National Geoscience Data Centre (NGDC): Drivers, experiences and lessons from seeking CoreTrustSeal certification

Garry R. Baker

Head of National Geoscience Data Centre (NGDC)

grba@bgs.ac.uk

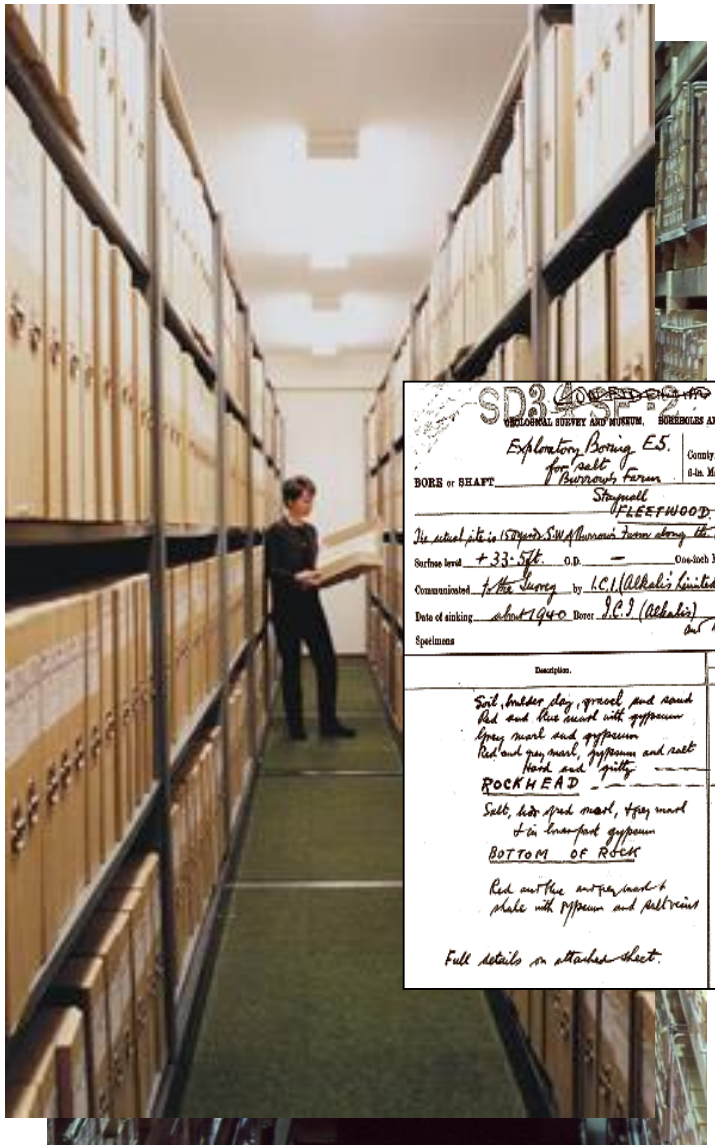
British Geological Survey
Keyworth
Nottingham
www.bgs.ac.uk

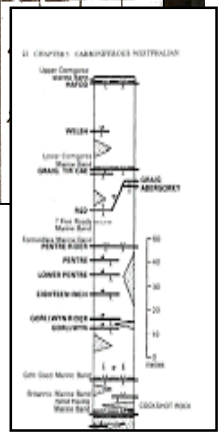
NGDC

- One of NERC Environmental Data Centres
  - National Geoscience Data Centre (NGDC)
  - Centre for Environmental Data Analysis (CEDA)
  - Environmental Information Data Centre (EIDC)
  - Polar Data Centre (PDC)
  - British Oceanographic Data Centre (BODC)

- Manage environmental scientific data and information from a range of differing geological or geoscience disciplines
- Holds historical/legacy observation data up to real-time geological processes, sensor networks and data streams
- Preserve this data for the long-term as evidence of existing of the current scientific/research record and to make it easily accessible for future re-use

