BUILDING RESILIENCE AT THE NATIONAL GEOSCIENCE DATA CENTER

Enhancing Digital Data Continuity Through Research Data Management Training

Jaana Pinnick

National Geoscience Data Centre
United Kingdom
jpak@bgs.ac.uk
ORCID 0000-0002-7718-5133

Abstract – The National Geoscience Data Center (NGDC) is the designated repository for the Natural Environment Research Council (NERC) grant-funded Earth science data and holds the CoreTrustSeal certification. The NGDC is hosted by the British Geological Survey (BGS), which co-funds post-graduate research students through the BGS University Funding Initiative (BUFI) program.

This paper describes the research data management training (RDM) course developed and delivered by the NGDC to help instill good data management practices in our students from early on, and to strengthen the long-term quality of research data they generate and deposit with the NGDC. It also looks at how RDM training fits into the wider context of the NGDC modular digital preservation program, currently under development.

This paper is aimed at data repository managers and research data managers who provide user training in data management best practice and digital preservation. It is also suitable for postgraduate students interested in digital continuity and preservation of their research data.

Keywords – Earth Science, Geoscience, Collaboration, Research Data Management (RDM), Post-Graduate Training.

Conference Topics – Collaboration: a Necessity, an Opportunity, or a Luxury?; Building Capacity, Capability and Community.

I. INTRODUCTION TO BGS AND THE NGDC

The National Geoscience Data Center (NGDC) is the designated repository for the Natural Environment Research Council (NERC) grant-funded Earth science research data and the guardian for many commercially funded datasets. The NGDC is hosted by the British Geological Survey (BGS), and as one of the NERC Environmental Data Centers [1], it is committed to supporting long-term environmental data management to enable continuing access to these research assets. The NERC data policy [2] requires that all environmental data of long-term value generated through NERC-funded activities must be submitted to the designated repository for long-term management and dissemination.

BGS has existed for almost 185 years and continues to hold some of its original early 19th Century notebooks and maps. However, in this digital age, we recognize that the persistence of digital data is much more precarious than that of these hardcopy materials. It is therefore in our interest that all Earth scientists collecting and generating digital data, which we may end up preserving in the long-term, have a solid understanding of data management processes and best practice.

II. BGS AND PHD STUDENTSHIPS

BGS co-funds postgraduate Earth and geoscience research students through the BGS University Funding Initiative (BUFI) program [3]. At least one of a student's supervisors is a BGS scientist, and the students often use the site science facilities for their PhD research projects. This may include access and reuse of data held by NGDC, and exposes them to the operations of the data center as part of their research activities.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

A. Research Management Training

BGS identified the need to bridge a skills gap in students' Research Data Management (RDM) training and application and decided to develop tailored guidance to support the data management element of their geoscience research projects. During their time at BGS, the students now participate in a one-day in-house RDM workshop, which has been run for BUFI students and another NERC student-funding program, Central England NERC Training Alliance (CENTA), since 2016. To date, a total of 120 PhD students have participated in the BGS workshop (March 2019). We have also been involved in discussions about delivering this training course to a third national geoscience PhD program in the future.

B. Aims of the RDM Workshop

The aim of the workshop is two-fold. Firstly, it aims to build up data management skills and capability amongst the students, and secondly, strengthens the quality of Earth science data that eventually is deposited at the NGDC. The repository is also expected to keep the data for at least ten years after the publication of the research that it underpins. However, the validity of Earth and geoscience data is usually much longer than this [4], and as a Place of Deposit under The UK National Archives, the NGDC is committed to looking after certain data in its care in perpetuity. The informational content of our research data underpins research long into the future because geoscience is an interpretative discipline. As such, the data do not often become obsolete, but the interpretation may change or be superseded as new methodologies and technologies become available.

