Report: UKRI JASMINx expansion: User need analysis

Prepared for UKRI

March 2022

Contributors: Victoria Moody, Tim Chown, Matthew Dovey, James Earl-Fraser, Andy Powell, Jeremy Sharp

With thanks to Robert Allen, David Hartland (Hapsis)

Contents

Executive Summary

Digital technologies play a vital role in supporting an inclusive and collaborative research and innovation ecosystem across the UK and internationally. Digital research infrastructure underpins workflow efficiency and helps researchers to maximise the public and economic benefit and impactful outputs of their research, supporting the UK research sector to remain resilient, sustainable, inclusive and collaborative.

The UK research policy landscape continues to change at pace. A culture supportive of team-based research and a need for greater multi-disciplinarity and impact in response to urgent societal challenges, grows in focus. Building equality, diversity and belonging into research and innovation is critical. A focus on reducing bureaucracy, international pressures on research funding and growing net-zero imperatives are increasingly important.

Acceleration of the need for digital infrastructure capacity in response to a wider range of, and more interconnected research domains, as well as increasing the scale, volume and complexity of data, available to and created as part of research, remains a key strategic driver.

JASMIN is a large data-intensive computing facility built to support the NERC environmental science community designed for potential expansion to service other communities within UKRI. UKRI contracted Jisc in February 2022 to support the evaluation of this potential expansion.

The project team developed a structured interview approach in order to enable comparability across interviews. The aim was to achieve a detailed understanding of research project workflows and derive an overview of the critical considerations, activities and facilities and the level of their coordination.

Focus was placed on the research course from hypothesis definition and team configuration and on to project completion and output management. Questions were posed which enable interviewees to detail their research, research leadership, funding and commissioning of infrastructure and the ways in which they coordinated and accessed data and digital infrastructure resources for research.

The project team was gratified by the response to their call for participation. Researchers, research infrastructure and facility leads, and research funders gave their time and engaged with the project.

The project team wish to acknowledge a very pleasing element of many of the interviews conducted, which was the openness and enthusiasm of the interviewees. They were generally passionate and highly motivated about research data, its management and analysis, and so were very keen to be involved in this piece of work. Where they had knowledge of JASMIN they felt it was a good model with strong cross-discipline possibilities. Many were also eager to see the results of this work and willing to be involved in future activity to assess the potential for expanded provision of research infrastructure.

Summary of findings

Research community representatives interviewed as part of this evaluation activity strongly supported the need to engage more with potential new audiences for existing infrastructure, for example JASMIN in order to define the technical governance, support and resource pathways required. Outreach and communication are necessary to expand the discussion across research domains.

There are significant opportunities to explore the expansion of existing digital research infrastructure to other research domains. Extending or expanding one investment, in this case JASMIN to cover the full range of possible research use cases would not be feasible. There is, however, the potential to focus on the pathways to evolution to big data which JASMIN area data took comparatively early and seek opportunities to mirror them across other research domains.

New approaches to data intensive research are being developed across a range of domains. It may be necessary to consider potential for extending an existing infrastructure such as JASMIN to mitigate against the potential for new technical infrastructure duplication or proliferation.

Decisions by researchers about which facility to use or infrastructure to implement were not always based on a full understanding of the best fit. There is a lack of comprehensive information available to researchers about the options that are available and most appropriate for them.

Multi-disciplinary research increasingly requires the creation of and access to linked data with a concomitant and significant risk of disclosure of personally, commercially or otherwise sensitive data, which requires a coordinated and efficient support layer across a more interconnected ecosystem.

Use of more powerful compute facilities for increasingly complex problems may need new skills in data organisation and the creation and governance of the use of algorithms.

There is an opportunity to test and iterate the scope for deploying elements of the JASMIN technical infrastructure across specific interdisciplinary research domain use cases to optimise understanding of the potential.

The opportunity lies in obtaining a fuller and more comprehensive understanding of cases where other research domains using or oriented towards environmental data could potentially utilise elements of the JASMIN facility to develop roadmaps and pathways to ascertain the financial, technical, governance, security, management, ethical and reproducible benefit.

Recommendations

1. Seek the input of a wider representation of research domains for interview. Commission a larger capacity online survey to understand at scale the needs of research domains from a wider set of respondents.
2. Test and by testing, iterate the scope for deploying elements of the JASMIN technical infrastructure across specific interdisciplinary research domain use cases to optimise understanding of the potential.
3. Define the technical characteristics of JASMIN as discipline agnostic and run scenarios for others to map the components to serve other research domains. Identify real use cases for testing threefold approaches:
    - 3.1. Spinning up discrete machine environments for test instances of data ingress, management and analysis of workflows and results
    - 3.2. Understanding how feasible it would be for high demand for GPU cases to deal with demand for enhanced data and process security

3.3. Understanding the capacity for utilising the comprehensive archival functionality of JASMIN data infrastructure, including the CEDA archive and cataloguing resources for exploring the expansion of the CEDA approach to other research domains

3.4. Understanding where other facilities may offer their facilities more widely (see 4.1-6 below).

4. Engage further with:

   4.1. Other strategic programmes focused on enhancing the efficiency, capacity and impact of digital research infrastructure such as UKRI and UKRI digital research investments across disciplines, and with Jisc

   4.2. Funding councils who would like to explore using JASMIN for their communities

   4.3. Groups looking across the landscape at federated access and data linkage such as PSDI and HDRUK to explore potential areas appropriate for including JASMIN

   4.4. The NERC digital solutions project at Manchester to understand how JASMIN can serve other disciplines, interdisciplinary research, and external users

   4.5. Data infrastructure such as ONS and the UK Data Service to learn from their experience, particularly with sensitive data, including the Five Safes and output management, data cataloguing, metadata and persistent identifiers, identifying collaboration opportunities

   4.6. Proactively explore with the above stakeholders the potential for data linkage and the potential for federated approaches.

5. As part of the above engagement, consider the skills and support requirement to effectively engage researchers from other disciplines including

   5.1. The technical capability to access and use the facilities in support of their research

   5.2. The technical capability of JASMIN support to address the use cases defined.

6. Develop an understanding of the potential, level of need for, and appetite to extend the JAMSIN functionality to other domains or simply to offer experimental machine instances which can be managed within the boundaries of the research collaboration alone, in what circumstances. Understand how and for whom, defining potential routes to utility in real research settings, addressing barriers and end-to–end lifecycle requirements.

7. Develop an understanding of the potential for specific use-cases where the JASMIN infrastructure model would suit interdisciplinary research of a type, at a scale of data volume and complexity appropriate to JASMIN capacity specifically, rather than for an innumerate and complex set of scenarios.

8. Understand the potential for the JASMIN compute, analysis, storage and archival models to be extended and offered to wider research domains with the support and service layer and governance frameworks provided by the existing expert investments.

9. Undertake awareness-raising on JASMIN combined with further consultation with other disciplines. It may be appropriate to choose a narrow set of disciplines that are more clearly suited to JASMIN's service, including disciplines that collaborate, or have data linkage, with NERC.

10. Data storage and compute facilities such as JASMIN that aim to expand, need to be able to demonstrate not only the utility of their service but also the cost implications and possible savings. Commission a review of costs incurred relative to potential savings.

Introduction

JASMIN is a large data-intensive computing facility designed and built to support the NERC environmental science community. It forms part of a larger UK investment in digital data infrastructure. UKRI sought to acquire information about how the next phase of JASMIN can be expanded and/or extended to meet the requirements of additional UKRI communities.

A new programme of work *JASMINx*, has been proposed to take the existing system through the next decade. This programme was costed to deliver against the known requirements of the environmental science community but also designed for potential expansion to service other communities within UKRI. The next step for a wider JASMIN remit, under the auspices of JASMINx is to examine the appetite in the community and amongst funders. UKRI contracted Jisc in February 2022 to support this evaluation project.

Requirement

UKRI wished to understand in more detail how this vision could be realised to better serve the UKRI community, which communities can and wish to exploit it, and how their research needs can be met by expanding the existing provision.

In this context:

- **Expanding** means adding more to the existing types of computing resource to meet greater demand
- **Extending** means adding new resource types (types of computing or storage) to meet different types of demand.
- **Provision** means addressing all the hardware, software, and human requirements.

The activity needed to address:

- **Research Needs:** What requirements flow from research needs to infrastructure? Who is involved?
- **Infrastructure Requirements:** How should those needs be delivered?
- **Support Requirements:** How are users supported, by whom?
- **Finance and Governance:** What models for ownership and governance might best suit a diverse constituency?

The activity involved interviewing 24 representatives from among the following five user groups:

- Key individuals involved in **scoping and defining a joined up UKRI digital research infrastructure**
- UKRI staff **responsible for data-intensive programmes** within their research council
- **Users** of existing data intensive facilities.
- **Providers** of existing data intensive facilities
- **Research staff in universities** and UKRI research centres with data-intensive computational requirements

- Comparator activities in the UK and Europe.

User stories were requested by JASMIN colleagues to be shaped around the following considerations:

"As a researcher I want to be able to store and manipulate my data, share it will colleagues, while exploiting data shared with me. I want to have my own programming environment and a reasonable level of service. I am unlikely to be able to accurately predict my analysis requirements in terms of amount of computing, but I might be able to estimate the volume of data that I will produce and need during manipulation."

"As a funder of research, recognising the growing role of data handling in our field, I want to expedite our work, facilitate inter-disciplinarity, reduce duplication of effort and costs, and ensure maximal value is got from data produced with our funding, now and in the future."

Objective

This report presents analysis of the objective to:

- Explore the potential for offering an expansion of the JASMIN platform to additional individuals, groups and communities who wish to share and analyse data from a wider range of disciplines.
- Enable JASMIN to understand how the infrastructure could be adapted to meet the requirements of researchers from a wider range of research domains, including supporting environments that support trusted research.

In scope

- The JASMIN platform technology, data, user base, outputs and processes and exploration of the potential for adaptation for support of a range of additional disciplinary use cases.
- Other infrastructures' platform technology, data, user base, outputs and processes and exploration of the potential for JASMIN adapting to meet these use cases.

Out of scope:

- Exploration of other infrastructure investments' potential for adaptation for support of a range of additional or multi-disciplinary use cases.

