a large proportion of German research is undertaken in the institutes of organisations such as **Fraunhofer-Gesellschaft**, the **Helmholtz Association**, the **Leibniz Association**, and the **Max Planck Society**, as well as those run by the Federal Government and the Lander.

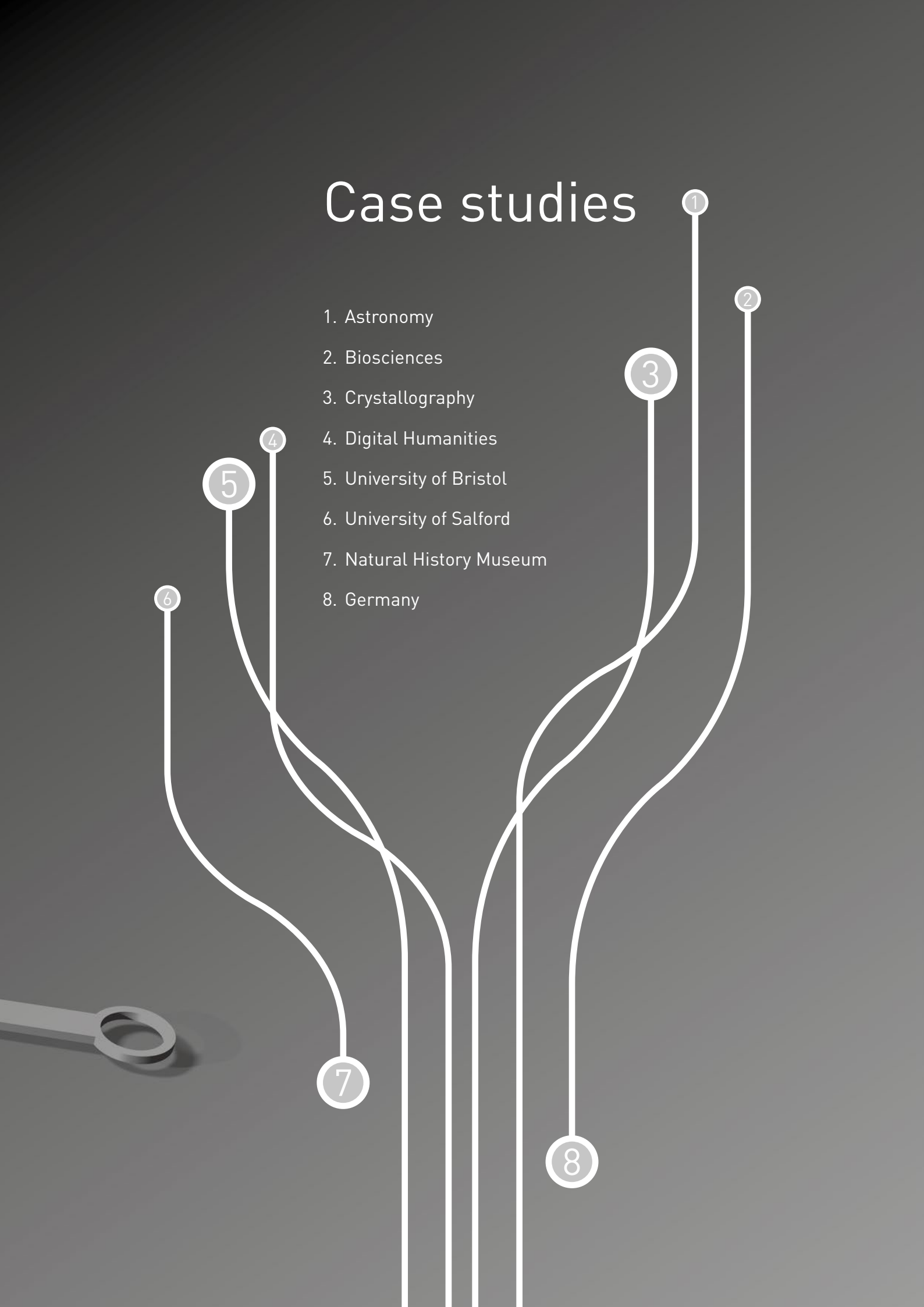
The Alliance of German Science Organisations and a range of other bodies have been active over the past fifteen years in producing statements and position papers on research data, and the Federal Government's **digital agenda** 2014-2017 calls for better access to research data as a goal. But a recent **report** suggests that there is a lack of strategy and co-ordination among project-based initiatives which tend to have a strong niche focus. Universities such as Bielefeld, Gottingen and Humboldt have adopted policies and principles to promote good management of research data; and some have also developed a range of data services to that end. Organisations such as the Max Planck Society and the Leibniz Association have also established repositories and related services. Some of the German Academies have also played a prominent role in developing data services, with particular support for digital humanities. German researchers and data specialists have also been active in a number of international initiatives, including the data infrastructure projects sponsored by the European Strategy Forum on Research Infrastructures, and in the development of the **European Open Science Cloud** (EOSC) and related initiatives such as **GO-FAIR**.

The **German Rectors' Conference** and a wide range of other organisations have called for the development of a distributed but co-ordinated National Research Data Infrastructure, with long-term funding mechanisms and strategies for training and skills, and closer networking with international organisations and initiatives. Many of the challenges in moving towards that goal are similar to those faced in the UK and other countries: the balance between desirable diversity and undesirable fragmentation; the sustainability of valuable bottom-up initiatives; the need to develop norms and standards that take account of the practices of different disciplines; and the balance between project-based and infrastructural funding mechanisms. The complexities of the German research landscape are different from those of the UK; but the challenges are very similar.

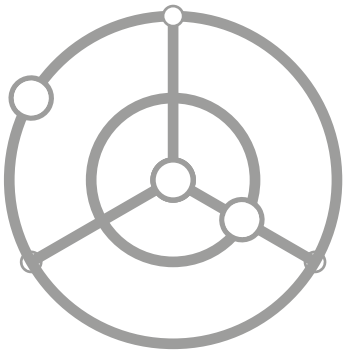


Case studies

1. Astronomy
2. Biosciences
3. Crystallography
4. Digital Humanities
5. University of Bristol
6. University of Salford
7. Natural History Museum
8. Germany



1. Astronomy



1.1 Background

Astronomy is one of a handful of disciplines with a long-established tradition of data sharing. Along with genomics and crystallography, it is an example of an area where community norms of behaviour towards research data are well-understood and entrenched. Astronomy databases incorporate star catalogues based on observations of the night sky going back to antiquity (Borgman et al 2016).

In more recent times, large, internationally-distributed, research infrastructures have become a feature of astronomy, notably in the form of telescopes. Astronomical data typically includes images, spectra, time-series data, and simulation data. Raw data are captured by scanning photographic plates or digital detectors recording objects or portions of the sky. They are transmitted to **data centres** which curate the data and make them available through **web-based catalogue services**.

The scale of this infrastructure, with correspondingly large, distributed and interdependent research teams, has fostered a highly-collaborative research culture. This supports the sharing not only of research data, but also instrumentation, research tools and services, enabling far more comprehensive access to astronomical knowledge than could be managed in any given local environment (Borgman 2010, Kitchin 2014). Thus international collaboration is a long-established feature of astronomy.

1.2 Data from international facilities

The **Sloan Digital Sky Survey** (SDSS), now in its fourth phase[1] is seen as one of the most ambitious and influential surveys as well as being the most successful and the most cited survey in the history of astronomy. It promoted openness from its inception: the first **Principles of Operation**, in 1989, stated that "a reliable and easily utilized data base [...] will be made available to the public".

1. Of SDSS's 26 fully affiliated organisations across the world, two are from the UK: University of Oxford and University of Portsmouth. In addition, SDSS runs a number of National Participation Groups, including one from the UK, consisting of the following universities: Liverpool John Moores, Cambridge, Edinburgh, Nottingham and St Andrews. **http://www.sdss.org/collaboration/affiliations/**

SDSS thus issues annual data releases, along with online data access tools, each suited to particular needs. Data are made available relatively quickly, with only a short proprietary period to clean them and prepare them for release (Sands and Borgman, 2016).

A recent study found that the SDSS community identified a number of benefits arising from open data, including;

- improvements in the efficiency of science;
- increases in the volume and quality of feedback from peers;
- alignment with the policies of funders; and
- improved engagement with amateur astronomers.

Engagement with the public is a feature of astronomy **recognised by the Royal Astronomical Society**; it is one of the few sciences where amateurs make significant contributions to research, notably through telescope observations. More recently, citizen science initiatives have made astronomical data – largely in the form of images – publicly available to enlist the help of amateurs with identification and classification of celestial objects. **Galaxy Zoo** is perhaps the best known of these initiatives, but there are many others, such as **Planet Hunter** and the **list of citizen science projects** run under the auspices of NASA.

The astronomy community's commitment to open data is also exemplified by the **Large Synoptic Survey Telescope** (LSST), a ten-year survey, due to start in 2023, which is expected to collect 60 petabytes of data over that period. "LSST will be like a giant 'search engine' of the sky, digitizing and making available in a non-proprietary database the locations, motions, and characteristics of 20 billion galaxies and 20 billion stars". Unlike SDSS, LSST expects to release data immediately, without a proprietary period for the project's investigators. However, funding conditions and the need to ensure returns on investment mean that it will offer different levels of data access determined by the partnership level of each contributing country and/or institution (Sands and Borgman 2016). Nevertheless, "LSST has been designed as a **public facility** from the beginning [...] [it intends] to develop research projects that can be done by students in classroom settings, at home, and via science museums with the public".

The largest facility of all will soon be the Square Kilometre Array (SKA), an international initiative with 11 member countries with headquarters at the UK's Jodrell Bank Observatory. The aim is to build an array or collection of radio telescopes in Australia, South Africa (and in due course, in eight other African countries) with around one million square metres of collecting area, designed to study the Universe with unprecedented speed and sensitivity. Construction is expected to start in 2018, and will take around five years, but with early observations from 2021. It is expected to **collect raw data** amounting eventually to around 62 exabytes, and it is not yet clear how access will be granted to such unprecedented volumes of data, or exactly how it will be archived and curated.

Deposition of data in data centres is generally a condition of access to any large facility in astronomy, and hence has become common practice across the discipline. But in some cases, researchers may be given a period of exclusive access to their data for a proprietary period. Thus the **European Southern Observatory** allows for an exclusive period usually of one year.

1.3 Data archiving and data centres

Ten years ago, in its publication '**Portals to the Universe'**, NASA explained the critical importance of data archiving and curation in the operation of its data centres, and by implication for astronomy in general. The principles that NASA set out remain true to this day: the imperative of long-term, sustainable preservation as part of the core mission of astronomy facilities; the capacity to accommodate a rapid turnover of scientific results and to provide rapid access to them; the importance of good metadata and documentation to ensure the long-term accessibility and usability; and a recognition that data curation and provenance are labour-intensive processes which are major challenges in their own right.

In response to the OSTP Memorandum on Increasing Access to the Results of Federally Funded Scientific Research, NASA published in 2014 a **Plan for Increasing Access to the Results of Scientific Research**. This includes principles and requirements on the management of research data, with the aim of "[extending] NASA's culture of open data access to all NASA-funded research".

In 2016 it inaugurated a **public web portal for research results**, which provides easy access to:

- its **Data Portal** (incorporating the **Open Data** site), a registry of datasets generated through NASA-sponsored research, as well as open source codes;
- **PubSpace**, the repository (hosted by PubMed Central) where all NASA-funded authors and co-authors are be required to deposit copies of their peer-reviewed scientific publications and associated data.

Astronomy data centres are numerous and widely scattered across the globe. They are too numerous to list here, but include well-established facilities such as the **Centre de Données astronomiques de Strasbourg** (CDS – Strasbourg Astronomical Data Centre), which has been collecting and distributing astronomical data and related information since 1972. The **re3data** registry currently (August 2017) lists a total of **147 such centres** worldwide in astrophysics and astronomy – nearly 8% of the total number of centres across all disciplines. Those located or managed fully or partly in the UK include:

- **CHIANTI**, an atomic database for spectroscopic diagnostics of astrophysical plasmas;
- **LEDAS**, the Leicester Database and Archive Service, which deals mainly with data from high-energy astrophysics missions. Leicester is also a partner in **ROSAT,** a German X-ray observatory;
- **UKSSDC**, the UK Solar System Data Centre, an STFC and NERC jointly-funded central archive and data centre facility for solar system science in the UK, based at Rutherford Appleton Laboratory; and
- the **World Data Center for Geomagnetism** dealing with digital geomagnetic data as well as indices of geomagnetic activity from a worldwide network of magnetic observatories, based at the University of Edinburgh.

The Royal Astronomical Society's **list** of astronomical databases and archives includes a number of additional facilities, including:

- **UK Astronomical Data Centre** (formerly RGO Astronomy Data Centre), part of the Cambridge Astronomy Survey Unit (CASU), with data from the UK's ground-based telescopes; and
- the **AAT (Anglo-Australian Telescope) Archive Database** with an online index to the data from the Anglo-Australian Telescope.

Much of CERN's data is also relevant to astrophysics and astronomy. Its **Open Data Portal**, launched in 2014, provides access to a growing range of data; and it disseminates the preserved output from various research activities, including accompanying software and documentation.

Such data centres are major sources for astronomical data, but discovery may also occur through publications from which data is linked (Henneken 2015). This underlines the importance of bibliographic databases, and notably the Smithsonian Astrophysical Observatory/NASA **Astrophysics Data System** (ADS), a service providing access to over 13 million records covering publications in astronomy, astrophysics and physics, including arXiv e-prints. As outlined above, NASA's Data Portal also serves as a registry, albeit specifically for the outputs of its own funded research.

1.4 Big data: challenges and opportunities

The rapid and huge increase in the scale of data collection has meant that the size of repositories has increased to petabytes and more. The challenges and exigencies of big data have in themselves made research collaboration indispensable (Zhang and Zhao 2015). The amount of collected data has also increased hugely: the **first annual release** of SDSS data, in 2001, amounted to just 2.8 TB, compared to 156 TB for the latest release, a more than fifty-fold increase. Moreover, astronomy big data is defined not just by its scale, but also by the speed of producing, transmitting, and analysing it. LSST will collect 15 TB of data every night, requiring a capacity to manage an enormous daily turnover of data.

The processing capacity will be even bigger for SKA, which has recently signed **an agreement with CERN**– another generator of huge volumes of data – for collaborative working in exascale computing and data storage.

The huge quantities of astronomical data pose problems with regards to the capacity of astronomers to analyse and synthesise it. The problems are less about capacity and storage than the need to develop tools and techniques for handling this rapidly-accumulating output. It has been suggested (Dillon, 2015) that the four biggest challenges are:

- data visualisation;
- creation and utilisation of efficient algorithms for processing large datasets;
- the efficient development of, and interaction with, large databases; and
- the use of machine learning methodologies.

This view is confirmed by the '**Open to all?' case study** on the VO, which suggests that checking, characterising and managing the increasing volumes of images collected have become major issues in astronomy. But it also points out that while large facilities and funders require researchers to make their raw data accessible (perhaps after an embargo), they do not impose the same requirement for derived data, which is much less commonly made publicly accessible.

Data mining plays a crucial role in enabling these processes, through a range of different tasks: summarization, classification, regression, clustering, association, time-series analysis, and outlier/anomaly detection. For each of them there are approaches, such as artificial neural networks, which themselves relate to applications that are specific to astronomy. These include, for example, spectral classification of stars, galaxies, quasars and supernovas; stellar physical parameter measurement; and special or rare object detection. Astro-informatics and astro-statistics have emerged as disciplines to help solve the complexities associated with big data in astronomy. But again the analyses may not always be made publicly available.

1.5 Standards and interoperability

Central to astronomy's collaborative culture is the **Virtual Observatory** (VO). This allows astronomers to interrogate multiple data centres and datasets in a seamless and transparent way, provides new powerful analysis and visualization tools within that system, and gives data centres a standard framework for publishing and delivering services using their data. The VO's practical embodiment is the **International Virtual Observatory Alliance** (IVOA), formed in 2002, which debates and agrees the technical standards for data description and access that are needed to make the VO possible (it has focused on standards more than on building data centres). IVOA brings together 19 national and 2 European transnational member organisations, of which **AstroGrid** is the UK participant.

Agreements on data standards were developed in the 1970s and widely adopted by the later 1980s as part of the transition from analogue to digital astronomy. Such standards are crucially important because the data from different telescopes or projects have their own formats, which can cause difficulty in integrating data from different sources for analysis. In general, each data item has a thousand or more features, which causes a large dimensionality problem. Developing an infrastructure of software and standards is therefore an essential underpinning for astronomical research, and openness has been a key characteristic of that work. Astrogrid and IVOA support this through the development of standardised data formats, analysis tools, resources, and registries that identify where these resources are located. Thousands of standardised resources have become available through VO registries.

Metadata is standardised around the **Flexible Image Transport System** (FITS) standard, originally developed in the late 1970s and now in widespread use. It encodes essential information about the instrument, conditions of observation, wavelength, time and sky coordinates in a standard data format. FITS has evolved over the years, encompassing more complex data structures arising from the use of new instruments, and providing support not just for images, but also other outputs including spectra, data cubes, text tables and binary tables (Hanisch, 2001). The more recent **VOTable** format is an XML standard for the interchange of tabular data. Both FITS and VOTable are open standards, maintained through community efforts.

1.6 Data publication

The Royal Astronomical Society set out its **publishing policy** in 2012. This does not explicitly address research data, but the document states in general terms that the Society "supports the free availability of peer reviewed results and supports its authors in distributing such results through open sources such as ArXiv when appropriate."

The top ten rankings in the current (August 2017) Google Scholar **list** of titles in astronomy and astrophysics with the highest h5-index demonstrate the importance of ArXiv for both authors and reader: six of the ten journals are arXiv preprint titles. But although there is a commitment to open access, these journals do not prescribe any policy on the publication and/or depositing of data. The remaining four titles – the **Astrophysical Journa**l, **Monthly Notices of the Royal Astronomical Society, Astronomy & Astrophysics, and Journal of Cosmology and Astroparticle Physics –** have somewhat divergent policies. One requires authors to publish their data immediately on acceptance of an article, while two 'encourage' authors to deposit data and provide DOI links, and the fourth makes no reference to depositing underlying data.

