Data First

2019 UCL Research Data Strategy

Report by the UCL eResearch Domain:
Co-chairs: Prof Phil Luthert and Prof Jonathan Tennyson
Strategic Research Coordinator: Dr Louise Chisholm

Executive summary

Our vision is for UCL to create an environment that optimises researchers' capacity to flourish using data. This includes external data created in a non-research context which is used as the starting point for research and data generated during the research process. Uniquely, data can be reused without being degraded and its reuse can create new value. Our aim is for UCL to be *the* leading institution for data-focused research and to inspire and empower our research community for the long-term benefit of humanity.

Effective harnessing of research data can lead to increased research funding, publications, citations, and socio-economic impact. This supports key institutional strategies including UCL 2034, UCL Research Data Policy (2018), and the Innovation and Enterprise Strategy (2016).

Responding to a new landscape

Technological advances have expanded the number of disciplines which use data and associated methodologies. More external organisations are collecting data which can be used as a basis of new research. Open Science encourages data to be shared as openly as possible and as closed as necessary. This impacts all faculties, the UCL's broad corpus of data can initiate new interdisciplinary research.

Research funders and the UK government promote the use of data in research by launching new funding opportunities *via* the Industrial Strategy Challenge Fund, Research Councils, and Alan Turing Institute. At the same time, the regulatory and legal environment is becoming increasingly complex for researchers to navigate, following GDPR and similar legislation. Data breaches could lead to serious fines and penalties from UK, EU, and international bodies.

Compared to the ways of data control, the governance and mechanisms to facilitate appropriate access to data are still being developed. These challenges drive two opposing behaviours: acting to increase data sharing and to increase the control and restrictions on sharing data. UCL needs to address both challenges simultaneously to promote behaviours that will effectively harness the increasing data-related opportunities available.

Current Challenges

Data-focused research approaches require researchers to successfully create, access, store, analyse, curate and share data. This includes data which originates from external sources and from UCL research activities. The four types of data (ultra-secure, highly secure, secure and open) involved will influence how these activities are implemented. UCL needs to ensure that data is used appropriately to maintain the public's trust and maintain UCL's reputation. We have identified challenges and risks associated with each of these activities. Risks include, 1) loss of research data to the research community through data deletion; 2) loss of research, collaboration and funding opportunities; 3) higher costs for purchasing multiple data licences and access to external infrastructure; 4) duplication of resources and effort to resolve common data issues; 5) data mishandling and regulatory penalties; 6) the lack of systematic institutional awareness of compliance regarding legal and regulatory obligations.

The 18 recommendations below aim to create an environment to enable researchers who use data to flourish by optimising UCL's environment and to mitigate hurdles for using these data.

Recommendations to improve coordination, governance and compliance

1. Create a UCL Central Research Data Office (CRDO) led by a single, accountable, senior academic, reporting to UCL's senior management structures. The CRDO will:
    a. Be responsible for the implementation of this strategy and refresh it as needed;
    b. Act as a hub sharing best practice developed by professional service teams, improving signposting and researchers' awareness of support available;
    c. Work with professional service teams to optimise internal processes to create, access, store, analyse, curate and share data.
2. To ensure that UCL's governance, ethical, legal and regulatory processes enable UCL researchers to confidently use research data, in particular;
    a. Ensure UCL provides ethical review and assurance regarding its numerous national and international legal, regulatory and ethical obligations relating to data;
    b. Work with HR to ensure that maintaining data security and confidentiality is a core requirement of all staff and students, with clear penalties for breaches.

Recommendations to improve access to external data

3. CRDO to develop a relationship management approach for strategic data providers, to explore appropriate internal governance mechanisms to widening access to the data.
4. CRDO to explore governance and grant-recharge mechanisms to increase use of external data sets with high-cost licences.
5. UCL Research Services and CRDO to share best practice on negotiating with external data providers. Research Contracts to expand its capacity to negotiate data licences.
6. UCL to actively engage with all partner NHS Trusts to establish a unified approach to use of data for research or service evaluation by UCL.

Recommendations to improve sharing and curation of data

7. CRDO to promote a data rich research environment promoting maximum appropriate access to UCL-generated datasets.
8. CDRO promote data curation and processing to maximise the impact of UCL's datasets.
9. CRDO to facilitate a joined-up service providing advice to researchers on how to comply with their data sharing obligations spanning contractual, ethical, legal, regulatory, IP, and funder requirements.
10. CRDO and Library Services to champion the recognition of data creation and sharing in the UCL academic promotion framework.
11. UCL to continue investment in data management, stewardship and FAIR data initiatives.
12. CRDO, Library Services, Research IT Services (RITS) to provide leadership and influence how relevant HE sector research data challenges are addressed.

Recommendations to improve infrastructure for data storage, sharing and analysis

13. UCL to continue investment in the UCL data repository led by RITS.
14. ISD to continuously review the computer network, and other infrastructure provision for data.
15. RITS to provide options, including commercial, for secure data storage and high-performance computing analysis that are not currently available at UCL.

Recommendation to improve the skill development and training for staff

16. CRDO and internal training providers to develop pathways for research staff to develop their technical skills (e.g. data science) and associated knowhow (e.g. ethical approvals).

Recommendations to improve the retention of highly skilled staff

17. CRDO to work with HR to ensure UCL professional service staff and researchers with specialist data skills have clear career paths within UCL.
18. HR to review the current salary for roles that require data specialist skills.

Table of Contents

Index

1 Introduction

Our vision is for UCL to be *the* leading institution for data-driven research and to inspire and empower the UCL research community to leverage the exciting opportunities created by recent technological advances within and across their respective disciplines. UCL should create an environment that enables researchers to flourish using data generated within UCL and externally. Due to the breadth of research, UCL is generating a unique and exciting corpus of research data across the faculties. If it is appropriately managed, it has the potential to give UCL a major advantage over its competitors.

Data are a unique resource as they can be used and reused without being degraded. The same dataset can be used in different ways, to support research, innovation, and teaching activities. Their value can be enhanced by the subsequent addition of new data, e.g. through linkage to different data sets. The adoption of the vision and recommendations below would demonstrate that UCL values research data and recognises the potential benefits there are for the society, institution, professional service teams, academic communities, individual researchers and students.

This report focuses exclusively on data in a digital format that can be used in research. This includes data which is created in a non-research context (e.g. governmental, healthcare, or commercial data), which can provide the starting point for research. It also includes data generated during research which provides the information necessary to support or validate a research project's observations, findings or outputs or used as input for further research. Data may include statistics, digital images, digital documents, sound recordings, sensor readings, genomic data, laboratory results or numerical measurements.

This report does not address data that is used to support UCL students, UCL staff or data used by professional service teams to manage research, such as Worktribe, Institutional Research Information System (IRIS), Research Fish, Research Publications Service (RPS), human resources, or student records.

Image 1: Terrestrial laser scanning analysis of sycamore trees used in the analysis of volume uncertainty in cylinder fitting from Wytham Wood. Credit to M. Disney, UCL Geography.

"Weighing trees with lasers: advances, challenges and opportunities" M. I. Disney, M. Boni Vicari, A. Burt, K. Calders, S. L. Lewis, P. Raumonen and P. Wilkes. Interface Focus 8 2018 https://doi.org/10.1098/rsfs.2017.0048

This project was initiated in response to discussions with representatives from several faculties, who expressed concerns regarding research data. The UCL eResearch Domain established a Working Group to articulate the high-level academic vision of how UCL's academic environment can be improved to support data intensive research and to make recommendations to achieve this vision. The Working Group includes representatives from across UCL faculties, Research IT Services (RITS), UCL Libraries, and Information Security (see Appendix 7.4).

The Group identified four themes to explore: data access; data curation and use; support for staff and researchers; institutional environment for data research. These themes were explored further in a world café style workshop and written submissions were also solicited from the community (see Appendix 7.4.1). The output of these activities has informed the development of this report. The draft strategy was open for consultation for 1 month. It received positive feedback during a UCL Townhall meeting (89 attendees) and *via* an online form (41 submitted).

The strategy was also discussed in meetings with Research Information and IT Services Group (RIISG), General Data Protection Regulation (GDPR) and Clinical Research Committee, the UCL AI network, UCL Library Services, RITS, GDPR Preparedness Programme and with the Pro-Vice Provost (Library Services), and Vice Provost (Research).

1.1 The report and its structure

This report was developed following wide engagement with the academic community and professional service teams. It has the following structure:

Section 2: Why does UCL need a research data strategy?
Section 3: Current Challenges
Section 4: Overview of the Central Research Data Office
Section 5: Risks of not acting
Section 6: Conclusion
Section 7: Appendix

Recommendations on how to address the challenges identified are within Section 3.

2 Why does UCL need a research data strategy?

Data-focused research occurs across the whole of UCL. Feedback from staff is clear that they are facing a range of challenges to access, store, analyse, curate and share data (see Section 3). There is the opportunity to improve the support for data-focused research, which will enable UCL to achieve its strategic aims. An effective environment for data-focused research is influenced by:

- External regulations and laws and by data providers' requirements (e.g. GDPR, contractual agreements, confidentiality requirements).
- UCL infrastructure which enables researchers to download or transfer data over the network, repositories to store data, and compute to manipulate data.
- UCL professional services staff and teams providing technical support to researchers such as data facilitation, data curation, data management, research software engineering, IT systems administration.
- UCL professional services staff and teams providing support for processes such as ethical review and data protection assurance, negotiation of data licences and data sharing agreements, purchasing data licences, legal advice, training, negotiation with data providers.
- UCL commitment and expectations regarding ethics, conflict of interests, research reproducibility and research integrity.
- External and internal data that is **f**indable, **a**ccessible, **i**nteroperable and **r**e-useable (FAIR) for research.
- The skills and expertise of researchers to create new knowledge and understanding from the data.
- The culture to share and reuse data in research. The implementation of the open data sharing principle of 'as open as possible, as closed as necessary' varies between disciplines and datasets.

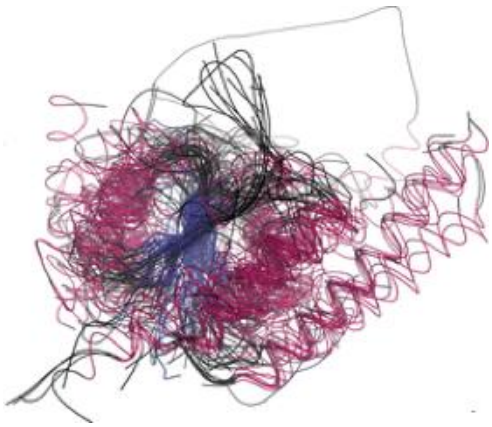


Image 2: Structural diversity in the Nitrogenase molybdenum iron protein domain superfamily (ID: 3.40.50.1980) obtained from CATH-Gene3D. This is a free, publicly available online resource providing information on the evolutionary relationships of protein domains through structural, sequence, and functional annotation data. Credit to N. Dawson, UCL Institute of Structural and Molecular Biology.

"CATH-Gene3D: Generation of the Resource and Its Use in Obtaining Structural and Functional Annotations for Protein Sequences." N.L. Dawson, I. Sillitoe, J. G. Lees, S. D. Lam, C. A. Orengo In: Protein Bioinformatics. Methods in Molecular Biology Wu C., Arighi C., Ross K. (eds), 2017 (155). Humana Press, New York, NY http://dx.doi.org/10.1007/978-1-4939-6783-4_4

Data-focused research is ongoing across UCL involving every faculty and many departments. Strategic coordination across this complex ecosystem, involving professional service teams, faculties and departments will be required to improve the academic environment for these researchers. Without an institutional level strategy there is the risk that multiple, and different, solutions will be developed to address the challenges. While these solutions may meet the needs of a particular department they might not suitable at an institutional level. There is also the risk that opportunities for cross-disciplinary collaboration will be lost.

2.1 Supporting UCL 2034

Technological advances combined with the UK government's objective to "put the UK at the forefront of the artificial intelligence and data revolution"[1] has created new opportunities. The 2019 UCL Research Data Strategy aims to create an environment that optimises researchers' capacity to take advantage of these opportunities. Data and data-focused approaches have the potential to transform how knowledge is created and shared, and the way that global problems are solved. Our recommendations support the following UCL 2034 Key Enablers:

A: Best student support. To enable access and use of data for dissertations and data to support students learning data science, artificial intelligence (AI) and other data-focused techniques.

B: Valuing our staff and delivering on equality and diversity. Highly skilled research and professional service staff underpin UCL's capacity and capability to use data in research. To create an environment for data-focused research which will attract and retain world-leading researchers; to enable staff to develop their knowledge and knowhow and to feel valued for their technical expertise.

C: Financing our ambitions. To improve the value for money for infrastructure and data licences; to reduce the risk of breaching data regulations and the associated financial penalties; to leverage existing UCL data resources to generate new research funding opportunities and impact, which will contribute to UCL's Quality-Related funding.

