UK Research and Innovation

18 December 2020

Freedom of Information request: FOI2020/00389

Thank you for your Freedom of Information request received on the 14 November in which you requested the following:

Your request:

Hi, i am a science student doing an assignment on the scientific storage of data in a work place and laboratory. I want to know what system is in place you use such as lims. Also i want to know the benefits and issues with storing large quantities of data and how you send and receive this data within the workplace/lab.

On 18 November we sought clarification from you informing you that "UKRI brings together the seven disciplinary research councils, Research England, which is responsible for supporting research and knowledge exchange at higher education institutions in England, and the UK's innovation agency, Innovate UK" We asked "Are you specifically asking about workplace and laboratory data systems generated by UKRI staff and scientists or are you interested in the wider storage and dissemination activities in funded research centres and facilities?"

You clarified on 19 November 2020 as follows:

Yes, please proceed. I am interested in all data storage and handling, workplace and laboratory data systems and funded research.

Our response

I can confirm UK Research and Innovation (UKRI) hold some of the information relevant to your request.

We have broken your request down into three separate parts to enable us to fully answer your request. They have been broken down as follows:

1. What system is in place for scientific storage of data in a workplace and laboratory, such as LIMS?
2. What are the benefits and issues with storing large quantities of data?
3. How do you send and receive large quantities of data within the workplace/lab?

I can confirm that we hold some information for question 1 and 3 of your request, please see the information you have requested in Annex A at the end of this letter. The information in Annex A relates to the following Research Councils: MRC, STFC, NERC, BBSRC, ESRC, AHRC and EPSRC and some of the Centres, Institutes and Facilities that they fund.

The Freedom of Information Act is a mechanism for the public to gain access to the records that a public authority, such as UKRI, hold. This means that any requests for an opinion cannot be answered unless this has been recorded. As such we are unable to answer question 2 of your request under the FOIA.

If you have any queries regarding our response or you are unhappy with the outcome of your request and wish to seek an internal review of the decision, please contact:

Head of Information Governance
Email: foi@ukri.org or infogovernance@ukri.org

Please quote the reference number above in any future communications.

If you are still not content with the outcome of the internal review, you may apply to refer the matter to the Information Commissioner for a decision. Generally, the ICO cannot make a decision unless you have exhausted the review procedure provided by UKRI. The Information Commissioner can be contacted at: http://www.ico.gov.uk/

If you wish to raise a complaint regarding the service you have received or the conduct of any UKRI staff in relation to your request, please see UKRI's complaints policy: https://www.ukri.org/about-us/policies-and-standards/complaints-policy/

Yours sincerely,

Marilyn Kamanyire
Information Governance
Information Rights Team
UK Research and Innovation
foi@ukri.org | dataprotection@ukri.org

Annex A: Data Storage Systems

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
Medical Research Council (MRC)		
MRC Institutes	- MRC Institutes use bespoke LIMS systems to store large quantities of research project data such as phenotyping and mass spectrometry data. - Cross-platform relational database to store sequences, plasmids, oligos and screens. - Animal Facility Management Software – used to manage resources in facilities - Cloud data storage – used to store scientific data.	The data is centrally held within these systems and data is managed via the system (entered, uploaded, edited, analysed, exported etc.) - rather than being sent/received. Users have specific access (password protected) to only their projects, results files and data. These allow the institutes to store all data in one place, which makes sample tracking easy, and lab operation reports straightforward. For the LIMS, these are flexible and allow the ability to build in new workflows and is scalable to accommodate increased sample throughput. The data is secured and accessible by the owners of the data (individual research groups/facilities) via their desktops/laptops. Other data is transferred in the following ways: - Very small quantities, (sub 10MB): email - Medium quantities (up to 10GB): Own cloud - Large quantities (up to 500GB): File Transfer Protocol (FTP) - Massive quantities (Up to 30TB): dedicated large file transfer services - Extremely Massive quantities (>30TB): Air freight of whole server.
Science and Technology Facilities Council (STFC)		
STFC's science activities, including particle physics and astronomy, generate very large quantities of complex raw data – up to 100s of terabytes. Each scientific project has a data management plan indicated how and where this data is stored and	- STFC use a combination of commercial software products and specialised custom tools for these purposes. Without such systems, the construction of very large-scale scientific instruments would not be possible - A high level overview of STFC's computing requirements (including that of data storage) is available in STFC's eInfrastructure strategy, available at: https://stfc.ukri.org/files/e-infrastructure-strategy/	STFC has no recorded information relevant to how we "send and receive data within the workplace/lab", however the movement of high volumes of data between STFC and its international partners presents many operational challenges and technical design choices – a good introduction to the field of data transfer and network performance tuning is available at: https://fasterdata.es.net/ STFC typically deploy local computing and specialised data processing facilities close to the

