Exploring digital preservation requirements: A case study from the National Geoscience Data Centre (NGDC)

Jaana Pinnick

British Geological Survey, Keyworth, Nottingham

Abstract

Purpose - This case study is based on an MSc dissertation research undertaken at Northumbria University. The aim was to explore digital preservation requirements within the wider NGDC organisational framework in preparation for developing a preservation policy and integrating associated preservation workflows throughout the existing research data management processes.

Design/methodology/approach - This mixed methods case study used quantitative and qualitative data to explore the preservation requirements and triangulation to strengthen the design validity. Corporate and the wider scientific priorities were identified through literature and a stakeholder survey. Organisational preparedness was investigated through staff interviews.

Findings - Stakeholders expect data to be reliable, reusable, and available in preferred formats. To ensure digital continuity, the creation of high quality metadata is critical, and data depositors need data management training to achieve this. Recommendations include completing a risk assessment, creating a digital asset register, and a technology watch to mitigate against risks.

Research limitations/ implications - The main constraint in this study is the lack of generalisability of results. As the NGDC is a unique organisation, it may not be possible to generalise the organisational findings although those relating to research data management may be transferrable.

Originality/value - This research examines the specific nature of geoscience data retention requirements and looks at existing NGDC procedures in terms of enhancing digital continuity, providing new knowledge on the preservation requirements for a number of national datasets.

Keywords Digital preservation, digital continuity, geoscience data management, data centre, digital repository

Article type Case study

Introduction and background

This article explores the requirements of the National Geoscience Data Centre (NGDC) to ensure that the long-term preservation and usability of its digital data are supported and aligned to the corporate aims. It examines the specific characteristics of geoscience and geospatial data and looks at the efficiency of the existing data management procedures in terms of digital continuity within the current challenging funding climate.

The NGDC is the designated repository for the Natural Environment Research Council (NERC) grant-funded geoscience research data and the guardian for many commercially funded datasets. It is hosted by the British Geological Survey (BGS) and responsible for ensuring the availability of the data as one of the NERC Environmental Data Centres. BGS corporate budgets and staffing levels have decreased during 2010-2015, whilst the volume of digital data has more than doubled. Data management needs to consider the existing organisational framework under the Research Councils UK and the Department of Business, Energy and Industrial Strategy (BEIS).

As a public sector organisation BGS is committed, on behalf of NERC as the legal entity, to look after certain geoscience data in its care in perpetuity (Bowie, 2010) and to make most of it openly available to a wide range of stakeholders, who in turn use the data to develop products and services as well as to inform their decision-making. This requires the organisation to monitor the ongoing condition of digital data and to take appropriate actions in collaboration with its stakeholders to ensure the usability, trustworthiness, and future interoperability of those data. These attributes can only be achieved if data remain both accessible and understandable for future users.

The geoscience data held at the NGDC includes a wide range of data types including but not limited to borehole, bedrock, hydrogeology, geochemistry, seismic, marine geoscience, oil and gas, airborne geophysical, and geohazards data. They have been collected and accumulated over long periods of time and are used by industry, manufacturing, construction, and transport sections, as well as the public sector and academia and researchers, to build UK infrastructure, develop insurance and other data products, innovate, build risk models, answer science questions e.g.

in climate change research, and to support many geoscience applications. Many stakeholders (40%) have been using the NGDC datasets for over 10 years. The use of numerous proprietary software packages (Vulcan, MicroStation) over the years, the lack of restrictions in used file formats in the past, and the occasionally incomplete contextual metadata means older digital data is not always easily accessible to current users if preservation actions are not taken at appropriate times. Past decisions made – or not made – by data creators and guardians at the ingestion phase have a direct impact on the data quality today.

An additional strategic driver for building a digital preservation programme is the plan to apply for a Trusted Repository Status under the Data Seal of Approval (DSA) and the International Council for Science World Data System (ICSU-WDS) certification (RDA, 2016), requiring the NGDC to provide evidence for its long-term preservation capability as a data repository. This includes having a continuity plan in place to ensure the ongoing preservation of data holdings, ensuring the integrity and authenticity of the data, and managing the long-term preservation in a planned and documented way. Although a lot of work has already been done, some of it is not documented consistently. The data centre currently holds over 275TB of data and, although it has considered the digital preservation aspect before, has no formal workflows in place to incorporate preservation actions within data management processes outside the Oracle relational database management systems. Maintaining the value and usability of the data means introducing these workflows has to be a priority in going forward.

Methodology

This exploratory case study started by reviewing literature in order to place the findings in the context of long-term preservation of digital geoscience data at a digital repository. The field work phase used a mixed methods approach employing both quantitative and qualitative data. The sequential design included quantitative and qualitative data to provide a comprehensive analysis of the case ‘by combining information from complementary kinds of data or sources’ (Denscombe, 2008). The iterative approach used ethnographic methods acknowledging the role of the researcher as part of the organisation studied (O’Reilly, 2012) and constant comparative analysis to create categories from raw data (Pickard, 2013).

Manipulation and interpretation of dataset access statistics, a literature review, and a macroanalysis of corporate documentation provided a framework on which an external stakeholder survey was based. The datasets and products were selected based on them being widely accessed via the BGS OpenGeoScience web service (http://www.bgs.ac.uk/opengeoscience) and included borehole geology, DiGMapGB, groundwater flooding aquifer designation, radon, GeoSure, and mining hazards data. These data are used by industry, academia and general public alike. Corporate and academic users were invited to participate in a small online survey. Transcripts from a purposive sample of staff interviews were compared with the findings. The aim was to cover the research data lifecycle from data management planning and data creation through to accession, archiving, preservation, and reuse.

Establishing trustworthiness of the qualitative research strands can be judged by demonstrating credibility, transferability, and dependability (Pickard, 2013). To increase credibility, this study used triangulation in the form of multiple data collection techniques and sources, which also 'reflects an attempt to secure an in-depth understanding of the phenomenon in question' (Denzin, 2012). Transferability was made possible by providing rich contextual data. To strengthen dependability, member checking was employed during the transcript phase.