D: Excellent systems. Act to optimise the infrastructure, services and processes that are required to create, curate, store, use, access, and share of data by researchers or technicians; to ensure that new initiatives are developed holistically across professional service teams and faculties.

F: Communicating and engaging effectively with the world: deliver public and societal benefits. To be transparent regarding UCL's ethical use of data in research to build trust with the public; to promote cross-disciplinary reuse of data to address societal challenges; to confidently and appropriately share data with the public and external organisations.

[1] The Grand Challenges: Policy Paper (2018) Department for Business, Energy & Industrial Strategy.

2.1.1 Maintaining and extending research excellence

UCL is distinctive in the scale and breadth of its research[2], which is mirrored by the rich and varied data sets that are generated or hosted by its researchers from all faculties. Technology advances have enabled a massive growth in the volume and variety of data types generated. 2.5 Quintillion bytes of data are created each day globally and this pace is expected to accelerate with the growth of the Internet of Things[3].

Encouraging the sharing of data can[4]:

- Lead to insights obtained from one dataset that might not be directly foreseeable to the researcher who created it;
- Facilitate the merging of two complementary datasets with the potential for more insight than possible through keeping them separate;
- Facilitate the use of data collected in one research area in another (e.g. transport data can inform healthcare research).

Facilitating and promoting the re-use of external and internal data sets would inspire and empower the research community to foster disruptive thinking and enable cross-disciplinary research. Currently, half of the respondents to the 2016 Research Data Management (RDM) Survey identified that they had reused someone else's data, however, only 23% did it regularly[5].

However, realising the value of data depends on governance and the ways in which the data are used. If data is unstructured, unlinked to other datasets and inaccessible to analysis (i.e. not machine readable or not annotated with metadata)[6,7], the scale and breadth of data generated at UCL will be underexploited and the value will not be realised.
Within UCL this massive data generation includes that arising from disciplines that are traditionally associated with data intensive science such as high energy physics, social sciences, or computational finance, as well as disciplines which have more recently adopted state-of-the-art data and associated methodologies such as biosciences (see image 2), digital humanities (see case study 1 and image 3), computational archaeology, architectural computation, and health informatics. However, data maturity, which is defined as the journey towards improvement and increased capability in using data[8], varies greatly between academic fields. UCL has the opportunity to build on expertise and develop data maturity across disciplines that have more recently started to use data and associated methodologies.

[2] UCL Research Strategy 2019

[3] Data Never Sleeps 5.0

[4] How does big data affect GDP? Theory and evidence for the UK. Goodridge and Haskel, Imperial College London Business School (2015)

[5] UCL researchers and their research data: practices, challenges & recommendations: Report on the 2016 RDM Survey Fellous-Sigrist, M; UCL Library Services (2016)

[6] Growing the Artificial Intelligence Industry in the UK, Hall and Pesenti, Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy (2017)

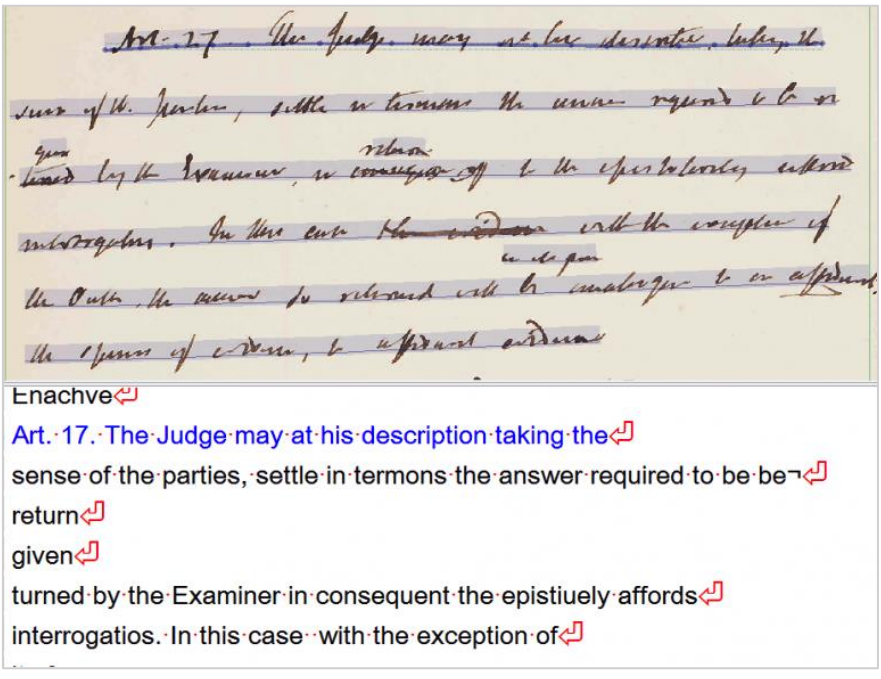
[7] The economic value of data: Discussion paper HM Treasury (2018)

[8] The data evolution project report Data Orchard and DataKind UK (2017)

Case study 1: Survey of English Usage
UCL English language and literature, Faculty of Arts and Humanities

The Survey of English Usage (https://www.ucl.ac.uk/english-usage/) carries out research in English language Corpus Linguistics. The Survey gathered samples of naturally-occurring language for the purposes of description and analysis. Recent major corpus projects have been digitised, transcribed, annotated and indexed on computers.

The digitised Survey is used in research to understand how language works and changes over time, or between speaker communities and to develop new ways to teach English grammar in secondary schools. The Survey is also used to create 4 software packages and 4 mobile phone apps for schools, researchers and the public.



Enachve
Art. 17. The Judge may at his description taking the
sense of the parties, settle in termons the answer required to be be
return
given
turned by the Examiner in consequent the epistiuely affords
interrogatios. In this case with the exception of

Image 3: An automated transcript of very difficult handwriting using Handwritten Text Recognition (HTR) technology (Transcribe Bentham Project, UCL Laws). Courtesy of UCL Special Collections.

http://blogs.ucl.ac.uk/transcribe-bentham/2018/11/28/project-update-automated-recognition-bentham-handwriting/

2.1.2 Maintaining and extending market share of research funding

Research data are increasingly the focus of research funding opportunities. Most funding organisations support projects where data are created, collected, and/or analysed. Changes in the funding landscape that have led to this increased emphasis of data include:

- The creation of UKRI which has increased the emphasis on interdisciplinary research, and data have the potential to be a cross cutting theme. For example, the cross-disciplinary Mental Health Research Agenda[9] was developed as a result of collective interest across the Research Councils in mental health research from a medical, biological, environmental, cultural, social, technical and historical perspectives.
- The Industrial Strategy Challenge Fund (ISCF) supports the UK government's Grand Challenge of AI and Data. In addition, data underpins the objectives of the other three Grand Challenges (Ageing Society, Future of Mobility and Clean Growth).
- UCL is a founding partner of multiple recently established data intensive national institutes. This includes the Alan Turing Institute for data science and AI, Health Data Research UK, the Rosalind Franklin Institute for life sciences and the Faraday Institute for electrochemical energy storage science and technology.

Further funding opportunities are expected.

2.1.3 More efficient use of research funding

There is an opportunity cost associated with not reusing data. Promoting the reuse of data could reduce the need to repeat (and fund) experiments that generate the same data. This could allow the proportion of the funding award to be reallocated to other types of activity such as enabling the researchers to address additional research questions, or to support impact activities. This behaviour is expected in some academic communities (such as astrophysics and crystallography), but not in others (such as history). Data reuse would enable the efficient use of research funding, supporting the UCL Research Data Policy (2018)[12.]

Researchers are likely to be aware of UCL's secure methods which can protect personal data, but there is a low uptake of the services e.g. 45% of researchers were aware of the DataSafeHaven but only 16% of researchers used the service (GDPR Survey 2018). 60% of respondents were unaware of the Research Data Management website and the Research Data Storage facility (RDM Survey 2016). Researchers procuring external services to protect personal data is prevalent (GDPR Survey 2018).

More efficient use of funding would also improve the value of the investment in time and resources used to apply for funding e.g. pre-award support from Research Services, departments, and OVPR research coordination offices, as well as researchers' time.

[9] Widening cross-disciplinary research for mental health AHRC, BBSRC, EPSRC, ESRC, MRC, NERC and STFC (2017)

2.2 Increasing impact of research

Big data is predicted to contribute more to economic growth in the period from 2012-2025 than typical contributions from R&D[4]. Improving the capability of researchers to share research data (using appropriate data licences and governance systems) has the potential to increase the impact from data created at UCL. External organisations provide opportunities for business, charity, policy and governmental partners to benefit London and beyond. This supports both the Innovation and Enterprise Strategy 2016-2021[10], and UCL Research Strategy's (2019)[2] third aim to deliver impact for public benefit.

The impact of research can be increased by encouraging researchers to submit data to repositories and continuing to develop related UCL data repositories. Data archiving can double the publication output of research projects[11], according to a study of 7,000 National Science Foundation and National Institutes of Health-funded research projects in the social sciences. Citation impact of research papers has also been shown to increase when data are made available – by as much as 50% in astrophysics and between 9-35% in gene expression microarrays, astronomy, and paleooceanography. This supports the 2018 UCL Research Data Policy's aim to maximise the impact of data[12].

2.3 National and international leadership opportunities

There are opportunities for UCL staff to engage in and support leadership across the UK and international research community in support of Aim 1 of the UCL Research Strategy (2019)[2].

2.3.1 Harmonising Funders' Data policies

It is a commonly held belief that the data produced by publicly funded researchers should be regarded as a public good. Shared principles related to research data have been agreed by stakeholders in the Open Research Data Concordat (see Appendix 7.2). However, there are, and notable, differences in the data management requirements and support provided by the seven UK Research Councils, Wellcome Trust, and Cancer Research UK (CRUK). With the creation of UKRI some of the policies may become more closely aligned in the future[13].
For example, EPSRC requires research organisations to maintain their own data catalogue, the ESRC provides a centralised UK Data Service alongside a network of Big Data centres[14] enabling access to administrative, business, and government data. The MRC has supported the Dementias Platform UK (DPUK) data portal (see case study 2)[15].

There are unresolved data management challenges in the sector including sustainable funding of the ongoing curation and managing costs required to share datasets beyond the duration of the research project and potentially the PI's position at UCL. EPSRC requires data to be preserved 10 years from the date of last access, and the MRC requires clinical data to be preserved for 20 years.

[10] The UCL Innovation and Enterprise Strategy 2016 – 2021

[11] The State of Open Data Report 2017 Science, Digital Report - Infographic. figshare. Paper (2017)

[12] UCL Research Data Policy 2018

[13] Research Data Infrastructures in the UK Open Research Data Taskforce with Michael Jubb, 2017

[14] ESRC Big Data Network

[15] Dementias Platform UK's data portal

Case study 2: Dementias Platform UK
UCL Dementia Research Centre, Faculty of Brain Sciences

The Dementias Platform UK (DPUK), which is a public-private partnership funded by the Medical Research Council. The UCL Dementia Research Centre is one of 20 university and industry partners involved. DPUK hosts data from multiple cohorts into a single, secure, environment – the DPUK Data Portal – the need to transfer increasingly large and complex data files between research groups is greatly reduced. By maintaining these data in an environment operating to the highest data protection standards, cohort participants and researchers can be reassured that the data are managed securely and responsibly; maintaining privacy whilst maximising scientific value.

The large number of individuals in our cohorts allows key research questions to be answered more rigorously and more rapidly than would otherwise be possible. To make the best use of these data resources, DPUK has established a number of collaborative initiatives with universities across the UK to enable discovery and testing of new disease concepts, diagnostics or potential treatments.

Case study 3: Consumer Data Research Centre (CDRC)
UCL Geography, Faculty of Social and Historical Sciences

The £11m ESRC Consumer Data Research Centre (2014- 2020) is led by UCL and the University of Leeds, with partners at the Universities of Liverpool and Oxford. The CDRC creates, supplies and maintains consumer-related data for a wide range of users. They work with private and public data suppliers to ensure efficient, effective and safe use of data in social science. The CDRC provides data with three different levels of access:

- Open data service: Over 10,000 open datasets are freely available to all for any purpose following a simple registration. Products include those created and acquired by the CDRC and products where the CDRC have added value.
- Safeguarded data service: Data with restricted access because of license conditions, but where data are not considered 'personally-identifiable' or otherwise sensitive.
- Secure data service: Controlled data which need to be held under the most secure conditions with highly restricted access.

CDRC research projects provide fresh perspectives on the dynamics of everyday life, along with better understanding of issues of economic well-being and social interactions in cities

The CDRC also provides a range of training and data analytics programmes for academic and non-academic researchers, ranging from introductory courses for postgraduate students through to advanced training for data scientists.

