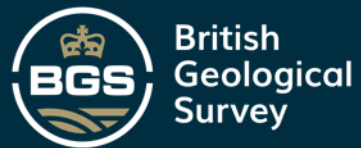




JAANA PINNICK

BGS science records at the National Geoscience Data Centre (NGDC)



ORIGINS OF OUR DATA

The context of BGS/ NGDC science records

British Geological Survey

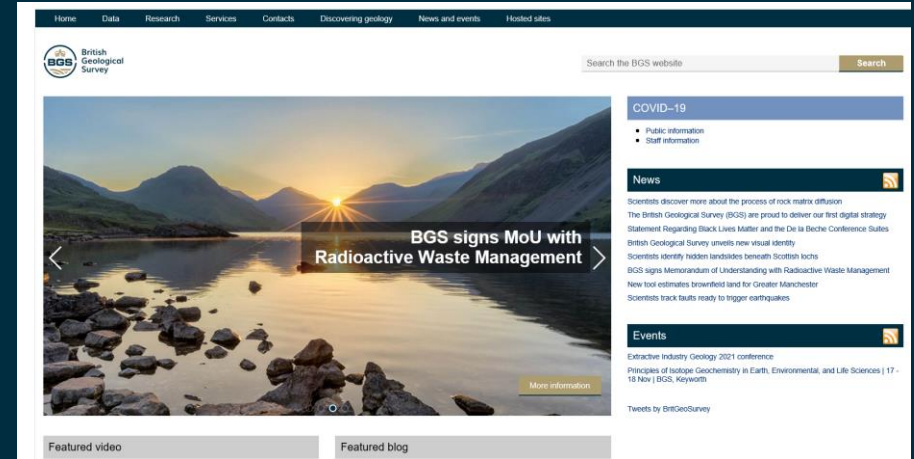
Founded in
1835 – one of
the oldest
geological
surveys in the
world



BGS:

Organisational drivers

- Place of Deposit under the PRA
- Under legal obligation to manage some types of data (e.g. Mining Industry Act of 1926, Water Resources Act 1991)
- Most data licences based on the UK Open Government Licence (OGL) or made available under EIR 2004
- Scientists need to have data available immediately in a crisis (e.g. landslides, foot and mouth outbreak, tsunamis)
- Combination of data and staff expertise a unique corporate asset



Implementing UKRI data management best practice:
data that by their nature cannot be re-measured or re-created [...] may often warrant ‘indefinite storage and preservation’

COLLABORATION

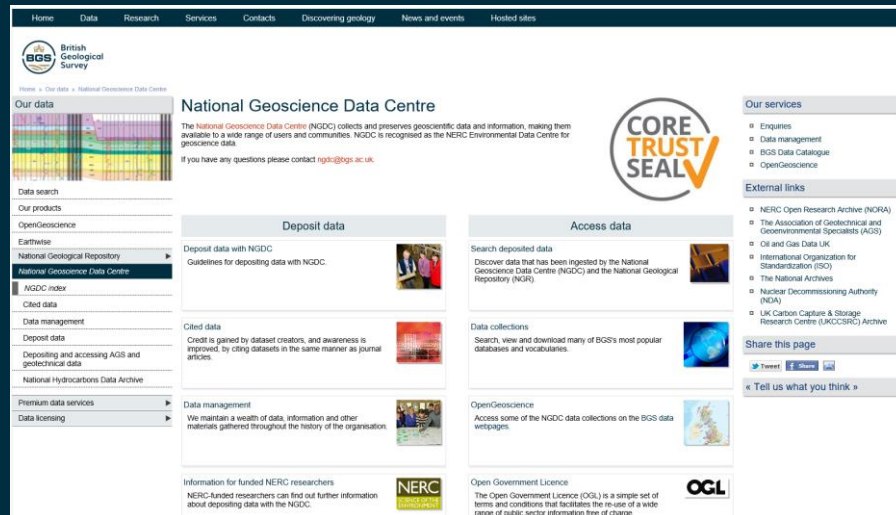
The BGS is a data-rich organisation with over 600 datasets in its care