Literature review

Digital preservation

The early report of the Task Force on Archiving of Digital Information 'laid foundations for most subsequent work in the field, and continues to shape the agenda even today' (Brown, 2013). Its first conclusion, still relevant, is one of the cornerstones of this research: 'The first line of defence against loss of valuable digital information rests with the creators, providers and owners of digital information' (Garrett and Waters, 1996).

Key areas in digital preservation research are wide-ranging and include, but are not limited to:

- Preservation planning, policy and strategy development *(Farquhar and Hockx-Yu, 2007; Becker et al., 2009)*
- Preservation metadata and standards *(BS ISO 15836, 2009; Lavoie and Gartner, 2013; Library of Congress, 2016)*

- Digital preservation infrastructure, models and toolkits *(Jones, 2006; Ruusalepp and Dobreva, 2012; Lavoie, 2014)*
- Trusted/institutional repositories *(Hockx-Yu, 2006, Ball 2010)*

The Digital Preservation Coalition (DPC) Technology Watch series has captured many salient topics (DPC, 2016). As recently as in 2012, Ross expressed 'an urgent need for a theory of digital preservation and curation' (Ross, 2012), listing nine themes (e.g. repository management, preservation as risk management, and preserving the context) agreed by the DigitalPreservationEurope (DPE) and providing a framework for digital preservation research.

To create a digital preservation programme, a solid understanding of corporate drivers is required. This may include collection development, use access, information reuse, legal and regulatory compliance, and efficiencies and savings (Brown, 2013). A recent survey by the Information Governance Initiative (IGI) preferred the term 'long-term protection and access' to 'digital preservation' and suggested there is no distinction between permanent retention and keeping digital content for at least ten years (IGI, 2016).

A UK report on research data management suggests greatest benefits are created by developing ingest and access activities (Beagrie *et al.*, 2010). The UK Data Archive argues that 'maximizing ingest processing efforts as early as possible in the process of digital curation ensures that long-term access is available at a lower total cost' (Woollard and Corti, 2014) and emphasises the importance of the researcher creating good documentation to support data preservation (Van den Eynden *et al.,* 2011).

The NGDC has recently put a lot of effort into developing its digital data ingestion processes and offers guidance to depositors on its data portal website. It employs a NERC-wide Data Value Check List to support the appraisal of data in the pre-ingest phase, captures metadata as part of the ingest process, and offers a list of preferred file formats. All this contributes towards creating better data documentation early in the data lifecycle, but more focused preservation planning and process automation is required as data volumes increase.

The draft BGS digital preservation policy indicates it will use the Open Archival Information System (OAIS) reference model covering ingestion, archival storage, data management, administration, preservation planning, and access functions (BSI, 2012a). The other high-level standard, BS ISO 16363 (BSI, 2012b), whilst extremely competent, is acknowledged to require too many resources to achieve at this point in time.

Repositories and research data management

Ruusalepp and Dobreva recommend the analysis of preservation 'from the point of view of the organisation's business processes and stakeholders' and state 'understanding the *specific* requirements and their implication on the preservation infrastructure is important' (Ruusalepp and Dobreva, 2012, *italics in original*). This chimes with McGovern and McKay who state '[o]rganizations cannot acquire ready-made, out-of-the-box digital preservation programs' (McGovern and McKay, 2008).

Kenney and Buckley studied the relationship between repositories and digital preservation and concluded 'insufficient attention was being paid to the organizational context of digital preservation programs' (2005, quoted in McGovern and McKay, 2008).

In the UK context the expectation is to preserve research data for ten years. The RCUK Data Policy states: 'Data with acknowledged long-term value should be preserved and remain accessible and usable for future research' (RCUK, 2015b). The accompanying guidance notes add: 'To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the use of public funds' (RCUK, 2015a). The NGDC, which receives its funding from NERC, considers the retention of most of its geoscience data to be longer than ten years due to its 'national good' value, unrepeatability, and the high expense of collecting/creating some of the data. It therefore carries even more responsibility for maintaining the long-term usability of its digital assets.

To oversee long-term research data management, several non-profit and open source digital repository solutions and frameworks, including Dryad (http://datadryad.org/) and Fedora (http://fedorarepository.org/), offer repositories the benefit of wider user communities but require an extra development layer, and a

resource need, to integrate the technology with existing infrastructure and to capture sufficient metadata. They are also generic data repositories for any types of research data, whereas the NGDC has the added benefit of in-house geoscientists to e.g. support the creation of descriptive metadata for even externally generated data and to provide guidance in the future use of data.

Risk assessment and preservation toolkits, such as DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) (DCC and DPE, 2015) and SCIDIP-ES (SCIence Data Infrastructure for Preservation with focus on Earth Science) (EC, 2014), assist repositories in identifying and documenting risks associated with digital research data. The collaboration of repository staff with geoscientists is essential to capture the preservation metadata and representation information.

Pryor suggests 'lower long-term costs for the preservation of datasets are perhaps the largest shared benefit' (Pryor, 2014) of a shared infrastructure. He also points out that the allocation of public funds in the RCUK data policy does not say explicitly how the costs are covered. NERC has allocated its environmental data centres, including the NGDC, a small percentage from a central grant budget for long-term management of standard grants data. For larger grants this is costed in the project plans.

Data repositories can be divided into institutional repositories based at universities; discipline-specific repositories such as the NERC data centres; and repositories attached to publishing houses, which are increasingly storing data underpinning articles in their scientific journals (Campbell, 2015; Jones, 2014).

In the UK '[m]uch could be done to consider digital preservation from the outset, to involve the authors and to embed digital preservation into repository workflow, which will ease the later preservation tasks' (Hockx-Yu, 2006). It is an ongoing challenge for repositories 'to balance the need for fixity in the datasets they offer with the fluidity of changing practices in their designated communities' (Daniels *et al*., 2012).

A key challenge in the data lifecycle is the transfer of data from the researcher(s) to the repository, which ought to be clear about its role and responsibilities to ensure high-quality metadata is created and ingested with the data for the purposes of discovery, accessibility, restrictions, user terms and conditions, and preservation,

and to provide tools to aid the researchers in the transfer process (Jones, 2014). Data preservation forms the end of the data lifecycle, but efforts to strengthen usability are best made earlier using effective data management planning and risk mitigation as tools to organise the data and capture preservation metadata.