There is the opportunity for UCL to influence the development and implementation of policies and infrastructure which enable data access, storage and sharing. This builds on UCL's existing leadership in hosting prestigious data resources (e.g. Consumer Data Research Centre (CDRC, case study 3), Cohort and Longitudinal Studies Enhancement Resources (CLOSER, case study 4), and provision of compute facilities (e.g. high-performance computing (HPC) facilities Grace and Thomas).

2.3.2 New infrastructure is needed

The UK government's Artificial Intelligence Sector Deal outlines the aim to develop data sharing frameworks such as the Data Trust[16]. These could provide third-party management and safeguarding of data.

There is a need to balance mechanisms that facilitate the benefits of data sharing, while importantly maintaining public trust and confidence.

There is the opportunity to build on UCL's expertise in the social implications of new technologies and in developing new approaches to data management. For instance, DRIVE (**D**igital **R**esearch, **I**nformatics and **V**irtual **E**nvironments) is a unique informatics hub to harness the power of the latest technologies to revolutionise clinical practice and enhance the patient experience[17]. It is both a physical and conceptual unit and is the result of a unique partnership between Great Ormond Street Hospital, UCL, NHS Digital and leading industry experts.

[16] AI Sector Deal Department for Business, Energy & Industrial Strategy (2018)

[17] New unit opening at Great Ormond Street Hospital set to revolutionise how technology is used in hospitals

Case study 4: CLOSER - Cohorts and Longitudinal Studies Enhancement Resources
Department of Social Science, UCL Institute of Education

In 2012, CLOSER – a partnership of 8 studies, the UK Data Service and the British Library – was established to maximise the use, value and impact of the UK's longitudinal studies portfolio. Over the past six and half years, CLOSER has played a vital role in facilitating collaboration across the longitudinal research community, promoting opportunities and best practice in comparative longitudinal research, raising the profile of the studies, and generating impact. This has been achieved through five areas of work:

- **Data discoverability:** a searchable metadata repository that offers detailed documentation on the data collected across these longitudinal studies, aiding data discoverability.
- **Data harmonisation:** CLOSER has been a driving force in retrospective data harmonisation. This brings longitudinal data together in a consistent format and enables researchers to compare data from different studies for the first time.
- **Data linkage:** CLOSER has played a critical role linking data held by government to survey data collected by longitudinal studies enables researchers to gain rich insights into how different aspects of people's lives interrelate. It has also been instrumental in helping longitudinal studies overcome the range of legal, ethical, social, practical constraints facing their linkage efforts.
- **Training and knowledge exchange:** CLOSER's programme of knowledge exchange events, workshops, resource reports, and its flagship introductory resource, the Learning Hub, complement the efforts of individual studies and support skill development across the longitudinal community.
- **Impact and policy engagement:** CLOSER works to ensure the value of longitudinal data and evidence for policy is recognised, by influencing government and Parliament, supporting the academic community to engage with policymakers, and funding research that addresses the biomedical, social, economic and environmental challenges facing the UK.

Data Resource Profile: Cohort and Longitudinal Studies Enhancement Resources (CLOSER), D O'Neill, M Benzeval, A Boyd, L Calderwood, C Cooper, L Corti, E Dennison, E Fitzsimmons, A Goodman, R Hardy, H Inskip, L Molloy, A Sacker, A Sudlow, A Sullivan, A Park, International Journal of Epidemiology, dyz004, 2019 https://doi.org/10.1093/ije/dyz004

2.4 Reduce risk of non-compliance of UCL's responsibilities

The external regulatory environment for data is complex and changes frequently, creating considerable challenges for researchers and potential liabilities for UCL. Researchers are not governed solely by UK requirements but also international bodies such as the National Institutes of Health (NIH). Breaches could lead to legal, financial (up to 4% annual global turnover penalty for infringing GDPR, and $1.5 million for the USA Health Insurance Portability and Accountability Act (HIPAA)) and reputational damage (see Table 1 for more examples). The regulation surrounding creation, sharing, and analysing data is becoming more challenging for researchers to navigate, which increases risks to UCL and creates a barrier to research activity.

Country	Key Legislation	Potential Penalties
Australia	Federal Privacy Act 1988 (Privacy Act) and its Australian Privacy Principles (APPs)	Fines of up to AU$420,000 for an individual and AU$2.1 million for corporations may be requested
Canada	Personal Information Protection and Electronic Documents Act ('PIPEDA')	Fine of not more than $100,000.
Europe	The General Data Protection Regulation (Regulation (EU) 2016/679)	Up to EUR 20 million or, in the case of an undertaking, up to 4% of total worldwide turnover of the preceding year, whichever is higher.
Hong Kong	2012 Personal Data (Privacy) Ordinance (Cap. 486) (Ordinance)	Fine of up to HK$1 million and imprisonment of up to five years.
Malaysia	Personal Data Protection Act 2010 (PDPA)	Fine of up to 500,000 Malaysian Ringgit (~ USA$156,850) and/or imprisonment of up to three years.
Philippines	Data Privacy Act of 2012	Imprisonment from three to six years as well as a fine of ~$20,000 - $100,000.
USA	California Consumer Privacy Act of 2018 (CCPA), effective January 1, 2020. The Health Insurance Portability and Accountability Act of 1996 (HIPAA)	Statutory damages between $100 to $750 per California resident and incident, An annual maximum of $1.5 million.

Table 1: Selected international data protection legislation[18]

[18] Data Protection Laws Of The World, DLA Piper

3 Current Challenges

We have identified six challenges that are preventing UCL from harnessing the maximum potential from research data and recommendations to address them.

Challenge 1: Coordination and support.

Challenge 2: Creating and Accessing data.

Challenge 3: Sharing and Curating data.

Challenge 4: Infrastructure for data storage and analysis.

Challenge 5: Skill development and training.

Challenge 6: Retention of highly skilled researchers and support staff.

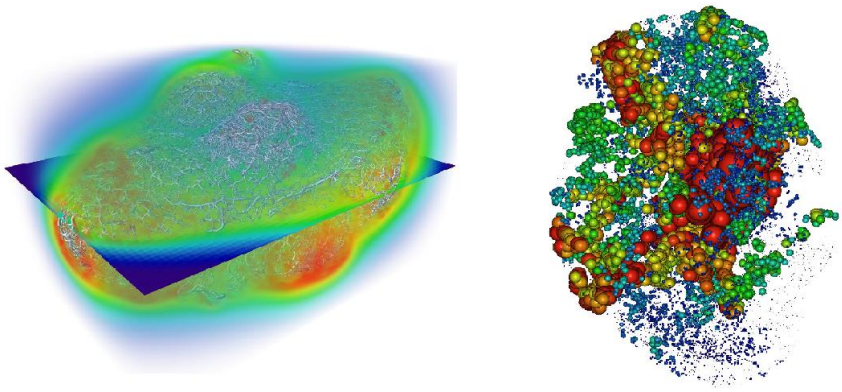


Image 4: Computational fluid dynamics with imaging of cleared tissue and of in vivo perfusion predicts drug uptake and treatment responses in tumours. Credit to Dr Simon Walker-Samuel, UCL Division of Medicine, Centre for Advanced Biomedical Imaging and Dr Rebecca Shipley, UCL Mechanical Engineering.

Computational fluid dynamics with imaging of cleared tissue and of in vivo perfusion predicts drug uptake and treatment responses in tumours Nature Biomedical Engineering, A. d'Esposito, P. W. Sweeney, M. Ali, M. Saleh, R. Ramasawmy, T. A. Roberts, G. Agliardi, A. Desjardins, M. F. Lythgoe, R. B. Pedley, R. Shipley & S. Walker-Samuel. Nature Biomedical Engineering 2018 (2), 773–787 https://doi.org/10.1038/s41551-018-0306-y

Challenge 1

Coordination and Support

"Our ability to do great things with data will make a real difference in every aspect of our lives."

Jennifer Pahlka, Founder and Executive Director for Code for America

3.1 Challenge 1: Coordination and support

Researchers stated that it is unclear to them how UCL's professional services teams' processes connect and interrelate with each other regarding data or where to go for support when issues cut across professional service teams.

3.1.1 UCL is complicated to navigate

The expertise to provide advice or support required to access, use, or share personal and/or sensitive research data spans several UCL professional service teams (Appendix 7.4). Staff find it difficult to identify the appropriate UCL services, training courses, or infrastructures for their needs. 70% of the respondents of the RDM Survey 2016[5] indicated that either no one was identified as being in overall charge of data management within their team or department or they didn't know who was responsible. Researchers often rely on their personal network to identify internal services; however, many are not aware of the support available (e.g. Research Data Management website and the Research Data Storage facility are unknown to 60% of the participants who responded to the 2016 RDM Survey). Researchers find UCL's internal environment complicated to navigate.

Professional Service teams adjust the scope of their activities in response to changes in policy, legislation or regulations (e.g. the introduction of Data Protection Impact Assessments). These changes will also affect the demands on professional service teams and increased resource may be required to support researchers. UCL is also implementing new projects and initiatives in response to the external changes (see Table 2). This includes trying to influence researchers' behaviour, for example encouraging them to share more data (e.g. Open Science) and to restrict data sharing (e.g. GDPR

compliance). Researchers may receive mixed messages on how to share and protect their data from the multiple ongoing change projects. This may lead to oversharing of data and breaching GDPR or other regulations, as well as, under-sharing of data which would reduce the opportunities for further research and impact. There is no central forum to share information about these changes across central services, schools and faculties.

Many academic communities are interested in data-focused research However, whether researchers are able to take advantage of the opportunities, depends in part on the data maturity of the field and department. The Data Maturity Framework[8] presents the five stages of progress in data maturity (Unaware, Nascent, Learning, Developing and Mastering) together with the seven key themes (Data, Tools, Leadership, Skills, Culture, Uses and Analysis). Currently, the journey for departments to improve their competence in leveraging data is based on trial and error. A framework tailored for academic disciplines or departments could help identify the type of support that is likely to have the greatest impact to enable researchers to achieve their ambitions. This approach would also provide insights into the future demands on UCL's professional services and infrastructure regarding research data. As more disciplines adopt data-focused approaches, the demand and the diversity of support required from Professional Service teams will increase.

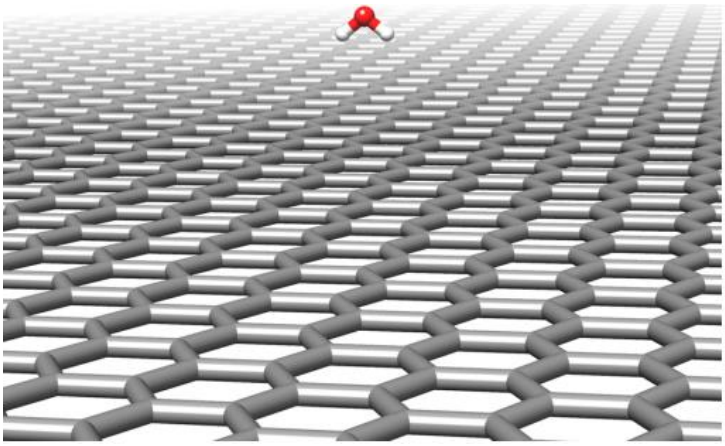


Image 5: Physisorption of Water on Graphene: Subchemical Accuracy from Many-Body Electronic Structure Methods. Credit to Prof Angelos Michaelides, UCL Physics & Astronomy and LCN and Prof Dario Alfe, UCL Earth Sciences.

Physisorption of Water on Graphene: Subchemical Accuracy from Many-Body Electronic Structure Methods Jan Gerit Brandenburg, Andrea Zen, Martin Fitzner, Benjamin Ramberger, Georg Kresse, Theodoros Tsatsoulis, Andreas Grüneis, Angelos Michaelides, and Dario Alfè The Journal of Physical Chemistry Letters 2019 10 (3), 358-368 DOI: 10.1021/acs.jpclett.8b03679

Initiative	Purpose	Lead
UCL GDPR Preparedness Programme	To ensure that UCL complies with the GDPR. This affects the way in which personal data can be collected, used, retained and deleted.	Legal Services
Devolved clinical research governance	To support Heads of Departments in SLMS to meet their research governance responsibilities.	Joint Research Office, OVPH
Data Safe Haven Service Re-Design	To support the storage and HPC analysis of sensitive (including identifiable) data which was generated at UCL or by an external organisation.	RITS, ISD
Developing UCL's safe room infrastructure.	To facilitate on-campus access to external datasets which require the researcher to be monitored by CCTV when they are using the data e.g. education, census and administrative data.	Centre for Longitudinal Studies; Institute of Healthcare Informatics, ISD, eResearch Domain
Open Science Office	To make UCL generated scientific research, data and publications accessible to all levels.	Library Services
Research Data Repository.	To provide the infrastructure and support for researchers to store and publish UCL generated open datasets launching .	RITS, Library Services,
'Value of Data' project	To determine how to recognise the respective patients', BRC Hospital's, and UCL's contributions to data generation, storage and curation in commercial agreements.	Translational Research Office, OVPH
HR strategy for RITS	To improve recruitment and talent management of highly skilled staff members.	HR, RITS

Table 2: Overview of selected projects impacting research data at UCL

3.1.2 Ethics, Governance, and Compliance

The ethics of data and data-focused technologies, such as AI, are mentioned frequently in the media and this is a rapidly emerging area. The Information Commissioner's Office's (ICO) report into the Cambridge Analytica scandal[19] highlighted that it is essential for higher education institutions to have the correct processes and due diligence arrangements to minimise the risk to data subjects and to the integrity of academic research practices. It also noted that while well-established structures exist in relation to the ethical issues that arise from research, similar structures do not appear to exist in relation to data protection. UCL has instigated several projects which are developing new data protection guidelines, training and governance structures (e.g. GDPR Preparedness programme, SLMS Devolved Governance see Table 2).