Data journals have emerged over the past ten years or so as a means of securing the peer-reviewed publication of datasets, rather than or as well as conventional research articles. Candela et al identified 116 such journals in 2013. Perhaps surprisingly, not a single one of them related specifically to astronomy or astrophysics. Nor do astronomy dataset feature much in data journals, such as **Scientific Data,** with more general coverage. **The Astrophysical Journal Supplement Series**, however, publishes "manuscripts containing extensive amounts of data or calculations with relatively little analysis or interpretations, or manuscripts of very specialized interest". It has been published monthly since December 1996, with all material over twelve months old available open access.

1.7 Key actors and roles

The key players in the development of data policies and open data have been the various large facilities on which the practice of astronomy depends, and the realisation of astronomers themselves that their research depends on international collaboration. And astronomy data centres have followed suit, with the development of standards such as FITS to ensure interoperability.

Funders are also key players. NASA's influence as a major research funding agency is outlined above. In the UK, the Science and Technology Funding Council (STFC), with its **Astronomy and Space Science Programme,** does not run any data centres of its own. But its **Research Data Group** supports STFC facilities and programmes with the management of research data. STFC's **Scientific Data Policy** incorporates the RCUK principles on data management and sharing, which state that "publicly funded research data .....should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property."

1.8 Conclusions

There is a long history of data sharing in astronomy, driven in large part by the need for collaboration in the use of large infrastructures for the conduct of research, and the scope for setting standards and rules governing their use. There has developed in parallel an impressive set of international infrastructures – in the provision of which the UK has played a significant part – to handle the huge amounts of data arising from astronomical research. Recent developments, however, including the SKA, are posing new challenges in terms of volumes of data, and the capabilities and capacities needed to handle and to analyse it.

It is notable also that data sharing may not at present necessarily imply immediate provision of open data: in some cases it does (as with the LSST), but for other large projects such as the SDSS the data is released openly in annual tranches; and for smaller projects, an embargo may apply before data is made openly accessible. It is as yet by no means the common rule that data associated with publications is made openly accessible.

References

Borgman, C. L. (2010). **Scholarship in the Digital Age: Information, Infrastructure and the Internet**. MIT Press.

Borgman, C. L., Darch, P. T., Sands, A. E., & Golshan, M. S. (2016). **The Durability and Fragility of Knowledge Infrastructures: Lessons Learned from Astronomy**. In Proceedings of the 79th ASIS&T Annual Meeting (Vol. 53). Copenhagen: ASIS&T. **https://www.asist.org/files/meetings/am16/proceedings/submissions/papers/31paper.pdf**

Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015). **Data journals: A survey**. J Assn Inf Sci Tec, 66: 1747–1762. **https://doi.org/10.1002/asi.23358**

Dillon, M. (2015). **Big universe, big data, astronomical opportunity**. The Guardian, 25/06/2015. **https://www.theguardian.com/science/across-the-universe/2015/jun/25/big-universe-big-data-astronomical-opportunity**

Hanisch, R. J. *et al* (2001). **Definition of the Flexible Image Transport System (FITS)**. Astronomy and Astrophysics, 376(1), 359–380. **http://doi.org/10.1051/0004-6361:20010923**

Henneken, E. (2015). **Unlocking and sharing data in astronomy**. Bul. Am. Soc. Info. Sci. Tech., 41: 40–43. **https://doi.org/10.1002/bult.2015.1720410412**

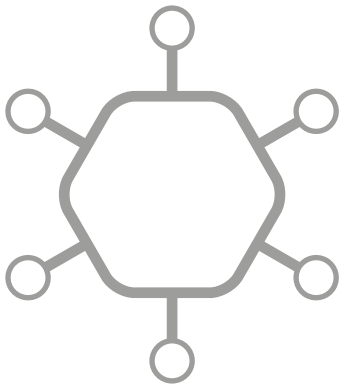
Kitchin, R. (2014). **The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences**. Sage. London

Pasquetto, I. V. *et al* (2016). **Open Data in Scientific Settings: from Policy to Practice**. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 1585–1596). New York, NY, USA: ACM.

Sands, A.E., Borgman C.L. (2016). **Open Data in Astronomy Sky Surveys**. Presentation to SciDataCon2016: Advancing the Frontiers of Data in Research. 11-13 September 2016, Denver, Colorado, USA. **http://www.scidatacon.org/2016/sessions/17/paper/272/**

Zhang, Y. & Zhao, Y. (2015). **Astronomy in the Big Data Era**. Data Science Journal. 14, p.11. DOI: **http://doi.org/10.5334/dsj-2015-011**

2. Biosciences



2.1 Background

There has been rapid growth in the volumes of data produced by researchers in the biosciences. This has been particularly evident in genomics, fuelled by the rapid fall in the costs and availability of DNA sequencing technologies.

But across many other areas of the biosciences – from imaging to phenomics – new technologies are making it easier to generate increasingly large volumes of data – with some projections suggesting that biological data will soon rival astronomical data in volume (Stephens et al., 2015). Because bioscience research tends to be more highly distributed than, say, particle physics, and the metadata more diverse, collecting and organising the data is a significant challenge. Moreover, bioscience researchers are also generating large numbers of small-scale datasets, and these have been increasing in numbers too. Many of these datasets have been collected or created manually, using diverse file formats.

Heterogeneity is a key characteristic of life sciences data. The so-called 'omics boom is not limited to genomics but embraces proteomics (protein structures and functions), metabolomics (chemical processes involving metabolites), transcriptomics (RNA molecules structures and functions), and metagenomics (genetic material recovered directly from samples). The increase in the volume of data across the whole spectrum of biology, leads researchers to integrate omics data of different types to inform hypotheses and biological questions, as well as combining with other data such as an organism's phenotype (observable characteristics) or phylogenetics (evolutionary history). Thus data from freely accessible biomolecular data resources are extensively reused for comparative studies, method development and to derive new scientific insights.

Data can take many different forms: raw and analysed data files, formalised results, models, software and tools, standard operating procedures and so on. For some data types, there are standard file formats (e.g. CIF format for crystallography), while in other cases, a variety of formats may be used (and these may change over the lifetime of a project). A wide range of analytical tools and techniques, algorithms and software may also be used, with a wide range of computational profiles. There are a wide range of metadata standards and ontologies for different kinds of data the consolidation of standards is patchy, which limits reuse and reproducibility.

In fields such as bioinformatics and genomics, data sharing has for some time been accepted as the norm, often with an expectation that data be shared openly rather than on a more restricted basis. In addition to requiring open access to research publications, most major funders now also specify that at least some kinds of data must be deposited in public repositories; and many publishers have followed suit. Several publishers have also established data journals to promote data access and use (Candela et al., 2015). Hence, progress towards open data has been more rapid than in many other subject areas; with a high degree of community support for infrastructure needed to enable this.

Services to support sharing of biomolecular information began in the 1970s, with the **Protein Structure Database** (PDB) which today contains over 125,000 structures. With the advent of DNA sequencing, nucleotide sequence databases

were also established, with the EMBL Data Library (now the **European Nucleotide Archive**) in Europe and **GenBank** in the USA. Shortly thereafter, the first protein sequence databases were established, which are now unified in the **UniProt** knowledge base at the EBI.

Since then, services have proliferated globally, reflecting the increased adoption of high through-put approaches and the range of technologies and data types. There are broadly three key types of service:

- In deposition databases, research communities have identified a collective need to collect experimental data using agreed standards to maximise citation and reuse.
- In knowledge-bases, various groups and organisations provide added value to the data in deposition databases to create new interfaces that facilitate browsing and discovery, thus saving researchers huge amounts of time.
- Tailored data management platforms for "in house" use.

The latest version of the **Nucleic Acids Research Molecular Biology Database Collection** includes over 1,600 databases, 54 of which were added in 2016. It identifies a hundred databases 'that have consistently served as authoritative, comprehensive and convenient data resources widely used by the entire community'. The majority are knowledge-bases, providing centralised access to data of many different kinds. Many are active in developing community standards such as controlled vocabularies, but the diversity in standards makes information harder to find and to use.

2.2 Key actors and their roles

The key organisations providing services in a coordinated fashion across the globe are the **National Center for Biotechnological Information** (NCBI) in the USA, and the **European Bioinformatics Institute (EMBL-EBI)** based in the UK. European initiatives including **ELIXIR**, **Euro-BioImaging** and EMPHASIS aim to consolidate services across Europe and have a significant footprint in the UK. Major international organisations like the Global Alliance for Genomics and Health (GA4GH) are seeking to develop common protocols to enable data sharing across the globe. Community grassroots organisations also play key roles - driving standards and promoting open data and data stewardship. Examples include the Computational Modelling in Biology Network (**COMBINE**) which coordinates the development of community standards and formats for computational models, the **Metabolomics Society**, and the **Genomic Standards Consortium**.

The heaviest concentration of openly-accessible databases and knowledge-bases globally is in the USA, followed by Europe and the Far East, with increasing numbers in China. Within Europe, the UK is by some distance the largest provider, ahead of Germany and France. The **re3data** registry records a total of 135 repositories in biology with UK involvement, and many are shared international enterprises. EMBL-EBI is a major focus of UK activity - running more than 100 such services, along with analysis tools, ontologies and other resources – such as the **Ensembl** genome database, **Array Express** and the **Gene Ontology** of gene functions and processes.

The EBI works collaboratively both in the UK and internationally – seeking to make data easily discoverable and usable via the web and cloud resources, with the use of APIs, scalable search technologies, and extensive cross-referencing between databases. These all present challenges as data volumes increase exponentially and new services are launched, with a continuing need for new storage and computational hardware. **EBI Search** has been developed as a scalable text search engine to resolve queries regardless of the volumes of data being searched. It makes use of the extensive cross-referencing and interactions between databases and knowledge-bases. The Resource Description Framework (RDF) **platform**, also allows a common query across different resources - enabling intuitive sharing of molecular data across different applications. Arguably even more important for deep scientific re-use are algorithms that can compute on data in structured formats – with the **BLAST** and **CLUSTAL** tools to align nucleotide and protein sequences being classic examples. As machine learning approaches gain traction, the need for large, clean and accessible datasets is essential.

For example, deep learning methods are revolutionising image analysis, but require large, carefully annotated datasets as training data. This gives biological images a dual role – both as raw data and as a resource to support development of analysis techniques.

In addition to the EBI, there are a variety of databases and services developed and operated by a rich variety of university-based research groups and institutes such as Rothamsted, Roslin and Earlham. Specialist datasets often arise from community developed resources close to research activities. Many collaborate with the EBI, use EBI archives or contribute to EBI knowledge bases. For example, **CATH**, a protein structures classification database operated and curated by University College London (see Box 1 below), makes a significant contribution to the **InterPro** protein sequence analysis and classification knowledge-base hosted by the EBI.

Building on its rich and diverse ecosystem of data resources and services, the UK plays a leading role in European initiatives that aim to coordinate and sustain life sciences data infrastructures. **ELIXIR** brings together 21 countries working together in a hub and nodes model to co-ordinate and build a single infrastructure for scientists to find and share life science data. It has three footprints in the UK.

- The ELIXIR Hub is based at the Hinxton Wellcome Genome Campus adjacent to the EBI and coordinates the work across ELIXIR and its national Nodes.
- ELIXIR-EBI node which coordinates ELIXIR's relationship with EMBL-EBI.
- ELIXIR-UK node, coordinated by the Earlham Institute, brings together 14 UK universities and research institutes outside the EBI to provide services, training and databases from the UK base.

The aim of ELIXIR is to coordinate and develop the collection, quality control and archiving of biological data across Europe; to improve the long-term sustainability of biological datasets; and to provide the services and capacity to ensure that they are FAIR (Wilkinson *et al.,* 2016). One of the key strands of work is to identify key data resources across Europe and support linkages between them, as well as between data and scholarly literature. Other strands cover computing resources to access, store and move data; tools to analyse it, development of standards and tools to enhance usability and interoperability, and training for researchers.

Other examples of major European initiatives with strong UK involvement include:

Euro-BioImaging - which provides open physical user access to a broad range of state-of-the-art technologies in biological and biomedical imaging for life scientists, and data support and training for infrastructure users and providers. It consists of a set of 29 geographically distributed Node Candidates (specialised imaging facilities) that can grant access to scientists from all European countries and beyond.

EMPHASIS - which seeks to provide comparable services to the plant and crop phenomics community. EMPHASIS is engaging 23 partner countries to help boost the exploitation of genetic and genomic resources available for crop improvement, with FAIR data principles at its heart.

2.3 Key issues and how they are being addressed

2.3.1 Data complexity and heterogeneity

The life sciences are evolving rapidly and data requirements are diverse and complex. Even for a single kind of experiment, data formats may vary widely, and require a range of open source and proprietary software. There are challenges in determining how to characterize reliably and reproducibly the details of specific environments where researchers collect data. Decisions on what data to keep and to share openly may too be problematic; with the relative value of different types of data not immediately apparent. In addition, a single project may generate data appropriate to multiple data repositories, each with distinct

metadata requirements, reporting standards and submission systems. Some large datasets can take weeks to prepare and validate for submission, and the effort required to follow best practices may present serious barriers to adoption.

2.3.2 Variety of repositories, databases and services

For data users, fragmentation of data across multiple sites, in multiple formats is a major barrier to re-use: 'if I can't find it or combine it with other data I can't use it'. Hence, there is overwhelming community support for the FAIR principles and the development of integrative tools to apply standards nationally and internationally.

ELIXIR has developed a process to identify a set of **core data resources** to help understand key data infrastructures, build trust among researchers, and move towards a sustainable funding model for these core resources (Durinx *et al.,* 2017). The process uses five categories of indicators aligned to the FAIR principles: scientific focus and quality; user community served; quality of the service; legal, funding and governance; and impact and translational stories. The initial list of core resources are predominantly EBI-based resources, although the UK-based CATH database is included (see Box 1).

More resources based outside of EBI will be added as the exercise is refined and repeated, and it is extended to include, for example, image data resources.

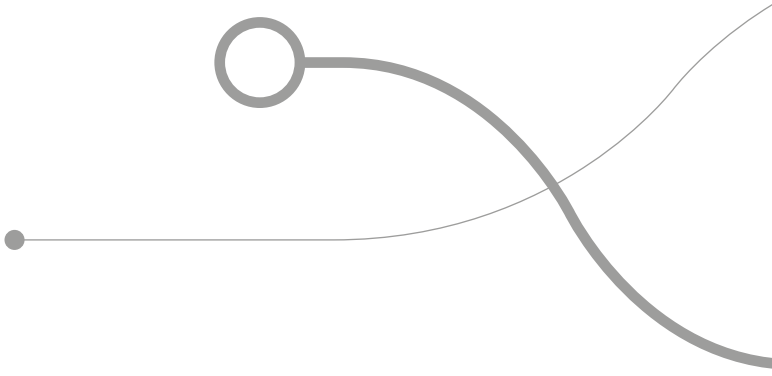
Box 1: The CATH Protein Domain Classification

The CATH Protein Domain Classification integrates protein data provided by a range of public resources to provide structural and functional predictions for proteins to the biosciences community.

With over 90 million entries, CATH is the most comprehensive protein domain classification resource in Europe. It is managed by a team led by Christine Orengo at University College London.

The added value of CATH is in application of computational algorithms to generate derived data on protein evolutionary and functional relationships. This allows reliable annotation of structural and functional properties from experimentally characterised proteins (less than 10% of known proteins are functionality characterised to uncharacterised proteins, to guide experimentation. CATH has also been used to rationalise antibiotic resistance and the impacts of genetic variations in driving cancer and other diseases. CATH depends on FAIR sharing of data from its key data sources (Protein Data Bank and UniProt) and its derived data complies with FAIR principles: it is widely disseminated via a dedicated website and through highly used international resources including the PDB and InterPro protein classification knowledge-base.

CATH is maintained and developed at UCL but is completely dependent on short term funding mostly from UK Research Councils. The rapid increases in the primary data mean that re-engineering of the computational platforms is constantly required. However, this type of activity is not supported by response mode Research Council funding and only very small pots of funds are available for increasing numbers of resources from focused calls. In addition, there is no proper career structure for the curators and software engineers needed to maintain and develop the resource, which makes it hard to recruit and keep a good team.



The many databases and resources that are not so far designated as core data resources are developed for a number of reasons, including:

- Specialist databases handling data whose structure could not easily be represented in the more general databases;
- Databases that store information about specific organisms or classes of organisms in a depth that could not easily be handled in general databases;
- Data intended for specific purposes, e.g. training sets for deep machine learning that combine image fragments with ground-truth annotations;
- Support or derivative databases that increase the value of the general databases by providing controlled vocabularies or enabling data aggregation and analysis
- Databases associated with other European intergovernmental organisations or communities ELIXIR has yet to fully engage with, such as bio-imaging resources.

Examples of such databases that operate in the UK include: the **Image Data Resource (IDR)**, to store and integrate and image datasets from published scientific studies; the **Pombase** database for fission yeast; the **Protein Circular Dichroism Data Bank**; **Phi-base** for pathogen-host interactions and the **Ligand Gated Ion Channel Database**.

The challenge of finding and accessing data amongst the plethora of services has stimulated several registry and discovery initiatives. Systematic identifier and naming systems are essential. The EBI's identifiers.org service and **Semantics as a Service** toolkit are widely used by academia and industry, for example in the **OpenTargets** application for target validation. **FAIRsharing** (hosted at University of Oxford) lists 1,029 data resources, cross-referenced with standards. In a search-based approach, ELIXIR sponsors the **Bioschemas** initiative which exploits the **Schema.org** standard for marking up and harvesting web content established by search engines such as Google and Bing. Using web-scale approaches for discovery and access is an increasing trend, including use of Cloud resources for hosting datasets, standard APIs and **Authorization and Authentication Infrastructure (AAI)** to allow single sign-on to services.

The Life Sciences pioneered the FAIR principles that have been strongly embraced by the European Commission, the European Open Science Cloud, the NIH Data Commons and the publishing community. International efforts are seeking to turn principles into practice, including development of a wide range of metrics to measure compliance (Bousfield *et al.,* 2016).

Seamlessly bridging "the last mile", from in-lab resources to international data infrastructures, helps drive curation practices "upstream" to the data source. Furthermore, project results scatter data across silos depending on their datatypes, discarding the experimental context from which they arose. Initiatives like **FAIRDOM** and **BioStudies** specialise in standards-rich cataloguing and collection making that span datasets silos and attempt to retain an integrated experimental viewpoint over different types of data, models and other kinds of results.