Our data is managed by the National Geoscience Data Centre (NGDC)

NGDC drivers

- Custodian of the nation's geoscience and subsurface data archive
- Data from BGS science activities
- Receives statutory, commercial, and voluntary data donations, and data from NERC-funded geoscience grants
- Reputation as a reliable and centralised data access point, data reuse and collaboration
- Data preserved and made accessible for 10+ years after completion of research, major projects 20+ years
- Long validity of geoscience data means permanent retention is often required

Part of NERC-funded
Environmental Data Service
(EDS) hosted by NERC
Research Centres



The screenshot shows the homepage of the National Geoscience Data Centre (NGDC). The header includes navigation links: Home, Data, Research, Services, Contacts, Discovering geology, News and events, and Hosted sites. The main content area is titled 'National Geoscience Data Centre' and features a 'Our data' section with a grid of data categories. To the right, there's a 'Deposit data' section with a 'Guidelines for depositing data with NGDC' link. Below that is the 'Access data' section, which includes 'Search deposited data', 'Data collections', and 'OpenGeoscience'. The footer contains information for funded NERC researchers and the Open Government Licence (OGL) logo.

Complex, diverse range of digital geoscience datasets

- 750TB+ on the SAN, over 150 TB on tape
- Oracle RDBMS with over 3000 objects (~20TB)
- Over 1 million open access borehole records with over 3.7 million associated scanned images
- 500,000+ scanned images containing site specific geological information (such as fieldslips, mine plans, maps etc..)
- 200,000+ digital geophysical well data logs and curves
- 150,000+ photographs and imagery e.g. core photos, 3D fossil scans
- 50,000+ spatial data files

Sensor Networks data

Amalgamated data warehouse objects containing 100's million rows of data, and growing

Social data

Logs of usage and access, social media feeds - iGeology: 45+ million rows of data



Geoscience data

- Borehole
- Bedrock
- Hydrogeology
- Marine geoscience
- Geochemistry
- Geophysics
- Engineering geology
- Mining and minerals
- Natural resources
- Natural hazards e.g. earthquakes, landslides, flooding and tsunamis
- Seismology
- Geomagnetism
- Earth characteristics
- Geological processes
- Rocks
- Sediments and soils
- Land contamination
- Energy
- Oil and gas
- Climate change
- 3D modelling

Designated community?

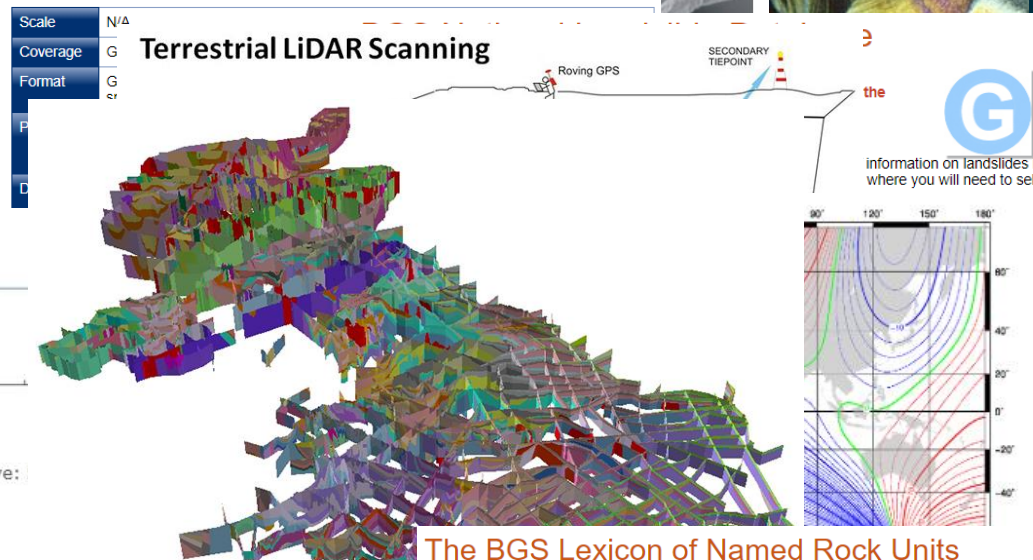
- Members of the public
- UKRI/ NERC/BGS staff
- NERC-funded Grant Holders (Principle Investigators)
- Academia – students, researchers
- Scientific communities
- Local and National Government
(Highways Agency, EA, DEFRA, OGA...)
- Commercial companies (insurance, civil engineers....)
- Geological or geoscience consultants
- Data resellers