The regulatory landscape is complex, for example, one research project could be required to fulfil obligations to NHS Digital, UK, and international legislation, and research funders' policies (UKRI, NIH) etc. Researchers are responsible for complying with their regulatory and legal obligations regarding data. The Academic Handbook[20] only requires Head of Departments to ensure that researchers and students are aware of UCL's arrangements for research governance and the associated procedures.
The UCL Research Data Policy (2018)[12] provides a framework defining the responsibilities of all UCL members regarding how to manage their data.

UCL relies on positive and proactive action from individuals to identify and comply with their obligations appropriately.

It is important that ethical standards and practices are developed by individual disciplines, to ensure that they meet the needs of a particular project e.g. ensuring that no harm comes to sources in ethnographical or healthcare studies and other disciplines where sensitive data is used. UCL SLMS proposes[21] to implement devolved clinical research governance system to promote best practice. Each department will nominate a Research Governance Champion, who will be a member of their respective Institute's/Division's Research Governance and Integrity Group (RGIG). The RGIG will produce a Research Governance Plan and an annual report for the SLMS Clinical Research Governance Committee.

UCL has world leading academic expertise in data governance, such as Professor Jonathan Montgomery and Professor James Wilson (both UCL Laws), Dr Jack Stilgoe (UCL Science and Technologies Studies) and Dr Madeline Carr (UCL Department of Science, Technology, Engineering and Public Policy). They may be able to advise on the development of new UCL procedures and policies. There is the opportunity for UCL to provide leadership and to engage with new initiatives such as the UK Government's Centre for Data Ethics and Innovation and the Ada Lovelace Institute which promotes 'ethical practice in the public interest' concerning data.

[19] Investigation into the use of data analytics in political campaigns Information Commissioner's Office (2018)

[20] 20 UCL Academic Manual, Chapter 12, paragraph 9 (February 2019)

[21] Implementing devolved clinical research governance in SLMS. Nick McNally, Janet Darbyshire, Susan Kerrison, Rajinder Sidhu – Joint Research Office February 2019

3.1.2.1 Using externally generated data

Research increasingly involves personal data obtained from third party sources. However, it will not always be ethically appropriate to accept data from external organisations. There is the potential to damage the public's trust in research and UCL's reputation. UCL Research Data Policy (2018)[12] does not include data generated by external organisations explicitly.

Organisations which provide data can request an audit to assure that their data is being governed and managed as expected. This requires the PI to develop and implement appropriate management processes. Individual researchers are responsible for this and often it is not clear to them who is responsible to give them support from the department, faculty or professional services (see section 3.1.1). If the appropriate paperwork is not in place to create the audit trail required, UCL is at risk of losing access to data and to contractual, legal and regulatory penalties.

3.1.2.2 Sharing UCL generated data

To share data with external partners with confidence, staff need to understand their data ethics, copyright and licensing options that are available to them. Researchers need to be aware of the dangers of 'dual use data', data which poses a risk to safety and security if misused, and data where participants have not agreed to their data being shared. These are important considerations with serious (including legal) implications. Following the investigation into Cambridge Analytica[19], Universities UK is working with the ICO to consider the risks arising from the use of personal data by academics in a private research capacity, and when they work with their own private companies or other third parties. UCL Innovation and Enterprise, UCL Consultants, Translational Research Office and other industry facing units need to be aware of the risks to ensure that all appropriate ethical and governance considerations have been fulfilled. The Department for Digital, Culture, Media & Sport's Data Ethics Framework[22] developed for public sector data could be used to inform UCL's approach.

[22] Data Ethics Framework Department for Digital, Culture, Media and Sport (2018)

3.1.3 Building and maintaining public trust

It is crucial that research participants' and the public's expectations and attitudes in agreeing priorities, research design, and decisions affecting data use are taken in to consideration. The value of data can only be unlocked if we maintain public trust in using data. Transparency is central to demonstrating trustworthiness. Currently, certain initiatives engage with the public with respect to the use of data such as the UCLH BRC project About Me, My Care and the Research Hospital.

The Academy of Medical Sciences[23] has identified that there are strong expectations from patients and the public for transparency around the use of data-driven technologies. The public want clarity about why and how data-driven technologies and associated patient data are being used, by whom, for what, and how decisions about these uses are arrived at. When sharing how data are being used with the public, researchers need to be mindful not to prematurely disclose their intellectual property.

Providing open data in a publicly accessible format (which does not require high levels of data literacy) will enable research insights to benefit to individuals, local, and global communities. Promoting and facilitating Citizen Science projects which collect data would also provide avenues to build trust, public engagement, and research.

Recommendations 1–2 to improve coordination, governance, and compliance

1. Create a UCL Central Research Data Office (CRDO) led by a single, accountable, senior academic, reporting to UCL's senior management structures. The CRDO will:
    a. Be responsible for the implementation of this strategy and refresh it as needed;
    b. Act as a hub sharing best practice developed by professional service teams, improving signposting and researchers' awareness of support available;
    c. Work with professional service teams to optimise internal processes to create, access, store, analyse, curate and share data;

2. To ensure that UCL's governance, ethical, legal and regulatory processes enable UCL researchers to confidently use research data, in particular;
    a. Ensure UCL provides ethical review and assurance regarding its numerous national and international legal, regulatory and ethical obligations relating to data;
    b. Work with HR to ensure that maintaining data security and confidentiality is a core requirement of all staff and students, with clear penalties for breaches.

[23] Our data-driven future in healthcare.
The Academy of Medical Sciences 2018

Challenge 2

Creating and Accessing Data

"Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom."

Clifford Stoll, astronomer

3.2 Challenge 2: Creating and accessing Data

Accessing data, whether it has originated from a UCL researcher or an external data provider, enables researchers to address questions that would not otherwise be possible.

3.2.1 Creating data at UCL

Planning, developing large datasets and collecting data can be very time-consuming for the researcher involved. This reduces the amount of time available for data analysis and writing publications. It is important that datasets receive a Digital Object Identifier (DOIs) and researchers are encouraged to cite the datasets that they use. The Academic Promotional Framework recognises activities that promote Open Science, and this includes creating and sharing data. This is important so that the time invested by an individual researcher is recognised appropriately.

3.2.2 Relationships with data providers

There is no institutional oversight of UCL's data licence agreements, and it is not clear how many people pay to access the same data or develop relationships with the same data providers.

Relationships with data providers are almost exclusively managed by individual researchers. Large strategic datasets, however, can be used by multiple disciplines across various departments and faculties, and can be accessed from multiple points of contact. This was the ESRC's motivation to create the Big Data network of centres to make strategic administrative, business and government datasets available to the social science research community. This model has been successfully used for consumer-related data at UCL for the CDRC (see case study 5).

Adopting an institutional relationship management approach with external organisations who provide strategic data sets for research, would strengthen UCL's influence while negotiating contracts, reduce the risk of losing access to the data if a researcher leaves UCL, and increase the external organisation's understanding of the academic perspective in negotiations. This would require data management and administration support UCL in maintaining the full licence for a particular data base. The internal governance would need to ensure that the data is controlled effectively in line with legal, contractual and regulatory obligations, and that it is managed ethically and with integrity. This is analogous to UCL Innovation and Enterprise's approach to managing strategic business partnerships. For example, Transport for London have asked for one point of contact at UCL to renegotiate data licence agreements, following the introduction of GDPR legislation, and that UCL accepts unlimited liabilities.

3.2.3 Contracts and data licenses

Contracts negotiated regarding access external confidential datasets currently tend to restrict use to a narrow definition such as a specific investigator for a particular project, and a new contract is negotiated *ab initio* when a new researcher is recruited to the project. This takes up valuable time of research contract negotiators (such as members of Research Contracts, Innovation and Enterprise, UCL Consultants) and delays the research.

There is *ad hoc* sharing of best practice on how to structure contractual arrangements or on governance processes regarding data access between researchers and contractual teams. Various solutions have been developed across UCL which could be shared more widely (e.g. CDRC, CLOSER, CALIBER, Smart Energy Research Lab (see case studies 3, 4, 5 and 6).

There is the opportunity for UCL to explore an institutional framework agreement with data providers with the aim to reduce barriers for researchers access to data and to reduce time spent during contractual negotiations and the overall cost to the institution.

Data licences are purchased by individual research groups or departments without any insight into whether UCL already has a licence. Also, there is no consideration of whether it is more appropriate to purchase a licence on a per project basis, or an institutional licence and recover costs via grant income. The amount recharged to grants would need to reflect the data licence, and internal data management, as well as administration costs required for UCL to provide access responsibly. For example, the Clinical Practice Research Datalink (CPRD) is a government service which enables access to anonymised NHS primary and secondary care records for public health research. CALIBER is platform within a research group in the UCL Institute of Health Informatics (see case study 5). It provides UCL researchers with access to "research ready" variables extracted from linked NHS electronic health records and administrative health data under licence from the CPRD. The Health Improvement Network (THIN) database provides non-identified patient data from UK General Practice (GP) clinical systems. It costs the UCL Department of Primary Care and Population Health ~£65K per year for the data licence. They are currently exploring whether to establish a consortium whereby other departments and institutes contribute towards the cost of the licence and internal data management and administration.

Case study 5: CALIBER
UCL Institute of Health Informatics, Faculty of Population Health Sciences

CALIBER is a collaboration-driven data sharing platform and prioritises collaborations based on scientific-added value, capacity development and sustainable funding. CALIBER provides access to "research ready" variables extracted from linked NHS electronic health records and administrative health data under licence from the Clinical Practice Research Datalink (CPRD). This includes data of primary care activity and links to secondary care (such as Hospital Episode Statistics inpatient, outpatient, A&E and diagnostic imaging dataset data) and national registry social deprivation and mortality data from the Office for National Statistics.

The CALIBER platform can support researchers from development of research proposals through to release of CPRD data in the CALIBER data safe haven. CALIBER has supported innovative research spanning rare and common clinical conditions and development of novel methods and tools, including machine learning to improve health and healthcare. It has led to 72 research projects and over 50 publications. Research findings derived from CALIBER have informed clinical guidelines and affected the health and healthcare of millions with cardiovascular disease.

Case study 6: Smart Energy Research Lab (SERL)
UCL Energy Institute, Bartlett – the Faculty of the Built Environment

The £6m EPSRC Smart Energy Research Lab (2017 –2022) will provide a secure, consistent and trusted channel for researchers to access high-resolution energy data. Data will be collected from households via SERL on a strictly voluntary basis and with the explicit consent. Additionally, SERL's strict governance framework will ensure that only accredited researchers will have access to anonymised data.

The portal will transform UK energy research through the long-term provision of high quality, high-resolution energy data that will provide a reliable evidence base for intervention, observational and longitudinal studies across the socio-technical spectrum.

3.2.4 Internal data access and infrastructure

The UCL's computer network is a significant barrier to being able to download large external datasets to desktop computers. Research groups downloading the same large dataset will use large quantities of compute storage, increasing the demands on the local and central infrastructure. The computer network also limits researchers' ability to access data internally (e.g. transferring large datasets from offices across Gower Street) and to external partners, with appropriate identity and access management systems.

3.2.5 Research data and teaching

Undergraduate and postgraduate students would benefit from access to data for dissertation projects. Data access, data protection training and ethics procedures can present a barrier due to the shorter life span of dissertation projects. Many postgraduate courses across BEAMS, SLMS, SLASH and IOE include modules on data-focused methods (such as machine learning and computational modelling) which would benefit from access to research data. There is the opportunity for the proposed CRDO to liaise with ARENA to develop opportunities which could enhance UCL's teaching and to contribute to UCL's Connected Curriculum.

Recommendations 3–6 to improve access to external data

3. CRDO to develop a relationship management approach for strategic data providers, to explore appropriate internal governance mechanisms to widening access to the data.

4. CRDO to explore governance and grant-recharge mechanisms to increase use of external data sets with high-cost licences.

5. UCL Research Services and CRDO to share best practice on negotiating with external data providers. Research Contracts to expand its capacity to negotiate data licences.

6. UCL to actively engage with all partner NHS Trusts to establish a unified approach to use of data for research or service evaluation by UCL.