Training to increase the data skills capacity in the UK's Life Science sector is recognised as essential to academia and industry. For instance, the **BBSRC/MRC Vulnerable Skills Report 2017** highlighted concern over data analytics and data stewardship skills. The ELIXIR-UK Node provides the **ELIXIR TeSS training portal** as well as **Data and Software Carpentry workshops** in partnership with the **Software Sustainability Institute**. At this stage, however, there is no proper career structure for the curators and software engineers who are vital for maintaining and developing data resources, and attempts to enable the primary data creators to undertake curation have yet to gain traction.

2.3.3 Big data

As researchers generate increasingly vast data resources, integrating and analysing these data demands complex software tools and increased computing power. For some kinds of research, bio-scientists now need access to the high-performance computing facilities that were until relatively recently used

mainly in physics; and require the tools and the skills to exploit that power to the full.

In the US, the NIH Data Commons programme aims to address these issues - facilitating broad use of biomedical big data, developing analysis methods and software, enhancing training for large-scale data analysis, and establishing centres of excellence for biomedical big data. In Europe, the ELIXIR compute programme is developing distributed cloud, computing, storage and access services, working closely with four use-case communities to ensure that technical solutions meet their specific needs.

All data services are available or becoming available as containerised deployments (using Docker, BioConda or similar technologies) on private clouds such as the EBI Embassy Cloud or and public clouds such as AWS. Data generation, access and analytics is becoming cloud based. The scale of data precludes the movement of data across networks and instead co-locates processing with data in the cloud.

2.4 Benefits

The benefits of bioscience data repositories and services arise at several different levels. First, the community itself benefit from access to data, and the possibilities of gaining more information and understanding about the molecular components of cells and their interactions and processes without having to carry out laboratory work. Second, that knowledge can be exploited for the benefit of human health and well-being, whether through medical, agricultural or environmental applications.

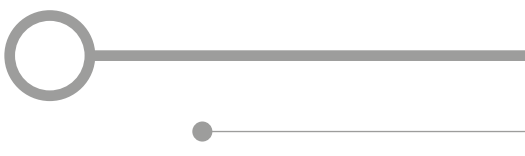
In a recent study of the value and impact of EMBL-EBI (Beagrie and Houghton 2016), more than half of survey respondents said that not having access to EBI services would have a 'major' or severe' impact on their research. It also estimated that the direct value of those services to users was between £270 and £320 million; and that the benefit in terms of making research more efficient was of the order of £1 billion, more than 20 times the direct operational cost of those services. Overall, the study estimated that use of the service 'contributed to the wider realisation of future research impacts conservatively estimated to be worth some £920 million annually, or £6.9 billion over 30 years in net present value'. It is anticipated that evaluations of other data services would show similar levels of return on investment, which is the basis of the ELIXIR's selection criteria for Core and Node services. A recent report (Garcia, Smith, Blomberg 2018) argues many SMEs in the Life Science sector would flounder operationally without public data resources.

While recognising the contribution of the EBI, it is vital to also acknowledge that much of the innovation, content, tools, expertise and know-how that feeds the work of EBI comes from the *"long tail"* of scientists and students in universities and research institutes. Large projects kick-start small ones and vice versa. And we will not have data scientists if we do not have academic groups producing them. We need a healthy ecosystem incorporating both large data centres and university-based resources and groups.

2.5 Costs and sustainability

It is difficult to get a reliable picture of overall costs for the various repositories, databases and services in the UK and across Europe. The ELIXIR survey showed annual costs for just over 150 databases amounting to around 35 million euros, and a total investment to date (since 2009) of just over 350 million euros.

Major data resources and services are funded by a variety of mechanisms, reflecting their individual histories. In most cases, the funding derives from national agencies, European and international funding, and major research charities. There are significant challenges relating to relatively short funding cycles, changing policies and priorities, and competition between supporting research infrastructures on the one hand, and research proposals on the other. The Open Microscopy Environment (see Box 2) provides one exemplar of the challenges of funding and sustaining a major UK-based data service.



Box 2: The Open Microscopy Environment

Since 2000, the Open Microscopy Environment (OME) has built open source interoperability tools for biological image data (**https://www.openmicroscopy.org/**)

It is led by Professor Jason Swedlow at the University of Dundee and has three main components:

- OME Data Model and OME-TIFF, an open metadata specification and file format for biological imaging (**https://docs.openmicroscopy.org/ome-model/5.6.2/);**
- Bio-Formats, a plug-in library for reading proprietary scientific image data and metadata into a common model (**https://www.openmicroscopy.org/bio-formats/);**
- OMERO, a client-server software platform for image data management and analysis (**https://www.openmicroscopy.org/omero/;** Allan et al, 2012).

OME's tools enable access to data, regardless of format, programming environment, or geographical location. They are used in thousands of academic and industrial labs worldwide. Bio-Formats is started →100,000 times/day worldwide and is incorporated in several open source and commercial products. Several commercial companies have adopted OME's open OME-TIFF file format and sponsored the development of new proprietary file readers. OMERO and Bio-Formats are used in several on-line public image data repositories, including the **JCB DataViewer**, the **CELL Image Library** and the **Image Data Resource**.

Like CATH, OME is dependent on a succession of short term funding awards from UK and EU agencies. To provide alternative support for OME development, OME has spun out a commercial arm, **Glencoe Software**, to provide support, services, and customisation for OME's software. An OME-licensed version of OMERO is the foundation for PerkinElmer's leading Columbus® image data management software. Glencoe's version of OMERO Plus provides data management solutions for customers in academia, biotech, pharma, imaging technology and scientific publishing.

The EMBL-EBI's Annual Report for 2015 shows that its incoming funds amounted to 67.2 million euros, of which a significant proportion went to fund the EBI's operational expenditure. Most of the EBI's funding comes from member states of the European Molecular Biology Laboratory (EMBL); but other major funders include the European Commission, the Wellcome Trust, BBSRC, US NIH and MRC. It also receives funding from industry partners and the Department for Business, Energy and Industrial Strategy.

The ELIXIR Annual Report shows an income in 2016 of 4.66 million euros, four-fifths in the form of contributions from member countries, and an expenditure of 2.92 million euros. It is important to note that ELIXIR does not fund the datasets, training, platforms or tools it counts as its services – these are supported by member states. The **Royal Society's snapshot of UK research infrastructures** estimates the annual operating costs of the UK Node as £6 million, which is met largely by UK funders. BBSRC, for example, provides funding for the **TeSS training portal** to the ELIXIR-UK Node as part of its commitment to ELIXIR, and co-funds the Node's other services (like CATH, and the Expression Atlases) in competitive bids.

Many smaller resources and services depend on the voluntary efforts and support from individual institutions and the best efforts of their investigators. These have a precarious existence. Thus although the overall number of databases recorded in the **Nucleic Acids Research Molecular Biology Database Collection** has remained relatively stable, each year several are noted as having ceased operation. An ELIXIR **survey** of database providers in Europe in 2009 found that of 200 respondents just under half depended on institutional support, 36% said that their funding was not assured, and 30% that it was assured for only one year. Only 6% had funding assured for five years or more; and just under a third said they were very concerned about their long-term sustainability.

The BBSRC extensively: supports data resources in the UK and at the EBI; supports the European Infrastructures; and is an advocate of open data and software. It recently initiated an analysis of its research portfolio with a focus on data intensive bioscience. BBSRC has two funding competitions dedicated to supporting data services: the Bioinformatics and Biological Resources Fund (BBR) and the Tools and Resources Development Fund (TDRF). The BBR Fund provides £6 million per annum – a level which has remained unchanged since it was established over a decade ago, despite increasing competition between new applications and renewals for existing resources.

The TRDF is a pump-priming fund that has defended itself from budget cuts. The Wellcome Trust funds an annual competition for Biomedical Resources and several of these awards fund UK data resources and software tools. Funding is also in principle available through response mode grant schemes, but there is a strong perception that data-oriented proposals do not fare well in competition with research grants. Given the importance of data-driven biology, the stagnation of funding for data resources and related analytic tools is a major concern. Furthermore, the annual nature of the key funding opportunities limits the ability of UK researchers to respond rapidly in this area.

2.6 Conclusions

In key areas of the biosciences, such as molecular biology, bioinformatics and 'omics research, open data and sharing have for some time been accepted as the norm. Over the past four decades, a huge variety of knowledge bases and deposition databases have been established to serve the needs of the community. Communities like the Metabolomics Society and the COMBINE consortium, organisations like the NCBI and the EBI, and more recently European Research Infrastructures such as ELIXIR and Euro-BioImaging, have provided a key focus for the development of policies, practice and the underpinning infrastructure for data sharing and open data; and they play an essential role in consolidating, integrating and enhancing access to the data that is being created in increasing volumes. These organisations are sustained in the main by Government and European funding, and studies have indicated the high levels of value they provide both to the research community and more widely.

The sustainability of all data services, even the well-established core resources recognised by ELIXIR, is a challenge. The EBI's presence in the UK is a great benefit to our open data landscape, but even there long-term sustainability of key data resources is not guaranteed. Moreover, there are many other specialist and valuable data services that serve UK communities and are key to the enrichment of major knowledge bases, and these are often even more vulnerable.

Contributors

Carole Goble (The University of Manchester, ELIXIR-UK); David Carr (Wellcome Trust); Helen Parkinson (EMBL-EBI, ELIXIR-EBI); Jason Swedlow (University of Dundee, Euro-BioImaging); Malcolm Bennett (University of Nottingham, EMPHASIS-UK); Christine Orengo (University College London, ELIXIR-UK); Gos Micklem (University of Cambridge, ELIXIR-UK); Nick Juty (The University of Manchester, ELIXIR-UK) and Tony Pridmore (University of Nottingham, EMPHASIS-UK).

References

Anderson, W et al (2017) Towards Coordinated International Support of Core Data Resources for the Life Sciences **https://doi.org/10.1101/110825**

Beagrie, N and Houghton, J The Value and Impact of the European Bioinformatics Institute **https://beagrie.com/static/resource/EBI-impact-report.pdf**

Durinx C, McEntyre J, Appel R et al. Identifying ELIXIR Core Data Resources [version 2; referees: 2 approved] F1000Research 2017, 5(ELIXIR):2422 **http://dx.doi.org/10.12688/f1000research.9656.2**

Ellenberg J, Swedlow JR, Barlow M, Cook CE, BPatwardhan A, Brazma A, Birney E. (2018) Public archives foir bioimagng data. **https://arxiv.org/abs/1801.10189**

A. Iudin, P.K. Korir, J. Salavert-Torres, G.J. Kleywegt & A. Patwardhan. "EMPIAR: A public archive for raw electron microscopy image data." Nature Methods 13 (2016). **http://dx.doi.org/10.1038/nmeth.3806**

Stephens Z.D., Lee S.Y., Faghri F., Campbell R.H., Zhai C., Efron M.J., Iyer R., Schatz M.C., Sinha S., Robinson G.E. Big Data: Astronomical or Genomical? PLoS Biol. 2015; **https://doi.org/10.1371/journal.pbio.1002195**

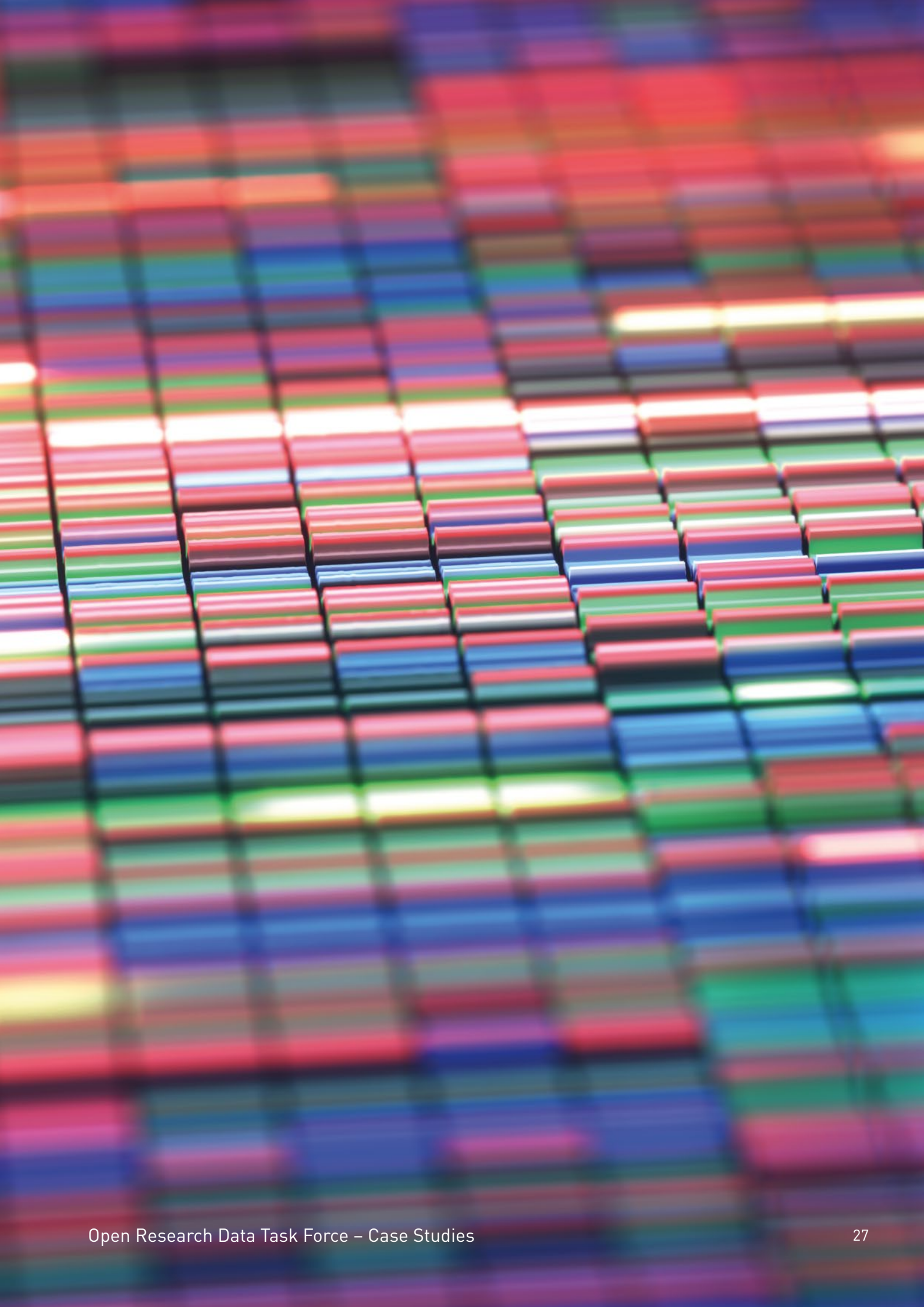
Wilkinson MD et al, (2016) The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3, DOI **http://dx.doi.org/10.1038/sdata.2016.18**

Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, Leo S, Antal B, Ferguson RK, Sarkans U, Brazma A, Carazo-Salas R, Swedlow JR. (2017) The Image Data Resource: A Bioimage Data Integration and Publication Platform. Nat Methods. 14:775-781. **http://dx.doi.org/10.1038/nmeth.4326**

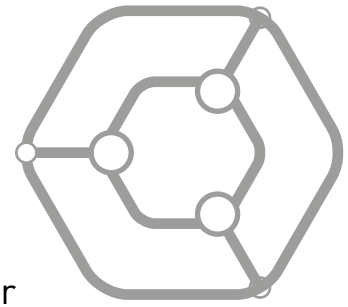
Bousfield D, McEntyre J, Velankar S, et al. (2016) Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. F1000Research. 2016;5:ELIXIR-160. **http://dx.doi.org/10.12688/f1000research.7911.1.**

Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015), Data journals: A survey. J Assn Inf Sci Tec, 66: 1747–1762. **http://dx.doi.org/10.1002/asi.23358**

Garcia PR, Smith A, Blomberg N (2018) Public data resources as a business model for SMEs. **https://www.elixir-europe.org/news/open-data-SMEs-report**



3. Crystallography



3.1 Background

In crystallography, research data use focusses on deriving the three-dimensional chemical and molecular structures of biological, organic, organometallic and inorganic compounds. This discipline of crystal structure analysis developed through the last 100 years or so.

It supports the advances in the bio and chemical sciences by enabling a better understanding of the compounds either to synthesise the new drugs, or to analyse an enzyme catalysis or understand molecular recognition. Crystal structures, including diffraction images and related data, are stored in a variety of databases across the world. Most of the results are open, but there are exceptions, which makes the analysis of structures in such cases limited to subscribers. Nevertheless, the crystallography community's **commonality of purpose in preserving and sharing its research data** is quite remarkable. The International Union of Crystallography (IUCr) is really the core part of this 'community driven institution'.

By John R Helliwell, School of Chemistry, University of Manchester, M13 9PL, UK
IUCr Representative to CODATA.

3.2 Organisations that drive the crystallography research data infrastructure

There are many facets to the Crystallography community activity for its research data. The main actors within this discipline are organisations such as:

- The International Union of Crystallography, which has an overarching role, coordinates discussions for standards and policies, and invests its financial surpluses from its journals for continuous sustainability and developments such as the automatic 'checkcif' (see section 3.4)
- The Cambridge Crystallography Data Centre (CCDC, a non-profit organisation that maintains and which supports the Cambridge Structural Database **in chemistry**)
- The Worldwide Protein Data Bank (wwPDB, funded by governments and research charities and oversees the Protein Data Bank **in structural biology**)
- The International Centre for Diffraction Data (ICDD, a non-profit organisation, ISO certified, that provides and charges for a number of services, including the largest database **for material science**, and provides bulletins and a peer-reviewed journal).

3.3 Where is the data?

The preservation of crystal structure analyses has been a core objective of crystallography, exploiting whichever digital storage medium of the age, and which started with supplementary data attachments to published articles such as the ones held in the archives of the IUCr. Indeed, the inventors of X-ray crystal structure analysis (the core method used to construct the three-dimensional images in crystallography), William Lawrence Bragg and his father

William Henry Bragg included the raw diffraction images and scans, their processed diffraction data details and the derived atomic coordinates in their first papers. The first crystal structure of sodium chloride was published by Lawrence Bragg in 1913 (W L Bragg 1913).