C. Benefits of the Workshop

As part of the collaboration with universities, BGS shares the latest research practices and methodologies with the earth science students, and acts as a link between cutting-edge Earth science research and the extensive data held by the repository. We see the whole UK geoscience research community benefiting from our collaboration with the students in the long run, cultivating relationships, developing mutual trust, exchanging knowledge, and providing professional guidance to them. This collaboration raises awareness of the long-term impact of RDM best practices and its role in the digital continuity and preservation of Earth science data within the research community, including the students and staff at universities, and also those at BGS and NGDC.

III. RDM Workshop Content

The RDM course informs students about the NERC data policy and advises them about the obligations and rights of NERC-funded researchers, such as the requirement to offer a copy of their datasets to a NERC Environmental Data Center, and the possibility to access and reuse existing data. The course includes the basic RDM concepts and gives students the practical skills necessary for them to manage their data in a way that both benefits them and supports the aims of their assigned NERC data center. In the following sections we describe the content of the one-day workshop and what the students can expect to take away from it to develop their individual research data management practices.

After a brief introduction to NERC data policy and its Environmental Data Centers, the *Managing Research Data and Metadata* module demonstrates the value of adding structure to the data, and shows how making data interoperable and discoverable through a number of data portals enables its reuse in the future with tools and by users yet unknown. It also includes guidance on developing robust file naming and versioning strategies, organizing data using a clear file and directory structure, and selecting appropriate file formats before depositing the data in a long-term repository.

The module *Data Management Planning – Completing a DMP* is always popular with the students, as this is often the first time that they have been asked to write a data management plan. By completing a practical exercise students are able to review their research project from beginning to end and to consider the impact of their in-project data management activities on the long-term storage and continuity of their data. It also requires them to consider different types of research assets they are generating, such as new digital data, websites, models, code, software, and so on. These all have their unique requirements that the students have often not thought about at this stage. The session immediately equips the students with data management plans for their own research and with new skills to employ in their research careers.

During the session on *Ensuring Data Quality and Preparing Data for Depositing* the students learn about the role each link in the PhD data management chain plays in data quality, whether they are the researcher, the data repository, or the project supervisor. We use real life examples from the data center and BGS to demonstrate cases of bad practice, to provide best practice guidance, and to show how to use repository resources to standardize data, to check data documentation and errors, and select what to keep.

In the module *Data Storage and Security and Long-Term Preservation*, we ask the students to consider their data storage and security requirements. We then talk about the main causes of data loss and about how to mitigate them. We also discuss the difference between in-project and long-term sharing and storage of data, why backup is not the same as preservation, and why they should care about preserving their data. In addition, they learn useful practical tips for future-proofing their data, such as using Open Office formats, and creating preservation-ready spreadsheets.

In 2018, we added a module on *Open Science and FAIR Data*, which was well received by the students. According to the feedback, students had not discussed this topic at their universities, and agreed with the benefits of Open Science for researchers, including the need for data to be Findable, Accessible, Interoperable and Reusable (FAIR) [5]. Using persistent identifiers for different elements of their research was also of general interest, with most the students registering for an ORCID unique identifier for researchers before the end of the session.

In the final session, *Data Retrieval and Reuse*, the students investigate research data repositories and data centers, learn to evaluate their trustworthiness, and search for data they may be able to reuse in their PhD projects. This session brings together many of the topics touched upon earlier in the day, including naming and organizing data in order to make it understandable and recoverable, and making it accessible by sharing it and depositing it at an appropriate long-term repository.

IV. The Benchmarking of RDM Training

A. Developing Feedback Process

To assess the quality of our RDM workshop, in 2017 we participated in an initiative led by Cambridge University to develop shared benchmarking metrics for RDM training courses delivered across the participating universities and research organizations [6]. The aim of the exercise was to agree on a minimal set of questions as benchmarking criteria to identify what works best for RDM training, and on which questions should be mandatory or optional. The participating members agreed to use six mandatory questions and a five point rating scale where (1) is the worst rating and (5) is the best [7]. We have used this feedback format on four of our workshops now and found it very useful when developing the workshop content and delivery further, establishing which modules and elements are the most useful for the students, and identifying any major gaps in the content.