Challenge 3

Curating and Sharing Data

"It's amazing how much data is out there. The question is how do we put it in a form that's usable?"

Bill Ford Jr.
Ford Motor Company

3.3 Challenge 3: Curating and sharing data

3.3.1 Knowledge and knowhow

Open access to research data can help speed the pace of discovery and deliver more value for funded research by enabling reuse and reducing duplication. UCL leads in aspects of open data, such as leading on developing resources for Research Performing Institutions to manage their research data[24] and UCL Library Services is developing an Open Science strategy which includes FAIR Data. UCL also participates in EUDAT Collaborative Data Infrastructure services, which enables European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment[25]. The UK Open Research Data Concordat[26] (see Appendix 7.2) aims to make data openly available for use by others wherever possible in a consistent manner. UCL's position on the Concordat is currently being reviewed.

Sharing data can be facilitated by implementing the FAIR principles of data management. The FAIR principles outline four broad requirements to make data: findable, accessible, interoperable and reusable. There are two major drivers for sharing data: 1) research integrity and reproducibility: availability of the data supporting the findings in research; and 2) the potential for reuse: availability of data for sharing with other users[27]. This will influence what type of data and level of curation is required to prepare the data before it is shared (e.g. raw, processed, supporting publication, synthesis, curated datasets). Researchers are encouraged to complete a Data Management Plan using a funder template or to use DMPOnline as this will make it easier to release data in

[24] LEARN Project (Paul Ayris, UCL Library)
[25] EUDAT project (Peter Coveney, UCL Chemistry)
[26] Concordat on Open Research Data (2016)

[27] What to Keep: A Jisc research data study. February 2019 Neil Beagrie (Charles Beagrie Ltd)

the most effective manner at the end of a project[28].

UCL could consider encouraging the reuse of data by highlighting high quality datasets that are available for reuse. For instance, open access data sets could apply for Open Data Certification. This is an independent quality mark from the from the Open Data Institute benchmarked against standards which assess the level of support and steps taken to make data reusable and discoverable.

Open Data Certification[29] aims to encourage best practice, and increase trust in the use of open data.

However, many UCL investigators still find it challenging to share datasets, particularly if they do not have access to a subject or funder-specific repository or a repository appropriate for sensitive data. shows only partial disciplinary coverage.[30]

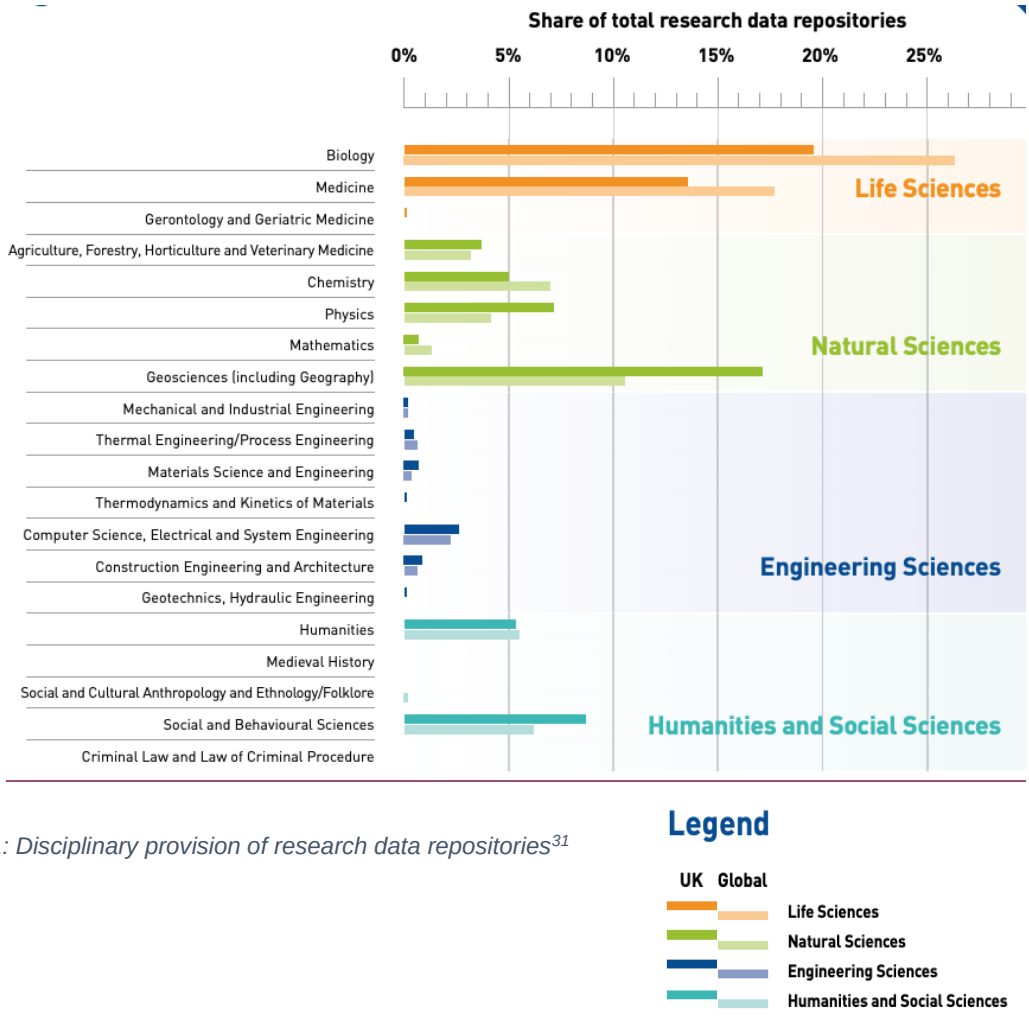


Figure 1: Disciplinary provision of research data repositories[31]

[28] What is FAIR data? UCL Libraries Research Data Management Blog

[29] Open Data Certificate, Open Data Institute

[30] Realising The Potential: Final Report Of The Open Research Data Task Force 2018

[31] Taken from the "Realising The Potential: Final Report Of The Open Research Data Task Force" July 2018 (page 21).

3.3.2 Processing and curating data

Data needs to be prepared, curated, and processed to unlock its full potential. Researchers need the knowledge of how to best prepare data for analysis to create the most suitable dataset and variables for the particular analysis, to enable them to act with integrity. This also involves creating the accompanying documentation to enable the data to be shared and the analyses to be reproduced.

If their academic discipline has not agreed metadata standards it is difficult for a researcher to determine the best way to organise and curate data prior to it being shared. To address this, Delft University of Technology has a team of data stewards to provide disciplinary support for research data management and sharing by their researchers[32].

Machine readability is a key issue which will affect how valuable the data set is to others. This involves structuring research data so it is machine readable and will greatly support the development and application of AI to more sectors and research disciplines. AI offers great potential to realise new value from that data, but only if training data sets are available. Hall and Pesenti**Error! Bookmark not defined.** recommend that publicly-funded researchers should publish research data in machine-readable formats.

Data anonymisation could facilitate sharing of data sets that would not be possible otherwise. Data protection law does not apply to data rendered anonymous in such a way that the data subject is no longer identifiable. Fewer legal restrictions apply to anonymised data, and this would permit wider sharing of data[33].

Data linkage connects pieces of information from different datasets that are thought to relate to the same person, family, place or event. Data linkage can enhance the quality and scope of academic study of the data set. There are a number of practical obstacles to data linkage, reflecting legal, ethical and social constraints, and different data sources having different access requirements and restrictions. Identifying and adopting appropriate strategies for obtaining consent and approval is key to developing linked datasets[34].

Processing and curating data increases its potential to be reused in research and to generate socio-economic impact.

3.3.3 Data sharing

Researchers need to appreciate the value of data when developing collaborations with industry or other partners. The value could be based on the potential to generate income, further academic research or public benefit. In cases were data has originated from a third party (e.g. NHS, industry, SMEs), they will also have an interest in how the data is re-used for further research or commercialisation. If data is collected in a hospital or other, but needs to be processed at UCL, this can involve a lot of regulations and practical difficulties, and not all departments and institutes have the capacity to provide support for the researchers in these activities. It is routine to clarify the ownership and rights of background and foreground intellectual property before initiating collaborations with third parties. To enable UCL to

[32] On a (cultural) journey towards FAIR data.. Marta Teperek 2018

[33] Anonymisation: managing data protection risk code of practice. Information Commissioner's Office

[34] Data Resource Profile: Cohort and Longitudinal Studies Enhancement Resources (CLOSER)' Dara O'Neill et al. February 2019

maximise the value of its data, researchers will need support to have similar discussions with third parties regarding the ownership and rights of data. This would clarify how data generated in a collaborative project could be reused for research and/or commercialisation purposes.

Funders request datasets to be made available, but the curation will always be required beyond the lifetime of the grant or even the PI's position at UCL. It is not clear how UCL, or the wider HE community, will address this issue.

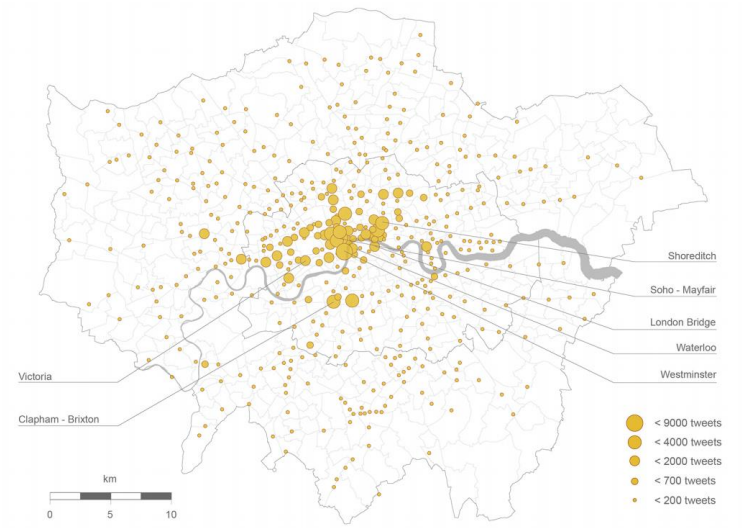


Image 6: Using Twitter data as a proxy for human activity in urban space (The Bartlett Centre for Advanced Spatial Analysis)

Sulis, P; Manley, E; Zhong, C; Batty, M; Using mobility data as proxy for measuring urban vitality. Journal of Spatial Information Science , 2018 (16). 137-162. DOI: 10.5311/JOSIS.2018.16.384.

Notably, these findings in UCL mirror responses to an international survey[35], where the key challenges to data sharing were identified as 'Organising data in a presentable and useful way', 'Unsure about copyright and licensing', 'Not knowing which repository to use', 'Lack of time to deposit data', and 'Costs of sharing data'. Infrastructure.

UCL does not have the infrastructure to share either non-sensitive or sensitive data with external organisations in the UK or internationally. Researchers rely on third party file transfer protocol (FTP) servers or sending hard drives by mail. Currently, investigators or departments contract independently with external providers and perform due diligence necessary to ensure compliance with UCL's legal responsibilities.

Infrastructure is needed to enable researchers to hold data securely on site and enable sharing of data under appropriate permissions and identity and access management systems with external partners and other data-intensive research groups within UCL.

Recommendations 7–12 to improve sharing and curation of data

7. CRDO to promote a data rich research environment promoting maximum appropriate access to UCL-generated datasets.

8. CDRO promote data curation and processing to maximise the impact of UCL's datasets.

9. CRDO to facilitate a joined-up service providing advice to researchers on how to comply with their data sharing obligations spanning contractual, ethical, legal, regulatory, IP, and funder requirements.

10. CRDO and Library Services to champion the recognition of data creation and sharing in the UCL academic promotion framework.

11. UCL to continue investment in data management, stewardship and FAIR data initiatives.

12. CRDO, Library Services, Research IT Services (RITS) to provide leadership and influence how relevant HE sector research data challenges are addressed.

[35] Practical challenges for researchers in data sharing Springer Nature Stuart, D et al (2018):. figshare. Paper.

Challenge 4

Infrastructure for Data Storage and Analysis

"Data is a precious thing and will last longer than the systems themselves."

Tim Berners-Lee, father of the Worldwide Web

3.4 Challenge 4: Infrastructure for data storage and analysis

The most common method identified for storing research data was on a personally owned computer (45% of responses to the 2016 RDM survey) and 20% did not know where they had archived their data or had no plans for long-term preservation after completing their research project.

It will not be feasible to store all of UCL's data in perpetuity due to the time and costs involved, as well as the physical limits of storage. UCL will need to prioritise the data that it will store long-term.

UCL will need to determine whether to continue to store datasets once minimum term for data retention, required by funders or law, is reached. The criteria to retain data could be informed from existing research data appraisal and selection processes that have been established by some domain repositories such as the social sciences and natural environment[27].