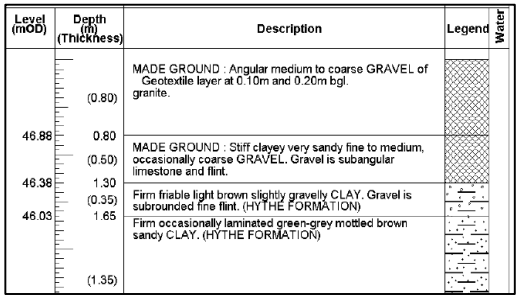
NGDC: Breadth of Data and Information

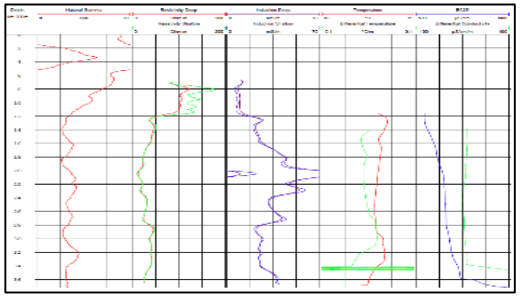


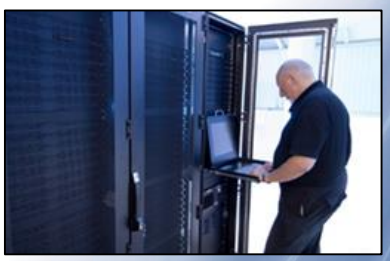


NGDC: Data Holdings

- Complex, diverse range of environmental / geoscience datasets including:
  - **900TB+ on SAN** (plus another 500TB in tape archive)
  - Oracle RDBMS with over 3000 objects
  - Near-real time data from sensors on monitoring sites and observatories
  - Over **1.5 million open access borehole records** with **3.7 million associated scanned images**
  - **500,000+ scanned images** containing site specific geological information
  - **200,000+ digital geophysical well data** logs and curves
  - **150,000+ photographs and imagery** (core photos, 3D fossil scans)
  - **50,000+ spatial data files (GIS)**
  - Business intelligence - logs of app usage, queries upon data stores, social media……
  - Data Warehouse built from sensor monitoring sites, sub-surface parameter data
  - …..plus….450km of core, millions of samples and 17 linear KM of paper records (Many differing analogue data types)





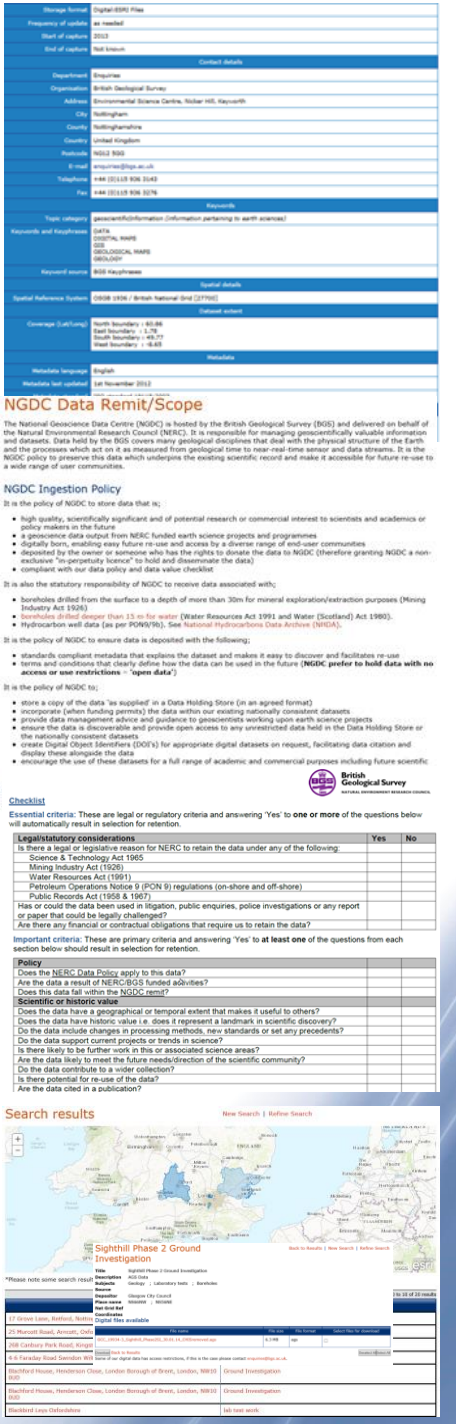


Drivers

- It's becoming a **crowded marketplace**, there's a need to stand out from the crowd, **demonstrate our professionalism at longer-term data & information management** (help demonstrate the value and benefits of the repository)
- We need to **consider the full lifecycle of the repository** from the core underpinning infrastructure, staff, policies, procedures, workflows and systems; starting from best practice guidance and donation through to delivery and ultimately re-use
- **Benchmarking our processes and service** against a recognised standard framework in CoreTrustSeal, provides evidence of the level of quality, compliance and impact from the repository
- The need to **provide confidence to our funders** on the investment they make within our repository
- **Alignment to scientific journals and publications** – they now expect to link to the data underpinning the scientific paper but only in reputable/certified repositories