A proposed set of criteria to measure metadata quality at repositories includes completeness, accuracy, and consistency (Park, 2009). In another study 39 tools for semi-automated metadata generation were evaluated, suggesting these are 'important considering the fast development of digital repositories and the recent explosion of data and information' (Park and Brenza, 2015). The NGDC is planning to identify and evaluate some of the available digital preservation tools to help automate its processes, ideally by integrating open source tools within its in-house ingestion and data management systems. Of interest are for example tools for extracting preservation metadata information and file format identification or validation.

The DCC provides resources and guidance for appraising research data and evaluating repositories (DCC, 2016; Whyte, 2015). Whyte's checklist contains three levels of proficiency, from basic to more advanced, and relates to the three-level framework endorsed by the European Commission (EC, 2016).

Geoscience and geospatial data

According to the NERC Data Policy, 'NERC requires that all environmental data of long-term value generated through NERC-funded activities must be submitted to NERC for long-term management and dissemination' (NERC, 2010). There is a pressing need to evaluate the preservation requirements of all NGDC geoscience data, a large part of which are geospatial.

Geoscience data are of numerous types and cover research areas such as climate change, earth characteristics, rocks, sediments and soils, seismology, marine geology, land contamination, geological processes including erosion and volcanic activity, natural resources, and many more. A specific characteristic of all geoscience data is their long validity, which considerably extends the usable lifespan of digital data assuming actions are taken to preserve them. Geoscience is an interpretive discipline working on previous data and hypotheses which must remain available for future research and interpretation. In addition, not all uses for scientific data are

known at the time of data creation and capture, as has been the case e.g. in the use of historical environmental and palaeontological data in modern climate change research or of basic borehole data in creating 3D-models.

As future research questions cannot usually be predicted, it is essential to document geoscience data well at the earliest opportunity to enhance their future usability and interoperability potential. This includes developing vocabularies relating to the data bearing the risk of changing semantics over time. For example, the BGS Lexicon of Named Rock Units dictionary (http://www.bgs.ac.uk/lexicon) provides previous and alternative names for terms currently in use to facilitate the interpretation of historical data.

Some geoscience data acquisition projects are too expensive to repeat, such as drilling deep boreholes to the depth of many kilometres costing tens of millions of pounds. These data must be managed particularly well in the long-term for future users to get the benefit of previous researchers having made the investment to capture the data in the first place. Seismic data originating from earthquakes is another example of unique and unrepeatable data which is useful for e.g. building hazard models.

Geospatial data is defined as 'conveying information about the Earth, the location of specific features, and attributes and properties of those geo-located features' (Library of Congress, 2014). They are data represented in various geographic coordinate systems, and to use the data it is essential to know which system was used to create them. Not all geospatial data is geoscientific; it may also refer to data used in town planning e.g. street information and post codes.

Geographic Information Systems (GIS) is often used as an alternative term for geospatial data, but strictly speaking it refers to geographic vector or raster data stored as layers and used in mapping. GIS data may also include remote sensing and geo-referenced satellite imagery data. The main GIS used at BGS is the proprietary Esri software, which is considered by staff to be an industry *de facto* standard and as such fairly stable and secure in terms of continuity. GIS data are combined and integrated to create maps and derived datasets, and therefore having accurate and sufficient metadata, such as the scale of the data, is a key attribute to their trustworthiness. In product development, it is important to maintain previous

versions of the master datasets which may get updated, in case of possible litigation action by third parties, should issues arise due to them using inaccurate out-of-date version of data.

Key preservation challenges include 'data versioning, file size, proprietary data formats, copyright, and the complexity of file formats' (Sweetkind-Singer *et al.*, 2006). The DPC Technology Watch Report, stating 'geospatial data inherits the preservation challenges inherent to all digital information', points out additional risks including the variety of data structures and the granularity at which the data are processed (McGarva *et al*, 2009). The National Digital Stewardship Alliance report adds to the mix frequently changing data and issues with scale and resolution of the datasets (Morris, 2013).

The US National Research Council (NRC) report 'Geoscience Data and Collections: National Resources in Peril' (NRC, 2002) discussed mainly physical data, but its recommendations apply equally on digital data (prioritisation of data difficult or impossible to replace; funding the collection of information about the data). Its criteria of inaccessible data also pertain: data thought lost, residing elsewhere, in proprietary formats, or not properly indexed and curated.

The US Geological Survey (USGS) has provided preservation guidelines for digital scientific data containing descriptions 'of different levels of increasing assurance that digital data will be preserved' (USGS, 2014). Lessons learned from the development of a digital geospatial archive include collecting data on behalf of organisations without a mandate to preserve them; not knowing future priorities and uses of data; the difference between storing data and preserving it; and preserving only the most important data (Erwin *et al*., 2009).

BGS has contributed to a more recent strand of geoscience data preservation, the SCIDIP-ES, a European Commission 7th Framework earth science project, as a geoscience expert. One of the outputs was an online survey of long-term data preservation practices (EC, 2012). A preservation toolkit is available on the project website, and NERC holds a copy at the Science and Technology Facilities Council (STFC) as a result of BGS's and STFC's joint contribution to the project.

The use of metadata and standards for archiving geospatial data (Hoebelheinrich and Banning, 2008) is vital because the multiplicity of data types and the variety of GIS (geospatial information systems), used to manipulate the data and possibly employing different coordinate systems, mean that combining data from various sources can be problematical depending upon the number of translation/projection stages.

The EU INSPIRE (Infrastructure for Spatial Information in the European Community) Directive came into force in 2007 and will be fully implemented by 2019. It is the main European framework for sharing spatial data across public sector organisations and applies on all environmental data. In the UK it is implemented via a pan-government UK Location initiative, in which BGS has participated by contributing towards the development of Geoscience Markup Language (GeoSciML), a data transfer standard for geological data. The Directive harmonises the vocabulary and aims to provide discovery metadata and increase interoperability of spatial datasets, so it will have a strong impact on preservation planning once implemented. BGS has made its own discovery metadata service available on its website (BGS, 2016a) and complies with the UK GEo-spatial Metadata INteroperability initiative (GEMINI2), the UK Government specification for spatial metadata.