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
processed, and the arrangements for retention of data and its disclosure. STFC science projects often form part of wider international collaborations and the task of storing and accessing the data may be shared amongst the international partners through a federated storage system. Most of the technical work required to handle this data and make it available to users around the world is carried out within our Scientific Computing Department. Preparations for data management make up a substantial component of the planning and design of new facilities, since cost-effective handling of data is a limiting factor in the scientific exploitation of experiments.	- STFC's Scientific Computing (https://www.scd.stfc.ac.uk/Pages/home.aspx) has a range of storage and data management solutions deployed for research communities. Specific details of the solutions used for STFC's facilities: ISIS, Diamond and CLF are available at: https://www.scd.stfc.ac.uk/Pages/Compute-and-Data-Intensive-Science.aspx - Leaflets (in particular CEPH, JASMIN and Tape Robots) providing details of some of the underlying disk and tape storage platforms used by the above data pipelines: https://www.scd.stfc.ac.uk/Pages/Leaflets-&-Flyers.aspx - More generally many of STFC's data intensive systems are closely integrated into National and International Digital Research infrastructures. STFC's Scientific Computing Department collaborates in many national and international projects of this kind – two relevant examples are: - The World Wide Lhc computing Grid (wLCG) https://wlcg.web.cern.ch - The European Open Science Cloud (EOSC) https://www.eosc-portal.eu	source of data, to make a first pass of data reduction before transfer to permanent storage. These 'data pipelines' can handle data rates of up to 100s of terabits per second, and are custom-designed for each experiment. • The choice of technical solution depends on many factors including: the volume of data, the rate of acceptance of data from instruments or partner facilities, the access rates required to process or analyse the data and the long-term retention needs of the data.
National Environment and Research Council (NERC) Information has been gathered from the research centres owned by NERC, British Geological Survey and the British Antarctic Survey, the Scientific Support Facilities that are funded by NERC (but hosted by a variety of Higher Education Institutions (HEIs) and the data centres that form the NERC Environmental Data Service.		
British Geological Survey (BGS)		
BGS and NERC run a number of laboratories each measuring differing parameters/characteristics in the course of our	Each lab uses the software most appropriate to the instruments they hold but we aggregate that data into a locally built LIMS system (at BGS). The data generated is sent to Project "Principle Investigators" who have commissioned/requested laboratory	The LIMS system contains integrity checks and data workflows which also streams the data back to our central BGS/National Geoscience Data Centre (NGDC) relation database management system (RDBMS) in which we hold the aggregated laboratory data in a number

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
funded environmental work	work (these projects will then upon their completion deposit their entire research datasets with an appropriate NERC-Funded Data Centre as mandated by the NERC Data Policy, for open access/e-use, there are five NERC supported data centres, the local one is the National Geoscience Data Centre located here at the BGS.	of differing data schemas depending upon the data types. There are plans to link these to application programming interface (API's) and web services for future easy access.

There are a considerable amount of instruments and parameters analysed in these suite of laboratories and the LIMS development is an ongoing process but one we have been and continue to prioritise. |
| **NERC Environmental Data Service (EDS)** The NERC EDS comprises five Data Centres – the Environmental Information Data Centre (based at UK-CEH), the British Oceanographic Data Centre (based at the National Oceanography Centre), the Centre for Environmental Data Analysis (CEDA) (managed by STFC in Harwell), the Polar Data Centre (based at the British Antarctic Survey (BAS)) and the National Geosciences Data Centre (based at the British Geological Survey). | With the exception of CEDA, the data centres are embedded within their host research centre and they therefore use their centre's IT systems. | |
| **Environmental Information Data Centre (EIDC)** | Online storage for the Environmental Information Data Centre (EIDC) repository is provided by a Storage Area Network (SAN) administered by a dedicated in-house IT infrastructure team. Data is backed up to Tape Libraries continually daily. The backups are managed using a commercial Storage Manager - tapes are stored off-site in a secure facility. | The majority of the data received comes in via a data uploader which is an online 'dropbox' tool where depositors can drag and drop or select files for upload (data and supporting documentation). Once depositors have added all the files they wish to upload they click 'finish' and the data is uploaded into a temporary secure location on EIDC's Storage Area Network. On receiving the data, they automatically receive a checksum so that EIDC can ensure data is not corrupted when it is |