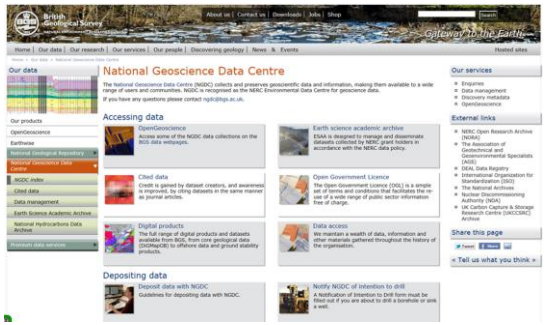
Experiences

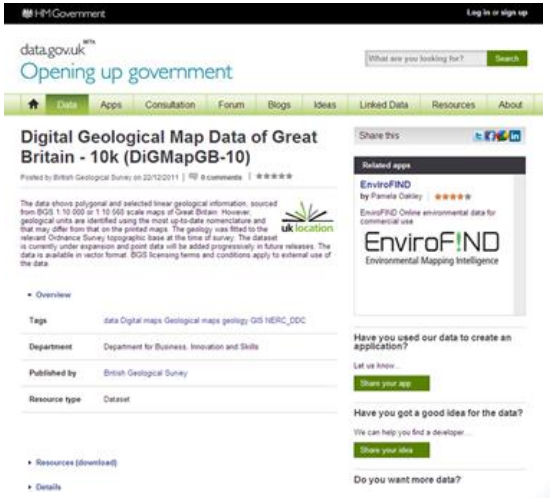
- Getting the team together was invaluable, proved a superb vehicle to enhance communications, discuss new ideas and agree priorities
- We had a range of policies, procedures and systems in place for engagement, donation, management and delivery
  - Some were robust and well thought-out
  - Some were okay but lacked clarity/not easily understood, or a little outdated
  - Some were implied, “knowledge in heads”
- Our community/global standards compliance is already strong, the CoreTrustSeal framework reinforced this
- We do have a diverse range of user types and communities (for both donation and data re-use) which presents challenges
- Clear reminder of the breadth of a modern data repository

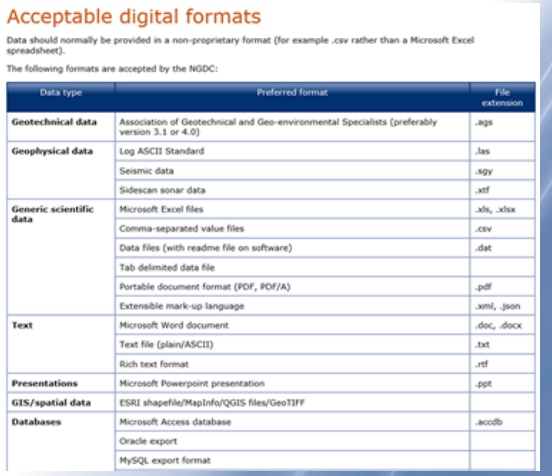


Lessons Learnt - 1

- Reminded us of the importance of keeping updated, accessible documentation
- Process of seeking CoreTrustSeal accreditation brought the NGDC team closer together, it created time for sharing knowledge, experiences and ideas (led to better communications and understanding)
- We identified some areas that needed improvement this included processes, workflows, clarity on roles, prioritising some future work activities and enhancements to our systems
    - *Data preservation, Disaster recovery, Terms & Conditions, access and use*
- Provides a vehicle to explain the business needs of the repository to managers and funders, especially as we undertake the continuous professional development aspects of the framework
- Encouraged us to enhance our expectations (requirements) from geoscience/environmental data donators
- Encouraged us to look more seriously at future stakeholder engagement activities and plans







Data type	Preferred format	File extension
Geotechnical data	Association of Geotechnical and Geo-environmental Specialists (preferably version 3.1 or 4.0)	.ags
Geophysical data	Log ASCII Standard	.las
	Seismic data	.sgy
	Sidescan sonar data	.xtf
Generic scientific data	Microsoft Excel files	.xls, .xlsx
	Comma-separated value files	.csv
	Data files (with readme file on software)	.dat
	Tab delimited data file	
	Portable document format (PDF, PDF/A)	.pdf
	Extensible mark-up language	.xml, .json
Text	Microsoft Word document	.doc, .docx
	Text file (plain/ASCII)	.txt
	Rich text format	.rtf
Presentations	Microsoft Powerpoint presentation	.ppt
GIS/spatial data	ESRI shapefile/MapInfo/QGIS files/GeoTIFF	
Databases	Microsoft Access database	.accdb
	Oracle export	
	MySQL export format	

Lessons Learnt - 2

- Both donators, funders and journal publishers really do like the robustness and reputation of certified repositories;
  - *NERC considering specifying CoreTrustSeal for their Data Service (as part of future commissioning processes)*
  - *BGS highlight the certification in discussions / presentations concerning their high-profile science facilities and observatories*
  - *Appreciated by key donators reinforcing the professionalism, quality, robustness, long-standing, public good aspects of the data centre*
- The role of training in disseminating best practice in data/info management and preservation to research or repository staff plus other external stakeholders (helps to mitigate future risks)
- Evidence generated from the CoreTrustSeal framework provides a sound platform from which its easier to assess and collaborate on FAIR, INSPIRE and other global data initiatives