Project management

- The Jisc team worked with the management team for JASMIN infrastructure:

  JASMIN Director

  JASMIN Infrastructure Manager

  JASMIN Operations Manager

- Regular update meetings were held throughout the project.
- A **draft report,** allowing for feedback was submitted on 23/03/2022 for review by UKRI before the submission of the final report.
- The **final report** was submitted on 31/03/2022, which received sign-off from UKRI on xx/xx/21.

Acknowledgements

We would like to thank Professor Bryan Lawrence, Director JASMIN, Philip Kershaw, Technical Manager, CEDA and Matthew Pritchard, Operations Manager, JASMIN, and representatives of the contributor organisations for supporting our exploration and analysis with their time and expertise. Contributors are listed in the interviewee table in section 4.

Disclaimer

We understand that the consultation timescales for delivery of the project placed constraints on the engagement phase and the Jisc team has planned for a representative inclusion of interviewees across a range of research domains and from a range of sector roles as far as possible within the timescales. A number of research domains representative were not able to be interviewed within the timescale. Further mitigation includes a summary of potential next steps for the JASMIN leadership team to consider in respect of expansion of the engagement process.

Section 1: About JASMIN

The Centre for Environmental Data Analysis (CEDA) was established in 2005 to incorporate the British Atmospheric Data Centre and the NERC Earth Observation Data Centre. Since April 2018, CEDA has been a component part of the NERC Environmental Data Service, which brings together the five NERC data centres into a single service commissioned by NERC as a National Capability.

JASMIN is the data intensive supercomputer which provides the infrastructure upon which the CEDA data, archives and services are delivered. Designed, built, and managed by STFC's Scientific Computing Department (SCD) for NERC, JASMIN is part supercomputer and part datacentre and provides a globally unique computational environment.

Increasingly, JASMIN provides flexible data analysis capabilities to a growing community, which benefits from high performance compute and a private cloud, co-located with peta-scale data storage. Environmental data are unique because the exact environment from which the data were derived cannot be reproduced. Environmental data will continue to grow, and projected demand for storage continue to increase.

The CEDA Archive provides access to thousands of atmospheric, climate change, and earth observation data, accessible via a file system from the shared science machines on JASMIN. CEDA is a trusted repository under the Core Trust Seal. Priority is given to data generated as a result of funding by NERC however, data generated through other funding sources will also be considered for deposit at the discretion of the CEDA Archive, the remit of which covers the following areas (see linked examples to some popular datasets):

- Climate - e.g. HadUK Grid, CMIP, CRU
- Composition - e.g. CCI
- Observations - e.g. MIDAS Open
- Numerical weather prediction - e.g. Met Office NWP
- Airborne - e.g. FAAM
- Satellite data and imagery - e.g. Sentinel

The aim of the JASMIN environment is to provide a platform for individuals, groups and communities to access, share and exploit data. JASMIN offers a design philosophy of "bring the compute to the data", collocating a range of different computing services with extensive storage for data. Differentiation is made between curated data (in managed archives) and active research data (held and shared within groups).

JASMIN has evolved over a series of discrete development phases: https://jasmin.ac.uk/about/evolution. A project board has had oversight over the technical direction of the system applying the latest innovation and available technologies to respond to user needs.

The main objective of the JASMIN environment is to provide a place where environmental data can be securely and persistently stored, managed, analysed and shared for a wide range of communities and applications while minimising the amount of, and number of copies of, data held on active media. JASMIN supports many modes of analysis co-located with a storage environment and a federation of compute systems and services sharing the same access and authorisation.

Section 2: JASMIN access management and eligibility, project governance and user support

Access management and eligibility

Current JASMIN eligible user groups include UK-based academics carrying out NERC-funded research or related environmental science projects.

They also include government or commercial organisations, which are undertaking research funded by a non-profit research organisation such as NERC, the European Space Agency or European research streams or participating in a non-profit research project with a UK-based academic partner which is eligible for a JASMIN account.

Projects must have UK research interest and the scientific remit must currently fall within the broad remit of NERC. There may be a cost to access JASMIN for some categories of project and or user, depending on the resources required.

Project governance

Resources on JASMIN such as storage and compute are currently allocated to science community “consortia”. Each consortium has a manager: a representative of that science community who is in touch with its major activities and understands the resource requirements for projects in that domain.

Representatives of individual projects discuss requirements with their Consortium Manager about the allocation of JASMIN resources within that consortium. Requirements can be documented using the JASMIN Projects Portal, but need to be approved by a Consortium Manager before being passed to the JASMIN Team for provisioning.

The overall allocation of resources to consortia is managed and reviewed periodically by the CEDA/JASMIN Board. Eligible individuals apply for a JASMIN account via the JASMIN Accounts Portal. This account is the user profile to which further access privileges are granted. In order to maintain a secure and reliable scientific infrastructure for its users, JASMIN restricts login access by maintaining an "allow list" of network domains.

User support

Users apply for the "jasmin-login" service, to enable them to connect to JASMIN machines using a ssh (Safe shell) key. JASMIN has a vast number of additional services, access to these is all managed in the Accounts Portal. Users search and apply for any services they require in the portal. In most cases, users will "belong" to a particular scientific project which may already have a presence on JASMIN, often in the form of a Group Workspace.

Some datasets on the CEDA Archive require specific agreements, to apply for access to these, users will need a CEDA account and to link a user's CEDA account to their JASMIN account enabling filesystem access to data on CEDA Archive.

The following services on JASMIN may then be accessed:

- Getting Started - details all the steps needed to get started on JASMIN. Most documents are linked to from the table above, but there may be some other useful information there too.
- Interactive Computing - introduces the resources on JASMIN available for interactive computing. This type of computing is the most common workflow on JASMIN for new users.
- SLURM Batch Computing on LOTUS - introduces the available resources on JASMIN for batch computing
- Software on JASMIN - information on running software packages within JASMIN
- Data Transfer - this category includes guidance on transferring data to and from JASMIN
- MASS - JASMIN has Read-only access to the Met Office MASS storage archive. This section explains how to get access
- Short-term project storage - this section introduces the concept of shared Group Workspaces and the different storage types on JASMIN. Group Workspaces (GWSs) are portions of disk allocated for particular projects to manage themselves, enabling collaborating scientists to share network accessible storage on JASMIN.
- Long-term archive storage - this section describes the long-term CEDA Archive which consists of thousands of atmospheric, climate change, and earth observation datasets. This is directly accessible as a file system from the shared science machines on JASMIN.
- For Cloud Tenants - JASMIN also provides a cloud computing service, this section describes this
- Workflow management - this category details the various tools available for managing workflows.

JASMIN user support is managed and provided by the CEDA team via the Help Beacon facility.

Section 3: JASMIN User base, outputs and impact

JASMIN supports well over 1500 registered users with well over 200 research projects, with a considerably larger number of users registered to access data from the curated CEDA data archive hosted on the infrastructure.

CEDA delivers Data Archive services for the National Centre for Atmospheric Science (NCAS) and the National Centre for Earth Observation (NCEO). In addition, CEDA delivers the NERC/STFC funded UK Solar System Data Centre (UKSSDC) and the IPCC Data Distribution Centre for the Intergovernmental Panel on Climate Change (IPCC) and recorded over 22,000 users in 2019-2020.

User surveys, user stories and case studies have been developed over the lifetime of the system to better understand patterns of usage and report back to stakeholders on how the system is being utilised. Regular seminars and training courses are held with the user community to engage with them and understand their needs. Examples from recent CEDA Annual Reports include:

Case Study: SPEEDING UP THE TIME TO SCIENCE: ANALYSIS-READY SATELLITE DATA: Ed Williamson, Philip Kershaw, Victoria Bennett: CEDA are supporting Defra and the Joint Nature Conservation Committee (JNCC) by providing access to processing and archival facilities to enable the creation of the new Sentinel Analysis Ready Data (ARD) for the UK. These derived data products, from Sentinel 1 and 2 imagery, have been produced to support land use applications, such as habitat mapping. Sentinel data provided by the Copernicus Programme works well for land use applications by virtue of the frequent revisit time, high spatial resolution and open access arrangements for the data. The concept of ARD has developed around recognition of the need to provide standardised pre-processed data products in a common form ready for analysis thus reducing duplicated effort and the potential for inconsistencies in data preparation. It is estimated that access to ARD products could save up to 70% of project time. The use of JASMIN has provided the computing resources to facilitate large-scale ARD production and lowers the barrier for user access by providing a large, centralised store to host and disseminate the data from the CEDA Archive. The data currently covers England, Scotland and Northern Ireland.

Case study: NEW EARTH OBSERVATION DATASETS ARCHIVED AT CEDA Ed Williamson, Steve Donegan: CEDA continues to support the National Centre for Earth Observation (NCEO) by adding new Earth Observation data to the CEDA Archive: these data are used in a wide range of research projects where they are processed into geophysical products and analysed to better understand the Earth system. The EO data our science community use originates from several international space agencies: CEDA transfers the NCEO's priority datasets into the CEDA archive where they can be accessed via the JASMIN computing infrastructure. This prevents duplication of effort, so that the scientists don't need to individually download and handle what can be up to several petabytes of data. This year, one of the key activities has been expanding the selection of Sentinel satellite data products that we archive - we now hold almost 8 petabytes of Sentinel data. For these large datasets, CEDA makes use of its Near Line Archive (NLA) system. This system moves older data onto tape to ensure there is space on the disk archive for latest data products. It also allows users to temporarily request data back to disk for processing.

Case study: SERVING TERABYTES OF GLOBAL OBSERVATIONS TO COPERNICUS USERS (GLAMOD): Ag Stephens, William Tucker: The Global Land and Marine Observations Database (GLAMOD) is a large climate dataset containing over 26 billion individual observations of temperature, rainfall, pressure, etc. The GLAMOD database, hosted by CEDA, is accessible via the C3S Climate Data Store (CDS) - whose aim is to provide an interface to historical and future climate datasets (Fig. 4). The Copernicus Climate Change Service (C3S) aims to deliver authoritative information about the past, present and future climate, as well as tools to enable climate change mitigation and adaptation strategies by policymakers and businesses. This work is an important part of the CEDA contribution to the C3S. Through multiple contracts, we provide climate data to a broad range of users in both the academic and decision-making sectors. In addition, we are acquiring new knowledge, and developing systems, for handling weather observations at a scale not found in our community.