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
		processed by the servers. More detailed instructions are sent to depositors prior to upload: https://eidc.ac.uk/help/depositing/uploadData. Data is also accepted from cloud locations (e.g. Dropbox, Google Drive or OneDrive) or an institution's file sharing service or File Transfer Protocol (FTP) Data can be accessed by users in a number of different ways. Whichever way they receive data they will always initiate the process from one of the discovery metadata records: https://catalogue.ceh.ac.uk/eidc/documents. Depending on factors such as the size, format or licensing of the data, it is made available in either a data package, by direct download or via an online order. Data available via data packager is accessed by clicking a button on one of the catalogue records. A package (zip file) is then downloaded onto the user's PC which includes the data, licensing and citation information and supporting documentation. For data available via direct access, upon clicking the button, the user is able to access the data directly from a directory on our file store. This allows users to pick and choose the files they require. Data accessed via online order manager system are commonly data resources with more complex licensing requirements, information products that are charged for, or where some form of selection is desirable (e.g. spatial clipping). When users click the button in the catalogue to access these resources they are taken to an order page, an order is submitted and if all is well, they receive the data via email a short time later.
Polar Data Centre (PDC)	The PDC's data are stored on BAS's on-premise central data storage system, managed by the BAS IT department. The data storage system has an Uninterruptable Power Supply to prevent uncontrolled loss of power and environmental control and monitoring (temperature and humidity) to keep	How data is sent to others depends on the size and format of the data. It could be via an email attachment, a direct download of the file from PDC's metadata catalogue, by FTP, or on physical media such as a hard drive. Data is received via: email attachments, FTP or other file transfer method and via physical media such as a hard drive, USB stick or SD card.

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
	these factors within the manufacturer's guidelines. Individual servers and storage systems within BAS's IT infrastructure have redundant power supplies, network connections and hard disks that ensure a single failure of any one of those high-risk components does not cause loss of data or interruption to service. The data storage system is backed up nightly, with any differences since the previous backup copied to a separate disk and tape storage system kept in a different fire zone. A copy of BAS's data is kept offsite on digital tapes, stored by a commercial data storage company in secure facilities. Data can be stored as files on the central storage system or in relational database systems. Files may be in a variety of different formats, but depositors are encouraged to provide data in open file formats. Please see https://www.bas.ac.uk/wp-content/uploads/2020/06/PDC_Data_Format_Guidance.pdf for the type of format data is requested in.	
British Oceanographic Data Centre (BODC)	BODC systems are an integral part of the National Oceanography Centre infrastructure, administered by a dedicated in-house IT team. We utilise a Storage Area Network (SAN), servers, systems and local/wide area networks including high speed access to the UK joint Academic Network (JANET). BODC data are held in a variety of systems depending on the type of data. Broadly, metadata are held in a relational database system whilst the associated data are held either in the database or on the SAN. Other digital data types such as photos and videos are also stored on the SAN.	BODC makes data available in a number of ways. Users can search BODC's data portal and download data of interest (see https://www.bodc.ac.uk/data/). Users can contact BODC by phone or email to ask the enquiries team to find the data of interest and then arrange to send the data over the network or via physical means (e.g. CD-ROM, memory stick). All of the metadata and documentation about a data set's provenance is supplied with the data when a user requests it, ensuring understandability of the data when being re-used by others. For data submission, BODC also accommodates a number of methods

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
		(see https://www.bodc.ac.uk/submit_data/) including by email, by standard mail on a DVD, CDROM or memory stick, or online via secure file exchange protocols. Where requested, the repository can mint a digital object identifier (DOI) using the DataCite service provided by the British Library. BODC can handle data in virtually any format, providing software to read it is readily available or that it is described in sufficient detail for the Centre to write the software. In all cases, an explanation of how the format has been used is required to give an understanding of what has been received. An automated data submission portal is currently being designed. This will allow depositors to easily upload their data and metadata. At present data deposits are handled manually (except for automated near-real time data streams). The BODC data management team are responsible for ensuring that the data and the metadata are complete before finalising ingestion into the BODC databases. Where the deposit is incomplete, the data management team will liaise with the depositor to ensure completeness. Data depositors are required to agree to data deposit conditions, which stipulate that they must have permission from the data originator to deposit the data.
Centre for Environmental Data Analysis (CEDA)	These are provided by JASMIN for the CEDA Archive. http://jasmin.ac.uk This is in the form of Parallel file systems, Scale out filesystems, objectstore and tape library systems.	Some users download data via http and FTP others have not direct access by using the JASMIN platform itself.
National Geosciences Data Centre (NGDC)	All of NGDC's metadata and most data is held within a spatially enabled relational database management system (RDBMS), some proprietary data is also held in the data formats in which they were supplied and stored upon a Storage Area Network with links to metadata and index information held within the RDBMS. NGDC hold some larger datasets at the CEDA large infrastructure.	Data is sent to NGDC by online data portal, email attachments, FTP or hard-drives by Principal Investigator (PI)/data donators. NGDC provide many access methods to their data for users including; • Data as supplied in search catalogues (in the original formats) • Open Geospatial Consortium (OGC) webserver Web Map Service