The Open Geospatial Consortium (OGC) provides open standards aimed at the global geospatial data community. ‘A goal of open standards is to ensure that "interoperability" (the ability to integrate datasets and related services of different types and from different sources) will minimize such costs and problems.’ (OGC *et al.*, 2015).

Findings

In this part of the paper, the first section discusses findings from the corporate documentation. The subsequent sections describe the outcomes from the stakeholder survey and staff interviews.

Organisational framework and resources

The corporate vision of BGS is ‘to be a global geological survey, working with new technology and data to understand and predict the geological processes that matter to people’s lives and livelihoods’ (BGS, 2014). The strategy points to the national

geological database as one of the core strengths of the organisation and states it 'will play a critical role in BGS in the next decade' (BGS, 2014).

The documentary analysis suggested a group of aspects which have an impact on research data management:

1) priority science areas
2) priority data and datasets
3) stakeholders, partnerships, user communities
4) corporate aims and drivers, change drivers
5) data management drivers and requirements
6) long-term outlook, data reuse, digital continuity
7) challenges and opportunities

The key science areas are Sustainable Natural Resources; Environmental or Climate Change; and Environmental Hazards. A fourth area, entitled Discovery Science, is a funding stream supporting 'excellent environmental research that is driven by curiosity rather than by NERC's wider strategic priorities' (BGS, 2016b). As some research projects and programmes are still in development, related corporate datasets may not be fully available to wider public yet.

The BGS has legislative obligations to manage some types of data, e.g. borehole data collected under the Petroleum Operations Notice 9 (PON9), or the Mining Industry Act of 1926. It accepts voluntary donations of other earth science data types and makes data available under the Public Records Act (PRA) 1958/1967, the Freedom of Information Act (FOIA) 2000, and the Environmental Information Regulations (EIR) 2004 (Bowie, 2010).

Corporate drivers affecting data preservation are diverse: the need to create efficiency savings; the drive for openness of publicly funded research data; the need to address societal and scientific challenges; increased collaboration between scientific disciplines; the need to update the skills and technologies employed by the organisation; the financial drivers to innovate and to look for new opportunities to contribute towards economic growth.

With fewer resources to maintain datasets, they risk eventually becoming unusable and not being further developed into data products. If data are not accessible, a loss of potential new business and product development, and of investment and

credibility in favour of other companies, may ensue. Building services using funds from various sources poses a complex task.

Scientists have little time available to condition data for archiving after project closure, as they are expected to quickly move on to the next project. Reducing central funds tensioned against strategic/science needs means limited funds are available to maintain existing datasets, so they may not be as well looked after as they could, reflecting one of the modern challenges of data and information management. However, the funders expect to receive high-quality data as a return for their investment, so a culture change is required of project managers to include the required resources as part of their project lifecycle and management.

The NGDC is interested in process automation for metadata creation, appraisal and ingestion, such as the tools and functions analysed by Ruusalepp and Dobrova (2012). Using tools created by the SCIDIP-ES project would support the NGDC in building its capabilities with minimal investment. The SCIDIP-ES Interactive Platform offers a range of services including Preservation Strategy and Certification Toolkits, whereas DRAMBORA provides a framework for assessing digital repository risk in eight categories, including acquisition and ingest, and metadata management. In addition, The National Archives provide a Risk Assessment Handbook and related tools on their website (TNA, 2011).

Development of data management

There has been a clear move from analogue to digital data at the NGDC over the last 20 years leading to considerable developments in data management. The introduction of the OGL has encouraged BGS to rethink the way it provides access to data and to consider data quality issues, whilst keeping in mind the legislation giving the tax-payers more access to public sector information.

The changes within BGS over the last two decades include a strong centralisation, standardisation, and deduplication of data. BGS integrated individual datasets into one Oracle database (Geoscience Integrated Database System or GeoIDS (Baker and Giles, 2000) and created a discovery metadata schema (Digital Geoscience Spatial Model or DGSM (Smith, 2000)) to support data discovery. Corporate scientific vocabularies and dictionaries completed the creation of BGS integrated data model. This improved the consistency of data but required enhancements in

pre-ingestion and ingestion processes and more knowledge from data centre staff, used to handling a more limited range of data types.

Being the custodian of over 180 years' worth of data makes decision-making about data challenging. Converting large amounts of data to new formats is a labour-intensive, expensive, and often a highly-skilled task. Data may need converting to modern languages before reuse, with a risk of losing original attributes or nuances in interpretation in the process. The organisation may no longer hold software to read all legacy formats, and the annual cost of available licences may be too prohibitive to justify their use.

It may be possible to use crowdsourcing to improve the availability of legacy data e.g. in digitising field notebooks and field slips, as these contain historical information that interests people and may encourage them to contribute. A data amnesty campaign might help capture data from staff who could provide metadata and descriptive information. Insufficiently documented data, with no reuse rights, should consequently be disposed of.

File format obsolescence and technological changes may lead to the loss of access to data. Old media such as DAT (digital audio tape) tapes and floppy disks have become unreadable, and recovery is a specialised task and not guaranteed. Deploying a technology watch and promoting preferred, open, and described formats would help mitigate against obsolescence.

As for creating full and accurate metadata, its importance and value cannot be overstated. Insufficient metadata hampers the discovery, interpretation, and preservation of data, decreases their value and threatens to make them unusable, and insufficient description may ultimately lead to false scientific conclusions.

The BGS discovery metadata, a mature and interoperable schema, provides a powerful tool for data discovery, preservation, interpretation, and validation across different levels (NERC, Data.gov.uk, EC INSPIRE Geoportal).Work is currently underway to build on this by developing a preservation metadata extension containing elements based on the PREMIS (PREservation Metadata: Implementation Strategies) Data Dictionary (Library of Congress, 2016a) and by configuring it to include key geoscience data preservation requirements.