Section 4: Project interview methods, process, analysis

Methods:

The project team developed a structured interview approach in order to enable comparability across interviews.

The team adapted a question set appropriate for each of the relevant user groups identified. User groups included funders and commissioners, national, regional and local facility managers and researchers from a range of disciplines. In addition, the team engaged with national projects on sharing infrastructure, and data providers from the central government sector.

Interview method: The team aimed to achieve a detailed understanding of research project workflows and the critical considerations, activities and facilities and their coordination.

Focus was placed on the research course from hypothesis definition and team configuration and on to project completion and output management. Questions were posed which enable interviewees to detail their research, research leadership, funding and commissioning of infrastructure and to detail the ways in which they coordinated and accessed data and digital infrastructure resources for research.

Question lines focused on a range of areas including:

- Diversity of researchers using infrastructure in support of a strong research culture and career opportunities
- Diversity of disciplines and interdisciplinary approaches
- The capacity to build and run own infrastructure
- Procurement of research infrastructure components and tools for the technical elements
- Efficiency and economies of scale
- Proliferation and duplication
- Connectivity and compute capacity
- Platform and software components
- Governance requirements, collaboration protocols across institutions or facilities, access and security concerns
- Data management, linkage, storage, analysis, archiving, cataloguing and preservation and who manages these workflows
- Volumes and complexity of data captured or generated, now and in the future
- Transfer of data between point of capture/generation and the compute
- IP and commercialisation needs
- Licensing issues
- Environmental considerations including strategy and targets in carbon footprint and net zero
- Funder requirements in terms of interdisciplinary research and data from other research domains
- What current additional challenges were noted.

Process:

For the interview phase a thematic analysis approach was used based on the interview transcripts. The project team identified stakeholders who were contacted via email, and positive responses were followed up with information on the project and consent and meeting times arranged.

Interviews were conducted online, primarily through Zoom, recorded and transcribed (where possible) for the purposes of analysis. The Dovetail application was used to transcribe and tag data, and identify themes and quotes, from which this report was based.

The thematic analysis approach involved examining extracts of the tagged and highlighted data to identify common topics, patterns and ideas, which were then organised into themes based around a combination of the original statement of requirements plus arising issues.

Organisations Interviewed:

Organisation
Office for National Statistics
Research Data Scotland
HDRUK
ADRUK
ESRC
BBSRC
EPSRC
AHRC
(SAIL) Databank
PSDI
UK Data Archive and **UK Data Service**
School of Advanced Study, University of London
University of Manchester
University of Southampton
Ulster University
Cardiff University
University of Cambridge
University of Oxford
The Royal Central School of Speech and Drama

Analysis:

The following presentation of the results from the interview analysis takes a multi-layered approach. At the top level are the key themes that stand out as the areas of greatest weight. Weighting was determined by a combination of the original requirements, the volume of data on that topic, and the importance assigned to it by interviewees.

Within each top-level theme is a set of findings. These include specific insights drawn from an interpretation of the data, usually from evidence noted in multiple interview transcripts. Attached to each finding is one or more quotes from interviews that provide evidence and illustration to the finding.

As the interviews were broad ranging, across a wide audience base, the focus of the content necessarily is stronger in some areas than others. This focus reflects both the background and knowledge of each interviewee, but also time constraints on the interview. Owing to the large amount of data produced across the 24 interviews, this report is only able to provide key selected highlights. Annexes of the full data will be made available to the commissioning organisation for a duration specified within the collection and data management agreement.

In this analysis we draw on the two user stories identified in the requirements to view the findings in two perspectives. These perspectives could be subdivided further, as described in the definitions below, and where appropriate in the analysis below, we refer more specifically to a single user type (for example funder). In other places, we simply refer to the researcher or community representative to reflect these two main perspectives. This process approach reflects the relatively small sample size and the broad scope of the data.

1. That of the researcher, who wishes to manage, analyse and share data. This group will include those with data and compute needs that are an apparent fit with JASMIN, and also those whose needs are largely met or expected to be met by local infrastructure.
2. The (research) community representative, who is in a position to provide, or influence the provision of, facilities and infrastructure. This group includes funders, facility providers, non-higher education sector providers (for example, government sector organisations including ONS and Research Data Scotland) and UK-wide investments (including ADRUK, HDRUK, UK Data Service, PSDI). Where appropriate, this report will distinguish them as: funder, facility provider, infrastructure investment and government sector organisation.

Theme 1: Data management

Data storage, archiving and preservation

Interviewees noted that there is a wide range of provision available for data management, storage, archiving and preservation at multiple levels, from individual PCs through to national and international infrastructure, with a very mixed level of support to researchers[i].

A significant amount of data produced by experimental science practice are not catalogued, discoverable, accessible and persistent in association with the published output or in an archive, limiting reproducibility. This lack of connected outputs is partly a cultural issue, aligned with the publishing model historically focused on journal article publication.

Even where data are catalogued, accessible and persistent in association with the published output, it may lack associated facilitation mechanisms for reproducibility such as a stable archival model associated with the computational research method or analysis code that would be necessary to fully access the research process, reproduce the data or build on it[ii].

Many widely available resources are being utilised to address this need for the association of an output with data, methods and code, for both experimental and curated outputs. Researchers report the use, for example, Zenodo, GitHub, OSF and also, a number of pre-print services (a number of which emerged in support of faster routes to sharing and reviewing outputs as a result of the pandemic) and the Open Archives Initiative. Octopus will provides a new approach to publishing outputs by type.

The availability of and access to data storage, archiving and preservation at the infrastructure level also presents a varied picture, with some research domains offering long-standing provision, including cataloguing, persistence and user support such as CEDA funded by NERC and part of the JASMIN infrastructure and the UK Data Service funded by the ESRC, which in particular supports persistence using Digital Object Identifiers for data collections.

Funder representatives have some awareness of strengths and weaknesses in equitable access to data management, storage, archiving and preservation provision, although some have prioritised compute over data storage and management. For example, as a funder the EPSRC notes its particular strength in compute[iii].

Volumes and complexity of data

Volumes and complexity of data captured or generated are expected to continue to increase. Some disciplines that historically have not been characterised as typically used high volumes of data, or highly complex multi-dimensional data, or have low to moderate storage requirements, are now beginning to create and access data on a larger and more complex scale, for example 'born digital' archives in the digital humanities[iv].

Volumes and complexity of data may arguably vary considerably across disciplines. Perspectives about what is large also vary, in some cases 10's of GBs is considered large while in other disciplines data is measured in terabytes or petabytes. These differences influence choice of or (actual or perceived) access to digital research infrastructure. Increasingly, multidisciplinary research is drawing on data from different domains, with varying sizes and complexity of characteristics. The wide variation of data volumes and complexity has implications across storage, archiving and preservation compute and connectivity, as well as methods, reproducibility and research ethics (as discussed later in this report)[v] [vi] [vii].

Data access and licensing

Facility providers and researchers report that they perceive issues relating to clarity over data "ownership" which can be complex when collaborating, including using external facilities where multiple parties may generate and process data resulting in a 'new' dataset convened from a range of experimental data supplied from other sources[viii].

Researchers noted that some data are restricted under licence so the ability to access and share the full dataset is defined by specific rights.[ix]

In a number of cases the licensing of the results of research and analysis was discussed, and, in particular, the types of licence that could be applied. Licensing was widely acknowledged by researchers and facility managers as an established, clear and effective capability in the social and economic sciences but considered to be available to a much lesser degree in other disciplines. One funder commented on the issue of multiple licence types adding to complexity in some research domains[x].

Researchers and facility providers reported that the level of data access may be determined by the resource intensity of recreating the analysed data and research results. If the research is experimental, the full, raw data may not be available, discoverable or accessible, by contrast costly longitudinal or discretely funded resource intensive and sustainably funded data are often made fully available in a well-governed archive[xi].

It is worth noting an increasing need to ensure that both data as well as infrastructure is managed and accessed securely and appropriately, accounting for regulatory, legal and governance aspects, along with the potential for a trust framework that provides a set of behavioural standards and a set of technical standards, an approach signalled in the UKRI infrastructure roadmap work.

Data integration and linkage

Multi-disciplinary research projects, or those expanding the data types available to the relevant research domain are increasingly focused on linking data, an activity which will expand significantly resource requirements for storage and compute as well as governance and management, as outlined in the work programme set out by DAREUK[xii].

Research community representative stakeholders and groups are focusing on new programmes for linking or integrating data across domains and sectors, particularly with respect to sensitive data. Some propose concepts of federated access to the new data products to support management and reduce proliferation and duplication of data, including:

- HDRUK - connecting health data with data from aligned or related domains including environmental data[xiii] [xiv] [xv] [xvi].
- NERC Digital Solutions Programme - linking health and social data to environmental for a range of public and voluntary sector organisations, connecting with JASMIN[xvii] [xviii] [xix]
- PSDI - connecting physical sciences data from various sources[xx] [xxi]

Managed data linkage is one of the main functions of ARDUK (Administrative Research Data), the ONS (Office for National Statistics) Secure Research Service and the Welsh service SAIL (Secure Anonymised Information Linkage databank). The investments note considerable progress in managed linking of some of the enormous range of data collected by government departments and, through the provision of analytical tools, the creation of new data resources[xxii] [xxiii].

For more information on data linkage: Developing standard tools for data linkage, ONS.

Data standards, curation and cataloguing needs

Curation of data (and expertise in such) is considered crucial by researchers from a range of disciplines, and facility providers in those disciplines, to ensure the stored data adheres to standards and is meaningful for subsequent users[xxiv] [xxv] [xxvi].

Standards are seen to be increasingly important across research domains, whether involving curation or as part of the research process or data deposit, curation, cataloguing and access management. Motivations vary, including researchers and facility providers wanting to follow good practice in data management, maximise re-use, provide sustainable longitudinal data infrastructure to meet Findable, Accessible, Interoperable, Reusable (FAIR) standards, ensuring good research conduct and culture, and meeting requirements for research integrity[xxvii] [xxviii] [xxix].