“A digital science record”?

- Core and thin section photographs
- Electron micrographs of fossils
- Hydrogeological Database, National Landslide Database
- Seismograms from earthquakes (-real-time seismic data)
- LiDAR scans
- Magnetograms
- 3-D models
- Geochemical maps
- Vocabularies and taxonomies



WellMaster hydrogeological database



Location index for chemical data in environm

This Google Earth project, and the generation of kml file from R scripts, is described in the [project repo](#). You will need Google Earth on your computer to use this application. [Download Google Earth](#)

Note. Locational data are based on British National Grid coordinates (Irish Grid in N Ireland) converted Survey maps or, since 2003, using GPS. Possible error in spatial locations are estimated at ±100 m.

England Scotland Wales N Ireland				
SOIL				
	KML (topsoil)	KML (deepsoil)	Info	
G-BASE				
FOREGS				
GEMAS				

The Lexicon of Named Rock Units database provides BGS definitions of terms that appear on our maps

Search the Lexicon

Rock Unit:

Computer Code:

Preferred Map Code:

Maximum Age of Rock Unit:

[Reset Defaults](#)

[Submit Query](#)

All Ages

- 185 years of legacy data in a variety of formats and media types
- Expense and difficulty of data creation and collection
 - Data from deep boreholes to the depth of many kilometres, costing tens of millions of pounds each to drill
 - Seismic data originating from earthquakes, unique and unrepeatable data
- Insufficient or unclear T&Cs and/ or metadata for data reuse/ repurposing, or re-interpretation
- Changing semantics and scientific vocabularies over time
- Prohibitive cost of annual licensing or expired licences