Today, most if not all the crystallographic research data is stored and archived in a series of databases. These have grown substantially, and proliferated, over the years, but I will describe the key ones that had the major impact in building and driving the infrastructure for this discipline:

The **Cambridge Structural Database** (CSD) is one of the oldest numeric scientific databases founded in 1965 under the leadership and initiative of the CCDC. This is the first database in crystallography, as well as the largest with over 900,000 entries. As Kennard 1997 observed, the preservation and access to data within this database was not limited by technology, as this was *"well able to keep pace with the exponential growth of information stored in the database"*, and its champions and developers indeed made sure that was part of the continuously improved operating model. The CCDC has been able to develop and maintain the database for more than 50 years now via user subscriptions.

The largest database for structural biology is the **Protein Data Bank**, launched in 1971 following CSD, holding more than 120,000 structures. Its sister databases are the Research Collaboratory for Structural Bioinformatics (RCSB) in the USA, in Europe - PDBe and in Asia - PDBj.

The largest database for materials science is the **International Centre for Diffraction Data's Powder Diffraction File**, with nearly 400,000 entries. This diffraction data archive is extensively used in materials identification for example in complex mixtures. This finds direct applications in industry and in forensic science. The database is not fully open and available to be purchased.

There are other crystallographic databases that have followed the leadership of the CSD and PDB, proving their model and methodology as the most sustainable. The largest being the Crystallographic Open Database (COD), for chemistry, with over 300,000 entries (**http://www.crystallography.net/cod/**). For full details see Bruno et al 2017 Table 1.

The CSD, the COD and the PDB are open data archives, i.e. researchers can access any individual structure within these databases, however only COD and PDB allow complete downloads. Complete downloads are harnessed for systematic analyses of more than one structure or for optimising crystallographic structures computationally (example: PDB-REDO project: **https://pdb-redo.eu/**). The CSD however does not make open its entire collection, hence structural systematic analyses are restricted to its subscribers, who also benefit from the software developed for the collection of the CSD's crystal structures.

The longevity of data preservation and sharing in crystallography is a highly notable achievement. This is testimony to its sustainability and the quality of the preserved data. The data has been made open wherever possible within the constraints of sustainability and data quality. As Olga Kennard emphasised though, there are *"new models that are evolving"*.

3.4 Ensuring archive data quality

The quality of the data is critical in crystallography.
Many checks need to be performed, automatically wherever possible, to ensure that first, you don't submit the same structure twice, and second, you are not submitting the wrong structure or incorrectly applied methodology. The IUCr led the development of the Crystallographic Information File ('cif') as a de facto standard to facilitate the growth of the CSD, and of the other databases. In addition, the cif approach standardised the description of crystal structures and thereby enabled the computerised automatic validation procedures to be implemented to enhance the quality of each of the crystal structure depositions (**http://checkcif.iucr.org/**). The PDB has launched in recent years a Validation Report based on checking the macromolecule cif, 'mmcif'.

In the case of chemical crystallography, the IUCr's journals for structural chemistry (Acta Cryst C and E and IUCr Data) also require making available the underpinning processed diffraction data and derived atomic coordinates for scrutiny by referees and editor. This sharing of data by authors with the referees and editor is in addition to the 'checkcif' procedure, with its 400 individual automatic checks, and allows the repeat of the authors' calculations if felt necessary. This very thorough approach based on openness and sharing of data seeks to guarantee data quality as true versions of record upon publication. Unfortunately, it is not replicated across all structural chemistry journals, where editors and reviewers evaluate the checkcif report alone.

In structural biology, as Helliwell (2017) identifies, the PDB offers a pre-validation service (**https://validate-rcsb-1.wwpdb.org/**) similar to the 'checkcif' described above. The macromolecular model is then deposited in the PDB. The structure is checked across all data bank entries of the wwPDB. When ready, the data files can be 'submitted', a formal step at which a PDB code is issued. A full validation report is then provided for the researcher to submit along with an article to a journal. Whilst a vital part of the refereeing process, unfortunately the PDB summary validation reports are not always sufficient to pinpoint the validity of an article's claims and molecular models based on specific electron density interpretations; this situation is exactly akin to the structural chemistry situation described above, although the biological macromolecules are naturally more complex in their 3D structure.

Any incorrect crystallographic database entries generate significant anxieties about the reproducibility of science within crystallography and are a real concern (Rupp et al 2016). Journal editors are the ones targeted for criticism because they do not follow thorough policies and guidance for checking the crystallography underpinning data for a submitted article. This is especially true for structural biology as described above. Thus, there is considerable interest within structural biology to replicate the peer review data validation model used in structural chemistry (see: **https://arxiv.org/abs/1704.08848**). This is not yet a mandated policy. The Wellcome Trust, for example, does have a stringent open access approach to publication requiring the release of research data upon publication. They do encourage sharing of data with referees (see: **https://wellcomeopenresearch.org/for-authors/data-guidelines** section 4.2.1). The International Council for Science (ICSU) recent report "Open Data in a Big Data World" (**https://www.icsu.org/publications/open-data-in-a-big-data-world**). fortunately also sees a role for journals to referee the data sets as well as the articles. The IUCr has provided a detailed Response to the ICSU report also emphasising both openness and quality of data (**http://www.iucr.org/iucr/open-data**).

Incorrect crystallographic entries can also be detected by analysing the overlapping structures among databases. A simple experiment of retrieving the entry for one compound from PDB and CSD can generate confusion when one database entry is correct while the other isn't. Such cases might be due to things like poorer diffraction resolution of protein crystallography compared to chemical crystal structure determination. There is now a concerted effort between PDB and CSD to remediate such errors; there is a CSD staff contingent based at the RCSB in Rutgers University working together on that. An example is given in Tanley et al 2016.

A further challenge within crystallography is the expansion of the concept of 'research data archiving'. The Diffraction Data Deposition Working Group (DDDWG) set up in 2011 and chaired by Professor John R. Helliwell, noted that it is *"increasingly important to deposit the raw data from scattering experiments; a lot of valuable information gets lost when only (processed) structure factors are deposited"*. Thus, new procedures and standards in archiving of diffraction data have been extensively investigated by the IUCr.

The dramatic improvements in digital storage capacity and various archiving options for crystallographers are making the preservation of raw data increasingly feasible as a new procedure for the IUCr community. These **raw diffraction data archives** complement the

existing databases, which have decided that whilst they can provide metadata details such as links to raw data sets via their DOIs, they cannot take on the costs of hosting the raw data sets themselves. The practicalities for crystallography have been summarised recently by Kroon-Batenburg et al (2017).

The ICDD is notable in having preserved raw powder diffraction data for some considerable time and has more than 10,000 such data entries. These they state can be useful e.g. in patent disputes as they can be more informative than one dimensional averaged profiles.

3.5 Costs and sustainability

Information on costs for running each database is rarely available. The two largest databases however have given a glimpse:

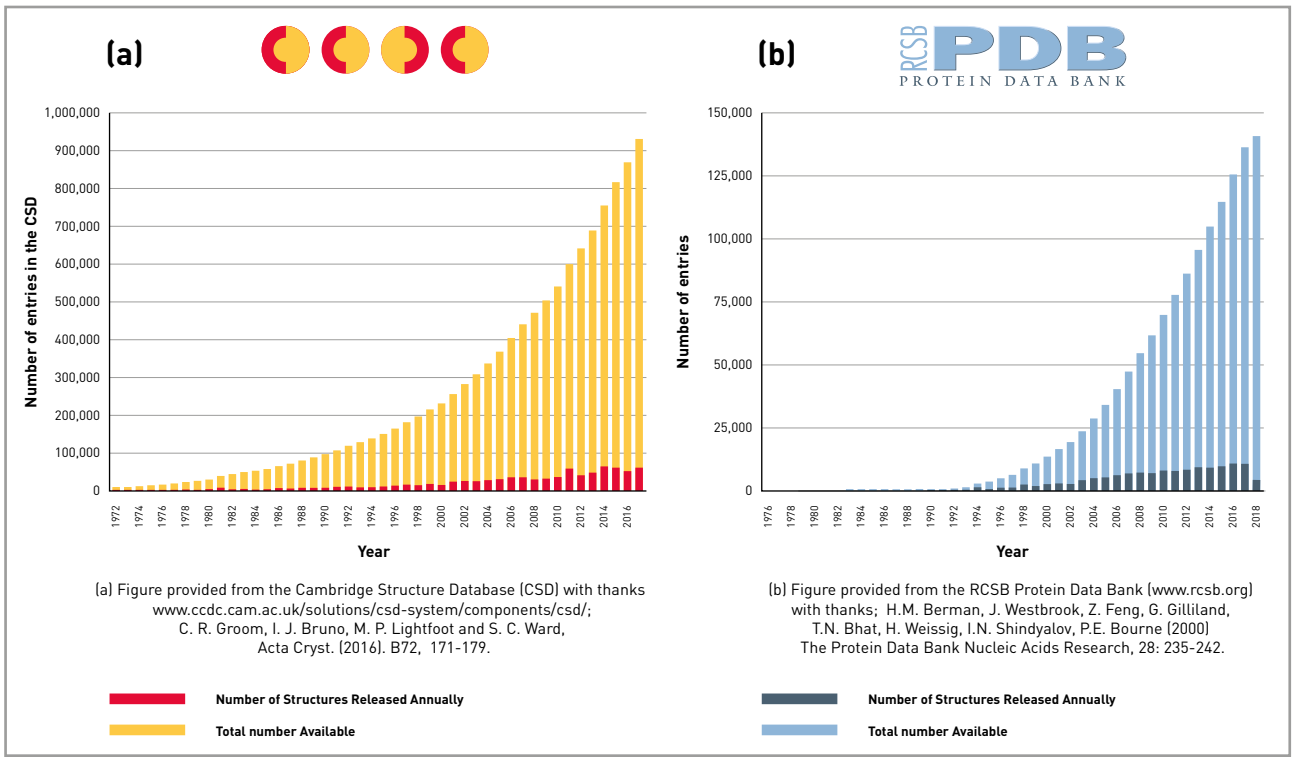
The RCSB-PDB's (the PDB in the USA) operational costs, including the costs of data creation and deposition, annotating and adding value to the data, and other expenses are stated as totalling $6.9 million per year. These costs are currently paid by granting agencies such as the NIH and The Wellcome Trust and there is no direct charge to depositors or to users (Sullivan et al 2017).

The CSD, which is run by the CCDC - a non-profit registered charity, is sustained by user subscriptions, but with an extensive outreach programme. The CCDC does not disclose its financial operation details to my knowledge, but it stated recently that it employs around 70 staff worldwide.

The current CSD and PDB rate of growth (Figure 1) indicates a likely increase in costs going forward unless more automation is feasible. However, most processes seem to have been automated already.

The ICDD finances the maintenance of its services from the sales of its powder diffraction data file products. It states that it is non-profit, hence it is expected that all proceeds are reinvested in the company.

Figure 1 Rate of growth of (a) the Cambridge Structure Database (CSD) and of (b) the Protein Data Bank PDB



(a) Figure provided from the Cambridge Structure Database (CSD) with thanks www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/; C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, Acta Cryst. (2016). B72, 171-179.

(b) Figure provided from the RCSB Protein Data Bank (www.rcsb.org) with thanks; H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) The Protein Data Bank Nucleic Acids Research, 28: 235-242.

3.6 Benefits and long-term impact

The benefits and impact of the crystallography databases are clear from their usage statistics:

The CSD states on its website that it is used by *"thousands of organisations in over 70 countries."* The USA RCSB-PDB estimates 295,465 users undertook 7 million sessions and 32 million page-views in 2016 alone (Sullivan et al, 2017). There are no estimates for its sister PDB in Europe and Japan, but similar figures could be implied.

3.7 Conclusions

Crystallography is a cross-disciplinary activity and this is also reflected in its range of databases (Bruno et al 2017). The wish for the preservation of crystal structures resulted in these databases, one of which now includes more than 900,000 entries. These entries are complementary to the corresponding articles in journal publications, very often underpinning them. Furthermore, there is a growing number of structural chemistry results that do not accompany a journal article but are directly deposited in the CSD, for example. The IUCr's launch of its open access journals Acta Cryst E and IUCrData are ways of retaining at least a short format description for every chemical crystal structure that is determined, hence driving the open data agenda further.

The discipline has made significant progress towards overcoming a range of key challenges: demonstrating proper validation services in advance of publication and deposit that are available both to the authors as well as the reviewers and editors; developing standards around the publication of raw data that are currently not available in most databases, and which will help with further ensuring reproducibility and developing improved methods in crystallographic science; and finally devising ways of tackling the increases in the volume of data and the need for continuous automated checks to ensure quality.

In terms of costs, examples include funding provided by the organisations like the NIH and The Wellcome Trust. The CCDC and ICDD are covering their operations and maintenance as well as their research and development through subscription costs and selling of services/ access to databases respectively. Discontinuing these databases would have a vast impact on multiple disciplines, a fact evident from the sheer number of unique users that visit the databases online and download the data.

References

Bragg, W.L. The Structure of Some Crystals as Indicated by their Diffraction of X-rays Proc. R. Soc. London, Ser. A 1913, 89, 248–277. **https://doi.org/10.1098/rspa.1913.0083**

Bragg, W.H. The X-ray Spectrometer Nature 1914, 94, 199–200. **https://www.nature.com/articles/094199a0.pdf**

Bruno, I., Gražulis, S., Helliwell, J. R. , Kabekkodu, S. N.,, McMahon, B. and Westbrook, J. (2017). Crystallography and Databases. Data Science Journal. 16, p.38. DOI: **http://doi.org/10.5334/dsj-2017-038**

Helliwell, J.R. (2017) New developments in crystallography: exploring its technology, methods and scope in the molecular biosciences Bioscience Reports Jul 04, 2017, 37 (4) BSR20170204; **https://doi.org/10.1042/BSR20170204**

Kennard, O. 1997 From Private Data to Public Knowledge In: Butterworth, I (ed.) The Impact of Electronic Publishing on the Academic Community. Portland Press Ltd. pp. 159–166. **https://doi.org/10.1017/S1062798700003057**

Loes M. J. Kroon-Batenburg, John R. Helliwell, Brian McMahon and Thomas C. Terwilliger IUCrJ (2017) Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements 4, 87–99. **https://doi.org/10.1107/S2052252516018315**

Rupp, B. , Wlodawer, A. , Minor, W. , Helliwell, J. R. and Jaskolski, M. (2016) Correcting the record of structural publications requires joint effort of the community and journal editors FEBS Journal, 283, 4452-4457. **https://doi.org/10.1111/febs.13765**

Sullivan, K.P., Brennan-Tonetta, P. and Marxen, L.J. Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank Rutgers Office of Research Analytics May 2017 available at **https://cdn.rcsb.org/rcsb-pdb/general_information/about_pdb/Economic Impacts of the PDB.pdf**

Tanley, S.W.M., Schreurs, A.M.M., Kroon-Batenburg, L.M.J. and Helliwell, J.R. Re-refinement of 4g4a: room-temperature X-ray diffraction study of cisplatin and its binding to His15 of HEWL after 14 months chemical exposure in the presence of DMSO Acta Cryst. (2016). F72, 253–254. **https://doi.org/10.1107/S2053230X16000856**

4. Digital humanities

4.1 Background

4.1.1 Data and digital humanities

Data is a problematic term in the humanities. Many scholars in the humanities would deny that they deal with data at all in the course of their research, while others would argue that if they do deal with data, it is in the form of texts in books and manuscripts, the images and other visual elements in drawings, paintings and photographs, and so on.

Digital humanities is a similarly problematic and contested term. Like researchers in other disciplines, the practices of all humanities scholars have been transformed by the digital revolution, but for the most part in much less fundamental ways than in, say, physics or engineering. But individuals and departments that describe themselves as operating in the digital humanities are typically engaged in the systematic use of digital resources and technologies, combining methodologies from traditional humanities disciplines with computational methods and tools.

Beyond a few areas such as literary and linguistic computing, data in the digital humanities, as in more traditional forms of scholarship, tends to be unstructured (as in most historical texts and images) or semi-structured (as in XML files), rather than structured data conforming to a clear data model and held in a relational database or spreadsheet. Much of the unstructured and semi-structured data comes in the form of digitised texts, images, video, sound and so on. Such data has different characteristics, and brings with it different challenges, from 'born digital' data. Where structured data does exist, it is typically extracted from existing texts, images and other sources, rather than from experiment or observation. And the extraction process can be complex and demanding; for the features and characteristics that scholars are seeking to analyse are seldom those around which those sources themselves are structured. The transformations and interpretations involved in the extraction process mean that data typically does not comprise discrete, fungible units whose qualities can be simply enumerated and which can be readily manipulated by computational means.

Many projects thus produce idiosyncratic or fuzzy data, and even when complementary projects have data regarding the same object (an image of a museum object or an inscription) it may have been described using very different metadata vocabularies or ontologies. Nor can it be assumed that someone else performing the same operations on the same data will produce identical results. Moreover, since it is commonly accepted that humanities scholars' individual perceptions and values are indelibly linked to the data and the scholarship they produce, there is much less interest than in other disciplines in notions of replicability and reproducibility.

For all these reasons, data sharing and open data, and the management and curation of data necessary to underpin them, have made much less headway in the humanities than in most other subject domains.

4.1.2 Data collections

Collections of data in the humanities take a number of different forms; and as we shall see, not all of them are open, or even readily shared:

Library, archive and museum collections

Over the past three decades, much effort has been put into digitising the texts and related materials in libraries and archives across the world, and producing digital images of museum collections. Funding has come from a number of sources including Jisc, the AHRC and the HLF in the UK, and the E-Content Plus programme across Europe. The resulting data – and associated metadata – takes many different forms, and it is rarely comprehensive: in libraries and archives in particular, vast swathes of their collections remain un-digitised. Thus the **Royal College of Music** has digitised a proportion of its collections and made them freely accessible online. Similarly, the British Library has made **one million images from scanned books** available on Flickr, though they represent but a tiny fraction of the BL's collections; the same can be said for the digital collections made available by the **National Archives**. The **British Museum**, on the other hand, has made its entire collections database available online and semantically. Some of the material digitised by or for cultural institutions has been aggregated into larger collections and is freely available. Examples include the **Perseus Digital Library** (built up over the last thirty years) of resources covering the history, literature and culture of the Greco-Roman world; and more recently Google Books and the **Hathi Trust**, both with a much wider remit covering the digitised collections of academic libraries. Metadata is aggregated by organisations such as **Europeana**. But much of the work of digitisation has been, and continues to be, undertaken by commercial companies which seek a return on their investment by charging for licences and access. Such data are thus not openly accessible, and licences and terms of access restrict the uses to which they can be put.