Researchers in some disciplines, for example social and economic sciences, and arts and humanities reported favouring a bespoke front end that catalogues data to a high level of quality and provides standardised and interoperable metadata.[xxx] [xxxi]

Theme 2: Analysis

Connectivity and compute; needs and capacity

Decisions by researchers about which facility to use (for example, where there may be a choice of national compute options) are not always based on the best fit to the capabilities of that facility. Such decisions can also be determined by a lack of comprehensive information available to researchers about the options that are available and most appropriate for them. Some funders require the use of a domain relevant and specified compute facility. Support, however, is needed to help research teams make informed decisions about the type of compute they need and how to access it, and the management of data throughout (and beyond) the research project [xxxii].

Compute needs vary considerably across different research areas, and even within disciplines. Many researchers with modest data needs rely on institutional file storage[xxxiii] [xxxiv]. Connectivity was not considered a major issue by respondents, however there was an awareness from researchers that transfer time was an influence on choice of storage or compute, with some research groups actively choosing to implement local storage and compute[xxxv].

Interviewees confirmed usage of High-Performance Compute (HPC) facilities, both national and local, to meet compute-intensive research, but we note that this usage currently tends to be focused on disciplines such as astronomy, geophysics, crystallography, and genomics for example,

It is noteworthy, however that an increasing number of research teams are commissioning HPC facilities across and within more research domains as compute capability becomes more accessible and research data volume and complexity grows and expands in utility for high-performance compute workloads. [xxxvi] [xxxvii] [xxxviii] [xxxix] [xl].

Analysis tools, platforms and software

A wide range of proprietary and open-source analysis tools and environments were utilised by research teams for analysis, dependent upon the specific research needs, the skills, custom and practice and the commonality of usage of available tools. Research teams also were in some cases routed to using bespoke code or open-source software, or proprietary software connected with specific equipment or experimental approaches.

Whilst machine learning functionality was not noted as a specific service need, its application may be inferred and is indicated here as a potential area for consideration in the expansion of infrastructure to multi-disciplinary research domains, both in terms of governance and standards but also in terms of meeting the requirements of different data types and associated responsible algorithm development. [xli] [xlii]

[xliii] [xliv] Local and national facilities (especially for secure data) can constrain what tools can be used, often in order to maintain research security and in some case open-source tools were not appropriate for secure research workflows.[xlv] [xlvi] [xlvii].

The following range of software and licences was mentioned during interviews including tools, formats, languages, packages, and repositories:

- Python
- C++
- GNU
- Apollo
- Google Sheets
- R
- QGIS
- SPSS
- GPL
- GITHUB
- Excel
- PDF files
- SAS
- SQL
- MATLAB
- MIT
- Pure
- CSV
- Stata
- MS Office

Software licensing issues

Licensing of software tools was perceived by researchers as an obstacle to using shared facilities, where prohibitive costs in terms of licences could restrict access to desired analysis tools, or represent an unsustainable or disproportionate resource requirement for the level of investment they may have access to and for what duration[xlviii] [xlix] [l].

Theme 3: Infrastructure

Infrastructure needs and provision

We noted some usage by researchers of partner facilities owing to lack of local infrastructure or access issues to it. Researchers from smaller institutions in particular confirmed that they would often seek research partners with access to appropriate infrastructure to support their data needs.[li] [lii] [liii].

Researchers and local infrastructure providers confirmed implementation of own provision, or access to partner-provided public cloud services such as those provided by Amazon Web Services (AWS) or Microsoft Azure. It may be notable that full cloud implementation is likely to increase, especially for new infrastructure, although on-prem data centres continue to be provisioned especially for research using sensitive data.

Interviewees reported a range of installation levels and types, for example compute, analytics tooling, including technology assisted computational techniques, or data storage. Respondents reported a high-level of support requirements to deploy their research on public cloud infrastructure.

Respondents' awareness of the limitations and benefits of public cloud provision in terms of the different types of workflows described varied and some asserted uncertainty about the cost-benefit ratio depending on workflow capacity needs[liv] [lv] [lvi] [lvii].

Virtual machines were used in a range of contexts, including by researchers, facility providers and research community stakeholders, in most cases of which, the aim of their usage was to provision secure data access and analysis environments with associated governance and security management wrappers. [lviii] [lix] [lx] [lxi].

Researcher respondents with modest data needs reported a reliance on local equipment for processing but recognised that workflow ‘time to process’ could be dramatically reduced with enhanced compute and analysis capability[lxii] [lxiii] [lxiv].

Barriers to using facilities

Researcher respondents reported a number of reasons why researchers are not using national facilities including:

1. Their needs are met by local infrastructure, they perceive that this is sufficient (for now)[lxv]
2. There is no national facility in their domain, or alternatively none is perceived to be available that would serve their needs for example relating to sensitive data[lxvi]
3. The facility doesn't hold data in their domain, and this is something they think is important, to co-locate with similar data[lxvii]
4. Lack of signposting to or lack of aggregated data catalogues covering a wide range of research domains and supportive of interdisciplinary research data[lxviii]
5. Outside of the assumed target community for each facility, there is little awareness of what is available, what they are entitled to use and how to access them[lxix] [lxx] [lxxi]

Sharing facilities

Sharing facilities and reducing duplication is generally seen as a good thing in principle by respondents across the use cases considered as part of this evaluation[lxxii]

National infrastructure initiatives report the need for enhancing skills and capabilities in data management, and computational data science techniques to enable the wider sharing of or integration with and by existing facilities and to support uptake, as well as the need to consider the skills and capabilities of the facility in supporting data from different research domains. Some research community representatives strongly supported the need to engage more with potential new audiences for existing infrastructure, for example JASMIN in order to define the technical governance, support and resource pathways required[lxxiii].

Environmental considerations and reducing the carbon footprint of research infrastructure provision

Environmental concerns were widespread with researchers, and to some extent smaller facility providers, but these were usually broad concerns about environmental issues and typically only acted on at the highest level such as the funding council or a national facility. Researchers commonly thought it was out of their control and focused on other environmental mitigations such as reducing travel[lxxiv] [lxxv].

Concerns were also raised by researchers and smaller infrastructure providers about the risks of being influenced by green marketing, but with minimal benefits. Delivering a good service to researchers was seen to be a priority. Overall, there was a perceived lack of good information on environmentally beneficial options and decision making, particularly by researchers but also from respondents across the range of interviewees[lxxvi].

Where more detailed environmental information was discussed, primarily with research community representatives, it included: sustainable software, preventing data duplication, data lifecycle management, energy costs (particularly cooling), costing and measuring the carbon footprint, recycling and printing[lxxvii].

A number of interviewees reported that although the environmental management of research was gaining in momentum, it was one that had not yet been addressed systemically. Typically, researchers and facility providers were able to point to a range of environmental activities that were not systematically recorded. The general feeling was that environmental management will become increasingly important across research management.

Theme 4: Governance, security and data management

Security and confidentiality

The collection, linkage, usage and management of personally and commercially and potentially nationally security area sensitive data is a matter of some concern, in many cases one which overrides compute or data storage needs. Without the means to securely manage sensitive data with the full confidence of infrastructure leads, funders, researchers and significantly, data providers, any infrastructure facility would be unusable by research domains which focus in any part on research using sensitive data. An activity only set to expand with the expansion of multi-disciplinary research.

With an increasing focus on multi-disciplinary research, it is inevitable that securely managing sensitive data in a range of research workflows will become an issue which requires expanded focus at a wider system level rather than supporting discrete use cases as presently. Security expertise (and not solely in relation to personal data disclosure) across the system will become an increasingly significant requirement as data linkage expands[lxxviii].

ADRUK (Administrative Data Research UK), the Office for National Statistics, the UK Data Service, SAIL and Research Data Scotland mange the usage by researchers of a significant, diverse and growing range of sensitive data. The data are derived from a range of government, non-governmental organisations and longitudinal investments, examples include health, census, education, judicial and survey data. They also provide secure compute and analysis environments and closely coupled support services.

Use of sensitive data in research requires the provision of a complex infrastructure of secure governance. A number of examples exiting or in the development phase might offer potential for extended review:

- implementation of the 'Five Safes' model;
- the initiative focused on the development of frameworks for authentication, authorisation and accounting (AAAI) which could ensure a high-level of information security between research collaborations;
- the use of managed and protected virtual machine environments that prohibit data being downloaded; data encryption; and disclosure control of analysis pre-publication.[lxxix] [lxxx].

Interdisciplinary research requires the provisioning of end-to-end secure management along with the support structures and sufficient resource investment[lxxxi].

Governance: IP and commercialisation

A range of IP and commercialisation management and optimisation needs were noted across a broad range of disciplines, sometimes in relation to the source data. In some cases, there was an awareness amongst researchers of commercial potential but no current process or support available for realising it[lxxxii]. End user licences can be developed for the management of access to data created by or available for research use to extend FAIR and in some cases to support the corporate social responsibility initiatives of commercial providers. Interviewees conversely, reported a perception that their research and the resulting data has no commercial value.[lxxxiii] [lxxxiv]

Data management

The costs involved in both the management of data and by extension the management of the results of its analysis were noted as being difficult to define at grant development stage. The full economic costing of projects with infrastructure costs, either at the institutional, collaborative or national infrastructure level was perceived as problematic to define across the research project lifecycle.

Perceptions of limitations between capital and operational expenditure in terms of licensing from third parties and payment for access to infrastructure was noted as potentially prohibitive, with research teams often developing local infrastructure to support the management of the research project.

Access to national facilities in the STEM area, where research domain aligned projects are approved for usage, is able to be costed more precisely at grant development stage, although the management of results may be less straightforward to cost unless deposit in the national facility is anticipated and costs are borne by that facility in such cases.[lxxxv] [lxxxvi].

An additional cost that was identified related to data collected and generated that was supplementary to the research results; whilst the researcher may only be interested using and storing the data that directly relates to their research findings and making the results reproducible, considerable volumes of data may require to be discarded, that could be of use to other researchers down the line, because of the costs involved [lxxxvii].

Funders in particular noted a concern in terms of data storage that commercial cloud-based services storage may be low cost, retrieval of data could be expensive. It is worth noting that use of express route and direct connect services can bypass egress limits but incur other costs. Tape is noted by JASMIN as a more environmentally sustainable approach to storing data.

There may be a concomitant need to focus on data curation and preservation in terms of what to keep and the need to explore innovative models for reproducibility to mitigate against perceptions of the need to keep everything[lxxxviii][lxxxix].

Funder requirements

Funders interviewed confirmed that they make some requirements of grant holders and researchers in terms of data management of research projects they fund. A number of funders expect data management plans to be created and for data to be made open or accessible within a repository after the publication process is complete. The ESRC, for example, requires deposit in the UK Data Service of data created through ESRC funded research.