As technology is rapidly changing, there is the danger that if data is archived for 20 years that it will not be possible to access or read. A challenge for the future is to ensure that researchers will be able to usefully reuse historical data using new systems, technologies and formats.

We suggest that data provision needs to be categorised into four levels (see Table 3). Open data has the most support for data analysis. External resources support open data researchers funded primarily by EPSRC, NERC, and STFC (Tier2 facilities, ARCHER, JASMIN and DiRAC) and at an European level PRACE (see Appendix 7.5).

However, there is no UCL, regional, or national resource that provides HPC for sensitive data (e.g. health data research funded by the MRC), so researchers purchase the capability from external providers. UCL relies on individual researchers to ensure that the solutions are appropriate for their data obligations and there is no sharing of best practice or preferred supplier.

Recommendation 13 - 15 to improve infrastructure for data storage, sharing and analysis

13. UCL to continue investment in the UCL data repository led by RITS.

14. ISD to continuously review the computer network, and other infrastructure provision for data.

15. RITS to provide options, including commercial, for secure data storage and high-performance computing analysis that are not currently available at UCL.

Data Type	Description	Example infrastructure
Ultrasecure	Data only to be held and analysed on unnetworked servers and computers in a highly secure and monitored environment	Jill Dando Institute Research Laboratory (see case study 7)
Highly secure	An environment that provides monitored access to data held in external DataSafeHavens e.g. sensitive data in Public Health England	Institute of Health Informatics' Safe Room
Secure	Data covered by regularity frameworks that can only be handled in prescribed fashion in controlled environment complying with ISO27001 standards	ITforSLMS DataSafeHaven
Non-sensitive	Data that is freely available and can be moved between computers, servers and networks without creating issues	Thomas, Grace, Legion, Myriad, desktop computers, UCL data storage facilities

Table 3: Four levels of data

Case study 7: The Jill Dando Research Laboratory (JDIRL)
UCL Jill Dando Institute (JDI), UCL Department of Security and Crime Science, Faculty of Engineering Sciences

The JDI has a research data laboratory (JDIRL) which enables the storage and analysis of official, official-sensitive and secret data within a secure university environment. It is a 'Police Assured Secure Facility' which means it is accredited as satisfying the stringent requirements and information governance provisions of the UK Government. It is located in a highly secure space with controlled access, secure wiring, and high levels of electronic screening to prevent eavesdropping and shoulder-surfing. The aim of the lab is to facilitate access to rich crime (and other) data sets that are currently unavailable to university researchers.

Since accreditation in 2015 a number of large-scale projects have used the JDIRL, including the ESRC funded Consumer Data Research Centre (see case study 3), an EPSRC funded multi-department project on crime, policing and citizenship and an ESRC project with the National Crime Agency to investigate labour trafficking. Three new projects making use of the lab will; investigate transnational human trafficking; evaluate interventions aimed at stalking behaviour, and; explore factors concerned with protection of national infrastructure.

We have a number of PhD projects that have been made possible through data sharing in the lab and an increasing number of students using data in the lab for their dissertation projects. This has been made possible through negotiated multiple-project access to key data sets from the police and other crime prevention stakeholders. These offer two-way collaborative benefits- students get experience of undertaking applied research with real world relevance and security and crime prevention stakeholders get access to skilled research analysts.

Challenge 5

Skills Development and Training

"Torture the data, and it will confess to anything."

Ronald Coase, winner of the Nobel Prize in Economics

3.5 Challenge 5: Skills development and training

Research excellence frequently requires state-of the art technology and consequently the academic community want to take advantage of new data and related approaches. However, some disciplines (e.g. biosciences, archaeology) have not traditionally included undergraduate or postgraduate level training in technical data management and analysis skills (such as data linkage or anonymisation) or associated knowhow (such as data ethics and research integrity).

Consequently, expertise is not distributed across the faculties equally. It can be difficult for researchers new to this field to identify training options and the most appropriate pathway to develop their skills in software development, acquiring data, hosting data, computational analysis and the safe use of data with disclosure control (e.g. Safe User of Research data Environments (SURE) Training by the UK Data Service).

Professional development for researchers in data management and analysis skills is highly fragmented and difficult to navigate. RITS provides training to use UCL compute systems and to learn coding languages (e.g. python). UCL Research Integrity is developing a training framework to ensure the integrity of their research (e.g. using the appropriate research methods, thorough research data management, consideration of ethical issues, etc.). Other relevant technical training providers include: ISD Digiskills, Centre for Applied Statistics Courses, Institute of Health Informatics Short Courses, UCL Research Data Management team, UCL Doctoral Skills Development Programme; Bloomsbury

Postgraduate Skills Network, lynda.com, UCL Research Integrity, eResearch Domain's career network, UK National Supercomputing Service (ARCHER), CDRC, CLOSER, Health Data Research UK London, UK Data Service, Administrative Data Research Network, Partnership for Advanced Computing in Europe (PRACE), and YouTube e.g. OpenCOBRA. Some training is only available for PhD students and not postdoctoral researchers or academics. It is not clear whether there are gaps in the training available or if there are multiple providers of the same training.

Reproducibility has been one of the major tools to establish the validity and importance of research findings. How research data is generated, stored, analysed, curated and documented will affect whether the results are reproducible[36]. Researchers also need the skills to critically assess the trustworthiness of datasets to decide whether they are appropriate for reuse. In some disciplines, there are a lack of standards to ensure data quality and researchers need to be aware of the potential harm of using poor-quality data[37]. Improving support from professional service teams and training for researchers in technical skills and the necessary knowhow will contribute to a culture of integrity at UCL[38].

Recommendation 16 to improve the skill development and training for staff

16. CRDO and internal training providers to develop pathways for research staff to develop their technical skills (e.g. data science) and associated knowhow (e.g. ethical approvals).

[36] Ensuring research integrity: The role of data management in current crises, https://doi.org/10.5860/crln.75.11.9224 Heather Coates 2014

[37] Data reusers' trust development, https://doi.org/10.1002/asi.23730 Ayoung Yoon 2016

[38] UCL Statement of Integrity 2015

Challenge 6

Retention of Highly Skilled Researchers and Support Staff

"I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding"

Hal Varian, Google

3.6 Challenge 6: Retention of highly skilled researchers and support staff

Departments and large data intensive projects require support from dedicated data managers, infrastructure services, systems administration, software development, statistical support, training and advisory services to be successful.

There is no clear career path within UCL for people with these skills. Professional service staff and researchers with this specialist technical expertise are challenging to recruit and retain due to public and private sector organisations outside academia paying higher salaries.

Recommendations 17–18 to improve the retention of highly skilled staff

17. CRDO to work with HR to ensure UCL professional service staff and researchers with specialist data skills have clear career paths within UCL.

18. HR to review the current salary for roles that require data specialist skills.

4 Overview of Central Research Data Office

In response to these challenges, we propose the creation of the Central Research Data Office. The Office would be responsible for improving UCL's research data environment. It would report to senior management structures within UCL. It could be led by a single, accountable, senior academic, reporting UCL's senior management structures. They would be supported by a coordinator and an appropriate budget. It would work closely with professional service teams and researchers to achieve its goals. It would have six key areas of activity.

A) Leadership and coordination

- Evolve UCL's Data for Research Strategy and be responsible for its implementation.
- Act as a hub to develop and share best practice for researchers and professional service teams, improving signposting and researchers' awareness of support available.
- Work with professional service teams to optimise internal processes to access, store, analyse, curate and share data.
- Develop a research data maturity framework to identify the type of support required by different academic communities, and to inform professional services of the demands expected in the future.
- Provide leadership and influence the development of consistent funder data-sharing policies, and address related higher education sector challenges working with UCL colleagues.

B) Governance and compliance

- Ensure that UCL's governance, ethical, legal and regulatory processes enable UCL researchers to confidently use research data.
- Ensure UCL's governance, ethical, legal and regulatory processes enable UCL researchers to confidently use research data and take advantage associated opportunities.
- Ensure UCL complies with its numerous national and international legal and regulatory obligations relating to data and coordinate processes that require the involvement of multiple professional service teams.

C) Access to external data

- Develop a relationship management approach for strategic data providers, to explore appropriate internal governance mechanisms to widening access to the data.
- Promote a data rich research environment by exploring governance and grant-recharge mechanisms to increase use of external data sets with high-cost licences.
- Share best practice on negotiating with external data providers.
- Engage with all partner NHS Trusts to establish a unified approach to use of data for research or service evaluation by UCL.

D) Curation and sharing data.

- Promote a data rich research environment promoting maximum appropriate access to UCL-generated datasets.
- Promote data curation and processing to maximise the impact of UCL's strategic datasets.
- Champion investment in data management, stewardship and FAIR

data initiatives led by Library Services.
- Work with other professional service teams to provide to provide researchers with joined-up advice on how to comply with their data sharing obligations spanning contractual, ethical, legal, regulatory, IP, and funder requirements.
- Provide leadership, with UCL Library Services and UCL RITS regarding research data

E) Infrastructure

- Champion continued investment in the UCL data repository and relevant compute infrastructure led by RITS, ISD.

F) Training and careers

- Work with internal training providers to develop training pathways for staff to develop their technical skills (e.g. data science) and associated knowhow (e.g. gaining ethical approvals)
- Work with HR to ensure UCL professional service staff and researchers with specialist data skills have clear career paths within UCL.
- Promote recognition of data creation and sharing within the UCL academic careers framework, with Library Services.
- Work with HR to ensure that maintaining data security and confidentiality is a core requirement of all staff and student contracts with clear penalties for breaches.
- Work with HR to review the current salary for roles that require data specialist skills and to determine if a market rate uplift is appropriate and to review the existing Market Pay Policy.

5 Risks of not acting

5.1 Loss of research opportunities or funding

Poor data management practices risk the loss of research data to the research community through data deletion. Academics are not motivated to share data or to explore how to implement FAIR principles in their respective disciplines. This reduces the scope to reuse the data in future research collaborations or funding proposals or for educational purposes (e.g. student projects). It also limits the ability of researchers to reproduce their findings, and to act with integrity.

Data needs to be processed and curated to unlock its full potential. By not acting, UCL is underusing its existing unique data corpus, which could be used to initiate novel, cross-disciplinary funding proposals. External organisations will collaborate with other universities if UCL's systems and processes are too cumbersome or do not meet their governance requirements. UCL could lose industry and research funding and the ability to participate in research projects.

Researchers may receive mixed messages on how to share and protect their data from the multiple ongoing change projects (see Table 2) This may lead to oversharing of data and breaching GDPR or other regulations, as well as, the under-sharing of data which would reduce the opportunities for further research and impact.

There can be a high barrier to access externally-generated data. Researchers find it difficult to gain an overview of what is involved, and who can provide help at the appropriate stage. This includes negotiating data licence agreements, data licence costs, establishing processes, gaining experience, accessing resources needed to acquire data and downloading external data. Access to external data often relies on academics with social connections with an external data provider (e.g. companies, charities). If individuals move institutions, UCL risks losing access to data sets, if they are not openly available. A huge volume of UCL research activity relies on access to externally-generated data. UCL Research Data Policy (2018)[12] only references data generated within UCL. This puts some projects at risk of failure, and of losing of collaborations and the ability to attract funding.

Delays in recruiting technical professional service staff will delay infrastructure projects and also reduce professional services teams' capacity. Difficulties in recruiting researchers with technical expertise will reduce UCL's capacity to perform world class research and secure grant funding. These risks can delay or derail research projects and impact UCL's ability to respond to successfully to funding opportunities.

5.2 Duplicating effort and resources

The difficulties researchers have in identifying support and navigating UCL's internal environment risks staff spending time to repeatedly devise solutions to a particular issue. If researchers cannot find the support they need, they will create or procure their solutions that are appropriate for the individual, but not UCL as an institution.

Researchers use grant funding to procure external services or infrastructure which exist within UCL. For certain datasets, it might be more cost effective to purchase an institutional licence, manage access and recover costs via grant income, than purchasing multiple project licences. Staff develop multiple solutions for the same issues, potentially duplicating costs.

5.3 Regulatory penalties

Low awareness or uptake of UCL services, training and infrastructure risks that the researchers will not comply with ethical or regulatory obligations. Researchers who develop their own solutions, increase UCL's risk of breaching legal, contractual or regulatory obligations. This is a systemic issue compromising the integrity of research at UCL which could lead to:

- GDPR breaches can result in fines of €20 million or 4% of UCL's global turnover. Non-European countries are strengthening their data protection legislation, which include financial penalties of ~USD$1 million or more for US, Australia, Philippines, Table 1).
- Damaging UCL's reputation nationally and internationally. UCL would lose trust from external organisations and the public, which would impact UCL's ability to access data or to secure funding.

UCL trusts that individual researchers will understand and implement the measures required to comply with regulatory and legal obligations. The lack of oversight regarding compliance increases the risks that UCL faces financial penalties (up to 4% of turnover for GDPR) and reputational damage. There is no guidance or a preferred suppliers list for UCL infrastructure that is available to researchers (such as HPC for sensitive data or to share data), instead UCL relies on researchers to ensure that the suppliers meet regulatory or legal requirements.