BGS/NGDC Integrated Data Model

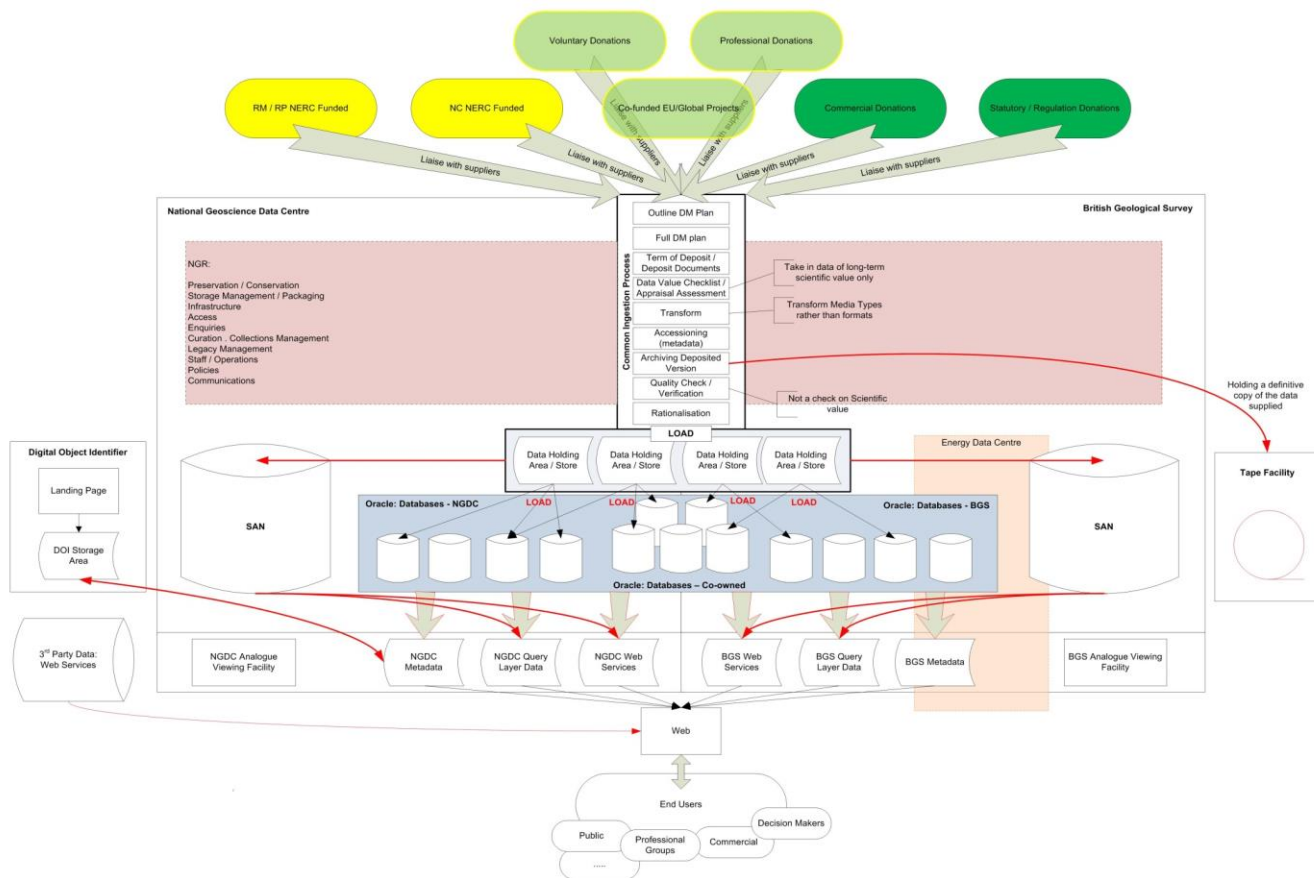
- Data management programme from ~1980
- 2000 ->: centralisation, standardisation and deduplication of data
 - Integration of various versions of datasets into one Oracle database
 - Creation of BGS Discovery Metadata to support data discovery
 - Data entry and delivery applications developed in-house
 - Corporate scientific vocabularies and data dictionaries
- Today, NGDC does data-level curation using BGS expertise to add value and enhance the content whilst preserving original data



FUTURE PROOFING OUR DATA

Research Data Management in 2020

FUTURE PROOFING: STREAMLINING DATA INGESTION



Data pathways into NGDC

1. NGDC Digital Data Deposit Application (<20 GB, ZIP files ok)

- Online guidance incl. T&Cs, Acceptable formats list, NERC data value check list, Data sharing agreements

2. BGS ShareFile (20 GB – 200 GB)

3. NERC Large Data Donation Portal (200 GB to multi-TBs)

- Ingestion of (meta)data to Original Data Store (Level 1)
- Processed and Accessioned: Donated Data Store (Level 2) and Oracle RDBMS/ National Databases (Level 3)

The screenshot shows a web browser window with multiple tabs. The active tab is titled "Thank you for depositing data with the NGDC". The browser's address bar shows the URL: "http://transfer.bgs.ac.uk/ingestion/deposit/5678851458CC71A1E039594ABC0F15C/confirm". The page content includes a confirmation message from the NGDC staff, contact details for Claire Shelley (Email: cshelley@bgs.ac.uk, Phone: 0115 936 3100), and a list of files deposited. The page is cluttered with various browser elements like address bars, tabs, and social media links. The BGS logo is visible in the top left corner of the page.