Funders noted that beyond principles and expectations, mandating very specific data management requirements would be unsustainable and represent an inefficient use of resource given the multiplicity of requirements that would need to be served. They note that data management is best supported at the project, consortia or infrastructure level where the expertise lies and developing a culture of good data curation and analysis represented a pragmatic and efficient approach, along with the implementation of FAIR practices.

It is also important to note that some funders do not provide national facilities and services with data storage and analysis expertise, making it unsustainable to require that data is stored and analysed though an approved facility. However, one funder did mention future plans for mandating data curation and archiving.[lxxxviii].

Theme 5: Disciplines and Communities

Disciplinary differences

There are a number of differences across disciplines noted in many of the themes discussed above. The social and economic sciences, as well and the arts and humanities are typically characterised as creating and using lower volumes of data in comparison to, for example meteorology or climate science. However, the dimensionality, complexity and types of data produced in and used by the non-STEM research domains can make them complex and therefore often demanding of corresponding analytical and compute complexity.

Data volumes are also growing in a number of research domains, including digital humanities, using born digital archives. Additionally in the social and economic sciences massive aggregations of administrative data, for instance, or data overlaid with GIS data longitudinally, to provide time series or real-time responses to for example disaster situations or public health emergencies are common approaches in an increasing number of research projects.

Researchers in predominantly STEM subject area domains and traditionally working with large data at high volume are more likely to have access to good provision for storage and increasingly compute, with HPC provision becoming more common at regional consortia level. They may also have more sustained access to technical and user support expertise, although funding phases can equal career precarity for technical experts. Researchers may also be funded to access national facilities for running their research projects.

Research domains typically characterised as creating and using lower volumes of data are also typically characterised as implementing or utilising local infrastructure. Researchers report that they may in some cases lack technical support. They also report low awareness of the potential of alternative options, particularly in small institutions. Centrally provisioned compute and analysis and in some cases curation, archival facilities and access management for reuse are lacking in particular for the arts and humanities. A need is noted for accessing hidden demand for users who don't know they have a need.

The expansion of data volume and complexity in research domains typically characterised as creating and using lower volumes of data indicates that corresponding need for advanced infrastructure is in some cases and will continue to exceed capacity and the opportunity to optimise for innovation.

In addition, researchers in the arts and humanities in particular note a need to work across many data typologies. Access is a necessary (albeit not sufficient) requirement to do this, and storage in local facilities can be a barrier to access.

Some research-intensive universities lead or host national infrastructure and have significant local capabilities, although they may be funded for specific collaborative research projects – it is unlikely that a single institution could provide the necessary skills and expertise across the broad range of disciplines where the data needs curation or is in specialist formats. There is also the question of co-location, with many researchers across disciplines preferring a "home" for their data that allows context-relevant front ends and searchability with other similar data.

Researchers and facility providers report that longitudinal studies in the social sciences for example have a long history of careful curation, use of standards, metadata and documentation, that are not replicated in other disciplines. Facility providers report that experimental science has large raw datasets that are typically not preserved with limited documentation on the research process to aid reproducibility.

Research access to sensitive data is typically characterised as a function of the social and economic sciences and health research domains. Interdisciplinary research however, is increasingly requiring the creation of and access to linked data with a concomitant and significant risk of disclosure of personally sensitive data, which requires a coordinated and efficient support layer across a more interconnected ecosystem.

Theme 6: Skills and capabilities

Skills and capabilities

A number of the findings above have strong links with skills and capabilities and it is worth bringing these together below, at the risk of some repetition of findings in other areas:

1. Lack of data curation skills and discipline specific expertise in facilities outside their research domain is one reason researchers don't use national facilities.
2. The need to present data in context, to ensure the data is meaningful and for ease of sharing, is often achieved through bespoke front ends, requiring specialist external skills.
3. Researchers are often unaware of what national facilities are available to them (except in their specific discipline) and have limited knowledge on how to choose or access the facilities.
4. Use of more powerful compute facilities for increasingly complex problems may need new skills in data organisation and the creation and use of algorithms, particularly in interactive research methods.
5. Research areas that could be considered to have historically low data volume requirements are in some cases beginning to expand to digital sources and combining datasets, leading to greater infrastructure need that they are not sufficiently informed to choose or access.

In a few cases there are discipline-based initiatives that aim to plug the skills development gap.[lxxxix]

Section 5: Conclusions and recommendations

Conclusions

Meeting complex and continually evolving needs

- There are significant opportunities to explore the expansion of existing digital research infrastructure to other research domains. Extending or expanding one investment, in this case JASMIN to cover the full range of possible research use cases would not be feasible.

- There is, however, the potential to focus on the pathways to evolution to big data which JASMIN area data took comparatively early and seek opportunities to mirror them across other research domains.
- The overriding opportunity lies in obtaining a fuller and more comprehensive understanding of cases where other research domains using or oriented towards environmental data, or which meet the methodological use case to make the JASMIN facility a viable destination
- It would be necessary to assess and identify specific use case which could potentially utilise which JASMIN functionality and how, specifically, to develop roadmaps and pathways to ascertain the economic, technical, governance, security, ethical and reproducible benefit.
- New approaches to data intensive research are being developed across a range of domains. It may be necessary to consider potential for extending an existing infrastructure such as JASMIN to mitigate against the potential for new technical infrastructure duplication.
- The wide variation of data volumes and complexity across research domains has implications for storage, archiving and preservation, compute and connectivity, as well as methods, reproducibility and research ethics. The risk of unnecessary duplication and proliferation of unmanaged data across the research system requires mitigation.
- Decisions by researchers about which facility to use or infrastructure to implement were not always based on a full understanding of the best fit. There is a lack of comprehensive information available to researchers about the options that are available and most appropriate for them.

Implications for governance, management, prioritisation and support

- Standardisation of technical and policy approaches to common research data management and support challenges, for example with respect to data usage from different domains, sensitive or linked data, could be explored and developed where feasible to enable a greater scope for interdisciplinary research and economies of scale in terms of meeting this potential.
- Utilising the significant collective expertise across UKRI investments to define any common approaches could offer faster routes to prioritisation for usage of a range of existing and expanding infrastructure provision, including JASMIN.
- A significant amount of research data remains to be catalogued, discoverable, accessible and persistent limiting reproducibility. The availability of and access to data storage, archiving and preservation at the infrastructure level also varies.
- There is a need to focus on data preservation in terms of what not to keep and to explore innovative models for reproducibility to mitigate against perceptions of the need to keep everything.
- Multi-disciplinary research projects or those expanding the data types available to the relevant research domain are increasingly focused on linking data, an activity which will expand significantly resource requirements for storage and compute as well as governance and management.
- Without the means to securely manage sensitive data with the full confidence of infrastructure leads, funders, researchers and significantly, data providers; any infrastructure facility would be unusable by research domains which focus in any part on research using sensitive data. An activity only set to expand with the expansion of multi-disciplinary research.
- Security expertise (and not solely in relation to personal data disclosure) across the system will become an increasingly significant requirement as data linkage expands.

- It would not be sustainable or efficient for JASMIN or any individual investment alone, to do all the heavy lifting. Synergies with other strategic programmes focused on enhancing the efficiency, capacity and impact of digital research infrastructure such as UKRI and UKRI digital research investments across disciplines, and with Jisc, should be considered.

Some technical management inputs to consider

- Securely managing sensitive data in a range of research workflows and computational approaches will become an issue which requires expanded focus at a wider system level rather than supporting discrete use cases as presently.
- Data volumes are also growing across research domains, including digital humanities, using born digital archives. Additionally in the social and economic sciences, massive aggregations of administrative data, for instance, or data overlaid with GIS data longitudinally, to provide time series or real-time responses to for example disaster situations or public health emergencies are common approaches in an increasing number of research projects.
- The dimensionality, complexity and widely varied types of data produced in and used by the non-STEM research domains can make them complex and therefore often demanding of corresponding analytical and compute complexity.
- While the movement of sensitive data must be kept to a minimum and duplication in local infrastructure avoided, research requirements which align well with the JASMIN model for colocation of compute with data indicate a need for a fast and resilient high bandwidth Network to connect researchers with research data infrastructure and research data infrastructures with each other.
- Fast and robust end-to-end network performance is required, and a key part of this provision would be the setting up of data transfer zones to facilitate the movement of large data volumes at high velocity into/out of systems based at Facilities, Data Hubs, HEIs, Public Sector and Industry. This approach will facilitate the routing of data securely for ingress into a facility for its effective management.
- Centrally provisioned compute and analysis and in some cases curation, archival facilities and access management for reuse are lacking in particular for the arts and humanities. A need is noted for accessing hidden demand for users who don't know they have a need.
- An increasing number of research teams are commissioning HPC facilities compute capability becomes more accessible and research data volume and complexity grows and expands in utility for high-performance compute workloads, adding potential proliferation and complexity to the landscape.
- Whilst machine learning functionality was not noted as a specific service need, it is indicated here as a potential area for consideration in the expansion of infrastructure to multi-disciplinary research domains, both in terms of governance and standards but also in terms of meeting the requirements of different data types.
- Respondents' awareness of the limitations and benefits of public cloud provision in terms of the different types of workflows described varied and some asserted uncertainty about the cost-benefit ratio depending on workflow capacity needs.
- A concern was noted in terms of data storage that commercial cloud-based services storage may be low cost, retrieval of data could be expensive. Additionally, environmental management across the research ecosystem will become increasingly important. Tape is noted by JASMIN as a more environmentally sustainable approach to storing data.

Recommendations

1. Seek the input of a wider representation of research domains for interview. Commission a larger capacity online survey to understand at scale the needs of research domains from a wider set of respondents.
2. Test and by testing, iterate the scope for deploying elements of the JASMIN technical infrastructure across specific interdisciplinary research domain use cases to optimise understanding of the potential.
3. Define the technical characteristics of JASMIN as discipline agnostic and run scenarios for others to map the components to serve other research domains. Identify real use cases for testing threefold approaches:

   3.1. Spinning up discrete machine environments for test instances of data ingress, management and analysis of workflows and results;
   3.2. Understanding how feasible it would be for high demand for GPU cases to deal with demand for enhanced data and process security
   3.3. Understanding the capacity for utilising the comprehensive archival functionality of JASMIN data infrastructure, including the CEDA archive and cataloguing resources for exploring the expansion of the CEDA approach to other research domains
   3.4. Understanding where other facilities may offer their facilities more widely (see 4.4 below).