Project data

Many of the projects conducted by digital humanities departments such as those at **Glasgow**, **KCL** and **Sheffield** make their data freely accessible online, though as the volume and complexity of such resources grows, it is not clear how they will be indefinitely sustained. In a few cases, institution-specific collections of this kind integrate into a single content management system data that may vary hugely both by subject matter and in technical terms; but this is relatively rare. Large-scale projects such as **Old Bailey Proceedings Online** have received grants from a number of sources to enable them to develop their resources and access to them, and its data warehouse and API are examples of good practice.

Scholarly editions and databases

A key element in the intellectual infrastructure of the humanities takes the form of scholarly editions, catalogues raisonnés, 'calendars' of information from archival resources, prosopographies and so on. Until relatively recently, such resources were typically produced in printed volumes; now they are produced in digital form, and/or as datasets. But again, there are significant differences in openness and accessibility. The online version of scholarly edition of **The Cambridge Edition of the Works of Ben Jonson**, for example, makes only a small subset of the content openly and freely accessible; for access to the bulk of the content, a licence must be purchased. On the other hand, the full set of records and the text of the volumes of some of the long-term infrastructural projects supported by the British Academy, such as the **Corpus of Anglo Saxon Stone Sculpture** published by OUP, are freely available online; in other cases, however, such as the **Corpus of Medieval Stained Glass in Great Britain**, only the images and location records are freely accessible, not the scholarly apparatus available in the published volumes.

Curated subject collections

A number of researchers and groups have created digital collections of data created by others as well as themselves, and selected and assembled for a specific purpose or tailored to the interests of particular research communities. In the networked information environment, such collections are likely to become increasingly important as means of organising otherwise scattered and diverse sets of data and of providing contexts for engagement with that data.

Such collections may vary hugely in scope and scale:

- national and international repositories and digital libraries provided by services such as **Text Grid** (now part of **DARIAH**) or the **Archaeology Data Service** (ADS) in the UK; and

- smaller thematic aggregations such as the **Digital Scriptorium** which provides access to images of manuscripts from the Middle Ages and the Renaissance along with bibliographic and physical descriptions; **CLAROS**, based at Oxford, which uses advanced technologies to bring together and provide access to scholarly databases relating to the art of classical antiquity; or the **Cuneiform Digital Library Initiative**, which provides access to images and texts of Assyrian cuneiform tablets from across the world.

4.2 Key issues and how they are being addressed

4.2.1 Cultures and incentives

Even more than in other disciplines, cultures and incentives in the humanities are not conducive to open data. As noted in Section 4.1.1, many researchers do not even regard the materials and sources they work with as data; and the very label 'digital humanities' is an indicator that those who do work with data and computational techniques are often regarded, and see themselves, as distinct from the mainstream of work in their disciplines. Moreover, even within the digital humanities, there may be limited value attached to the scholarly efforts involved in creating datasets out of often intractable primary sources.

There is as yet little sign that across the humanities, such work attracts the same kinds of scholarly credit as attaches to the publication of major monographs and journal articles. Hence despite the advice given about submitting various kinds of outputs to the REF and its predecessors over many years, of the 39k research outputs submitted to the REF 2014 across the arts and humanities, just 53 took the form of databases or datasets; and it is rare for authors of scholarly books or articles to make the data underlying their publications accessible to others. This is unlikely to change until researchers feel confident that their peers will give the same levels of credit to datasets and other digital outputs as they give to more traditional publications. Thus the incentives for researchers to manage their data effectively, and to ensure that it is preserved and made accessible to others, remain significantly less than they might otherwise be.

4.2.2 Data skills

Skills in the handling of data are less widely and deeply distributed in the humanities than in most other disciplines. The **British Academy** suggested in 2012 that deficits in such skills were more evident in the UK than in some other countries, and that this posed a risk to the health and international standing of UK research in the humanities. Some efforts have been made by universities, funders and learned societies to address these issues, but it will take many years before they are fully resolved. And in dealing with them, it is important to take full account of the unique features of humanities data noted in Section 4.1.1. What works in the natural sciences may not work in the humanities, and services that support community-building in the humanities (as for example through the **Digital Classicist** wiki) are clearly important. The work of subject-specific interest groups of the Research Data Alliance, as in **history and ethnography**, and **linguistics**, may also have some impact. But as in some other disciplines, while there are often suggestions that digital and data specialists should be placed 'upstream' in the research process to work with humanities scholars on research design, data creation and curation, there is sometimes scepticism about this among digital humanities scholars. There is clearly a balance to be struck between ensuring that domain researchers have the capabilities as well as the capacity to handle the key aspects of data creation, management, analysis and curation on the one hand, and providing trained professionals in dedicated roles on the other.

4.2.3 Discoverability and accessibility

The evidence is clear that a large proportion of research data in the humanities is not effectively

curated in ways to ensure discoverability and accessibility. Most scholars who spend time in libraries and archives accumulate large personal collections of digital images that could be added to the online collections of the institutions that hold the original texts, images and so on, or to curated subject collections. But even if the scholars were willing to add them to institutional collections, there are no mechanisms to allow them to do so; and few incentives to undertake the work necessary to add them to existing subject collections. Moreover, given the complex nature of much humanities data, high-quality metadata – often at a highly-granular level – is essential; but even curated collections have hugely-variable standards of metadata. Some progress has been made, in the development, for example, of tools to create or edit metadata records for data being deposited in specialist collections (as in the **Manuscriptorium** digital library), and in normalising and providing central storage for metadata from different collections (as in the **Text Grid** project); and the **Virtual Manuscript Room** at Birmingham provides a syndicated RSS feed for all the metadata it creates.

Within specialised fields such as epigraphy, specialist metadata schemas (EpiDoc) have been developed and widely adopted; while in papyrology, where metadata about the same papyri in different collections may vary significantly, the **Integrating Digital Papyrology (IDP)** project is providing a model of data integration. At a more generic level, the **CIDOC CRM** ontology developed by the International Council of Museums and the International Committee for Documentation provides a structure for describing concepts and relationships used in documenting the cultural heritage; but adoption has been patchy. Varying metadata standards thus remain a huge issue. In archaeology, for example, varying descriptive practices have made the development of large data repositories especially challenging, as well as hindering discoverability. Similar problems exist across many humanities disciplines; an RDA **Working Group** is seeking to develop good practice standards across the empirical humanities.

There is also much to be done to develop and sustain integrated collections with explicit and transparent rationales and collection development policies; high-quality descriptions of collections, not least to help users to assess the authenticity and representativeness of the collections; active management and curation; clear and user-friendly access mechanisms; and measures to facilitate interoperability and easy integration into users' workflows. Action is required on all these fronts if aggregation services are to create resources with value added beyond the value of individual items or collections.

4.2.4 Usability and interoperability

The complex nature of humanities data brings challenges for usability as well as interoperability. The scholarly judgements and interpretation involved in the creation of many datasets means that it is imperative for information about such interpretations to be recorded alongside the data. Creating a scholarly digital edition of a historical or literary text (which may have many manuscript sources or editions), often involves thousands of scholarly decisions about the reading of individual words: about which text variants to highlight and which sources are considered more reliable than others. Similarly, in the creation of datasets from unstructured (and often intractable) sources the scholarly interpretations involved must be encoded and stored alongside other types of descriptive, technical or administrative metadata. And curators need to understand how the data is likely to be used by scholars in order to support active curation and to help plan for future use of the data.

Various standards have been developed for different humanities disciplines and the digital humanities more broadly, such as ArchaeoML for archaeology and the Text Encoding Initiative (TEI), which has been widely adopted for textual data. But in the humanities as elsewhere, the specific practices and needs of researchers in particular subject areas have brought a proliferation of standards. Even in archaeology (where data management and sharing is more firmly embedded than in many other humanities disciplines), it seems unlikely that a single standard is likely to be adopted by all researchers in the near future.

4.2.5 Big data

The Arts and Humanities Research Council (AHRC) has supported a small number of 'big data' projects in collaboration with Arts Council England and NESTA to explore how cultural and arts organisations can use the data they hold to extend their reach and develop new strategies. But relatively little has been done as yet to exploit the potential of large and diverse datasets across a subject field, in order to address new kinds of research questions with new kind of methodologies. Some have suggested that there is a tension between 'small data digital humanities' and 'big data digital humanities', and that the latter is for the present relatively rare (Kaplan 2015).

Most of the data created by researchers in the humanities – scholarly editions, linguistic corpora, image databases and so on – involves time-consuming scholarly effort. Machine learning and AI may in the future help with this; and automated methods are now being used to some extent in converting library catalogues and the like to digital form, and in the tagging of large data sets. But for the present, producing large sets of data requires huge amounts of manual work, and this is not scalable. Moreover, work of this kind has not fundamentally challenged or transformed traditional modes of scholarship. Moving to big data on the scale now seen in some other disciplines will pose new challenges, and require new methodologies and new approaches to research, not least in relation to whether, and if so how, such work will be made fully open.

4.3 Key actors and roles

As in other subject areas, the drivers for innovation and change come bottom-up from key individuals in the community and top-down from funders and policy-makers, with interplay between the two. The demise in 2008 of the Arts and Humanities Data Service (AHDS, see Section 4.4) resulted from lack of support from the community for a service funded by the AHRC in concert with Jisc. The success of the one part of the AHDS to survive and flourish (the ADS) stems from the support it has received from its particular community in archaeology. Scholarly journals in the humanities – even those specialising in digital humanities – have so far played little active role in promoting or stimulating data sharing or open data.

4.4 Costs and sustainability

The **re3data** registry of data repositories suggests that there are over 500 repositories worldwide with humanities data. Around 50 of those are based in the UK, though their sustainability is not known. This includes over 30 institutional repositories, and a slightly smaller number of subject or discipline-based collections, such as the **Romani Morpho-Syntax Database** (relating to the Romani language and linguistics) based in Manchester or the **Archaeology Data Service**. (But the registry is not comprehensive: it does not include some services mentioned above, such as the **CLAROS** service based at Oxford, the **Virtual Manuscript Room** at Birmingham, and other services and initiatives based overseas but with UK participation.)

The **Archaeology Data Service** (ADS) is in one sense the key remnant of the Arts and Humanities Data Service which was established in 1996. It was funded by Jisc and the Arts and Humanities Research Board (the precursor of the AHRC), but the service closed in 2008, after the AHRB decided that low levels of engagement from the arts and humanities research community meant that it did not represent value for money. The service operated mainly through five subject-based centres, for archaeology, history, literature and language, the performing arts, and the visual arts. Although four of the five remain in existence, three of them do so exiguously, with no funding other than some support from their host universities. The ADS, however, has continued to receive funding from a range of sources, and actively collects data from university-based researchers and more particularly from field archaeologists working for a wide range of public, voluntary and commercial sector organisations. Its business model is based upon charges to depositors based on a **costing model** that takes account of the time spent in processing the deposit. A recent study (Beagrie and Houghton 2013) suggested that the ADS received grants from funders and fees paid by depositors amounting in 2012 to c£1.2m a year.

Services other than the ADS depend either on support from their host institutions (for institutional services) or from project funding combined with some institutional support for subject-based services.

4.5 Benefits

There is a widespread view in the digital humanities that digital resources have inherently democratising potential; and that they present an opportunity for the humanities to demonstrate their relevance and value to society at large, in a context where many humanities disciplines feel under threat, but also where the public at large is seen as a key audience for scholarly work. Readily-accessible data can be more closely integrated and linked to published interpretations of that data in popular works as well as in scholarly publications and reports. But data and data services are valuable only when a significant user community attaches value to them. Sustaining data services thus depends on integrating them as fully as possible into teaching and research, as well as outreach into wider communities.

Relatively little work has been undertaken on the costs and benefits of data curation and open data specifically in the humanities. But a study of the impact and value of the ADS (Beagrie and Houghton 2013) indicated that it represented good value for money. As the accredited repository for the great majority of archaeological data in the UK, it has a strong user community who are aware of the value of the services it provides. *"For funders and depositors, the ADS is important for dissemination, impact, reaching the widest possible audience, and ensuring a long term legacy for their work."* In economic terms, the value was calculated at five times the costs of operation, data deposit and use. Whether such conclusions would apply beyond the specific circumstances of the ADS is less clear.

4.6 Conclusions

Data in the humanities tends to take more complex and heterogeneous forms than in many other disciplines; and much of it is not only unstructured, but idiosyncratic, fuzzy, and subject to a range of meanings and interpretation, many of which are contested. The provision of open data remains patchy at best. Most of the data with which humanities scholars work has been produced by or on behalf of a range of libraries, archives, museums, galleries and related institutions, rather than by researchers themselves, either via digitisation or through the creation of catalogue and similar kinds of data; and access to significant proportions of that data depends on the payment of subscriptions. Where researchers themselves create datasets, it requires at present considerable scholarly effort; and there are as yet few signs of cross-fertilisation between those working in this way and those producing the larger-scale catalogue and similar kinds of data in libraries, museums and so on. And the metadata required for effective discovery and re-use of highly-specialist data is itself highly complex, but metadata quality is often variable, and hindered by the use of competing standards.

There remains a key distinction in the humanities (and within the digital humanities) between data produced in the form of a resource for others to use in further research on the one hand, and data gathered or created in the course of research that leads to a scholarly article, essay or monograph on the other. In the former case, the data may be made available in accordance with FAIR principles (as with Old Bailey Proceedings, or Perseus) or it may not (as with the Works of Ben Jonson, or the digitised resources created via commercial partnerships). In the latter case, very few datasets outside archaeology are made openly accessible. The concept of open data has thus made relatively little headway in the humanities.

The provision of data services specifically catering for the needs of digital humanities scholars is also patchy, and take-up of those services is similarly variable. Other than in archaeology, provision depends either on generic services provided by universities, or on subject-specific services supported in large part by the voluntary efforts of individual groups of interested scholars. The scope for further growth of such services is limited, and their sustainability is subject to much doubt.

References

Beagrie, N and Houghton, J (2013) The Value and Impact of the Archaeology Data Service **http://repository.jisc.ac.uk/5509/1/ADSReport_final.pdf**

British Academy (2012) Society Counts: Quantitative Skills in the Social Sciences and Humanities: A Position Statement **http://www.britac.ac.uk/sites/default/files/BA%20Position%20Statement%20-%20Society%20Counts.pdf**

Kaplan F (2015) A map for big data research in digital humanities. Front. Digit. Humanit. 2:1. **https://doi.org/10.3389/fdigh.2015.00001**

5. University of Bristol[1]

5.1 Background

Bristol is a research-led university in the Russell Group, and achieves high rankings in the various world university ranking lists. Sustaining and improving its *"world-leading reputation for research"* is part of the vision in the University's **Strategy**, and two of its current strategic objectives are to "build capacity in world-leading research and to *"establish a limited number of specialist research institutes in which Bristol has the potential to sustain world-leading research of scale"*.

Bristol employs nearly 2,800 academic staff. It received in 2015-16 £149 million in research grants and contracts, and a further £47m in QR block grant[2]. Together these represented 34% of its total income. Over 1,100 Category A staff were submitted to the REF 2014.

1. The assistance of Zosia Beckles of the RDM Service at Bristol in collating information for this case study is gratefully acknowledged.

2. Figures from the University's Annual Report and Financial Statements for 2015-16; and from HEFCE final allocations of recurrent grants 2015-16.

5.2 The University's engagement with research data

The University first began to engage in a co-ordinated way with issues relating to research data management (RDM) when it secured funding for a series of projects funded by Jisc:

- the **CAiRO project**, in 2009-11, focused on the needs of the practice-as-research community in the performing arts (where Bristol is strong), with the aim of raising awareness, developing RDM skills, and developing standardised practices where appropriate
- a joint project in 2009-11 with Leeds and Southampton focusing on policies, guidelines and infrastructure for palaeoclimate research.
- the data.bris project, in 2011-13 aimed to create a research data repository, initially intended to cover the arts and humanities; but in the course of the project, its scope was extended to provide a repository service for the whole university.

5.3 Policy

The purpose of the University's **research data management and open data policy** is to provide guidance and support on the responsibilities of the University and its staff in managing and preserving research data. It covers all research conducted by staff and PGRs, and describes the responsibilities of 'data stewards' (normally PIs) on issues including the ownership of data (particularly when research involves external partners); the need for data management plans and for costs to be built

into research proposals; secure storage; licensing; the protection of research participants' interests; and preservation and access. The policy states that *"research data that a [researcher] feels underpins a published research output or will be of wider use to the research community should be deposited in the University's Research Data Repository (or other repository) in a form suitable for long-term retention and, where possible, wider publication"*. The policy also encourages the publication of data in non-proprietary formats wherever possible, and the recording of significant datasets in the University's research information system. But it does not amount to a mandate for open data. The University makes commitments in the policy to provide storage and repository facilities, but also advice and support, training and guidance. The policy includes explicit links to related policies on research governance and integrity, research ethics, and information security; to the RCUK Common Principles on Data Policy; and to guidance on data protection and freedom of information.

5.4 Key actors and their roles

5.4.1 Development of policies and services

Bristol was in a position to exploit the opportunities presented by the RDM programme established by Jisc a decade ago because it had already invested in data storage, and it subsequently appointed a senior research data librarian. Later developments, including the establishment of a centrally-funded research data service and the development of the RDM policy, stemmed from the active engagement of the PVC Research at the time.

Since 2015 the repository and the services surrounding it have been centrally funded by the University. The Jisc-funded projects were originally run by IT services, but RDM services are now located primarily in the Library, which has close working relationships with IT services (itself responsible for technical operation of the repository). It is notable that development of these services preceded the establishment of the University's RDM policy, which was drafted and approved by the Senate in 2015. The policy was thus based on practical experience.

5.4.2 Staffing and management of RDM services

RDM services sit in the Library as part of the research support team under a 0.8 FTE assistant director. There are 3 FTE staff (4 posts) in the library, plus a technical development post based in IT services. The senior research data librarian has worked on digital humanities and cultural heritage collections; and he took the lead in the CAiRO and data. bris projects mentioned in Section 5.2.

The services are overseen by a Research Data Storage and Management Executive chaired by a senior academic and with representation from academics as well as IT services, the Library, and the Research and Enterprise Development (RED) Office (which provides services on governance, contracts, programme management, research funding, policy and commercialisation among other things). There is also an operational committee with representation from key members of staff. There are said to be good levels of engagement from senior academics and other members of staff.