The participating students often also work in wider NERC-funded research programs. Providing them with the best practices for long-term data management reaches a larger number of our end users because the students share this knowledge with their supervisors and research partners at universities, helping us disseminate the funder requirements for good quality data at the creation stage rather than at the point of deposit. As these practices become a staple part of their professional practice from early on, early career scientists will benefit from these skills over their entire career. This will lead to better transparency and reproducibility of their science and enhance their collaboration opportunities. Earth science discipline as a whole benefits from more robust science and data, which contributes to the development of the national data collections.

B. Student Feedback Received

The feedback received from the RDM training course since the introduction of the shared benchmarking metrics indicates that the more tailored the content of the course is to match the needs of the students, the more they benefit from it. BUFI students stated that the course met their expectations on a level of 4.2/4.2 out of 5, whereas CENTA students' rating was slightly lower at 3.2/3.8 respectively. When asked if they would recommend the workshop to their peers, BUFI students gave the course a rating of 4.0/4.3 and CENTA students 2.6/3.7. The figures show that using the feedback loop to enhance the course content and delivery has led to higher satisfaction by the students.

Areas where the course was felt to be particularly useful were learning new data management skills, completing a data management plan, considering what aspects of data management may enhance the continuity of digital data (selection of file formats, providing robust metadata and data documentation alongside the data), and learning consistent file naming and versioning strategies which enable wider data reuse in the future.

Suggestions for improvement included providing more interactive activities and discussions, and spreading the training sessions over a longer period of time. This indicates that it is a challenge to achieve the right balance between providing enough information and skills on one hand, and time for students to practice the learning on their own research projects on the other. To us, collaboration with and communication between universities, students, and their supervisors, is the key to better data management practices, and consequently, to more robust data quality in the long-term.

V. BUILDING DIGITAL PRESERVATION CAPABILITY AT BGS AND THE NGDC

Delivering the RDM training workshop is one of the contributing elements to the development of a sustainable and modular digital preservation program for BGS. In the following paragraphs, we give an overview of some of the work we have planned and undertaken following an initial investigation of the NGDC digital preservation requirements [4].

A. Strategic Framework

The BGS digital preservation policy, first introduced in 2017 [8], states that our overall approach is to develop a scalable preservation program, which will be further detailed within the internal preservation strategy (currently under development). Promoting best practice and delivering staff training were identified as key components of our preservation framework. We further evaluated some of the different implementation options available in our internal business case, which also emphasized the role of training and raising awareness of digital preservation.

As a public sector organization, we do not have a large budget to spend on commercial solutions. However, we have extensive in-house data management and developer skills to support the integration of new workflows and procedures as well as training. We therefore decided to pursue a modular solution which allows us to be flexible with our development and implementation, and started by reviewing and enhancing our existing procedures, infrastructure, and digital skills to implement our preservation framework.

We will use our existing discovery metadata schema to create a digital asset register, adding preservation and technical metadata elements to build a complete picture of our digital objects. To develop our digital preservation action plan we will use findings from stakeholder surveys, interviews, and risk assessments. The asset register will provide us crucial information required to make fact-based decisions on our preservation priorities and updates to our data management procedures and strategies.

We will conduct a digital preservation capability assessment to identify the gaps and where our resources are best employed. Implementing our top preservation priorities will be done in collaboration with both the BGS and NGDC data center staff, and with the senior management and the end users. Our digital preservation policy and a flexible strategy – organic yet controlled – will be tailored for the organization and its designated community. Our aim is to use our resources wisely, integrate new relevant digital skills into our existing workflows and practices, and focus our thinking on the long-term preservation and continuity of earth science data.

B. In-House Data Infrastructure

Our data and information infrastructure is largely built and developed in-house. Our corporate digital research data holdings, which are stored on the storage area network (SAN) and the corporate tape archive, currently exceed 1,200 TB.

The key datasets on the SAN are backed up at two other geographical locations. In addition, we hold a legacy magnetic media archive of over 5,600 items on different data cartridges, reels, tapes, CDs/DVDs, cassettes and other media. However, we do not have in-house access to the older technology or the resources needed to rescue most of these legacy data, or to enable informed decision-making on which content to migrate onto new technologies. Even with the necessary resources, we may not have sufficient contextual and rights metadata to allow appropriate reuse of these data.