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
		(WMS), Web Feature Service (WFS) • Application Programming interfaces (API's) • Suite of online applications and viewers • Via BGS enquiries when sub-sets of data to match specific queries is requested and existing systems are not able to provide the user an answer (using emails and FTP)

NERC Facilities

NERC funds a number of Science Support and Facilities (S&F). A full list of the facilities that NERC funds can be found at: https://nerc.ukri.org/research/sites/facilities/list/). As each of these facilities is hosted by different Higher Education Institutions (HEIs) or research centre the storage and transfer of data varies in each facility. It should be noted that if the users of the NERC facilities are NERC grant holders then under the NERC data policy, they should make their data available via one of the Environmental Data Centres. Responses from the facilities have been tabulated below.

Research Council/Facility	Question 1	Question 3
Geophysical Equipment Facility (GEF-Edinburgh)	GEF-E staff will often assist users with data processing on return, but the data volumes concerned are such that a laptop and backup to external drives are adequate in the field.	Users are responsible for storing and archiving data obtained in the field with our equipment, so there isn't a need to send or receive large data volumes.
Geophysical Equipment Facility (GEF-Leicester)	All data is on networked storage and is transferred internally via the University's network.	This allows for rapid secure access to the data and regular secure back-up. The downside to this is the need to maintain the physical infrastructure.
Geophysical Equipment Facility (GEF-Durham & Southampton)	Data is provided to scientists at the point of acquisition with the scientist responsible for transfer and storage within a NERC data centre.	Data is provided to scientists on hard drives.
Ion Microprobe Facility (IMF)	The facility does not use LIMS. Data is stored in a secure data store server and backed-up and managed by the University of Edinburgh.	Data is provided to users (who visit the facility) on USB stick or external hard drive. Data is sent to non-visiting users via email attachment. If larger files are required this could be done via FTP transfer.
National Environment Isotope Facility (NEIF-Bristol)	Data is shared in open access formats for the purposes of sharing/publishing. Besides local storage on instrument computers and University computers, the University of Bristol has extensive facilities available for the management and sharing of data. The Research Data Storage Facility (RDSF) provides secure, long-term storage for research data providing nightly backup of all data, with further resilience provided by three geographically distinct storage	The data.bris Repository offers a means for NEIF-B researchers to openly share non-confidential research data, without the need for external data users to undergo any form of authentication. Each deposit is accompanied by appropriate metadata and is assigned a unique Digital Object Identifier (DOI) via the DataCite scheme. All data published by the Repository is available under a permissive re-use license. The RDSF provides secure data storage for a minimum of 20 years, and the Research Data Repository can publish data and

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
	locations. A tape library is used for backup purposes and for long-term, offline data storage.	ensure it is openly available also for a minimum of 20 years.
National Environment Isotope Facility – British Geological Survey (NEIF-BGS)	All project data (metadata, reduced and interpreted data) are stored in a Storage Area Network and backed up regularly. Sample information and some reduced data are stored in corporate systems at BGS.	
National Environment Isotope Facility – Scottish Universities Environmental Research Centre (NEIF-SUERC)	SUERC NEIF data are stored locally following collection on a dedicated server that is backed up to the University of Glasgow.	The various laboratories can access the data through different databases. Once data are passed to PIs and progressed to publication/thesis/abstract etc. the data are stored electronically with the publication body and PIs are encouraged to store data with the NERC data centre, including relevant metadata to allow interpretation.
NERC Environmental Omics Facility (NEOF-Liverpool)	Samples are tracked from receipt to completion with a LIMS.	Data are transferred locally across our gigabit network and stored locally on a storage network with backup. Long term and public accessible archive use the European Nucleotide Archive or similar platforms under ELIXIR.
Field Spectroscopy Facility (FSF)	Operational data is managed through commercial systems at the University of Edinburgh. Field-collected data is owned by the user. FSF collected data is stored on a storage network, disseminated through secure FTP server and will be distributed through CEDA where appropriate.	Large datasets are sent through FTP server.
Atmospheric Measurement Observation Facility (AMOF)	AMOF uses JASMIN as the main repository for experimental data collected in the field. Final datasets, with appropriate metadata, are archived at CEDA. For further details see https://sites.google.com/ncas.ac.uk/ncasobservations/home/data-project?authuser=0	Data can be accessed via CEDA (please see Centre for Environmental Data Analysis entry above)
NERC Earth Observation Data Acquisition and Analysis (NEODAAS)	NEODAAS use various computer systems depending on the type of data stored. The large archive of satellite data are stored on servers with key datasets produced by the facility backed-up or archived to tape. Some products are archived at NERC data centres for long term storage (e.g., airborne data archived at CEDA).	The large quantities of Earth Observation data stored by NEODAAS underpin all the work carried out by the service. Typically, data are received from outside the lab via the internet and transmitted around the building via ethernet.