5.5 Profile of data services

5.5.1 Advice and guidance

The University provides an impressive set of online guidance material on a broad range of RDM issues. Effective linkages between the RDM service and the University's Research and Enterprise Development (RED) Office mean that the guidance starts from the general - 'how to prepare a good research bid' (in three broad subject areas) – before moving onto specific data issues including

- detailed guidance on meeting the **data management plan requirements** of each of the Research Councils and other major funders, and templates for using the Digital Curation Centre's DMPOnline service;
- **software management plans**, drawing on the work of the Software Sustainability Institute(SSI), and including advice on possible commercialisation;

- **identifying the costs of data management** through the lifecycle from creation and analysis to curation, dissemination, preservation and use;
- **research data evaluation**, with criteria for reaching judgements on the potential re-use value of data, and thus what should be preserved;
- **storing and using data**, including advice on **file organisation**, **automatic back-up**, and **use of the University's storage facility**;
- **sharing research data**, with model data access statements for use in publications, an interactive tool for the creation of bespoke statements, and more detailed guidance on sharing **data concerning human participants**;
- **dealing with sensitive data** relating to people or to animals, data generated or used under restrictive commercial research funding agreements, and so on; and
- the possibilities of **data commercialisation** and the protection of intellectual property.

Much of the guidance has also been built into an online interactive 'boot-camp' tutorial, first developed as part of the Jisc projects mentioned in Section 5.2, and drawing on material from other universities including Edinburgh and Oxford.

5.5.2 Training

The RDM service runs regular workshops on open research (working with the Open Access team), on the data requirements of funders such as EPSRC and AHRC, and on sharing ethically-sensitive data. It also participates in training events run by other services, including those for PhD students, and on grant-writing. It also provides bespoke training for faculties and departments on request (for instance at a recent Faculty of Arts retreat).

5.5.3 Using data

The RDM service works with experts in the **Jean Golding Institute** to provide advice on data science methods in research; and the **Advanced Computing Research Centre**'s HPC service can also provide data analysis support. IT services also provides database services for research projects using Oracle, Microsoft SQL Server and MySQL database platforms.

5.5.4 Data storage

The University provides a number of file-stores for data, including Departmental File-stores for administrative and work-related files; and Microsoft OneDrive for files that do not need to be shared with others, including transient research data. But the main storage facility for research data is the Research Data Storage Facility (RDSF), which is formally owned by the PVC Research and operated by the Advanced Computing Research Centre (ACRC).

The RDSF provides storage for research data that meets both legal and regulatory frameworks for particular types of research and is in line with the policies of external research funders; and supports researchers' compliance with the University's RDM policy. **Policy** and **terms of use** documents cover issues such as data protection, freedom of information, other legal and ethical matters, technical issues, data ownership, security and access, costs, and data sharing. These and other issues must be covered when PIs, after applying to be a data steward, then register a project, requesting use of the RDSF and access for relevant members of their research group. Data stewards can have up to 5TB of storage free of charge. Beyond that, the charge is £750 per additional TB, charged up front; this funding model meets the costs of storage for up to 20 years. Since data is replicated in two separate locations, the free storage limit is in effect 10TB. From 2018 data stewards will be allocated 50TB of tape storage for archived data in addition to the 5TB on spinning disk.

The RDSF is set up to support both Windows and Linux users, and data can be accessed as a Windows or Mac shared drive or via a network file system on Linux.

5.5.5 Data repository

Depositors have 1TB data publishing space per project in the **data.bris** Research Data Repository. Once a deposit record form describing the data is verified, DOIs are allocated

and the deposit record becomes publicly available. Datasets are associated with a text file providing

- an inventory of the major parts of the dataset, so users can identify any missing parts
- details of any software and/or operating system required to make use of the data
- information about any other dependencies (e.g. particular libraries) required to make use of the data
- for tabular data, descriptions of column headings and row labels, any data codes used and units of measurements

Detailed **guidance** and support are provided to take researchers through the deposit process.

Data can be deposited under various access restrictions where there are legal, ethical or commercial reasons for so doing, though the RDM service encourages researchers to seek secure ways of making their data accessible (through anonymization, the use of embargoes and so on) rather than restricted access. The repository has two levels of restriction: restricted and controlled. Access to **restricted data** is provided by application to authenticated researchers whose host institution agrees to a data access agreement specifying how the data may be used. **Controlled data** is made accessible only after access requests are approved by the University's Data Access Committee, and the applicant's host institution has agreed to a data access agreement.

The repository currently contains 800 GB of published data, with some 400 datasets, just under two-fifths in the form of metadata records, with the data held in a range of external repositories including the UK Data Archive, NERC data centres, Dryad and figshare. This is around 1% of the active research data stored in the RDSF. These figures should be set in a context where SCOPUS records indicate that authors from the University published over five thousand articles in scholarly journals in 2016. More than three-quarters of the overall total of deposits come from the science and engineering faculties, with much smaller numbers from biomedical, health and social sciences and from the arts. The datasets are in a range of formats, with text accounting for by far the majority, but also including application/octet stream, video and PDF.
Only sixteen have restricted access, and six controlled access. Thus over 97% of the datasets in data.bris may be regarded as open data.

5.5.6 Take-up of services

The EPSRC mandate has had an important impact on awareness and take-up across the University, along with development and promulgation of the University's own policy. RDM staff find that researchers undertaking large-scale projects in areas like astronomy or genomics tend to need little support: they know what they need to do without asking. Engagement with those undertaking smaller projects tends to be more problematic, and RDM staff are aware that the penetration of their services across academic departments is patchy. Academics in the arts, humanities and social sciences tend to be the heaviest users of the services relating to data management plans, since neither AHRC nor ESRC provides dedicated services for grant applicants. The AHRC's technical review process means that applicants' data management plans are under special scrutiny, and require a level of technical expertise which many researchers do not have.

The RDM staff are trying to generate more awareness and take-up through the distribution of leaflets and posters, and a programme of visits to research groups and departments. They also find that there tends to be strong interest from research students and ECRs.

5.5.7 Relationships with external services

The RDM and related services at Bristol have drawn on lessons and materials from a range of services at other universities and elsewhere; but they have also taken a leading role in a number of areas including data publication and procedures for sensitive data. The GW4 Alliance, which brings together Bristol, Bath, Cardiff and Exeter universities, has an active Data Services Working Group that shares experience, expertise and projects such as the **RDM triage tool**, as well as running training events.

With regard to the repository the RDM staff are clear that for most purposes a subject repository is preferable to an institutional one; and as noted above a high

proportion of the content in data.bris takes the form of links to data in other services. These are gathered with the aim of providing a record of Bristol-produced datasets.

Harvesting the metadata from services such as the UK Data Archive, Dryad and other services is currently manual and time-consuming because of the limitations of search (many repositories do not allow for search by institutional affiliation); and a central harvesting service would be a great boon. There is not at present any comprehensive system for picking up from the RED office when research projects are coming to an end, and chasing for any data. Nor is there any systematic checking via those researchers who are making use of the RDSF during the course of their research. Nevertheless, the RDM service does seek to follow up EPSRC and NERC-funded projects where researchers claim funding to meet article processing charges, in order to check that authors have met funders' requirements relating to data access; and this has led to some deposits in data.bris.

5.6 Evaluation: strengths and areas for development

The RDM service submits a quarterly report to the Research Data Storage and Management Executive, which oversees the service and plans for further development. A set of metrics for the service has been agreed with the Operational Planning Group and they will be further refined with the Research Data Storage and Management Executive, which will seek to provide KPIs as well as information about costs and benefits.

The RDM team is seeking to develop further advice and training programmes on issues including digitisation and IP (stimulated in part by the University's drive towards e-theses). They are also planning to do more relating to commercially-sensitive data, and the provision of restricted access to it, in order to help academics balance funders' data sharing and open data requirements on the one hand, and commercial imperatives on the other. They are also considering the possibility of seeking accreditation for the repository, but no decision has yet been made.

Nevertheless, it is clear that Bristol is further along in the development of its RDM services than many other universities, since it started earlier, and has had more time to build up its portfolio. Links with other parts of the University - academic as well as support services - are strong; and much of the RDM service's guidance has been developed in collaboration with other services such as RED. Training, advice and guidance are the core of the services, and they are proud of what they've been able to provide in terms of guidance and training on issues such as sensitive data and on software (where they worked with the **SSI**). Evidence suggests, however, that take-up of services, and the deposit of datasets in data.bris or in external repositories remains patchy, given the scope and scale of research activity at the University. Whilst the University's policies and services provide effective pathways to support and promote open data, it cannot as yet be said that open data is embedded as common practice across the University.

6. University of Salford[1]

6.1 Background

The University of Salford is a teaching-led institution. Its research activities are modest, and its **Financial Statements** for 2015-16 show research income amounting to £10.3m, or 5.4% of total institutional income; research grants account for £6.2m (of which £1.4m from Research Councils) and the HEFCE recurrent grant for research makes up the remaining £4.1m [2]. Over 200 staff were submitted to the REF 2014, and SCOPUS records indicate that staff published some 800 articles in scholarly journals in 2016.

Over the past three years, the University has been developing its policies on research data, management (RDM) including expectations relating to open data and data sharing. The key statement is the **Research Data Management Policy**, effective from January 2016.

Its purpose is to ensure best practice in RDM, and it is informed by the requirements of research funders: it specifies that research data must be generated, stored, deposited and made accessible in line with funders' requirements and expectations. It applies to all research data created by academic staff and postgraduate research students, but not postgraduate taught students or undergraduates. The policy states that *"research data selected for archiving must be made openly available, where appropriate, with as few restrictions as possible"*. Archiving should take place either in an appropriate external data centre, or in the University's own repository. The policy also states that researchers must use open data formats where possible to reduce data obsolescence and increase re-use potential. It sets out a range of factors that may restrict openness, including ethical or legal requirements, reasonable rights of first use, and so on.

The policy refers to a wider framework of research policies and information governance, notably:

- The **Information Framework** sets out seven duties to ensure that information (including data) is captured, stored, used and shared in ways which enable the University to meet its objectives and the requirements of its members. This mirrors principles of good RDM, and the Framework references Data Management Policy. Two of the duties relate specifically to managing information; and to sharing it and making it available.

1. We are indebted to David Clay, the University Librarian, and Bill Ayres, Research Data Manager, for their help with this study.

2. By comparison, the University of Manchester's research income of £342m accounted for almost 35% of total institutional income during the same period; see Facts and Figures 2017 - **http://documents.manchester.ac.uk/display.aspx?DocID=31312**

3. This is a recent document, effective from June 2017, and at the time of writing, it is not yet available on the University website

- The Research Code of Practice[3] is articulated around a set of principles, one of which relates to openness and candour. The Code states that *"subject to legal, ethical and commercial constraints, there should be open and transparent reporting of research methods, and of the collection, analysis and interpretation of data. Research findings [...] should be made widely available"*.

- The **Code of Practice for the Conduct of Postgraduate Degree Programmes** requires supervisors to provide guidance to research students on ensuring that data is stored in accordance with the Research Data Management Policy; and to ensure that requirements over data protection and open access are explained to them.

- The **User Guide on Good Practice in Authorship of Research Publications** calls on authors to ensure that their publications meet open access and open data requirements.

- The University has a Policy on **Open Access**, but although this briefly mentions the depositing of data, open data lies beyond its scope.

Not all these policies refer explicitly to open data, but principles of good RDM practice are at least implicit in all of them; and two of the documents refer to the Data Management Policy which, as outlined above, explicitly addresses open data. Complementing these formal policies, the University sets out comprehensive RDM guidance on its **website**, including a page on publishing and sharing research data. This makes clear that *"research data are increasingly considered as a valuable research output, equivalent to communicating research results through journal articles and monographs"*. The guidance lists the benefits of data publication, among which is the building up of academic reputations; the guidance states that *"making data openly available facilitates discovery and re-use, and is associated with increased citation rates"*.

6.2 Drivers and actors for research data management and open data

In 2014, the University initiated a project to develop an RDM service, for which the Library has taken the lead. Central to the service was the appointment of a full-time Research Data Manager. Initially, the drivers for this were top-down: Salford has tended to receive much of its research grant funding from EPSRC, and its requirements acted as a strong incentive and justification for the project. The RDM service is not focused specifically on open data: openness is seen as an integral part of good practice, but is not addressed through any specific initiative or mechanism, other than the data repository (see section 6.5). In developing its approach, the Library worked closely with interested members of the University's research community, its IT team and its Research and Enterprise (R&E) Department. These players see good RDM practice as something more than just meeting funders' expectations; dialogue between them has provided opportunities to consider the challenges that they each face. But in practice their priorities have been less about open data than issues around RDM as part of the research process.

The drivers have evolved over time, as more researchers have begun to approach the Research Data Manager to seek advice on best practice and on how RDM fits with open access and open research. The number of data-aware researchers is growing, partly thanks to the work undertaken by the Research Data Manager and the influence of disciplinary bodies, and also because more researchers recognise that good RDM practice can help to build their careers. Nevertheless, the demand from researchers remains patchy, with variation between disciplines, communities and individuals; some researchers still fail to recognise the value of research data.

The disciplines where researchers appear most data-savvy include biological sciences, physical sciences and – more specifically at Salford – acoustics. But often, enthusiasm or the reverse is more about individuals than disciplines, to levels of support or otherwise from colleagues and/or managers, and more broadly to cultural influences: the extent to which individuals have evolved in environments characterised by readiness to share data, and where there is a collaborative

rather than competitive research culture. Characteristics such as age or career stage do not seem relevant to variations in attitude.

6.2.1 Support for researchers

The Library has taken the lead in RDM issues because its staff – notably the Research Data Manager – have the right skills, and work in close collaboration with the wider academic and IT communities at the University. The Library has also secured buy-in at a senior institutional level from the then Pro Vice-Chancellor, who at the outset agreed to sponsor and commit to the project to develop the RDM service, and has been consistent in his support.

Engagement with the key stakeholders, along with the Library's advocacy, have helped to facilitate the embedding of RDM into research processes and practices. For instance, data management planning is now understood to be an integral part of the grant application process and RDM is threaded through the University's new research Code of Practice. The Library is thus recognised as a partner in the research enterprise.

The Library is currently reviewing its research data training provision for early career researchers (ECRs), and how to make the research data service more visible. This will provide further opportunities to build relationships within the various faculties and on that basis, and the Library hopes to encourage the emergence of research data champions.

To date, training has tended to be offered as ad hoc events, such as a 90 minute **training session** on open research, covering OA publishing and RDM, including matters relating to open data. The Library hopes that in future there will be a more joined-up institutional approach to training, tailored to the needs of researchers at different career stages. It could then provide training explicitly tied to the University's broader support for career development, and incorporated in a structured training programme targeted at research students, ECRs and possibly more established researchers.

Salford's relatively small size is both an advantage and a disadvantage in fostering good practice. Relatively small numbers of research-active staff make it easier to develop contacts and relationships with key individuals. Conversations with researchers are more straightforward than in larger universities. Institutional structures and hierarchies are less complex than at the nearby Universities of Manchester and Liverpool, and getting institution-wide engagement is therefore less onerous. Conversely, with lower research income than Russell Group institutions, the effort (including the financial effort) needed to implement good RDM practice always risks outweighing the benefits.

6.3 Costs, value and sustainability

The cost for provision of the research data management service is expected to amount to £108k in 2018-19. The main costs relate to the Research Data Manager's salary and to the systems and services that sit behind the service, for instance Figshare, Syncplicity, GitHub Enterprise licensing. This expenditure is met from the Library's core budget; but it does not cover RDM activities embedded in other services and workflows. For example, data protection queries are dealt with by staff in information governance, IP by the relevant team in R&E, and other RDM issues on occasion by the research funding team. This makes life simpler for researchers and helps to keep costs down. But the spreading of RDM support functions across different departments makes it difficult to ascertain overall institutional RDM-related costs.

The infrastructure, which has been piloted as part of the RDM service development project, is designed to deliver an end-to-end research lifecycle service. In order to ensure value for money, and to justify and sustain the service, the Library had to consider carefully what it could afford, and what it could reasonably expect to deliver. A constant dialogue between the Research Data Manager and colleagues in related areas such as scholarly communications, institutional repository management and metadata standards ensures awareness of what each is doing

and thus helps to optimise staff time and service delivery. Predicting costs associated with research data infrastructures can also be challenging; for instance, academics tend greatly to overestimate their data storage requirements.

6.4 Evaluation and impact

The Library is aware of the need to demonstrate value for money and returns on RDM investments, for instance by deploying activity metrics. There is a similar need to evaluate the impact of open data and good research data practice. The Library has thought about this too, and is looking at how other institutions address the issue.

Building on its interest in researchers' individual attitudes to good RDM practices (see section 6.2), the Library wishes to know more about the contrasting perceptions of researchers and of the University at corporate level. For instance, some humanities academics may see value simply in the long-term preservation of their data, whereas the University as a whole takes the view that over and above preservation, openness is the more important because it facilitates citation and re-use. For the Library, trying to reconcile such differences in perception remains an outstanding challenge.

The capacity to undertake impact evaluations is limited since there is only one member of staff dedicated to RDM. Resources are insufficient to develop fully-fledged evaluation methodologies.

6.5 Relationship with other infrastructures

Salford has recently appointed a new Chief Information Officer; and it is going through a digital transformation programme, with a new Digital Strategy[4]. It is important to ensure that the RDM service is aligned with the new Strategy, which has as a key aim the creation of an integrated technology platform. Hence the institutional and data repository solutions were chosen in part because they could be readily integrated into that platform.

The institutional and data repositories are cloud-based and distinct, with the former focused on published research outputs. A separate data repository provides more flexibility with regards to tools, integration and user experience. Conceptually, users find it easy to distinguish between the two repositories.

4. The Digital Strategy was approved in July 2017 and, at the time of writing, isn't yet publicly available

6.6 Conclusions

The University of Salford has, over the past couple of years, put in place mechanisms to promote and encourage good practice in RDM, including open data, notably through the establishment of an RDM service and the appointment of a Research Data Manager. Other than in the case of the data repository, data openness is not the prime focus of these mechanisms, although it features as an integral part of good practice. The RDM post is located in the Library, which takes the lead on this issue and provides a means of addressing information and data issues across the University. The Library has played a key role in helping to develop Salford's framework of relevant policies and strategies, and has done so by reaching out proactively, working in partnership with different institutional stakeholders, including researchers themselves. Such partnerships are critical to the success of the RDM service. The establishment and development of a dialogue with these players, along with the emerging policy framework, appear to be positive outcomes of the service. The increased tendency for researchers to approach the Library for guidance and advice is also positive.