Providing Best Practice and Guidance

NGDC Data Remit/Scope

The National Geoscience Data Centre (NGDC) is host to the Natural Environment Research Council's (NERC) geoscience data and datasets. Data and the processes of depositing data to the NGDC policy to provide a wide range of user access.

Good data deposit: Acceptable digital formats



Good data

Data should normally be provided in a non-proprietary format (e.g. spreadsheet).

The following formats are acceptable:

NGDC Ingestion

Your data should:

It is the policy of NERC that:

- high quality, suitable for policy makers
- a geoscience (or other) data
- digitally born, or converted to digital
- deposited by the data owner
- exclusive "in-house" data
- compliant with the Data Value Checklist



Good data donation

When packaging up your data to deposit:

It is also the policy of NERC that:

- boreholes drill data should be grouped appropriately
- Industry Act 1 data should be grouped appropriately
- boreholes drill data should be grouped appropriately
- Hydrocarbon data should be grouped appropriately

It is the policy of NERC that:

- standards conform to the Data Value Checklist
- terms and conditions of use are clear
- access or use is appropriate

Contacts

It is the policy of NERC that:

- store a copy of the data
- incorporate (when funding permits) the data with the data management advice and guidance
- provide data management advice and guidance
- ensure the data is discoverable and provide open access to the nationally consistent datasets
- create Digital Object Identifiers (DOI's) for appropriate display alongside the data
- encourage the use of these datasets for a full range of projects and within information products or decisions

Contacts

For further details, please contact ngdc@bgs.ac.uk.

Data type

Geotechnical data

Geophysical data

Generic scientific data

Text

Presentations

GIS/spatial data

Databases

NGDC Data Value Checklist

Purpose and scope

The Data Value Checklist is a National Geoscience Data Centre tool.

The data value checklist is a guidance on assessing the value of data.

General guidance on the checklist

Selection of data should be based on a contribution to the scientific international context.

1. **RELEVANCE TO MIS:** Is the data aligned with the NERC Data Value Checklist? Consideration should be given to compliance with the Environmental Information Regulations and long term management arrangements.

2. **SCIENTIFIC OR HISTORIC:** Is there, or could there be potential for, a contribution to the scientific international context? Is the data of potential importance? Is it difficult but consideration should be given to its value to the scientific community?

3. **UNIQUENESS:** Is this the primary and most up-to-date version of the data? Have there been any updates? Are there any updates? Are there any updates? Are there any updates?

The NERC Data Centre will accept data that is:

Checklist

Essential criteria: These are legal or regulatory criteria and answering 'Yes' to **one or more** of the questions below will automatically result in selection for retention.

Legal/statutory considerations	Yes	No
Is there a legal or legislative reason for NERC to retain the data under any of the following:		
Science & Technology Act 1965		
Mining Industry Act (1926)		
Water Resources Act (1991)		
Petroleum Operations Notice 9 (PON 9) regulations (on-shore and off-shore)		
Public Records Act (1958 & 1967)		
Has or could the data been used in litigation, public enquiries, police investigations or any report or paper that could be legally challenged?		
Are there any financial or contractual obligations that require us to retain the data?		

Important criteria: These are primary criteria and answering 'Yes' to **at least one** of the questions from each section below should result in selection for retention.

Policy	Yes	No
Does the NERC Data Policy apply to this data?		
Are the data a result of NERC/BGS funded activities?		
Does this data fall within the NGDC remit?		
Scientific or historic value	Yes	No
Does the data have a geographical or temporal extent that makes it useful to others?		
Does the data have historic value i.e. does it represent a landmark in scientific discovery?		
Do the data include changes in processing methods, new standards or set any precedents?		
Do the data support current projects or trends in science?		
Is there likely to be further work in this or associated science areas?		
Are the data likely to meet the future needs/direction of the scientific community?		
Do the data contribute to a wider collection?		
Is there potential for re-use of the data?		
Are the data cited in a publication?		