To avoid this issue from reoccurring in the future we have developed a standardized ingestion and accession procedure, requiring the data depositors to submit all the necessary discovery and rights metadata to accompany their deposits, so that future users have all the information they need to be able to reuse the dataset with confidence. We will build our data preservation capabilities further by adding the function of checksum value creation at ingest and fixity checking for key datasets, and selected PREMIS metadata fields to be maintained alongside our corporate discovery metadata schema.

C. CoreTrustSeal Certification

Reviewing the capabilities of the NGDC was also part of our self-assessment for the CoreTrustSeal (CTS) certification, which we gained in January 2018 [9]. The NGDC wanted to gain the CTS certification to build stakeholder and end use confidence in the repository, and to help benchmark our processes against a recognized methodology.

As part of the CTS submission, we confirmed that all BGS and NGDC staff have access to a comprehensive learning and development program, keeping them up to date with the latest data management techniques. Including PhD students in our training offering is part of this strategy. Engaging with all of our end users from an early stage in their research project lifecycle will help us identify where we can make improvements for the

users and streamline the processes to facilitate data management workflows for them.

We are working toward further enhancing our capabilities within the CTS schema, and this will form part of our submission for continued future certification. We see the collaboration with our end users as the way forward to ensure that this certification delivers benefits for both the data center and the users. To achieve this, we will monitor their most up to date requirements and share expertise and experience with other memory and preservation organizations.

VI. CONCLUSION

Training early career scientists to manage their research data with a view to its long-term preservation accomplishes many important objectives: it raises the students' awareness of digital preservation; it builds their digital preservation capability and professional RDM skills; and it enhances the quality of data and data management skills in the whole Earth science community. To achieve this, NGDC staff must make data preservation relevant to early career scientists, and as easy as possible and automated where this is feasible. We need to communicate our aims and strategies to the next generation of researchers in a way that is useful to them and proves that RDM and digital preservation are a key part of their career progression and wider skills.

ACKNOWLEDGMENT

The author would like to thank Helen Glaves and Barbara Yarusso for their invaluable input and editorial advice.

REFERENCES

[1] Natural Environment Research Council. *Data Centres*. Available at: https://nerc.ukri.org/research/sites/data/ (Accessed: February 26, 2019).

[2] Natural Environment Research Council. *Data Policy*. Available at: https://nerc.ukri.org/research/sites/data/policy/ (Accessed: February 26, 2019).

[3] British Geological Survey. *BGS University Funding Initiative (BUFI)*. Available at https://www.bgs.ac.uk/research/bufi/home.html (Accessed: February 26, 2019).

[4] Pinnick, J. (2017) 'Exploring Digital Preservation Requirements: A Case Study from the National Geoscience Data Centre (NGDC)'. *Records Management Journal* 27(2), pp.175-191. doi: 10.1108/RMJ-04-2017-0009

[5] FORCE11. *The FAIR Data Principles*. Available at: https://www.force11.org/group/fairgroup/fairprinciples (Accessed March 02, 2019).

[6] University of Cambridge Office of Scholarly Communication. *Unlocking Research*. Available at: https://unlockingresearch-blog.lib.cam.ac.uk/?p=1723 (Accessed February 26, 2019).

[7] RDM Training Benchmarking (2017). Available at: https://osf.io/pgnse/ (Accessed: May 31, 2019)

[8] British Geological Survey. *BGS Digital Preservation Policy*. Available at: https://www.bgs.ac.uk/downloads/start.cfm?id=3173 (Accessed February 26, 2019).

[9] CoreTrustSeal. *National Geoscience Data Centre Implementation of the CoreTrustSeal*. Available at: https://www.coretrustseal.org/wp-content/uploads/2018/01/National-Geoscience-Data-Centre.pdf (Accessed: February 26, 2019).