4. Engage further with

   4.1. Other strategic programmes focused on enhancing the efficiency, capacity and impact of digital research infrastructure such as UKRI and UKRI digital research investments across disciplines, and with Jisc.
   4.2. Funding councils who would like to explore using JASMIN for their communities.
   4.3. Groups looking across the landscape at federated access and data linkage such as PSDI and HDRUK to explore potential areas appropriate for including JASMIN.
   4.4. The NERC digital solutions project at Manchester to understand how JASMIN can serve other disciplines, interdisciplinary research, and external users.
   4.5. Data infrastructure such as ONS and the UK Data Service to learn from their experience, particularly with sensitive data, including the Five Safes and output management, data cataloguing, metadata and persistent identifiers, identifying collaboration opportunities.
   4.6. Proactively explore with the above stakeholders the potential for data linkage and the potential for federated approaches.

5. As part of the above engagement, consider in particular the skills and support requirement to effectively engage researchers from other disciplines including:

   5.1. The technical capability to access and use the facilities in support of their research,
   5.2. The technical capability of JASMIN support to address the use cases defined.

6. Develop an understanding of the potential, level of need for, and appetite to extend the JAMSIN functionality to other domains or simply to offer experimental machine instances which can be managed within the boundaries of the research collaboration alone, in what circumstances. Understand how and for whom, defining potential routes to utility in real research settings, addressing barriers and end-to–end lifecycle requirements.
7. Develop an understanding of the potential for specific use-cases where the JASMIN infrastructure model would suit interdisciplinary research of a type, at a scale of data volume and complexity appropriate to JASMIN capacity specifically, rather than for an innumerate and complex set of scenarios.
8. Understand the potential for the JASMIN compute, analysis, storage and archival models to be extended and offered for wider research domains with the support and service layer and governance frameworks provided by the existing expert investments.
9. Undertake awareness raising on JASMIN combined with further consultation with other disciplines. It may be appropriate to choose a narrow set of disciplines that are more clearly suited to JASMIN's service, including disciplines that collaborate, or have data linkage, with NERC.
10. Data storage and compute facilities such as JASMIN that aim to expand, need to be able to demonstrate not only the utility of their service but also the cost implications and possible savings. Commission a review of costs incurred relative to potential savings.

Appendix A: Quotes from interviews

Theme 1: Data management

Data storage, archiving and preservation

> [i] *"All the big central facilities tend to have data infrastructures that are sustainable and well-established, and organised and understood... then you have these small facilities that are aware of these issues, have some things in place, but actually there are some key tools missing, a lot of that's around recording process and making results available to users or the wider world. And then at the far end of the scale, individuals in their own research groups that could do with any kind of helping hand."*

> [ii] *"a facility might hold the raw data, but you're just presenting the results to a user. And then a user goes off and analyses them in whatever way they do and then comes to some conclusion and publishes it. And actually, the links through to the raw data are lost. So what we're working on in crystallography at the moment is understanding when you need to keep the raw data, when you might need to make the raw data available, there are times when you don't need to make the raw data available. "*

> [iii] *"What we've noticed with the EPSRC and the EPS communities in general is that actually we're a bit behind some of the other communities in terms of our data structure and data storage"*

Volumes and complexity of data

> [iv] *"You need … really large quantities of textual data before you're starting to hit really serious storage and computing problems. But now that we have born digital archives that are multimedia and starting to be at a scale that humanities research has hadn't worked with before then, these kinds of problems are really becoming very apparent."*

> [v] *"it's a sort of ingrained cultural thing, largely driven by the way, we currently publish things. The PDF that conveys a result is a screen or a cover for what can be terabytes worth of data underneath that never makes it to the surface. And this kind of simulation data is particularly bad because much of the experimental data in physical sciences is not huge in terms of volume. And historically, a lot of this gets shoehorned into supplementary information in the publishing process."*

[vi] *"When I say large data, I mean in essence anything up to about 20 megabytes, up until that size, that is possible to use email. But it's quite common with our workflows that we would have something between say 20 and 50 megabytes, sometimes we might have things that would be in the one to five gigabytes sort of range, but probably no bigger than that, it's very variable"*

[vii] *"...mainly lightly hierarchical or lightly relational datasets that come in several files from a data provider, which are moderate in size in their own. But they can get very big when you put them all together. And the geographic data, the background data layers can be quite substantial, for example, looking at the entire road network for all of England, with quite a lot of attributes"*

Data access and licensing issues

[viii] *"There's another subtle issue as well. And that's who owns the data. So, someone in a completely different university has generated the sample and sent it to the facility. The facility does some stuff. It keeps the raw data. It gives back the result to the originator and away you go. And the question then becomes one of who owns the data. Sometimes those services are paid for which complicates the issue of ownership and other times then there's collaboration networks and that kind of thing. So that's a slightly complicated issue."*

[ix] *"It could be sensitive …we don't know what's in there, but the reason for (the restriction) is to protect publisher investments in income streams... legislation under which the web is archived does not allow the re-publication of that data, …the trade-off for being allowed to do it is that you can't just make it open to anyone. (For example, opening) the archive of the whole of an online newspaper would mean people wouldn't subscribe to it."*

[x] *"There are so many licensing options out there. It's not something that we mandate - which type of license - we would normally direct people to, for example, the software sustainability Institute, for some advice on that or from within their local organisations, what their research offices or IP departments might decide is best for the academic's data."*

[xi] *""The founding principle should be that all science should be reproducible. So the experiments that created that raw data should be reproducible. There's a caveat to that, some experiments take two years to perform. If you lose that raw data, it would be a disaster. So, there are some situations where we keep the initial set of raw data. There are others where there are several steps in the journey towards a set of data that you can analyse properly. And so, I'm thinking mostly the stuff that comes from image data, there'll be the initial set of image data that is quite large. Then there's some processed data that comes from that that's even larger. Then I think there's another final set of data that is what's used for analysis, and that middle chunk can be deleted because it's possible to recreate that from the first chunk. So, there are certain circumstances where it's easy to identify chunks of data that can be deleted because they're easily reproduced from other sets of data."*

Data integration and linkage

[xii] *"...our project is connecting roughly 30 petabytes of data, and that's constantly growing with a whole range of other social economic, environmental, and health data across the whole of the UK."*

[xiii] *“"We have a pan UKRI remit. We cover all the UKRI Research Councils and we are looking at how to create the next generation of analytic environment for sensitive data, and how to do a cross domain, cross jurisdictional, linkage of data in a secure and trustworthy manner with the public.”*

[xiv] *“We're also looking at linkage with non-sensitive data... We'll have data, let's say pure environmental data to look at how pollution affects people's health. Well then suddenly the link data becomes sensitive, so you have to be mindful of how you handle that.”*

[xv] *“The new trend that we're all trying to go to … is research environments that sort of turn that on its head. So, we keep the data, the sensitive data in the trusted environments, instead of having people download the data.”*

[xvi] *“We're looking at federated data access and federated analytics. So, in a sense, can you send the workload, the analytics to where the data resides instead of having to ship data back and forth between different trusted research environments? At scale, shipping data around, it's not going to work quite frankly, the bandwidth is not there. And it's too expensive to stand up temporary data environments for petabytes of data? So, you want to somehow do this in a distributed computing environment and while still doing so you need to have trusted research environments, where they have a shared trust between each other so that they can actually operate on data across two different locations. Potentially a lot of them, the crux of this to us, the way we see it, it's actually governance related.”*

[xvii] *“We've got JASMIN sat behind all of this data so that you can run analysis models to support your decision making. So, what is kind of interesting from my perspective is the fact that we're going to be at the end of this, connecting non-academic users to all of this data. And also, all of just potentially being able to use JASMIN compute power.”*

[xviii] *“You potentially could be running your own flood risk model for the whole of Scotland say, and rather than taking days for that to run on your machine in the Scottish government, you will be able to learn that on JASMIN. Now at the moment, JASMIN’s not really set up for that kind of use”*

[xix] *“What we're doing is looking at the relationship between human health and the natural environment. So, if you want to run some kind of epidemiological model on the whole of the UK population to say, let's say we want to analyse the extent to which respiratory illness is linked to traffic pollution, someone could run that model on JASMIN. So were connecting all of those things together.”*

[xx] *“What we're looking at is the concept of doing federated analytics, where you can basically send the workload into two disparate environments and then marry the results afterwards, that would work if they don't have to talk all the time, or at least if you can limit the crosstalk at the amount of packets that you have to send between the two.”*

[xxi] *“One thing is federated data access, where you have basically consented data access across multiple domains, the other one, or joint perhaps, is the concept of federated analytics where the workload can be sent to multiple domains. And I think the data custodians are the people responsible for the data. They need to have full insight and clarity in terms of how their data is used. So, it is challenging if they start shipping it round to a lot of different places.”*

[xxii] *".. if you combine different data sets together, that new linked data can be asked to answer a myriad of research questions. So those linked data assets can serve multiple purposes across government and outside of government. And if you start from the principle of which data would it be useful to link together, which thematic assets can we create that are of value, then we can get far more value because once you've linked the data and all the reuse of those data costs pennies and deliver significant returns."*

[xxiii] *"we're also working with DCMS who hold the national data strategy, which pushes very hard for common data standards, common metadata standards. So, we want to advance that through the integrated data service. The metadata of these different datasets are consistent as well, so that you can see that you're comparing like for like"*

Data standards, curation and cataloguing needs

[xxiv] *"It's very dependent on people. If you like the knowledge base, ...it does rely on the presence of the team, we have a group called geodata, which is like a faculty level specialist, group of researchers and people working the enterprise unit within university. And they are programmers have done a lot of the heavy lifting once we've worked on specific projects"*

[xxv] *"My staff are able, because they've been employed on that basis, to deal with the specific types of data, which social scientists use and longitudinal data is a case in point where you're using time as a third dimension in terms of the data. So, you're looking at change over time and very few of the disciplines create data in this format."*

[xxvi] *"We want it to be specific and considered at the point of collection and we want it to be curated at the point of release. So, it's not about just dumping all of the data out there"*

"Having archivists who are not just the dumping ground, but active champions of making our data more visible is extremely important."