5.4 Institutional leadership

UCL will be required to conform to new policies and infrastructure requirements regarding data. If we do not influence their development, these will be better tailored for other organisations.

6 Conclusion

UCL has the opportunity to be the leading institution for data-driven research and to inspire and empower the UCL research community to leverage the exciting opportunities created by recent technological advances within and across their respective disciplines. By improving coordination, processes and infrastructure UCL has the opportunity to harness the value from the research data by creating new research and impact outputs.

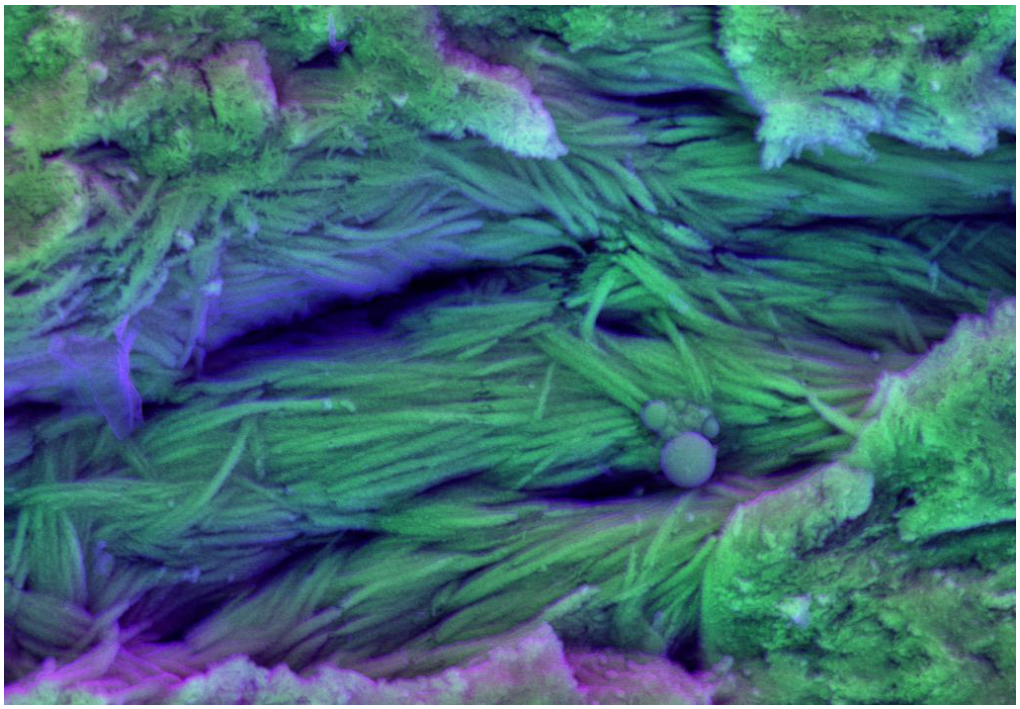


Image 7: A dinosaur fossil with calcified collagen fibres. Credit to Dr Sergio Bertazzo, UCL Medical Physics and Biomedical Engineering. https://www.mineralomics.org/

7 Appendix

7.1 Appendix I: Abbreviations

AI	Artificial Intelligence
ARCHER	Advanced Research Computing High End Resource (computer infrastructure)
CDRC	Consumer Data Research Centre, UCL Geography
CLOSER	Cohort and Longitudinal Studies Enhancement Resources, Social Science - IOE
CPDR	Clinical Practice Research Datalink
CRDO	Central Research Data Office
CRUK	Cancer Research UK
DiRAC	Distributed Research utilising advanced Computing (compute infrastructure)
DOI	Digital Object Identifier
DRIVE	Digital Research, Informatics and Virtual Environments, UCL Computer Science
EPSRC	Engineering and Physical Sciences Research Council
ESRC	Economic and Social Research Council
FAIR	Findable, Accessible, Interoperable and Reusable
FTP	File Transfer Protocol
GDPR	General Data Protection Regulation
GP	General Practice
HIPAA	Health Insurance Portability and Accountability Act (USA)
HPC	High Performance Computing
ICO	Information Commissioner's Office
IP	Intellectual Property
IRIS	UCL Institutional Research Information System
ISCF	Industrial Strategy Challenge Fund
ISD	UCL Information Services Division
JRO	UCL Joint Research Office
MRC	Medical Research Council
NERC	Natural Environment Research Council
NHS	National Health Service
NIH	National Institutes of Health (USA)
OVPR	UCL Office of the Vice-Provost Research
PRACE	Partnership for Advanced Computing in Europe (compute infrastructure)
RDM	Research Data Management
REC	UCL Research Ethics committee
RGIG	Research Governance and Integrity Group
RICS	Research Impact Curation & Support
RIISG	Research Information and IT Services Group

RITS	UCL Research IT Services
SLMS	UCL School of Life and Medical Sciences
STFC	Science and Technologies Facilities Council
SURE	Safe User of Research data Environments training
THIN	The Health Improvement Network
UKRI	United Kingdom Research and Innovation

7.2 Appendix II: Concordat on Open Research Data 2016

UKRI is strongly committed to opening up research data for scrutiny and reuse, to enable high-quality research, drive innovation and increase public trust in research. The Concordat (Concordat on Open Research Data (PDF, 178KB)) was developed by a UK multi-stakeholder group to provide expectations of best practice reflecting the needs of the research community. The four original signatories (the former Higher Education Funding Council for England, UKRI (the former RCUK), Universities UK and the Wellcome Trust) have now been joined by The Natural History Museum, Cancer Research UK, Sheffield Hallam University, Scottish Funding Council, The Higher Education Funding Council for Wales, and University of Glasgow. The concordat sets out ten principles with which all those engaged with research should be able to work:

1. Open access to research data is an enabler of high-quality research, a facilitator of innovation and safeguards good research practice.
2. There are sound reasons why the openness of research data may need to be restricted but any restrictions must be justified and justifiable.
3. Open access to research data carries a significant cost, which should be respected by all parties.
4. The right of the creators of research data to reasonable first use is recognised.
5. Use of others' data should always conform to legal, ethical and regulatory frameworks including appropriate acknowledgement.
6. Good data management is fundamental to all stages of the research process and should be established at the outset.
7. Data curation is vital to make data useful for others and for long-term preservation of data.
8. Data supporting publications should be accessible by the publication date and should be in a citeable form.
9. Support for the development of appropriate data skills is recognised as a responsibility for all stakeholders.
10. Regular reviews of progress towards open research data should be undertaken.

7.3 Appendix III: Overview of UCL professional services involved

This is adapted from the 16th November 2017 RIISG paper "Research Data Governance – Risks and Opportunities" by G. Rees, C Gryce and T Peacock.

Data Protection Office (Legal Services)

Responsible for managing UCL's compliance with Data Protection, Freedom of Information legislation and GDPR. Support researchers by registering their research projects, publishing guidance and clarity on anonymisation, data storage, international transfers and other overseas research. This includes the GDPR workstream.

Research Contracts

Responsible for negotiating and managing all of UCL's research-related agreements, including Data Access and Data Sharing Agreements. Authorise contracts as UCL signatory.

The Joint Research Office (JRO)

JRO between UCL/UCLH, provides support for research management, biostatistics, finance, contracts, regulatory affairs, commercialisation of research and negotiation of **MTAs for clinical trials.**

ISD Information Governance (IG) Advisory Service

Closely linked to the Data Safe Haven, supports researchers in complying with the SLMS IG Framework and NHS Digital IG Toolkit.

Research Impact Curation & Support (RICS - OVPR)

Supports dissemination, engagement and impact of research.

ISD Research IT Services (RITS- ISD)

Develops, delivers and operates IT services to assist UCL researchers in meeting their objectives at each stage of the research lifecycle including High Performance Computing platforms, a Data Safe Haven environment, collaborative research software development services, Research Data Storage and Repository services, specialist support and training.

ISD Architecture

Develops computer network systems across UCL.

Various Departmental IT teams which provide support for research

e.g. Computer Science, Physics and Astronomy.

Records Office (Library)

Provides help and advice to staff on information management issues, produces policy in line with legislation and promotes good governance.

Research Data Management (Library)

Provides guidance to manage research data and find best practices to plan ahead for data management and sharing.

UCL Research Ethics Committee (REC)

All research proposals involving living human participants and the collection and/or study of data derived from living human participants which requires ethical approval to ensure that the research conforms with general ethical principles and standards. The UCL REC has the following delegated authority from Research Governance Committee (RGC).

IOE Ethics Committee

The UCL Institute of Education maintains its own REC. At the Institute, all research projects by staff, students or visitors which collect or use data from human participants including secondary data analysis, systematic reviews and pilot studies, are required to gain ethical approval before data collection begins.

Research Integrity (OVPR)

The UCL Statement on Research Integrity (May 2015) sets out the standards expected by all those involved with research at or in collaboration with UCL, including adherence to the

Code of Conduct for Research and the principles of integrity set out in the UCL statement on research integrity.

UCL Business (UCLB)

The UCLB team is responsible for approving, negotiating terms and signing all incoming and outgoing material transfer agreements on behalf of UCL.

Information Security

The UCL Information Security Group exists to support the University in its management of information risk, providing strategic guidance, advice, and support to both staff and students. The team also co-ordinates the handling of security incidents within the college.

ISD Digitskills

ISD Digitskills are part of the Digital Education team within ISD. Providing online, classroom and blended IT learning opportunities to staff and students at UCL. Courses are available to students and staff. Example courses include: An Introduction to R with RStudio, Data, Visualization in R with ggplot2A Quick Introduction to UNIX, Introduction to Matlab, Introduction to Research Programming using Python.

ITforSLMS

ITforSLMS provides training and guidance on information governance responsibilities and control, how these affect research activities and how the SLMS Information Governance Framework can help researchers meet their legal obligations.

7.4 Appendix IV: Research Data Working Group

The members of the Working Group involved were Le Wei, Richard Mott, David Beavan, Irene Petersen, Andrew Hayward, Trevor Peacock, Phil Luthert, Jonathan Tennyson (chair), Louise Chisholm, David Osborn, Kate Walters, Harry Hemingway, Spiros Denaxas, Rosalind Raine, Ruth Gilbert, Claire Gryce, James A J Wilson, Amitava Banerjee, Jessica Sheringham, Emiliano De Cristofaro.

The Group met twice to identify the challenges to explore further. We invited the UCL community to share their experiences, opinions and ideas regarding data intensive research at UCL and what the research environment should look like in 10 years' time. In either an open workshop or via written submissions. The focus of the consultation was on:

1. The need for UCL to improve its environment for data intensive research
2. Improving access to data at UCL
3. Improving data curation, use and reuse at UCL
4. Improving support for researchers and staff at UCL

Written submissions were received from:

- Research Data Management Group (UCL Library)
- Richard Mott (Professor, Biosciences)
- Ruth Lovering (Principal Research Associate, Institute of Cardiovascular Science)
- Stephen Jivraj (Lecturer, Institute of Epidemiology and Health Care)
- Hallgeir Jonvik (Data Manager, Institute of Neurology**)**
- Paul Longley (Professor of GIS, Director of Consumer Data Research Centre, Geography)
- Aida Sanchez (Senior Data Manager, Centre for Longitudinal Studies, IoE)
- Matteo Carandini (Professor, Institute of Ophthalmology)
- Isaac Bianco (Fellow, Biosciences)

7.4.1 Summary of the data strategy workshop

Attendees were Chris Seers, Hallgier Jonvik, Paul Hammond, Tau Cheng, Pia Hardelid, Paul Ayris, Fabrice Ducluzeau, Phil Luthert, Richard Mott, Paul Longley, Trevor Peacock, Clare Gryce , Jonathan Tennyson, Mary Rauchenberger, Jenny Bunn, Daniel Van Strien, Christine Orengo, James A J Wilson, Clare Thorne, Linda Wijlaas, Barry Mant.

The attendees selected which topics they wanted to discuss and they were guided by their facilitators to discuss the a) current situation, b) challenges and barriers, c) potential solutions, and d) benefits for UCL. The discussions were captured on flipchart paper.

The following topics were identified from the previous Data Strategy working group discussions:

- UCL Institutional Response
- Data Access
- Researcher Support
- Data Curation

After 30 minutes, the attendees then discussed a second topic in the same manner. The facilitators remained at the same topic. All of the attendees then came together after the second session and the facilitators shared an overview of what had been discussed on each topic. A summary of what was captured is below.