External stakeholder requirements

BGS data is utilised by the national and local government, academia, industry and commerce. Amalgamating the contradictory requirements of all stakeholders needs to be considered in the business and strategic planning of the organisation.

The small snapshot stakeholder survey (38 respondents, 27% response rate) explored the NGDC user requirements. The results reflected the access statistics with core borehole data being accessed by most. Groundwater and geohazard datasets were used by several participants, indicating that apart from general geoscience, BGS data are used to investigate environmental geohazards. This aligns with and supports the strategic priorities in the list of current key science areas.

The main access to data was via the website, although data are also licensed or shared within collaborative projects. Data were also found using the Discovery Metadata service or by contacting BGS staff directly. 40% have used the data over a 10-year-period, and 80% of users describe the data quality as good, as needed by their own internal systems. Over 70% use raw data, and 25% transform them into other formats including .las (Log ASCII Standard), Oracle spatial, GIS shapefiles, or AGS (Association of Geotechnical & Geoenvironmental Specialists Standard) data.

The main uses of NGDC data are business re-use, decision- or policymaking, research, and science projects. The main benefit of the NGDC is perceived to be the provision of a centralised access point to geoscience data, with reliable long-term availability of data and the provision of an opportunity for data reuse being other advantages. All participants believe the NGDC is a trustworthy repository.

Suggestions to improve long-term data usability range from providing richer extended metadata to complying with industry standards. Standardised web access and data downloads, and data management and preservation training for users, are also suggested. The main concerns highlighted are file format obsolescence and the lack of required preservation skills and resources, followed by the lack of metadata, the creation of digital silos, and increasing volumes of digital data.

The data centre and research data management

The repository requirements include the need to minimise the cost of ingestion and to prioritise the data to be made immediately available, and the creation of sufficient

metadata which complies with industry standards and facilitates data discovery and reuse. The NGDC has the ambition to pull together community user data both within Europe and globally, providing services to other geological surveys and aligned or commercial organisations. This requires an infrastructure resilient enough to cover the expected upturn in data volumes and sufficient resources to manage that potential increase.

The capability to answer research questions quickly, fully, and accurately using reliable and up-to-date data is essential if BGS is to maintain its reputation as a scientific and archival organisation. A possible legal liability, caused by old versions of datasets or their provenance not being available to customers, can be mitigated against by maintaining a comprehensive, up-to-date data collection, by providing full and accurate metadata record and version control for all datasets, and by using digital object identifiers (DOIs) to provide persistent links to different versions of datasets.

Clear lines of responsibility govern BGS data management: the Informatics Directorate leads on strategy and implementation; data providers ensure the data quality by selecting appropriate formats and providing sufficient metadata when depositing their data and providing quality science data; NGDC staff ingest, store, help deliver and preserve the data, and provide guidance to users. Science projects use additional measures, such as in-project quality assurance procedures and metadata reports. Information professionals educate scientists, provide guidance on data formats and documentation, and advice on file naming conventions.

Ingestion processes have been digitised and standardised resulting in the creation of an online Digital Data Deposit Application (http://transfer.bgs.ac.uk/ingestion). Updated procedures are already creating efficiency savings, and automation of metadata creation and an increased use of open formats would further facilitate data collection. The archival storage is good and the cost per unit has been decreasing. Data retrieval and discovery facilities include good external search tools. Web services providing an interface for users to copy raw data to other formats or to embed it within their own systems are increasingly popular.

The management of digital data faces challenges, such as predicting future societal and scientific requirements, rapid technological changes leading to increase in real-

time monitoring and sensor data, and the dependence on new technologies vis-à-vis creation, management, storage and retrieval of the data. Corti *et al*. list some advantages specialist data centres offer here, including having appropriate access controls in place and acting as a point of contact between the data and its users (Corti *et al*., 2014). The NGDC offers limited embargo periods for data creators, metadata-only publication for commercial data, and data licences for businesses who wish to use data in their own products. For this category, metadata may also be hidden.

The long validity of geoscience data means permanent retention is often expected. Adding descriptive and contextual metadata on datasets to ensure their longevity becomes essential, but convincing scientists to provide them remains challenging. With growing volumes of data coming in (175% increase between January 2014 and February 2017), more automated processes, infrastructure and resources are required to manage them.

Drivers influencing data management are the cost of data creation and collection, the support of data publication agenda, the need to create operational efficiencies and to maximise the cost-benefit ratio. As a national repository, the NGDC has a unique selling point in its position of providing open and centralised access to the nation's geoscience data while securing its validity and reliability. It benefits from its connection with the internationally recognised BGS brand, associated with its scientists' ability to answer complex science questions using the available data and knowledge.

The monetary value of data is difficult to measure, although it has been attempted (BGS, 2003). Geologists add value to BGS data e.g. by creating derived datasets and national good science, available to all under the Open Government Licence (OGL) and accessible via the OpenGeoScience website. Links to the underlying data are increasingly a requirement when submitting research papers for publication, and the NGDC provides the DOI service for this purpose.

As data users are increasingly taking the easiest option in sourcing data for their research, the NGDC will need to increase its visibility. Increasing demands from research funders for data to be stored in trustworthy repositories have led the NGDC

to capitalise on its current reputation through collecting evidence for the Data Seal of Approval repository certification.

The corporate information culture has evolved considerably. Initially data management procedures were enforced through the appraisal procedure, but a shift towards corporate data holdings and a general acknowledgement of the value of metadata has occurred. More recently, attitudes towards data sharing and reuse and data collection processes have also changed.

The NGDC needs to continue to enhance its present practices and implement best practice and comply with the latest standards and legislation; create robust metadata; appraise and ingest data efficiently, creating added value; minimise the costs and maximise the long-term usability, accessibility and availability of information. All earth science data must be backed up securely, stored and delivered to a diverse group of users.

If data are to be reused and retained for longer periods, preservation planning becomes critical. If unique or irreplaceable geoscience data are stored in the long-term, a technology watch needs to be in place monitoring risks and changes and keeping research data safe. The value of data can only be realised if the data remain usable and accessible to people who can achieve that value and create data products. All future uses of data are not known today, and even ‘useless’ data may become important.