[xxvii] *"We haven't quite got to the stage where we have the common standards, but we would like to move to a model where our repositories are interoperable and that they would share common standards and common frameworks for access, but we haven't actually got down to the specifics of looking at that. We will be in the next few months"*

[xxviii] *"We are moving inexorably to some Exoscale machine. But we are generally pretty well provisioned for compute cycles. That's not really the problem here. Storage is quite a "low hanging fruit" solvable problem. What we really need is to be able to share the results and for that you need standards and probably a little bit of culture change as well."*

[xxix] *"a lot of this has to do with open standards, interoperability that has to be there for people to sort of tap into a remote TRE (Trustworthy Research Environment). So, you need APIs that are well defined or for data access across domains, but also to perhaps send workloads over. A key thing is how data is described. So, the richness of the metadata, and then to make the structure of the data known to someone who wants to use it from outside through an API, if you make the schema known and what data attributes are available, that would be helpful, challenging for unstructured data, less so for structured data. But these are our areas that we're looking at."*

[xxx] *"Through something like CERN and some of the more generic data repositories, every data set is separate. Whereas our data sets exist in relation to each other. So having the capacity for a landing page that will eventually let us, and then each data set has its own code book and more information about it. And there you can download the data."*

[xxxi] *"The facility to get bespoke front ends to datasets written, published, which perhaps may be when you are dealing with something where, which could be purely mechanically set up "*

Theme 2: Analysis

Connectivity and compute; needs and capacity

[xxxii] *"How can you open it up to other disciplines that people can actually tap into this environment? ... it's the computational literacy, which comes into play a little bit here, ... If you then want to submit it to schedule and run it in a cluster environment, you have to learn how to do that... there is this activation energy barrier for people to go from interactive work, which they can fully control working through a scheduler, or you're sharing a large environment with multiple users. But if you do that, if these people who are running through that hub today have access to those environments and learn how to do this job submission, they can potentially run much faster"*

[xxxiii] *"It's extremely patchy in terms of what's out there currently and ranges from big, national level support in a well-structured environment, all the way through to individuals kind of floundering without support at all, and quite a few in-between bits as well. "*

[xxxiv] *"If you do a lot of AI work, which is of course popular today, you tend to rely on co-processors notably GPUs, to accelerate the calculation. Compute clusters that follow GPUs at scale becomes very expensive very quickly. So, you wouldn't set that up, if you had one problem you've got to solve and then never use it again. ... So, either you could tap into a public resource funded by the government, or potentially set up in a cloud environment where you might access those kinds of systems, basically renting them for that time. And then at least you don't own them afterwards and build an expensive infrastructure."*

[xxxv] *"You want it to look at environmental data, we call it meteorological data, which is really big. And you wanted to analyse that in the context of a smaller health data set. You're not going to take that Mr. Mountain environmental data and ship petabytes of that over to the health data that is of limited size... they had a hundred terabytes of health data and 50 petabytes of environmental data. Very likely the health data environment is not going to be able to hold all the data that you would need to ship over. And it will take you a very long time to do so. Like a truck with hard drives would be faster than sending across the wire. So, it isn't really realistic, I think, to do that. "*

HPC needs and usage

[xxxvi] *"...it's definitely the case that for high performance compute the biosciences need GPU enabled high performance computing. And there's very little in the UK for them to be able to access, that they are aware of it. I think that's a definite need. And because most of the high-performance computers in the UK, I think DiRAC at the STFC is one, but they don't access that, there's not great access and it's quite heavily used."*

[xxxvii] *"We typically run some of the stuff on a dedicated high-spec machine with some GPU cards in it, but we will run other things on our supercomputer operation ...which is a distributed virtualised operation with GPU clusters and other stuff in it."*

[xxxviii] *"There are lots of other disciplines around the university that generate tons of data and they put very high demands on HPC services. So, we were finding that the biology departments were sort of regarded to be second class citizens compared to astronomy and geophysics and crystallography and so on. So that was one of the challenges for us was that it took so long. And as I said, we had developed our HPC infrastructure long before UIS [University Information Service] started providing it as a central service. So now that we have it, and people are aware of that, if we suddenly just let that little lapse and I switched to using UIS, we would all of a sudden have this queuing system to deal with. And then that's a big disincentive. And I think for our researchers where every second counts, they'll just count you on paying for maintenance of this local system."*

[xxxix] *"Compute in the UK is fairly well organised, but it is a pretty discrete community, especially if we're talking about HPC. I mean, it's been opened up to a larger group of people more recently, as personal computers get more powerful and you can run some codes now on personal computers that 10 plus years ago you would have needed a supercomputer, but the supercomputing community is discrete and separate. And certainly, at PSDI, we, we spent a lot of time engaging with that community and what they don't want is yet another infrastructure, but what they do need is the ability to connect and bridge to the experimental scientists. So, we don't need masses of infrastructure for doing the computer itself, but what we need is to make those connections between experimentalists and computational folk or bridge the divide such that experimentalists can access supercomputing."*

[xl] *"There would be some domains that are focused on say high-performance computing type workloads. Say if you do a lot of molecular data analysis, like genomics for argument's sake, where the data sets are really big and a lot of data points, and you need to analyse them in a parallel fashion, let's say you want a cluster environment that won't be true for every theory, some theories can probably operate with far less horsepower and do all the research that they, that they need to do. And so, so I wouldn't say that every TRE needs a cluster form backing it up, some will not. And the type of computers that you will need in the HPC context are varied. Do they need a lot of memory? Do they need a lot of GPUs or not? It's going to matter very much from what kind of research you need to do. So, there that's scenario where either running in the cloud natively, or if you can burst into the cloud for specific use cases, that might be an avenue to do that where you can then leverage, you know, Amazon, Google, or Microsoft, let's say for specialised computers."*

Analysis tools, platforms and software

[xli] *"We use a range of open-source packages and then we build with them. Then there is also bespoke code that goes on top. So, we're kind of like, I think there are things, a lot of, for example, Plotly under the hood, but it's all freely available software that we then adapt to our needs. And we're also releasing these, the visualizations we're working on packaging up some of the bespoke software that we've written to release it open source as well."*

[xlii] *" You're looking at windows, you've got office 365 on there. ...there's a power BI. So, if you want to link your data to be presented in a dashboard format, Python RSQL, and there's, there's probably about 20 open-source project products on there as well, which are just a response to specific user requests...we decommissioned SPSS about two years ago, and then we had this little flow of users for whom it's SPSS or nothing."*

[xliii] *"We've got software, which we developed in-house, which we're in, in a second version of which effectively processes that library to produce small area estimates against a particular target type."*

[xliv] *"The data that's initially generated is often in a format that is proprietary. So, it can only be read and viewed with the equipment. … companies often would provide software that would enable you to read the data rather than analyse the data, but it's different depending on the different platforms."*

[xlv] *"Anything that's not deemed by the university to be a security risk, so downloading open-source software and using it within the university environment can be challenging. And also working with external partners with access shared access to data is almost impossible because of security restrictions."*

[xlvi] *"But if they're specialist software, we would have to purchase licenses for them to use on our systems. And there's not a lot of leeway to do that. And also, because it's very difficult to, to make sure with 100% confidence that the, the software is secure because it sits within a secure environment. If the software has been altered in some way to make changes to the way in which the file is created, then that that could be a security risk."*

[xlvii] *"If you're on trusted research environments it's even harder because you have to trust the software that people want to bring in so that it doesn't do something bad."*

Software licensing issues

[xlviii] *"Licensing for instance, that can be a challenge… if you have 'bring your own' software packages to another computing environment where they don't offer that at all, that would be problematic. If it was commercial code, I would imagine they would have to work something out there. But if you're running an HPC environment and you want to open it up to other users, you obviously would have to provide at least the most common tools, but that whatever that might be. And in some cases that would include promotional packages, I would argue like MATLAB … you need to offer those kinds of packages in the HPC environment, and then people can come in."*

[xlix] *"To expect 200 means that we need 200 Microsoft office licenses. It means we need 200 SPSS licenses and so on. And we negotiate some of these, we negotiate directly with the software vendor. And sometimes we negotiate through the university because it's because it's cheaper, but any research data infrastructure that is providing access to tools to users needs to be clear about the licensing issue of software and, and making sure that they are budgeted for"*

[l] *"We take data, code and software into the repository and in terms of licensing, the options are to choose from various creative commons licensing when it comes to data. And then we have software specific licenses for when people are sharing software code"*

Theme 3: Infrastructure

Infrastructure needs and provision

[li] *We don't have access to the kind of infrastructure that you might get in a larger and more diverse university…We don't have a research computing department, but the IT support services available are around general desktop management and so on. And anything that pushes that in a research direction, we really don't get any support for it …The collaborative projects will tend to nearly*

always be a partner who has a better infrastructure and can provide hosted data hosting and RSC [Research Services Computing] input."