- 3 local copies:
 - Original Data – Active Data – Delivery Data
- 3 geographically separate copies of key datasets:
 - BGS Nottingham – BGS Edinburgh – University of Nottingham
- Large multi-TB data at NERC Large Data Archive hosted by CEDA, linked to BGS Discovery Catalogue
- Storage on shared network drives (SAN):
 - W:\drive for active 'live' project data
 - S:\drive or Databases for corporate data
 - V:\drive for data to be archived
 - Tape archive

FUTURE PROOFING: BGS DISCOVERY METADATA

Field	
Dataset contacts – who is responsible and how to contact them	
Permission to deposit – your role	
Dataset title and description/ abstract	
Methodology used to collect the data	
Keywords	
Collection date range	
Description of geographical extent	
Spatial reference system	
File formats	
Access and use restrictions, copyright statement	
Data embargo dates	

Standards compliant:
ISO19115 / 19139
INSPIRE & UK Gemini v2.3

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<gmd:MD_Metadata xmlns:gmd="http://www.isotc211.org/2005/gmd" xmlns:gco="http://www.isotc211.org/2005/gco" xmlns:gml="http://www.isotc211.org/2005/gml" xmlns:gss="http://www.isotc211.org/2005/gss" xmlns:gts="http://www.isotc211.org/2005/gts" xmlns:srv="http://www.isotc211.org/2005/srv" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.isotc211.org/2005/gmd http://inspire.ec.europa.eu/draft-schemas/inspire-md-schemas/apiso-
  <gmd:fileIdentifier>
    <gco:CharacterString>085bd761-742f-3315-e054-002128a47908</gco:CharacterString>
  </gmd:fileIdentifier>
  <gmd:language>
    <gmd:LanguageCode codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resources/cod