The pressure on resources are encouraging BGS to think more strategically. It needs to consider its data management strategy and the skills required to implement it. The data ingested and created need to be fit for multiple purposes and of sufficiently high quality to justify the use of public funds on their preservation. Maintaining data consistency faced with growing data volumes and heterogeneity of data types and formats needs to be tackled to make the best use of the resources.

On a strategic level, the critical role of data management at BGS has been acknowledged by the senior management, who aim at a wider reach of BGS data, forming global partnerships. This necessitates data interoperability and the use of common standards to enhance the longevity and applicability of the data. For this purpose, even scientists agree that basic data management must be seen as critical. This requires high-level corporate support, raising awareness of digital preservation,

and providing appropriate guidance. As for prioritising data types for preservation, those more expensive to create and collect or completely unique are a priority. Preferred and open formats and consistent file naming should be considered on the outset to enhance data longevity for born-digital data.

It is impossible to recollect or re-create long time-series datasets and one-off observations, and some data are too expensive to re-create, e.g. deep boreholes costing tens of millions of pounds to drill. Historical datasets are used to monitor long-term trends and to inform Government public safety decisions in areas such as groundwater flooding and radioactive waste storage. Knowledge on archived datasets may be lost before they are reused. If data are needed urgently, as is the case in emergencies, properly archived and preserved data are available for decision-making immediately.

The impact and benefit of using corporate resources on making data more widely available needs to be demonstrated. The cost of the investment must be offset by a future financial or scientific gain as in any other business, possibly even more so when using tax-payers' money. Recreating lost data bears a huge cost and is not always feasible. Using scientists' time to search for data is not a cost-effective option. Decreased funding may lead to weaker quality data and science; however, this very much depends on the strategy basis of using supplied funds. Lack of resources for data management and digital preservation can eventually diminish the value of data as a corporate asset. Raising the awareness about digital continuity to ensure sufficient resources are available may go some way when deciding on corporate priorities and budgets.

Data centres have a key role in providing a facility for accessing and reusing scientific data. The NGDC is well placed within the earth science community to provide this service – as a national repository it has little UK competition and can guarantee the quality, validity, and provenance of the geodata in its care.

Conclusions

The users of NGDC geoscience data come from a variety of backgrounds and have varying requirements. This has wider implications on the need to maintain accessibility and usability of borehole, groundwater, and geohazards data, which are often used for ten years or longer by the same stakeholders. The interpretation of

what digital preservation means depends on the role and experience of the users. People may know and understand the data very well without fully grasping what actions they could take to improve its long-term usability. It is essential first to develop a common understanding of the concept across the user community, to agree on key messages, and to communicate them to all stakeholders in order to develop the digital preservation culture.

If the aim is to provide high-quality data and efficiency savings with fewer resources, it will be critical to integrate research activities better with data management procedures. This will ensure some progress is made in enhancing the long-term usability of the UK geoscience data archive for future generations and in maintaining the BGS's ability to answer research questions not yet known. The geoscience data archive is a national asset, which support, BGS in its core mission of providing objective and authoritative geoscientific data, expert services, and impartial advice to all sectors of the UK society. Merely having and storing the data is not sufficient for this aim but a long-term strategy and active, working preservation procedures are required to maintain the value and continuity of the archive.

The development of the NGDC data management procedures and discovery metadata schema, as well as the updated digital ingestion processes, have enhanced the digital continuity of data. However, by making some preservation activities mandatory would further strengthen their reusability. The inspection of current issues with legacy data suggests that by documenting data more extensively at the ingestion stage some of these issues could be avoided or at least be mitigated against – this is a key component of the data deposit portal.

A starting point to implement preservation methods would be to survey all different data types and risks relating to their long-term availability. This could be followed by the creation of a technology watch and a priority list to monitor risks for mitigation and response as and if they are realised. The experience of staff shows that such a function, implemented earlier, would have alerted them to migrating at-risk file formats before them becoming obsolete.

The stakeholder survey indicated that the user priorities with regard to BGS data are well aligned with the current key science areas and could be used to inform the preservation priorities to some extent. It also showed that in general the users

consider the NGDC to be a trusted geoscience data repository, a characteristic which should be further enriched by the acquisition of the DSA accreditation and a widening use of the DOIs.

Training and raising awareness amongst all the stakeholders is another key aspect of digital preservation and will be addressed in the coming months and years. This point is validated by a comment from a survey participant:

> *'The biggest threat to the preservation of digital research data is the lack of understanding by those outside the field'.*

In the digital era, there should be nobody left 'outside the field'. The NGDC has always had an interest in the longevity of its data, but in this digital age there are new challenges, risks and opportunities for the data archive which spans over 180 years. This research looked at the current organisational and data management framework of the NGDC, and the findings indicate that the groundwork to manage digital data to a high standard has been done over the last 20 years. Now is the time to take a long-term outlook for data reuse and digital continuity and to develop a functional preservation work plan to make the geoscience data work for the good of the nation.

Slightly paraphrasing William Kilbride, the Director of DPC, digital preservation doesn't do itself (Kilbride, 2013). It requires a broad range of skills, a strategic outlook, far-sightedness, long-term planning, financial resources, and above all, perseverance. To conclude in the words of one of the interviewees, in an organisation such as BGS, the custodian of data assets covering almost two centuries:

> *'We will be judged in the future by how good we tackled data preservation and continuity, it's not something that we can ignore. It will become another pillar that underpins our well-founded data centre.'*

Notes

The article is based on the author's MSc dissertation 'Exploring digital preservation requirements: a case study from the National Geoscience Data Centre (NGDC)' (2016) for Northumbria University.

Acknowledgements

The MSc research was supported by the British Geological Survey.