Salford's position as a smaller institution has made this dialogue easier than in some larger universities, although some researchers are more resistant than others to good RDM practice. The RDM service is still relatively new, with limited means, and more work is needed to enhance its visibility and demonstrate its relevance to the University's entire research community. In developing the service, the Library remains committed to deepening its dialogue with stakeholders and embedding good RDM practices in the institutional research landscape.

7. Natural History Museum[1]

7.1 Background

7.1.1 The Natural History Museum and its collection

The **Natural History Museum** (NHM) collections form one of the most important natural history collections in the world, collected over the past 300 years. Some 350 scientific staff are involved in the care and interpretation of the collections, and also in analysing and interpreting natural history materials more broadly.

The specimens in these collections carry with them information about their identification, and when and where they were collected.

They therefore constitute a critically-important foundation for research into biodiversity, and the many policy issues and obligations relating to it. And not least the collections also attract, inspire and educate millions of visitors a year.

1. The assistance of Dr Vince Smith of the NHM in preparing this case study is gratefully acknowledged.

7.1.2 The NHM and other museums of natural history

The NHM sees itself – alongside other museums – as playing a crucial role at the forefront not just of taxonomic but also biodiversity research. And the importance of its collections and of its research activities put it at the centre of efforts to gather and curate research data, and to make it openly accessible. But its collections form part of much wider sets of collections both in the UK – where there are more than forty other museums with significant natural history collections – and the rest of the world. Estimates of the total numbers of items held in the hundreds of such collections across the world range from two to three billion.

7.1.3 Biodiversity research

Over the past three to four decades, there has been an increasing realisation both of the value of biodiversity and of the threats that it faces. This has given rise to a number of international policy initiatives, including the **Convention on International Trade in Endangered Species of Wild Fauna and Flora**, the **UN Convention on Biological Diversity** and the **EU Habitats Directive**, and the **Global Strategy for Plant Conservation**. Biodiversity research aims to document diversity, and to identify the factors that generate and sustain it. Its key components include taxonomy; the geographical distribution of taxa past and present; and the relationships and interactions of organisms. The NHM is active in all these fields, and in the growing use of informatics in biodiversity research.

7.2 Cybertaxonomy and biodiversity informatics

7.2.1 Taxonomy

The names by which organisms are known are the basis for communication about them. When due allowance is made for problems such as synonyms, variant spellings, and so on, taxonomic names offer a near-comprehensive system of unique identifiers for past and present references to organisms, thus enabling scientists to index and organise biodiversity information.

It is a central tenet of the codes of nomenclature that all acts that affect names should be published (and in that sense made open); and most of the 15,000 to 20,000 new species descriptions and thousands of other acts published annually are now in digital form. Specialist taxonomists have consolidated this information in taxon-centric databases containing descriptions of the nomenclatural acts and relevant bibliographic citations. Many such databases are openly accessible on the Web. Two of the largest are:

- **LepIndex**, a computerised and updated version of the NHM's card index of the scientific names of the living and fossil butterflies and moths of the world; and
- **Systema Dipterorum**, based at the Natural History Museum of Denmark, which provides authoritative information about the names of two-winged insects.

Together, these two databases cover almost half a million species of butterflies, moths and flies, representing nearly 25% of all currently-described biota.

As the technical barriers have reduced, efforts have turned to cover more obscure taxa; and these are being integrated with larger databases in order to compile more comprehensive lists. The **Global Names Architecture** (GNA) is a system of web-services which helps people to register, find, index, check and organize scientific names and interconnect on-line information about species. It has the capacity to link taxonomic names and concepts to multiple classifications. The foundation for this effort is the **ZooBank** database, a community-led effort to compile nomenclatural information to ease the transition into electronic-only publications.

7.2.2 Species descriptions

Taxonomy depends – arguably more than any other science – upon historic literature. But the early literature is often rare, with limited distribution across the globe. Hence the major taxonomic libraries worldwide – including the NHM – joined forces through the **Biodiversity Heritage Library** (BHL) to coordinate the digitisation of out- of-copyright literature, which is then indexed to enable searching via a central portal. Most of the 38 million pages scanned to date cover species descriptions published before 1923 and in English, although subsidiary projects are now covering European, Arabic and Chinese literature. More recent material requires complex identification and negotiation with rights holders. And in order to make the content more readily and easily usable, it will have to be converted into structured databases. Standardised mark-up is a prerequisite for this, and **TaxPub** provides a tagset for that purpose, while the **TaxonX** schema was developed within the community to streamline the process of mark up. Again, the NHM has played a crucial role in these efforts, for example data-mining the scientific literature to build a picture of where biodiversity has been lost, and how that loss relates to environmental changes.

7.2.3 Standards and platforms

Data sharing, and more specifically open data, is essential to enable the collaboration and large-scale analysis necessary to address many of the issues relating to biodiversity. But early efforts to share taxonomic data, along with data relating to geographical distributions and to interactions between species, revealed problems arising from diverse data structures, and the lack of shared vocabularies. The Taxonomic Databases Working Group was established in the 1980s and developed data dictionaries and exchange standards across a range of fields. It now covers all organism groups and has extended beyond the taxonomic community; hence it has changed its name to **Biodiversity Information Standards** (TDWG). Much of its recent focus has been on protocols for exchange of data over the internet, XML schemas,

and how to achieve semantic and structural descriptions for domain-specific data.

There has also been in recent years a focus on the development of biodiversity information platforms, centred around central or distributed data stores, and seeking to cover activities from field work and description through to publication. They provide the interfaces needed to use external software and tools, and data from external applications can be integrated and processed in the platform environment. They thus promote collaborative working and sharing of information. The NHM hosts one of these platforms, **Scratchpads**, which provide an online virtual research environment, allowing anyone to share data and create their own networks. They also support communications with members and visitors via blogs, forums, newsletters and a commenting system. Sites can focus on specific taxonomic groups, biogeographic regions, or other aspects of natural history. Other platforms provide different types and ranges of tools, and there is scope for further integration and interoperability betweeen them.

7.3 Digitisation of biological collections

Detailed information can be extracted from the specimens in the NHM and most other natural history collections only by visiting the host institution. But the NHM and other museums are creating digital representations of a growing proportion of these specimens and the associated metadata, thus making it possible for digital surrogates to be accessed via the Web. Although digital surrogates are not always satisfactory substitutes for physical specimens, they are often sufficient to make taxonomic decisions, particularly where two-dimensional images can capture the salient features necessary to identify organisms, such as botanical specimens mounted on card, or lepidoptera, which are normally pinned flat.

Until relatively recently, digitisation focused on the tiny proportion of specimens with an applied commercial, medical or veterinary use, or with major cultural or historical value. Recent advances in technology have made more comprehensive programmes possible. An early mass digitisation programmes was at the **Naturalis Biodiversity Center** in Leiden, which between 2010 and 2015 digitised 7-8 million specimens in detail and a further 30 million at lower resolution. Such programmes showed that major efficiency gains could be made when working at scale: the larger a digitisation project becomes, the lower the unit cost.

7.4 NHM Data and Digital Collections Programme

7.4.1 Policy

The NHM has committed itself to open access and open data principles **'predicated on EU, UK Government and funder expectations and in line with good practice within the global research community'**. The default approach for NHM research and collections data is thus now immediate release under a CC0 licence for data and CCBY for images, unless the data meet a valid exception to the default rule. The exceptions cover matters such as assets that have a potential commercial value, third party rights, data that may be sensitive on legal or ethical grounds (for example relating to locations of red-list species), donor or funder conditions, and confidential documents. There is also provision for embargoes on data that may affect the research competitiveness of the museum and its staff. The museum has established procedures relating to all these exceptions, which may result in withholding the data, or an embargo, or release under a restrictive licence and/or copyright. It is also developing policies relating to data management planning for its scientific projects.

7.4.2 Digital collections programme

The NHM's digital collections programme began in 2014. The aim is to **'collate, organise and make available to the global scientific and public audiences one of the world's most important natural history collections'**; and the ambition is to digitise 20 million specimens by 2025. In order to so so, the NHM is developing the policies and protocols, workflows, people and skills, technical infrastructure, and partnerships with other organisations. In this sense digitisation is part of a wider programme involving activities across most of the museum.

Pilot projects are helping to establish high-throughput workflows for all the major collection types. The first of these covers British and Irish butterflies and moths, as a pilot for the digitisation of all pinned collections. The process is complex, involving

- removing the labels from beneath the specimen;
- photographing the specimen;
- entering data from the labels into the collections database using a custom-built data entry interface;
- georeferencing the locality data on the labels to a geographic centroid that can be mapped (thus showing the distribution of the collections and revealing collecting trends since the mid-nineteenth century); and
- rehousing specimens and labels (with bar codes) in new purpose-built entomological drawers.

Other pilots include the Mesozoic vertebrate collections; 70,000 plant specimens stored on herbarium sheets at the NHM and the Royal Botanic Gardens (RBG); and a selection of the microscopic slides collection.

Each pilot involves testing and refining methodologies for further work. For the lepidoptera, digitisation for each specimen currently takes an average of 2.9 minutes, and it has taken a year to capture all the specimen-level data. Work is in hand to investigate ways of automating the processes further, using informatic pipelines and computer assisted image recognition. For the herbarium collections, high-speed conveyor-belt imaging technology was used in collaboration with a commercial provider; and plans are now being developed to create a digital 'Open Herbarium' of the 11 million specimens in the NHM and RBG.

Three-dimensional specimens present further challenges. A wide variety of Computed Tomography (CT) techniques are now being used to obtain cross-sections which can be combined into virtual models of specimens without damaging the original. This approach is particularly suited to paleontological material, where a matrix of surrounding material may obscure much of a specimen. For some organisms like protists, videos of live specimens are probably more effective, since specimens cannot be readily stored using conventional methods. In other cases images may have little value, since the specimen's metadata is its greatest asset. This particularly applies to mineralogical specimens where mining industries are interested in the chemical analysis data associated with mineral samples.

These pilots are thus exploring ways to meet the challenges associated with the ambitious aims of the digital collections programme: digitisation on a massive scale; extraction of data via transcription, OCR, georeferencing, and image recognition; making use of the data by linking to archives and literature, analytical tools, visualisation, and search; and developing end-products in the form of apps and digital exhibitions. And the challenge is huge: as of 2015/16, around 4.5% of the NHM's collections had been 'digitised'.

7.4.3 Crowd sourcing and citizen science

Digitisation and open data open up many possibilities for citizen science. In some cases digital surrogates have been mobilised on the Web to enable many more people to transcribe specimen label metadata. Thus the **Herbarium@home** project of the Botanical Society of Britain and Ireland, crowd-sources the task of capturing plant label information from digital photos (including the herbarium sheets at the NHM), to collect structured textual data rapidly. Another of the NHM's pilots is using the zooniverse platform to enable volunteers to transcribe information from the labels of microscope slides of marine fossils; and a second batch of slides has recently been released. More recently the museum has worked in partnership with a range of US institutions (funded in the main by the NSF) on the **Notes from Nature** platform as its primary means of engaging with citizen scientists interested in transcribing museum records. And the Data Portal (see section 7.5) enables people outside the Museum to contribute to its collection records and databases.

7.4.4 Imaging and species recognition

Open data in the form of large reference sets of online specimen pictures have the potential to facilitate rapid identification

of species through automated identification systems which work in much the same way as the forensic databases of fingerprints or images of suspects. They offer the prospect of automatic online identification of species for biologists working in the field as well as in museums. They thus have the potential to transform practice for field biology as a whole. The NHM is active in the development of new image processing mechanisms to detect, sort and process specimens and the data from labels. It has also developed – in partnership with the Manchester Museum and the University of Sheffield - innovative 3D scanning of bird beaks, and used citizen scientists to help examine their evolution, linking morphology to phylogeny.

7.5 Data Access: NHM Data Portal

As digital natural history collections grow, it has become crucial to create repositories and portals to store, manage and access this information. Some major European collections use the **Europeana** portal to facilitate integration of digital collections. Another approach adopted by the **Global Plants Initiative** uses the **JSTOR** platform, a not-for-profit initiative initially funded by the Mellon Foundation for creating digital archives of scholarly resources.

The NHM's **Data Portal** has been developed to provide a one-stop access point, and to encourage innovation by sharing the Museum's data with the scientific community. The portal is designed so that people, projects and publications can be aligned around it. And one of the key aims is that the Museums's datasets will be enriched as scientists and the public contribute additional information, or correct errors. The portal provides access to over 8 million records, including images, sound, video and 3D; and it provides visualisations of the data, including global distribution maps and statistical overviews to help identify patterns and trends in the data. Each dataset is given a DataCite DOI, so that each can be readily cited; and there are traffic light indicators of data quality. The entire dataset is accessible through an API, and 2.4 billion records have been downloaded since the Portal went live in 2015. More recently, the Portal has also begun to provide access to some external data aggregators such as the Global Biodiversity Information Facility.

7.6 Key actors and their roles

The NHM is organised into three main Directorates, for Science, Public Engagement, and Corporate Services. Dr Vince Smith leads the NHM's science informatics activities, and has played a key role in many of the activities outlined above. The museum's more recent shift to an 'open by default' policy was strongly influenced by the appointment of new Directors of Science and of Public Engagement, and by meetings with Sir Nigel Shadbolt. The process of developing that policy across the museum, and ensuring adoption and buy-in from staff, was time consuming and complex. Strong support from senior management has been crucial, since the programe essentially involves changing cultures: supporting policies with new technologies; training and supporting staff; working across boundaries; showcasing high-impact outcomes; and gathering and deploying evidence to keep track of progress.

7.7 Monitoring and evaluation

The NHM is is using business intelligence systems to track the use and the impact of its digital collections, with the aim of gathering actionable intelligence on levels and profiles of usage, engagement through social media, data about citizen scientists and their activities, and so on. The aim is to support better decision-making on priorities, on how to improve systems and processes, and on what works well and less well.

7.8 The NHM and international initiatives

The NHM is also seeking to develop and sustain digital cultures not only within the museum but also across peer institutions. It is taking a lead in developing the vision of a global digital museum as a knowledge platform for the seven thousand institutions and agencies that hold natural history collections across the world. It is thus seeking via an ESFRI project (**DISSCO**) to unify European natural science collections, and to develop a new research

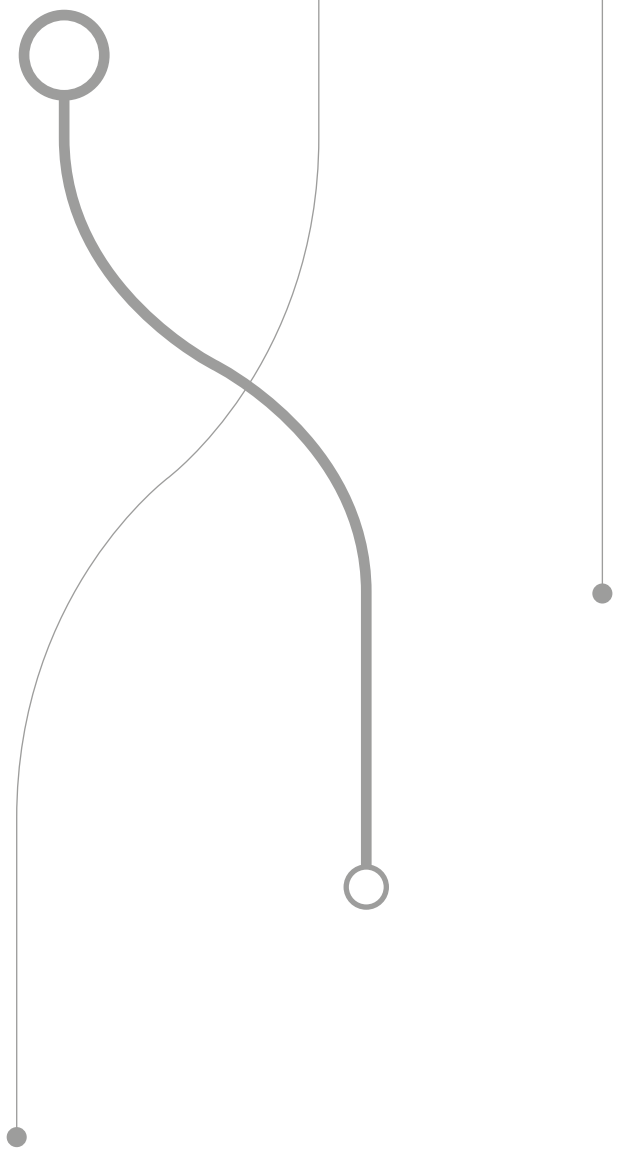
infrastructure providing access to reliable data, using a linked open data approach. But more generally it is working with major natural history museums across the world to promote the development of sustainable policies for digitisation, data curation, open access and exploitation of their collections and the data surrounding them.

7.9 Conclusions

The NHM houses one of the most important natural history collections in the world. Its collections and its scientific staff play a critical role in taxonomic and biodiversity studies in the UK. But it also operates as part of a wider group of institutions, and a wider programme of research, across the world. Digital technologies, and mass digitisation projects, are transforming taxonomic practice and biodiversity research, and also the leading role the NHM plays in such work, enabling it to build closer relationships with other institutions and with the scientific community more generally. Open data is a central part of that strategy, and it involves changes in practice – and in culure - across the musuem.

Key challenges arise from the sheer scale of its collections, as it seeks to shift focus from an essentially analogue world to a digital one. As a national museum, it has funded its digital collections programme and related work on research data in the main from Government funding; but it has also benefited from working with technology partners including Google. Such partnerships are likely to be crucial for the future in moving from projects to programmes and platforms in order to cope effectively with the vast scale of digitisation, the demands for large scale infrastructure, and progress in data extraction and interpretation.

At a scientific level, there is also the challenge of establishing closer links between the world of species data and genomic data, as genome sequencing increases the scope and the scale of its impact on biodiversity research.



8. Research Data in Germany[1]

8.1 Background

8.1.1 The research landscape in Germany[2]

There are significant differences between the research landscapes in Germany and the UK. There are, for instance, no precise German analogues for the UK Research Councils and Funding Councils. The major funding organisations are the **Deutsche Forschungsgemeinschaft** (DFG), the **Federal Ministry of Education and Research (BMBF)**, and intermediary organisations such as the **Alexander von Humboldt Foundation** and **German Academic Exchange Service** (DAAD, which supports the exchange of both students and researchers).