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
Culture Collection of Algae and protozoa (CCAP)	CCAP currently uses databases to store and manage the bulk of its data. These are stored on resilient and redundant systems with a full back up regime (onsite and offsite). Third party suppliers of data storage and services are vetted both by our IT department, but also in line with a number of our policies including: GDPR, Information Security and Computer operations policy.	Databases allow for the efficient storage and querying of the data, whilst allowing for multiple access pathways. For example, the data is usually entered via local computer terminals, but is then accessible via other local computers or via public facing portals such as our website. The Tools to manage the data on local computers include a LIMS as well as standard database (DB) interface.
The British Ocean Sediment Core Research Facility (BOSCORF) NERC also funds large Research Infrastructure, one of those, The British Ocean Sediment Core Research Facility (BOSCORF) is the United Kingdom's national deep-sea sediment core repository. BOSCORF is responsible for the long-term storage and curation of sediment core samples collected through UKRI-NERC research programmes. In addition, the facility operates a suite of analysis and imaging systems specifically designed for the acquisition of geophysical and geochemical data from geological core samples.	BOSCORF manages the physical samples - including all sample metadata and the analytical data and images collected within the facility's laboratory. Until recently, sample and data management systems had been developed in an ad hoc manner as the facility evolved over time. In the last few years, a more systematic approach has been applied with the ultimate aim of developing a bespoke and fully integrated system for the facility's sample and data management. At present samples and data are managed and stored using a multi-component system: (i) The physical sample repository where samples are stored in optimal conditions for long-term preservation of the materials. (ii) A spreadsheet-based sample inventory that spatially maps the physical samples and relates this to sample metadata. (iii) A data repository on a centralised server administered by the National Oceanography Centre, Southampton, where analytical data are archived within a collection of directories. (iv) Data management software, specifically designed for geological core data management, provides a central and integrated data catalogue of sample metadata, analytical data and core images that can be	Large data transfers to remote facility users are via FTP, or datasets are transferred to hard drives during a researcher's visit to the facility as part of the sample analysis process.

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
	rapidly searched and visualized. Furthermore, BOSCORF is currently constructing a bespoke relational database to replace the spreadsheet-based sample management system. The new database will incorporate the sample inventory with sample metadata and records of sample processing, laboratory analyses and data availability.	
Biotechnology and Biological Sciences Research Council		
BBSRC is not responsible for Institute or lab data management systems or UKRI systems and does not hold any relevant information.		
Economic and Social Research Council (ESRC)		
ESRC does not directly operate research facilities/laboratories nor do we hold any data generated by those researchers we fund thus do not use LIMS or other systems for managing large amounts of data.	Publicly available information on data handling is available from UK Data Service at: https://ukdataservice.ac.uk/ and the European archive at: https://www.cessda.eu/ Additionally, ESRC funds the Administrative Data Research UK (ADR UK) programme. Information about the Trusted Research Environments within ADR UK is available here: https://www.adruk.org/our-data/our-data-services/	
Arts and Humanities Research Council (AHRC)		
	AHRC do not own labs or institutes and as a workplace AHRC does not store or use large quantities of data.	
Engineering and Physical Sciences		

Research Council/Facility	Question 1 - What systems are in place for data storage, for example Laboratory Information Management System (LIMS)?	Question 3 - How do you send and receive large quantities of data within the workplace/lab.
Research Council (EPSRC)		
	EPSRC does not directly operate research facilities/laboratories nor do we hold any data generated by those researchers we fund thus do not use LIMS or other systems for managing large amounts of data.	