[lii] " We *can do it to a limited extent on SharePoint, but anything that really involves real time collaboration, and we just can't manage it all. And it's always a big performance to get other people allowed onto our local network."*

[liii] *" It's a big barrier at the start of most of those collaborative research projects, finding out which of the partners has got the least restrictive environment, so to enable collaboration."*

[liv] *"We were encouraged by industry to put in place some governance as well. AWS, we have regular meetings with them and they are advising on best practice…we are trying to centralise as much as we can, the provision of cloud."*

[lv] *"Generally, that would be harvesting Twitter data locally, and then uploading it to Amazon web services. In that instance, I think that's how it works. But the other cloud-based option is that you never actually have the data anywhere locally. It's entirely on a cloud-based service. That's provided by an archive library or third-party service affiliated with an archive or library. And that means that you are restricted in the kinds of analysis you can do by what's possible in browser."*

[lvi] *"Because of the size of the datasets that we have to move around, the logistics of moving these very large data sets to and fro between cloud services, the network doesn't support it at the moment. It's a simple, practical matter like that. I'm well aware of the value that AWS offers, just the logistics and also the cost of the cost of rapid turnover of data in AWS is it's not completely clear that it's massively more economical than, than having local infrastructure."*

[lvii] *"I think a lot of people think that cloud computing is easy. You just call up Amazon and they'll figure it out. When you run it yourself, you don't have a tech, you don't have to hire system administrators, that's not really true. You need people who are cloud aware and you have to hire staff that know how to do this. It is not necessarily cheaper to run anything in the cloud compared to building in house. But it's a calculus, often you're shifting money from Capex to Opex by, by going to the cloud. And I would say that if you're running computers at very high utilization, yet you're probably better off at home than to do it in the cloud. "*

[lviii] *"We're using VMs for all of that kind of analysis stuff that processing workflows they're all done in VMs, so they can be quickly created spun down again. We are also using VMs in our general IT infrastructure and service as well. But the VMs that are used for data processing are all built within the high-performance compute infrastructure."*

[lix] *"We do buy from our IT services, some virtual machines and run stuff. And we're actually thinking though, we may buy that next time, with their blessing, from Microsoft, because we've got something that's not necessarily very large, but it needs to work across multiple institutions. And there we have a problem because either we have to make them all visitors to Southampton or worry about an open port all the time, or the suggestion was since it wasn't too enormous amounts of data… not too many people and not too much data at this point, that may be just renting appropriate database service from Microsoft and building something on top of that. Well, it depends how much we have to build it on top of it, but then you, the access is controlled by us and is not putting the university at risk."*

[lx] *"Sciverse for instance, is one where you spin up a virtual environment, it's got lots of tools loaded up and you can import your data and, and access things. You can integrate data sets and analyse data. And you know, that software is going to work because it's all ported within a virtual environment. I think that's something that will be very useful."*

[lxi] *"Once the user and the project come together, what the user will get is the virtual machine. When you look at it, it looks like windows, it's just a windows front end, this thin client connection. And then they can go in all their dates will be there for them. They can go in, they can, they've got every tool imaginable STATS, R, Python, SPSS. They go in and interrogate the data, but nothing ever leaves"*

[lxii] *We could often be doing runs where we would be setting up something reasonably complicated and then leaving it be for six days to get the answer. We're not doing something which is time sensitive in the sense that we need our answer immediately. But if you were wanting to take it to an outside world operational environment, then you would be looking at a much more powerful computer environment, because you could weave through that in a very short space of time and do much more near real time. So that's, that's usable and talked about, but not implemented."*

[lxiii] *"The processing takes a very long time when an analysis on the Twitter archive that we've put together... you start it off and set it running, and it, it goes overnight and you come back 24 hours later and it would have completed."*

[lxiv] *"If I'm using my laptop and it takes an hour to run a program, then do I go back to my computer centre, the university and say, this is too slow. Or do I plug myself into somewhere else that can do this taking into consideration the transfer of transfer of data, if that's necessary and taking into consideration that you can't do this with data that is sensitive or data that is restricted in the access in some way."*

Barriers to using facilities

[lxv] *"we're not using, and should be using, a central compute facility. We have done this in a variety of ways by effectively doing very informal parallelization of the task and setting it up so that we're using different. ...faculty level computation tools to be able to split up different age groups in the population, being run different groups in a modelling scenario, might be splitting out different age groups of the population and running them in parallel, which works and then bringing it all back together again."*

[lxvi] *"There's not something that's readily accessible to most humanities researchers, as far as I know."*

[lxvii] *"When we went to look where we would put our data, one of our questions was just simply, where do we put it in? I mean, CERN in Switzerland is one of the biggest data repositories, but then when you search for something like 'dance', you get a handful of biomechanical movement analysis projects that are nowhere near our field. So, in terms of searchable findable data, because I think that's one of our questions is; how can folks reuse the data that we're creating, but it needs to be findable? And I'm just not sure a scientific that type of scientific data repository is the best*

home for our work. Something else that was important to us is also that it's fine linkable that we build a series of data sets."

[lxviii] *"When we went to look where we would put our data, one of our questions was just simply, where do we put it in? I mean, CERN in Switzerland is one of the biggest data repositories, but then when you search for something like 'dance', you get a handful of biomechanical movement analysis projects that are nowhere near our field. So, in terms of searchable findable data, because I think that's one of our questions is; how can folks reuse the data that we're creating, but it needs to be findable? And I'm just not sure a scientific that type of scientific data repository is the best home for our work. Something else that was important to us is also that it's fine linkable that we build a series of data sets."*

[lxix] *"I had never heard of JASMIN before, but I read about it yesterday."*

[lxx] *"..it would take reaching out to people who are, I would argue, somewhat computationally aware, at least, in the other Councils that are not currently using the resources and would take outreach. It will take education outreach, see where you can perhaps again, do cross domain type work."*

[lxxi] *"I look with envy, for example, at Danish colleagues, you have a cultural heritage and computing cluster. That's connected up to their national library where you can take your data and work with library data as well. And that gets around the restrictive access questions as well. When you have the hosting library archive as a partner, and there's not anything equivalent in the UK that I'm aware"*

Sharing facilities

[lxxii] *"I really hate this server sprawl, where people are running computers in an inefficient manner, because they want to have their own server. They will run at 30% capacity instead of sharing servers set there, I run it 80 or 90% capacity. So, I think that shared facilities like JASMIN is it's a really smart play from a national perspective… in the end, I think that we'll come up with, a limited set of TREs across the UK that are well-funded and set up in a sustainable fashion."*

[lxxiii] *"if you want to bring people in the door that typically do not go to JASMIN to do that work, it would take reaching out to people who are, I would argue, somewhat computationally aware, at least in the other Councils that are not currently using the resources. It will take education outreach, see what you can perhaps again, blue cross domain type work. So, you can have people who are used to using JASMIN to research with people who typically would not use a resource like that and perhaps bring them into the fold."*

Environmental considerations in terms of the carbon footprint of the research infrastructure provision

[lxxiv] *"It is a concern. There's not a great deal that one can do about it apart from trying to reduce demand, which will reduce the environmental impact."*

[lxxv] *"Everyone that does biomedical research knows that it uses energy and, and there are, there are issues in terms of other environmental issues like recycling and so on and so forth. But there are*

no particular concerns for ourselves, you know, beyond like the normal, you know, the general concerns that one would have"

[lxxvi] *"A couple of my team are leading on the development of a toolkit around environmentally sustainable digital humanities research as part of that, that coalition. And we are looking to cost everything we do now with a view to environmental sustainability. It's quite hard finding information about that."*

[lxxvii] *"I know the biggest single cheque the university writes every year is this electricity bill."*

Theme 4: Governance

Security and confidentiality

[lxxviii] *"If the algorithm fits the environment, it doesn't matter if it's environmental data or health data that you're looking at. As long as the computational infrastructure works, the challenge comes down into the sensitive space and where you have to treat the data in a special manner because of what kind of data it is. If there were no concerns about sensitivity of the data, you're set up in a compute environment that fits what you're trying to solve, what equations you're trying to solve effectively, right there, data relationships you're trying to solve. So, it might be data-driven, it might be an environment where you do say simulations that you would do in physics, where you need lots of CPU power, but maybe not so much data, really, but they're going to run for long periods of time. And it has conveyed a lot of crosstalk between CPUs. That's a very different type of work than you would see, probably in analysing most healthcare data."*

[lxxix] *It's a very rigorous process that researchers need to go through before they're allowed to have remote access. (Assured organisational connectivity). And it's essentially the head of security of the organisation in which you work and other senior people need to sign off ....you have to be working on a laptop or computer that's owned by your organisation. .....your organisation has to take responsibility for your behaviour and the equipment and the setup of the machine has to be done in a certain way as well. "*

[lxxx] *"...each project is, is vetted for the risks to privacy versus the public benefit or perceived public benefits from that. And that's done independently of the researcher at the, the organisation providing the data."*

Governance: IP and commercialisation

[lxxxi] *"...in terms of kind of sensitive data, we do have quite a distinctive view from some of the other councils and for me sensitive data is not necessarily about data that involves legally protected information. Personal data for us, it's much more about commercially sensitive data for the most part. So it could be relatively innocuous kind of research or analytical data that's generated by instrumentation, but it's in the context of the commercially sensitive Environments."*

[lxxxii] *"I would expect that some of the data would have a commercial use... we haven't typically addressed or how data custodians are handling their data, whether or not they want to make it commercially accessible as well as freely accessible. But I can't see that that would be a hindrance apart from the fact of course, that if you're working with personal data of people, that becomes a*

challenging part. If you open up NHS straight out to big pharma, for instance, to do that under safeguards."

[lxxxiii] *"The kind of research that I do and that a lot of my colleagues do is never going to result in anything that will be commercialised, but there isn't support ... There are never any conversations around IP and commercialisation. And I don't think anyone in the university would know what to do if that was suggested to them. ... We do have a standard contracts with collaborators about shared IP and so on, but it's never, never turned into, certainly in my research career, there's never been anything where that's been necessary."*

[lxxxiv] *"That wouldn't apply to, probably the overwhelming majority of people working in the humanities and particularly the historical humanities, because they would not be dealing with data that would have that kind of financial value."*

Data management and analysis

[lxxxv] *" And currently I don't have to write that cost into a grant because (the University) can provide most of it. If I were buying it from elsewhere, I'm going to have to put it into a grant and I'm going to have to have a good reason for the research council say, okay, yeah, we'll fund that. Then you add some money and not that huge for what we do,"*

[lxxxvi] *"That's another challenge we have internally is it's very, very difficult to get costings for, for those kinds of services, because we don't have a research computing department. So, we then get into kind of endless conversations about whether we're allowed to get quotes from external suppliers"*

[lxxxvii] *"There are potentially cost associated with archiving that raw data in a repository as well. So, there may be a massive pool of data in between that's sitting uncurated and the researcher would need to take time to assess how valuable is this data. So, what to keep, what to disregard."*

Funder requirements

[lxxxviii] *"We're moving towards a model where we mandate our grant holders to deposit their research data in a designated repository. And we're looking to create a federated infrastructure in order to enable that we're not there yet, we hope to be there partially, at least, from next year."*

[lxxxix] *"...we have things called High-end Compute Consortia, these are hosted by an institution, but they're basically community groupings of research. They have access to a certain share of Archer 2 and tier two services. They are like a brokering mechanism run by the community for the community. They can connect people to skills and resources. They have some money for skills development, networking, travel and code development, that kind of thing." ......"That's a good model training. We need to provide infrastructures and we need to support institutions, but actually, just matching people with someone who knows what they're doing. It could be serving a disciplinary community or an institutional community."*