Overall, around a third of funding for research and development is provided by the public sector, both by the Federal Government and the sixteen Lander (which act independently of each other with regard to research funding). Other sources of funding include, as in the UK, industry, charitable foundations such as the **Volkswagen Foundation** and the **Stifterverband** (an association of companies and foundations which supports research and education).

Much of the research in Germany is of course conducted in universities: there are some 400 HEIs in Germany: 110 universities, 230 universities of applied sciences, and 60 art and music colleges. Around 100,000 of Germany's 360,000 researchers work in HEIs and university hospitals. But a significant proportion of German research is also undertaken in the centres and institutes of a series of non-university research organisations including the **Fraunhofer-Gesellschaft**, the **Helmholtz Association**, the **Leibniz Association**, and the **Max Planck Society** (these four organisations together run over 250 research institutes and centres, and employ some 70,000 researchers). There are also some 40 federal research institutions, such as the **Federal Institute for Materials Research and Testing** and the **Robert Koch Institute** for biomedicine; some 150 research organisations run by the Lander; and a network of around 100 industrial research associations from various sectors of industry in the **German Federation of Industrial Research Associations**.

8.1.2 Research data policies, strategies and position statements

The Alliance of German Science Organisations (which includes the non-university research organisations mentioned above,

1. The assistance of Dr Stefan Winkler-Nees in preparing this case study is gratefully acknowledged.

2. This section draws on information in the Federal Ministry of Education and Research's Research in Germany website: **https://www.research-in-germany.org/en/**

plus the **German Academy of Sciences Leopoldina**, the **Council of Science and Humanities** and the **German Rectors' Conference**) adopted in 2010 a statement of **Principles for the Handling of Research Data**; and an Alliance working group on research data subsequently published **a position paper** in 2015. Indeed, since 2010 there has been no shortage of statements from organisations including the Council of Science and Humanities, the **German Initiative for Network Information** (DINI, an association of libraries, media and computer centres), the DFG and the Rectors' Conference. And the Government's **digital agenda** 2014-2017 calls for better access to research data as a goal. All the statements call for better management, accessibility and preservation of research data; but all note the key challenges of sustainable funding for research data infrastructures, the need to enhance skills and training, and the development of policies and commitments, but also better legal frameworks.

The most recent report from the Council for Scientific Information Infrastructures in 2016 was highly critical. It argued that despite several good examples of research data management (RDM) in Germany, there is an overall absence of coordination, and that current efforts often take the form of parallel, project-based initiatives. Universal access to RDM services is lacking, as the key players at present are individual institutions and organisations; and sometimes individual researchers with project funding and an excessive niche focus. Their efforts often suffer from high staff turnover, with the loss of valuable know-how. Moreover, the range of services is impaired by the absence of governance mechanisms to impart greater strategic direction. In addition, there are unresolved issues relating to quality assurance, legal compliance, data privacy, and data security. The Rectors' Conference and other organisations have endorsed the report's findings and recommendations, which include the need for

- long-term funding mechanisms;
- a collaborative and distributed, but co-ordinated National Research Data Infrastructure, composed of disciplinary consortia/bodies, and to be developed over time;
- good practice guidelines for researchers covering such issues as quality assurance, legal frameworks, and monitoring and evaluation;
- a training and skills development strategy;
- closer networking with international organisations and initiatives; and
- active management of the transition.

There are ongoing discussions between the Federal and Lander Governments, plus the key science organisations, about implementation of these recommendations. However, they are facing challenges in navigating between scientific requirements on the one hand and issues of governance and funding on the other; and regional and Federal elections in Germany add to the complexities. Nevertheless, there is some optimism that additional and sustainable funding to implement the recommendations may be provided in the coming year.

8.2 Data policies

The **principles** adopted by the various organisations in the Alliance of German Science Organisations are broad and aspirational, relating to

- the value of research data,
- preservation and accessibility,
- disciplinary differences,
- scholarly recognition,
- training and support,
- standards, and
- development of infrastructures.

Over the past few years, however, a number of the organisations and some disciplinary communities have gone further in developing their policies, although there is little evidence of mandates requiring open data. The **DFG's** guidelines, for instance, specify the data-related issues that must be considered in submitting applications for funding; that data should be made accessible 'as soon as possible' (so long as it 'does not conflict with the rights of third parties'); and that data should be archived for at least ten years (a period already defined by the 1999 White Paper (revised in 2013) "**Safeguarding Good Scientific Practice**"). The DFG also provides advice and guidance, funding to meet the costs of RDM and preservation in existing infrastructures, and specific funding schemes to help researchers develop new infrastructures. But it recognises that disciplinary practices and norms – and kinds of data – vary significantly; and it is also working with some disciplinary communities – in biodiversity and educational research, and more broadly in the social, behavioural and economic sciences, for example – to develop subject-specific guidelines.

The Helmholtz Association published a **position paper** on research data management in 2016, pledging that it would play a leading role in setting up and helping to co-ordinate the national infrastructure; and that its Centres would all have established guidelines by the end of 2017. Again, however, it recognised that discipline-specific guidelines will take some years to formulate. The Leibniz Association has established a research data working group to address the challenges posed by research data; and in 2015 it adopted **guidelines** on good practice (based on those adopted by the DfG in 2013) which prescribe that 'data must be stored in an accessible format for a minimum of 10 years'. The Max Planck Society's **rules** state that 'data as a basis for publications must, as far as possible, be stored for at least ten years on durable, secure carriers' and that access must be granted to those 'with a justifiable interest'.

A few universities, such as **Bielefeld**, have also adopted principles requiring staff to

- treat research data according to appropriate subject-specific standards;
- provide a data management plan;
- make their data widely available and preserve it for the long term to facilitate re-use, while balancing the need to protect intellectual property, personal data, and obligations to third parties; and
- promote high-quality RDM, with subject-specific training.

Humboldt University has a similar set of principles, emphasising the need to document the complete research lifecycle, including tools and procedures; but leaving to researchers the decision on when and on what terms data may be accessed. **Gottingen** has adopted a rather longer set of principles designed to ensure that RDM, curation, and preservation are all in accordance with recognized standards, meet high expectations and fulfil legal and ethical obligations. The policy leaves untouched 'regulations that relate to an assessment of research data according to the German employee invention act and specific contractual agreements'. It is not clear, however, how many universities have adopted similar policies.

8.3 Data services

The re3data registry records (August 2017) some 300 repositories in Germany (as compared to some 250 in the UK). Many of them represent, as in the UK, German involvement in international initiatives such as the **International Centre for Global Earth Models** at Potsdam, the **World Data Center for Remote Sensing of the Atmosphere**, and the **International Mouse Phenotyping Consortium**. Germany is also heavily involved in a number of EU data initiatives, particularly those included in the roadmaps of the **European Strategy Forum on Research Infrastructures**, including the **German contribution** to **Digital Research Infrastructure for the Arts and Humanities** (DARIAH) and to the **CLARIN** initiative with

historical text corpora. Most of the Helmholtz Association research centres now run their own repositories; and the Max Planck Society operates its own centralised repository (**Edmond**), with more than 12,000 items.

8.3.1 Federated initiatives and services[3]

A number of federated initiatives have been established in recent years, often with support from one of the non-university research organisations such as the Leibniz Association. Thus the **Generic Research Data Infrastructure (GeRDI)** project is seeking, with three initial pilot data centres, to develop an infrastructure to enable scientists – especially those with small amounts of data – to share their data across disciplinary boundaries. The aim is to implement the model as part of the national infrastructure envisaged in the Council for Scientific Information Infrastructures 2016 report. The **RADAR Research Data Repository** project led by five universities and institutes, along with the National Library of Science and Technology (TIB), has similar aims. It has developed a 'starter' package of services targeted at researchers and institutions in the 'long-tail', and a more advanced package aimed at researchers who are more interested in open data and re-use. The business model involves one-off payments for depositors, at levels depending on data volumes and storage periods. Other projects, such as the **SowiDataNet** , are discipline-specific, in this case covering social sciences and economics with a web-based repository and a focus on application scenarios for institutional RDM.

Other initiatives, such as the **German Federation for Biological Data (GFBio)** aim to bring together the data archiving and curation expertise of several national archives and data centres and to serve as an authoritative, national contact point for all issues concerning the management and standardisation of biological research data. It thus provides educational and training materials; central services for the submission of such data, and an integrated open access data portal. Advisory and technical development services and initiatives are also provided by established organisations such as **Technology, Methods and Infrastructure for Networked Medical Research (TMF)** and the **Council for Social and Economic Data**.

In some cases, initiatives have been led by various of the German Academies. Thus the **Berlin-Brandenburg Academy of Sciences** has established itself as a centre for developments in digital humanities, with the **Electronic Life of the Academy (TELOTA**) initiative which began in 2001. It provides advice and support to researchers to enable them to exploit digital technologies, and makes extensive collections of resources available to researchers and to the general public. It also provides a node, for example, to the EU CLARIN initiative mentioned above. The **Academy of Sciences and Literature, Mainz** has also established a Digital Academy for digital humanities, with a focus on cultural heritage from a digital perspective.

8.3.2 Museums and related bodies

In Germany as in the UK many museums and galleries have been active in digitising their collections. Thus the museums, universities and other institutions with natural history collections have established **German Natural Sciences Collections** with the aim of creating a federated infrastructure (DCOLL), to make the collections openly accessible over the web. And in a rather different subject area, the **DigiPEER** project is bringing together three museums and the Leibniz Institute for Spatial Social Research to digitise spatial plans and technical drawings to enhance research on the concepts and practices of spatial planning.

8.3.3 Big Data

Big data has been another focus of activity, with **Smart Data Innovation Lab (SDIL)**, for instance, providing access to a variety of big data technologies as part of a collaboration between industry, researchers and IT providers with the aim of boosting access and use of big data in key priority areas. The **Fraunhofer Big Data Alliance** similarly operates as a big data process chain adviser, providing technological support and training programmes in an industry-research collaboration.

3. A Knowledge Exchange study of federated data services in a range of EU countries is currently nearing completion.

And the Helmholtz Association is developing a **Helmholtz Data Federation** of HPC centres with a focus on big data. At a more subject-specific level, the **Novel Materials Discovery (NOMAD) Laboratory** is developing an encyclopaedia and big-data analytics tools for materials science and engineering, and is building big data services to help advance those disciplines.

8.3.4 University services

A relatively small number of individual universities, including **Bielefeld**, **Göttingen**, **Heidelberg**, **Mannheim**, **Munich** and **Tubingen** provide repositories and data services, with varying scope but usually including advice on RDM, metadata, preservation, access and publication, and legal and regulatory issues. In some cases, Lander governments have supported the development of collaborative services. Thus the ten Hessian universities are **cooperating** in the construction of a sustainable infrastructure to coordinate the organizational and technological processes for RDM. This includes not only a repository, but also advice and other services. In Baden Wurttemberg the Ministry of Science, Research and the Arts is sponsoring **a series of projects** to develop a distributed and shared digital research infrastructure, with a twin focus on RDM and virtual research environments. **In North Rhine Westphalia** the emphasis is on strengthening awareness of RDM issues and on the sharing of experience between different institutions.

8.4 Conclusions

The German Rectors' Conference conclusion that **"Germany is a developing country when it comes to information infrastructures"** is probably too harsh. A considerable amount has been achieved, and there is no shortage of initiatives at national, regional and institutional levels, many of which are supporting open data. Most of the challenges that Germany faces are common to other countries, including the UK: the balance between desirable diversity and undesirable fragmentation; the need for more co-ordination; the sustainability of many initiatives, especially those that have developed bottom-up; the balance to be struck between competing priorities; project funding vs long-term infrastructural funding; the need for work at disciplinary level to develop norms and standards that take account of the specific practices of those disciplines; and so on.

But there are features of the German research landscape that – even while they may be recognisable in a UK context – make for some significant differences in practice. First, the range of powerful and semi-independent non-university research organisations and funding organisations, each with its own distinctive culture, makes for significant difficulties in co-ordination. Thus there has been little in the way of effective and co-ordinated follow up to the statement of principles on the handling of data issued by the Alliance of German Science Organisations in 2010.

Such difficulties are exacerbated when one takes into account also the very significant roles in research funding of Government Ministries such as the Federal Ministry of Education and Research, and the Ministries of the sixteen Lander. This complicated landscape - including changes at political level - is a great challenge in responding to the recommendations of the 2016 report of the Council for Scientific Information Infrastructures referred to in Section 8.1.2.

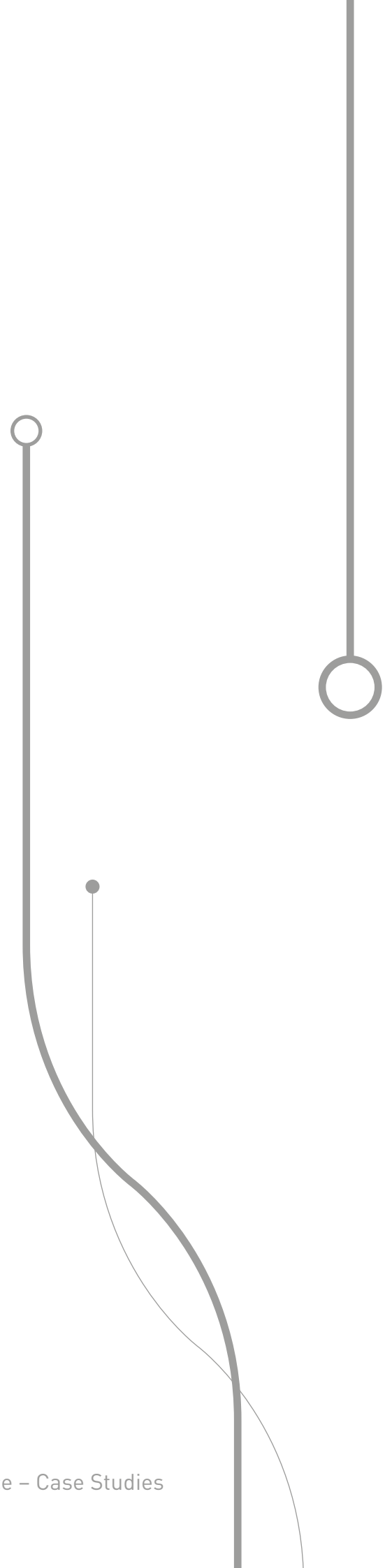
Second, the complex landscape outlined above means that there is no ready source of comprehensive information about the funding of research data infrastructures or projects, or of their costs and benefits. And setting up funding frameworks encompassing sources at federal and Lander levels can be problematic; in some cases there may be a reluctance to fund investments for cross-cutting initiatives.

Third, the university sector in Germany has been facing a number of well-publicised problems relating to (shortage of) funding (which is in the responsibility of the Lander governments), rising student numbers, and the implementation of 'Bologna' principles and structures for Bachelors' and Master' degrees. This may explain why relatively fewer German than UK universities seem to have been active in developing research data policies and services.

Fourth, Germany has more well-established systems for linkages between research and industry, notably but not only through the Fraunhofer-Gesellschaft and its centres (as noted above in the Fraunhofer Big Data Alliance, for example). But this can bring tensions when it comes to open data.

Fifth, freedom of science is written into Article 5 of the German constitution, and this may have made some funding and research organisations reluctant to be too prescriptive in setting out and/or seeking to enforce requirements for researchers relating to RDM, data sharing and open data.

Despite these differences, however, it is notable that the re3data registry indicates that Germany shares its involvement in research data repositories and services with the UK more than with any other single country other than the USA. The commonalities are perhaps as important as the differences.

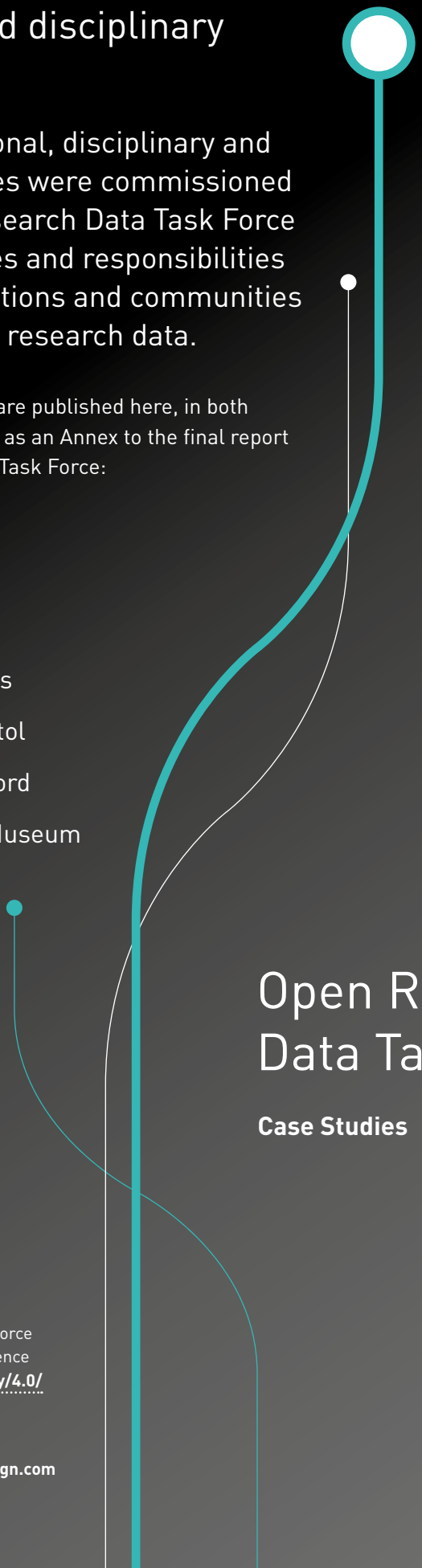


Institutional and disciplinary case studies

A series of institutional, disciplinary and national case studies were commissioned by the UK Open Research Data Task Force to illustrate the roles and responsibilities ofdifferent organisations and communities in the move to open research data.

The following case studies are published here, in both summary and full versions, as an Annex to the final report of the Open Research Data Task Force:

- Astronomy
- Biosciences
- Crystallography
- Digital Humanities
- University of Bristol
- University of Salford
- Natural History Museum
- Germany



Open Research Data Task Force

Case Studies

Report Design: www.crayfishdesign.com