    </gmd:language>
  <gmd:hierarchyLevel>
    <gmd:MD_ScopeCode codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resources/cod
    </gmd:hierarchyLevel>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:organisationName>
        <gco:CharacterString>British Geological Survey</gco:CharacterString>
      </gmd:organisationName>
      <gmd:contactInfo>
        <gmd:CI_Contact>
          <gmd:phone>
            <gmd:CI_Telephone>
              <gmd:voice>
                <gco:CharacterString>+44 115 936 3100</gco:CharacterString>
              </gmd:voice>
            </gmd:CI_Telephone>
          </gmd:phone>
          <gmd:address>
            <gmd:CI_Address>
              <gmd:deliveryPoint>
                <gco:CharacterString>Environmental Science Centre,Keyworth</gco:CharacterString>
              </gmd:deliveryPoint>
              <gmd:city>
                <gco:CharacterString>NOTTINGHAM</gco:CharacterString>
              </gmd:city>
              <gmd:administrativeArea>
                <gco:CharacterString>NOTTINGHAMSHIRE</gco:CharacterString>
```



► Add your review

Future proofing: RDM training

RDM training course for NERC-funded earth science PhD students from 2016 (165 students/ 9 courses to date)

- NERC data policy
- Active data/metadata management
- Writing a DMP
- Data quality and depositing data
- FAIR principles and open science
- Data storage and preservation
- Data repositories

“Building resilience at the National Geoscience Data Centre: enhancing digital data continuity through research data management training” (iPres 2019)