7.4.1.1 Workshop discussions

UCL Institution response

Current Situation

- Organic and unstructured approach
- Lack of institutional strategy to influence funding councils
- Library – open access data
- ISD focus on hardware
- Training limited for researchers
- Missing Legal Services
- GDPR

Challenges

- Different types of data
- Lack of solutions (hardware, metadata)
- Guidance is often generic
- It is unclear who is responsible when
- Management of data across institutions and of legacy data
- Training demand
- Duplication of open source key data sets, repeatedly downloaded from external sources

Solutions

- UCL Data Conference
- Share practice
- Data scientist maps of expertise (for specific research fields)
- Lack of capacity in expertise in some areas

Create a Data Office →
To develop UCL's own agenda of good practice beyond funders requirement.

- To Data licencing agreements at institutional level and awareness of the agreements to prevent duplication
- Develop data governance framework
- New Roles : Data officer, data acquisition officer. Liaising with data champions across faculties / departments.
- Needs to be connected with existing UCL infrastructure/services.
- UCL Library
- Respective research governance committees
- Ethics committee. Note: Could the chief data officer to sit on committee and update the committee of new UK and international legislation.
- Data 'type' champions
- Data led hardware development
- Legals and contracts
- Information security team

The changing landscape:

- SLMS: radical shift in research towards data in the last 20 years
- Funding data led opportunities opening up in disciplines that were not traditionally data focused.
- Important to maintain in UCL's offering as a destination for research talent
- Larger datasets are becoming available but it is difficult to down load them (open access) due to network capacity

Risks of not acting:

- Financial
- Reputational
- Dominance in research areas
- Talent retention/ attraction /loss: infrastructure

Benefits of investing in this area:

- UCL Leadership in data use
- Improve trust of data use
- Influence government and wider community
- Golden Standard

Support and Training for UCL researchers

Current situation

- Open access is good
- Wide data types → custom training needed
- Expertise not fully used
- Special courses needed
- Legion: not enough instructions and priority is for groups
- Data is not well recorded / archived
- Not a data sharing culture
- Research data services are not well known.

- Funding requirements are not met.
- Bio data – what is identifiable?
- Data not open enough between UCL researchers
- Repeat of experiments due to not having access to older data
- Fragmented advice / confusion

Challenges / Barriers

- JRO – could publicise best practice
- Time consuming processes
- Research application data
- Applying for external data
- National computers are not secure enough for sensitive data
- Data repositories are not used / known
- Unstructured data
- 10 year data saving is difficult and is a requirement from research funders
- Making data organised / searchable
- Funding to make central databases
- Bio images, distributed data

Solutions

- Specific training
- Actual software solutions and libraries
- Domain specific Data vice-deans / pro-vice provosts
- Standard data management plans (epsrc etc require them)
- Legion training and introductions
- Encourage data sharing / access
- UCL data repository / training / template
- E lab books
- More access to back up
- Bio data training
- UCL data sharing framework agreements that can be used as the starting point for many subsequent contracts/projects with the same data provider.
- Better central data management and awareness of data assets in UCL
- Archiving programmes
- Encourage culture of data logging
- Saving data in transferable formats (Accessed by anyone)
- More legal advice / support for sensitive data
- “one stop shop” for advice
- Automatic data tagging for bio imaging?

Benefits:

- Students better equipped to use legion etc
- Don’t repeat work / funding
- Easier access to previous data
- Safer data storage
- Easier access to UCL generated data
- Don’t repeat experiments

Data Curation, Use and Reuse

Current situation

- Different between access and use
- Once a project is finished → what is the “end of life care” for data? Who is responsible?
- Secure drives – datasafehaven
- Each researcher works differently (e.g. A&H)
- EPSRC pressure for institutional data centres
- Currently have citation metadata → difficult when we start to cross disciplines
- Difficult to find out what data is available across UCL
- Standard data types eg physics and astronomy vs dew data types eg bioscience

Challenges and barriers

- Terminology ‘archive’
- Sensitivity of data
- Long term stewardship and curation of data at UCL
- Economic sustainability of data curation / stewardship?
- Who is responsible for data once a project ends?
- Interaction with external organisations
- Capturing meta data
- Integrate new data types
- Support for standardisation across research communities, beyond UCL e.g. neuroscience,
- Data safe haven is not suitable for use with high performance computing
- Funders do not understand the practical implications of data storage / sharing
- Data sharing of anonymised sensitive data:
- Could be interpreted incorrectly
- Will it be used responsibly?

- Understanding of what research data is?
- Includes: Photographs and recordings
- No culture of data sharing
- Process for tendering for a long-term data repository?

Solutions

- IDS capture data from different groups
- Research councils could drive adopting uniform technology approach
- standardisation of lab note books and protocols
- Can we earn from good practice in departments across UCL?
- Incentives to share data and for good practice→ could relevant points be included in promotion criteria
- Journal publication funding
- Track citations and link to REF
- Request contributions on grants for data curation
- Make curation easier to use
- Data review (like literature)
- Who else would have the data
- Funding for research that reuses data / data from different disciplines
- Dedicated data manager (new roles)

- Advice on how to anonymise data -> funded by research councils ?
- Involve UCL Stakeholders – EU Office ? H2020?
- Training on how to produce metadata -> Publicise
- Educational activities on data sharing and a cultural shift.
- Redesigning the datasafehaven to work with a supercomputer
- Funding of undergrads to use existing data sets

Benefits

- Metadata for data stored in other archives
- Cheaper research.
- Better transparency
- Impact as people use your data
- Understanding that data exists at UCL combined research catalogue
- Education (use data in masters projects)

Data Access

Current

- Data is held by different providers
- No institutional access or coordination to strategic data sets
- Very expensive to purchase licences
- Long ethics process (can need 7 ethics approvals from different external committees)
- There is little support from within UCL, so the wheel is repeatedly being reinvented and it relies on personal contacts with other researchers
- A lot of the challenges in accessing data are outside of UCL's control but no knowledge / best practice is shared to how to overcome the challenges in a changing landscape
- GDPR is creating new changes and will lead to a changing landscape as well
- Datasafehaven lacks computational power. Need to pay externally for secure access to secure data compute.
- Data exchange mechanisms – dropbox is frequently used for non-sensitive data or the datasafehaven or external secure facility (cost implication)
- Data linkage challenges
- Little understanding of data available and what can be accessed
- Personal informal networks are used to sharing data

Challenges / Barriers

- Network infrastructure (it is not possible to give external partners access, without becoming an honoury researcher)
- GDPR – there are many unknowns and it could be used as an excuse to not do or use opportunities. We need to overcome our and our collaborators reactions / difficulties / misconceptions
- Secure processing of data. Labs independently buy high computation power hardware but it is stored in inappropriate rooms
- Bandwidth connection leading to slow downloads, the speed and ease of access
- Challenges regarding the scale of data
- Data linkage challenges
- Changing regulation

    - Scale of datasets slow to download and to store locally. If local staff are lost, who has responsibility.
    - Run out of storage internally, need external providers
    - Use of OS and other standards
    - Distributed geography of UCL causes challenges with the network and variable speed of access to download data and access to repository,
    - Local systems are used as the fall back, as the space in central services is not available
    - Researchers will find work arounds in the absent of guidance → Risk

Solutions

- Training – for researchers and for professional services (contracts, research facilitators, JRO)
- Provide forms / guidance on how to get ethics approval from X (possibly from the JRO / research facilitators)
- Learning from success stories
- FTP solution to share data with collaborators
- GDPR: UCL to act as a leader in how to adapt to the new legislation, so it is not used as an excuse not to collaborate. Possibly with external validation?
- Legal advice: need to understand the law and what is reasonable and ethical within this. It is currently unclear for researchers
- Legal advice regarding IP and the involvement of honoury researchers' contracts required for collaborators to access data.
- Legal advice to clarify what the funder requirements are. Researchers are unlikely to go into the details of this and it is unclear what is permitted.
- Make contracts easier to add extra researchers without having to go through the process again. This needs documentation, processes and governance and awareness
- Institutional access to strategic data sets
- Data stores matched with suitable UCL compute facilities
- Access to internal data and the ability to manipulate it
- Process to confirm that you can be trusted with the data
- Clear guidance on the data within UCL is needed

- Is it anonymous (sliding scale)
- Is the any PID / commercial value
- Who is the owner and what are the funder requirements
- Access restrictions / requirements
    - Metadata – to describe datasets
    - Have a central facility which would provide secure access / safe rooms for all researchers. It would also provide an opportunity for networking and sharing best practice between those present
    - Publishing open access metadata and rules of engagement
    - Integrate compute and large scale storage (larger scale than current)
    - Universal access system (compatible with different operating systems, software)
    - Learn from external organisations – Public Health England do it well.
    - Look at other universities eg Harvard, Heidelberg (H-box = Heidelberg dropbox)
    - One stop shop for UCL Data storage: quantity network,

Benefits

- New research opportunities (AI etc)
- Faster Delivery
- Security

7.5 Appendix VI: Overview of compute resources available

Resource	Brief Description	Community it is available to
UCL Grace	UCL's for large multinode parallel computing workloads, that require more than 32 cores.	UCL researchers
UCL Legion	UCL's general-use cluster suitable for serial or parallel computing workloads	UCL researchers
UCL Myriad	UCL's high-I/O, high-throughput cluster. It contains nodes of a few different types including GPUs. It runs jobs that will run within a single node rather than multi-node parallel job	UCL researchers
UCL DataSafeHaven	UCL's Data Safe Haven has been certified to the ISO27001 information security standard and conforms to NHS Digital's Information Governance Toolkit. It can provide desktop data analysis.	UCL researchers
EPSRC Tier 2 facilities	CSD3 facility (led by University of Cambridge): supports large scale simulation and next generation data analytics capability for researchers across a broad range of disciplines The Materials Modelling Hub (also called Thomas and is led by University College London) supports small to medium sized capacity computing focusing on materials and molecular modelling. JADE the Joint Academic Data Science Endeavour (led by University of Oxford) is the largest GPU facility in the UK supporting world-leading research in machine learning. HPC Midlands Plus (led by Loughborough University) UK Tier	EPSRC researchers

	2 Regional supercomputer for the Midlands and wider region. Isambard the GW4 Tier 2 HPC service (led by University of Bristol) provides multiple advanced architectures within the same system in order to enable evaluation and comparison across a diverse range of hardware platforms. Cirrus (led by University of Edinburgh) provides a flexible, state-of-the-art High Performance Computing system that provides an ideal platform for users to solve their computational, simulation, modelling, and data science challenges.	
ARCHER	ARCHER provides a capability to allow researchers to run simulations and calculations that require large numbers of processing cores working in a tightly-coupled, parallel fashion. It supports NERC and EPSRC researchers.	EPSRC and NERC researchers
JASMIN	JASMIN provides the UK and European climate and earth-system science communities with an efficient data analysis environment, and supports NERC researchers	NERC researchers
DiRAC	DiRAC provides distributed HPC services to the STFC theory community	STFC researchers
PRACE	PRACE provides access to 7 leading-edge HPC systems (supercomputers) to researchers and scientists from academia and industry through a peer review process.	European researchers

7.6 Appendix VII: Overview of external data repositories

7.6.1 Funder repositories

NERC supports five data centres covering a range of discipline areas:

- British Oceanographic Data Centre (Marine)
- Centre for Environmental Data Analysis which includes:
    - British Atmospheric Data Centre (Atmospheric)
    - NERC Earth Observation Data Centre (Earth observation)
    - UK Solar System Data Centre (Solar and space physics)
    - Environmental Information Data Centre (Terrestrial and freshwater)
    - National Geoscience Data Centre (Geoscience)
- Polar Data Centre (Polar and cryosphere)

UK Data Service Repositories. The UK Data Service is a comprehensive resource funded by the ESRC to support researchers, teachers and policymakers who depend on high-quality social, historical and economic data.

7.6.2 External repositories

Re3data.org offers detailed information on more than 2,000 research data repositories, re3data has become the most comprehensive source of reference for research data infrastructures globally.

Zenodo accepts all research outputs from across all fields of research and in any file format (as well as both positive and negative results). Zenodo assigns all publicly available uploads a Digital Object Identifier (DOI) to make the upload easily and uniquely citeable.

Figshare is an online digital repository where researchers can preserve and share their research outputs, including figures, datasets, images, and videos. It is free to upload content and free to access, in adherence to the principle of open data

7.7 Appendix VIII: Professional service and academic teams that provide support to UCL researchers

	Training	Data sharing agreements advice	Infrastructure	Data sharing	Ethics and governance advice
Data Protection Office (Legal Services)					
Research Integrity					
Research Contracts					
The Joint Research Office (JRO)					
Research Impact Curation & Support					

(RICS - OVPR)					
ISD Research IT Services (RITS- ISD)					
ISD Digitskills					
ITforSLMS					
ISD Architecture					
Various Departmental IT teams which provide support for research					
Records Office (Library)					
Research Data Management (Library)					
UCL Research Ethics Committee (REC)					
IOE Ethics Committee					
UCL Business (UCLB)					
Information Security					
eResearch Domain					
Centre for Applied Statistics Courses, Institute of Child Health					
Institute of Health Informatics				s	