References

Baker, G. and Giles, J. (2000) 'BGS Geoscience Integrated Database System: A repository for corporate data'. In: *Earthwise,* 16, pp.12-13. Available at: http://www.bgs.ac.uk/discoveringGeology/newsAndEvents/earthwise/downloadSearch.cfc?method=viewIssues (Accessed: 10 August 2016)

Ball, A. (2010) *Preservation and Curation in Institutional Repositories.* Digital Curation Centre Technology Watch Report. Available at: http://www.dpconline.org/advice/technology-watch-reports (Accessed: 7 February 2016)

Beagrie N., Lavoie, B. and Woollard, M. (2010) *Keeping Research Data Safe 2*. Available at: http://www.beagrie.com/publications/ (Accessed: 7 February 2016)

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hofman, H. (2009) 'Systematic planning for digital preservation: evaluating potential strategies and building preservation plans', *International Journal of Digital Libraries* 10, pp.133-157.

Bowie, R. (2010) *BGS Geological Survey – the Legislative Framework*. Available at: http://www.bgs.ac.uk/services/NGDC/records/policy.html

British Geological Survey (BGS) (2003) *The economic benefit of BGS*. Available at: http://www.bgs.ac.uk/about/economicbenefits.html (Accessed: 9 August 2016)

British Geological Survey (BGS) (2014) *Gateway to the Earth: Science for the next decade.* Available at: http://www.bgs.ac.uk/about/strategy.html (Accessed: 26 January 2016)

British Geological Survey (BGS) (2016a) *OGC Catalogue Service for the Web (CSW).* Available at: http://www.bgs.ac.uk/data/services/csw.html (Accessed: 2 June 2016)

British Geological Survey (BGS) (2016b) *Summary of BGS highlights for November to January 2016*. Internal BGS report. Unpublished.

British Standards Institution (BSI) (2009) *BS ISO 15836 Information and documentation – The Dublin Core Metadata Element Set*. London: British Standards Institution.

British Standards Institution (BSI) (2012a) *BSI ISO 14721:2012 Space data and information transfer systems. Open archival information system (OAIS) Reference model.* London: British Standards Institution.

British Standards Institution (BSI) (2012b) *BS ISO 16363:2012. Space data and information transfer systems. Audit and certification of trustworthy digital repositories*. London: British Standards Institution.

Brown A. (2013) *Practical Digital Preservation: A How-to Guide for Organisations of Any Size.* London: Facet.

Campbell, J. (2015) 'Access to scientific data in the 21st Century: rationale and illustrative usage rights review', *Data Science Journal*, Vol.13, pp.203-230.

Corti, L., Van den Eynden, V., Bishop, L. and Woollard, M. (2014) *Managing and sharing research data: A guide to good practice*. London: Sage.

Daniels, M., Faniel, I., Fear, K. and Yakel, E. (2012) 'Managing fixity and fluidity in data repositories', *Proceedings of the 2012 iConference,* Toronto, ON, Canada, February 7-10, 2012. New York: ACM, pp. 279-286. doi: 10.1145/2132176.2132212.

Denscombe, M. (2008) 'Communities of practice: a research paradigm for the mixed methods approach', *Journal of Mixed Methods Research* 2(3), pp.270-283.

Denzin, N.K. (2012) 'Triangulation 2.0', *Journal of Mixed Methods Research* 6(2), pp.80-88.

Digital Curation Centre (DCC) (2016). *How-to Guides and Checklists.* Available at: http://www.dcc.ac.uk/resources/how-guides (Accessed: 16 August 2016)

Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE) (2015) *Digital Repository Audit Method Based on Risk Assessment* (*DRAMBORA*). Available at: http://www.repositoryaudit.eu/ (Accessed: 8 August 2016)

Digital Preservation Coalition (DPC) (2016) *Technology Watch Reports.* Available at: http://www.dpconline.org/advice/technology-watch-reports (Accessed: 16 August 2016)

Erwin, T., Sweet-Kinder, J. and Larsgaard, M.L. (2009) 'The National Geospatial Digital Archives – Collection Development: Lessons Learned', *Library Trends* 57(3), pp.490-515.

European Commission (EC) (2012) *SCIDIP-ES D15.1 Report on the survey of technologies, policies, metadata, semantics and ontologies.* Available at: http://www.scidip-es.eu/wp-content/uploads/sites/6/2014/11/SCIDIP-ES-DEL-WP15-D15-1.pdf (Accessed: 16 August 2016)

European Commission (EC) (2016) *European Framework for Audit and Certification of Digital Repositories.* Available at: http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html (Accessed: 3 June 2016)

Farquhar, A. and Hockx-Yu, H. (2007) 'Planets: Integrated services for digital preservation', *International Journal of Digital Curation* 2(2), pp.88-99.

Garrett, J. and Waters, D. (1996) *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. The Final Report of the Commission on Preservation and Access and the Research Libraries Group. Available at: http://www.clir.org/pubs/reports/pub63watersgarrett.pdf (Accessed: 9 February 2016)

Hockx-Yu, H. (2006) 'Digital preservation in the context of institutional repositories', *Program* 40(3), pp.232-243.

Hoebelheinrich, N., and Banning, J. (2008) *An investigation into metadata for long-lived geospatial data formats*. National Geospatial Digital Archive project, available at: http://www.ngda.org/reports/InvestigateGeoDataFinal_v2.pdf (Accessed: 28 April 2016)

Information Governance Initiative (IGI) (2016) *The governance of long-term digital information: IGI 2016 benchmark.* Available at: http://iginitiative.com/reports/ (Accessed: 6 June 2016)

Jones, M. (2006) 'The Digital Preservation Coalition', *The Serials Librarian* 49(3), pp. 95-104.

Jones, S. (2014) 'The range and components of RDM infrastructure and services', in Pryor, G., Jones, S. and Whyte, A. (Eds.) *Delivering research data management services: Fundamentals of good practice.* London: Facet, pp.89-114.