https://ipres2019.org/static/pdf/iPres2019_paper_12.pdf

Future proofing: Digital data survey

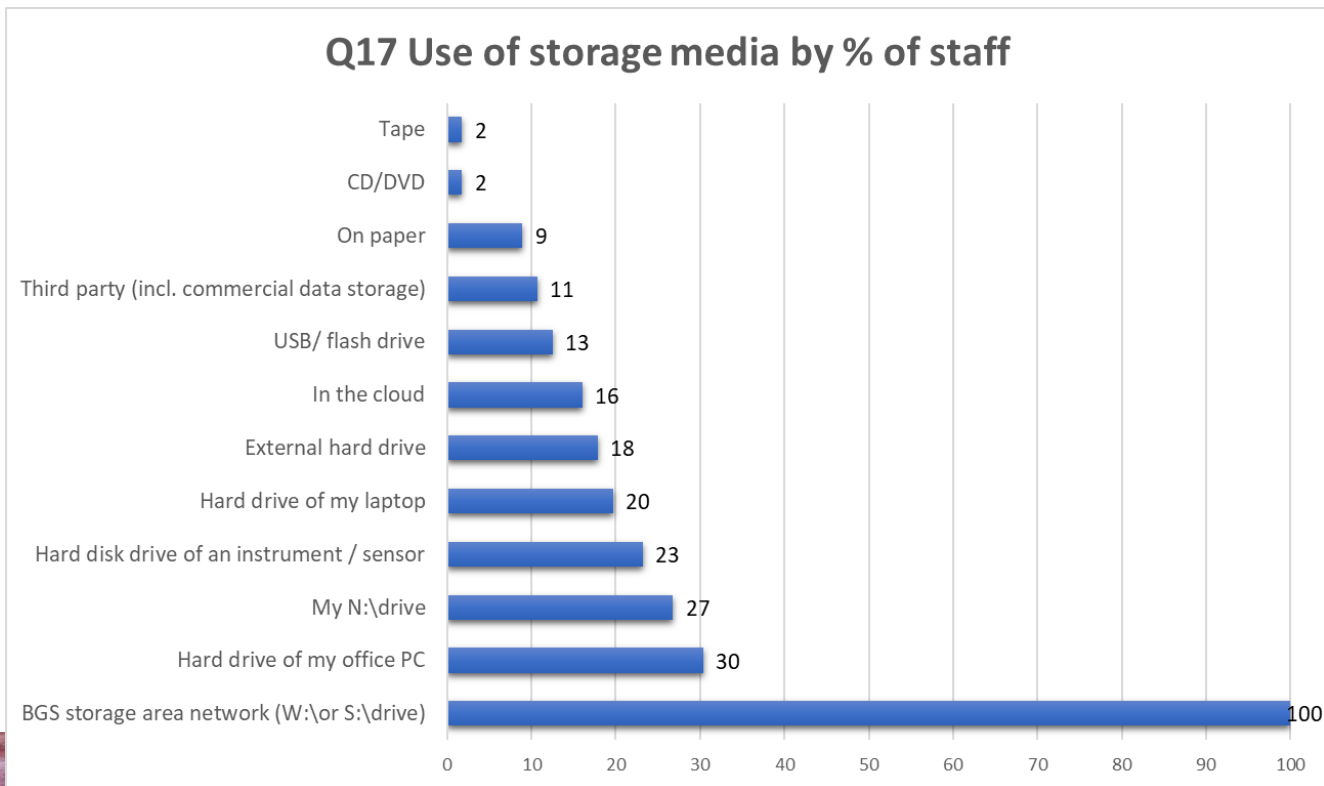
BGS Digital Research Data Survey
2019 (based on DAF methodology)
explored:

- Skills and resources
- Ingestion and storage
- Sharing and security
- IPR and data ownership
- Discoverability and reusability
- Archiving
- Digital preservation
- Internal report underway 2020

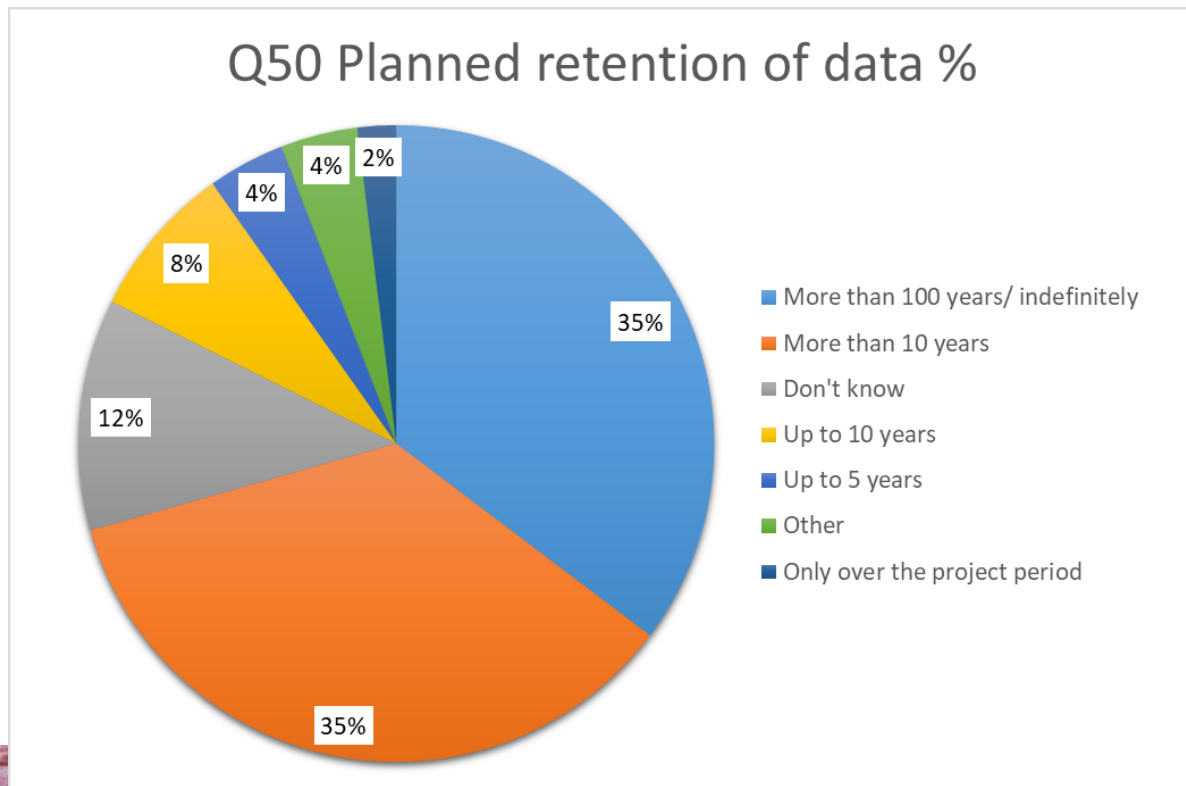
The overall objectives of the survey were to:

- Evaluate what impact researchers' data management practices have on the long-term usability and resilience of NGDC data holdings
- Evaluate and enhance corporate research data management workflows and processes in response to reasonable user needs
- Enhance the long-term accessibility and usability of our data (FAIR data)
- Enhance the long-term digital continuity and preservation of our research data assets (TRUST)

DIGITAL DATA SURVEY: USE OF STORAGE MEDIA BY % OF BGS STAFF



DIGITAL DATA SURVEY: PLANNED RETENTION OF SCIENCE RECORDS BY % OF BGS STAFF



NGDC aim:

to maintain the long-term reusability and accessibility of authentic born-digital and digitised geoscience data objects as long as required, as evidence of the current UK plc scientific/ research record

OUR MODULAR DIGITAL PRESERVATION PROGRAMME

BGS Digital Preservation Policy 2017

Business case 2018

CoreTrustSeal (CTS) certification in 2018

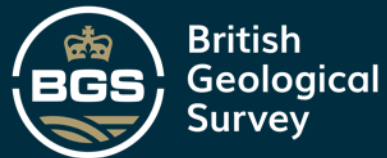
DP Strategy in 2019

DP Policy and strategy being updated in 2020

(aligned with 2020 BGS Digital Strategy)

CTS certification being updated for 2021





THANK YOU

Any questions?

jpak@bgs.ac.uk