Kilbride, W. (2013) 'Getting Started in Digital Preservation: what do I need to know?' [PowerPoint presentation.] *Getting Started in Digital Preservation: Extra Stop in London*. Available at: http://www.dpconline.org/events/previous-events/1148-getting-started-in-digital-preservation-extra-stop-in-london (Accessed: 9 August 2016)

Lavoie, B. (2014) *The Open Archival Information System (OAIS) Reference Model: Introductory Guide* (2nd edn.). DPC Technology Watch Report. Available at: http://www.dpconline.org/advice/technology-watch-reports (Accessed: 7 February 2016)

Lavoie, B. and Gartner, R. (2013) *Preservation Metadata* (2nd edn.). DPC Technology Watch Report. Available at: http://www.dpconline.org/advice/technology-watch-reports (Accessed: 28 February 2016)

Library of Congress (2014) *Introduction to Geospatial Resources and Formats.* Available at: http://www.digitalpreservation.gov/formats/content/gis_intro.shtml (Accessed: 24 May 2016)

Library of Congress (2016a) PREMIS Data Dictionary for Preservation Metadata, Version 3.0. Available at: http://www.loc.gov/standards/premis/v3/index.html (Accessed: 20 March 2017)

Library of Congress (2016b) *Recommended Formats Statement.* Available at: https://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf (Accessed: 22 August 2016)

McGarva, G., Morris, S. and Janée, G. (2009) *Preserving Geospatial Data.* DPC Technology Watch Report. Available at: http://www.dpconline.org/advice/technology-watch-reports (Accessed: 28 February 2016)

McGovern, N.Y. and McKay, A.C. (2008) 'Leveraging short-term opportunities to address long-term obligations: a perspective on institutional repositories and digital preservation programs', *Library Trends* 57(2), pp.262-279.

Morris, S. (2013) *Issues in the appraisal and selection of geospatial data. An NDSA Report.* Available at: http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_AppraisalSelection_report_final102413.pdf (Accessed: 28 February 2016)

National Geoscience Data Centre (NGDC) (2016). Available at: (http://www.bgs.ac.uk/services/ngdc) (Accessed: 12 August 2016)

National Research Council (NRC) (2002) *Geoscience data and collections: National resources in peril.* Available at: http://www.nap.edu/catalog/10348/geoscience-data-and-collections-national-resources-in-peril (Accessed: 28 February 2016)

Natural Environment Research Centre (NERC) (2010) *NERC data policy.* Available at: http://www.nerc.ac.uk/research/sites/data/policy/ (Accessed: 21 February 2016)

Open Geospatial Consortium (OGC), The International Organization for Standards (ISO) Technical Committee 211 Geographic information/Geomatics, and the International Hydrographic Organization (IHO) (2015) '*A Guide to the Role of Standards in Geospatial Information Management*'. Available at: http://ggim.un.org/docs/Standards%20Guide%20for%20UNGGIM%20-%20Final.pdf (Accessed: 1st June 2016)

O'Reilly, K. (2012) *Ethnographic methods*. (2nd edn.) London: Routledge.

Park, J.-R. (2009) 'Metadata quality in digital repositories: a survey of the current state of the art', *Cataloging & Classification Quarterly* 47(3-4), pp.213-228. DOI: 10.1080/01639370902737240

Park, J.-R. and Brenza, A. (2015) 'Evaluation of semi-automatic metadata generation tools', *Information Technology and Libraries*, September 2015, pp.22-42.

Pickard, A.J. (2013) *Research Methods in Information* (2nd edn). London: Facet Publishing.

Pryor, G., Jones, S. and Whyte, A. (Eds.) (2014) *Delivering research data management services: Fundamentals of good practice.* London: Facet.

Pryor, G. (2014) 'Options and approaches to RDM service provision', in Pryor, G., Jones, S. and Whyte, A. (eds.) *Delivering research data management services: Fundamentals of good practice.* London: Facet, pp.21-40.

Research Councils UK (RCUK) (2015a) *Guidance on best practice in the management of research data.* Available at: http://www.rcuk.ac.uk/RCUK-prod/assets/documents/documents/RCUKCommonPrinciplesonDataPolicy.pdf (Accessed: 17 February 2016)

Research Councils UK (RCUK) (2015b) *RCUK Common Principles on Data Policy*. Available at: http://www.rcuk.ac.uk/research/datapolicy/ (Accessed: 17 February 2016)

Research Data Alliance (RDA) (2016) *DSA–WDS Partnership Working Group Catalogue of Common Requirements.* Available at: https://rd-alliance.org/system/files/DSA%E2%80%93WDS%20Catalogue%20of%20Common%20Requirements%20V2.2.pdf (Accessed: 12 May 2016)

Ross, S. (2012) 'Digital preservation, archival science and methodological foundations for digital libraries', *New Review of Information Networking* 17(1), pp.43-68.

Ruusalepp, R. and Dobreva, M. (2012) *Digital Preservation Services: State of the Art Analysis*. Available at: www.dc-net.org/getFile.php?id=467 (Accessed: 8 February 2016)

Smith, I. (2000) 'The Digital Geoscience Spatial Model: The shape of the BGS of the future'. *Earthwise,* 15, p.8. Available at: http://www.bgs.ac.uk/discoveringGeology/newsAndEvents/earthwise/downloadSearch.cfc?method=viewIssues (Accessed: 10 August 2016)

Sweetkind-Singer, J., Laarsgaard, M.L. and Erwin, T. (2006) 'Digital preservation of geospatial data', *Library Trends* 55(2), pp. 304-214.

The National Archives (TNA) (2011) *Risk Assessment Handbook.* Available at: http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/risk-assessment/ (Accessed: 10 August 2016)

United States Geological Survey (USGS) (2014) '*USGS Guidelines for the Preservation of Digital Scientific Data'*. Available at: http://www2.usgs.gov/datamanagement/documents/USGS%20Guidelines%20for%2

0the%20Preservation%20of%20Digital%20Scientific%20Data%20Final.pdf
(Accessed: 4 May 2016)

Van den Eynden, V., Corti, L., Woollard, M., Bishop, L. and Horton, L. (2011) *Managing and sharing data: Best practice for researchers.* UK Data Archive. Available at: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf (Accessed: 7 June 2016)

Whyte, A. (2015) '*Where to keep research data: DCC checklist for evaluating data repositories' v.1.1* Edinburgh: Digital Curation Centre. Available on: http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data (Accessed: 4 May 2016)

About the author

Jaana Pinnick works in geoscience data management and digital preservation at the British Geological Survey. She is also the Chair of the Midlands Branch of the Information and Records Management Society (IRMS). She can be contacted on jpak@bgs.ac.uk.