Research Data Preservation Guidelines

2023. 5.

KIGAM Korea Institute of Geoscience and Mineral Resources
Geoscience Data Center

Table of Contents

List of Tables

1 Overview

【 Purpose 】

☐ The Presentation of research data preservation guidelines applicable to data preservation at the Geoscience Data Center of the Korea Institute of Geoscience and Mineral Resources (KIGAM).

☐ Select a durable format and follow procedures for submitting data to a repository for long-term preservation.

【 Target 】

☐ Administrator of Geoscience Data Center who want to preserve deposited research data.

【 Scope of Application 】

☐ Applies to research data generated through in-house research activities and research data donated by external organizations and individuals.

【 Application 】

☐ Matters not specified in these guidelines may be subject to the research data management guidelines of the National Research Council of Science & Technology (NST) and KIGAM Data Management Regulations.

2 Concept of Data Preservation

【 Meaning of Data Preservation 】

☐ A set of management activities taken to ensure the long-term viability and continued accessibility of research data.

☐ Long-term refers to a period of time long enough to be concerned about the loss of integrity of digital information held in a repository, including damage to storage media, changes in technology, support for old and new media and data formats, and changes in the user community.

【 Necessity of Data Preservation 】[1)]

☐ Digital data preservation should be a key aspect of any research project. Some research data is unique and cannot be replaced if destroyed or lost. However, referencing verifiable data can be enough to determine that a study is sound.

☐ Effective documentation of data.

☐ Storage media may degrade, or data may be lost.

☐ Data may not be readable if software file formats change in the future.

☐ Data may be difficult to understand if there is no documentation left for the data file.

☐ Data files may become unintelligible or unreliable when opened with new software to the extent that research cannot continue.

☐ The preserving period of data is stipulated to be permanent according to Article 16 of Chapter 5 of the KIGAM Data Management Regulations.

1) University of Leicester. (n.d.). Why preserve digital data?. Retrieved November 20, 2019. Available: https://www2.le.ac.uk/services/research-data/keep-data/lterm-pres

【 The Goal of Data Preservation 】[2)]

☐ Data management: Ensuring that digital records can be managed through inevitable changes.

☐ Accessibility: Ensure that data is easy to find and accessible.

☐ Availability: Ensure that users can work with data the way they need to.

☐ Data documentation: Help users understand what the data is and what it is about.

☐ Integrity: Ensure the reliability of data throughout the Data Lifecycle.

【 Data Management Plans and Preservation 】

☐ The Data Management Plan should specify the following retention-related matters [3)]

- ◦ The administrator who is responsible for Data Preservation.
- ◦ The data format description to be produced.
- ◦ The size of the dataset to be produced.
- ◦ Where the data will be stored.
- ◦ State if a data repository for the research field or institution exists and explain if it will be utilized.

【 Data File Organization and Description 】

☐ Data preservation is a set of management activities taken to ensure the long-term viability and continued accessibility of research data, and, therefore, includes data file organization and data description.

☐ The format of data files should follow non-proprietary and open standards to the extent possible, given the ongoing access and potential reuse of data.

☐ Metadata and documentation should be used to describe the data to be preserved.

---

2) BGS. (n.d.). BGS Digital Preservation Policy Retrieved August 7, 2019. Available: https://www.bgs.ac.uk/downloads/start.cfm?id=3173

3) USGS. (n.d.). USGS Data Management Plan Checklist. Retrieved August 7, 2019. Available: https://prd-wret.s3-us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/atoms/files/data-management-checklist.pdf

□ <Table 1> provides guidelines based on the type of material.

<Table 1> Data Type and Content

Data Type	Guide Content[4]
Data File	• Data in a machine-readable form, i.e., in a state that allows the software to view the individual content or internal structure of the data or to process it, such as modify, transform, extract, etc.[5] • It is recommended that the dataset or data file intended to be deposited be provided in a widely accepted format for future reuse or in a specific format that is universally accepted by the domain area community.
Documentation File	• Metadata describing the contents of the data file should be provided along with the data files. • Examples of documentation files may include codebooks, data collection instruments, summary statistics, project summaries, and lists of data-related publications. • In addition, it may include:[6] • Project background and objectives • Information about the methodology • Sources used • Relevant studies • Sampling procedure • Content and structure of the dataset • A description of the data and a list of file names • Tools or software needed to work with or read the data • A description of any known errors or weaknesses in the data • References to publications related to or resulting from the project • Documentation of records, data transformations, or format changes.
Metadata	• Metadata describing the contents of the data file must be provided with the data file. • Metadata includes project title, principal investigator name, summary, distributor, keyword, geographic scope, temporal scope, and depositor.

4) ICPSR(2019). Retrieved November 12, 2019. Available: https://www.icpsr.umich.edu/icpsrweb/deposit/
5) Article 12 Registration of Research Data of chapter 4 Registration and Management of Research Data, Research data management regulations,
6) ESSA(n.d.). ESSA guidelines for depositors. Retrieved

3 Selecting and Evaluating Data to Preserve

【 The Need to Choose Long-Term Preserving Data 】[7]

☐ Even if data storage is not costly, there are reasons to select data for long-term preservation rather than storing all data, including:

- ○ The rapid growth of digital data makes storing everything unaffordable.
- ○ Digital preservation methods are not sustainable without proper mirroring and backup systems, and ultimately, backup and mirroring increase the cost of preservation, which means that storage costs at least double.
- ○ Storing all data can require additional effort to determine which data are relevant to a search, which can be reduced by selectively storing data.
- ○ Since a lot of data management and preservation costs are required, the cost of creating and managing preservation metadata and the preservation cost of the data to be preserved must be considered.

【 Criteria for Selecting Long-Term Preserving Data 】[8]

☐ Due to data storage resource limitations, long-term preservation of all data is not possible, so the criteria listed in <Table 2> can be used to select data with a long-term preservation value.

☐ <Table 2> shows the criteria for selecting long-term archival data.

<Table 2> Criteria for Selecting Long-Term Preserving Data

Category	Content
Legal considerations	• Is there a legal reason to retain the data? • Is the data used or could be used in a lawsuit, public inquiry, police investigation, or a report or paper that could be legally challenged? • Is there a financial or contractual obligation to retain the data? • Was the data used to write the paper also used to register its performance?
Scientific or historical	• Does the data have a geographic or temporal scope that makes it useful to others? • Does the data have historical value (e.g., can it be presented as a landmark

7) DCC (2014). 'Five steps to decide what data to keep: a checklist for appraising research data v.1'. Edinburgh: Digital Curation Centre. Available: http://www.dcc.ac.uk/resources/how-guides

8) BGS. (n.d.). NGDC Data value checklist. Retrieved August 23, 2019, Available: https://www.bgs.ac.uk/services/ngdc/documents/DVCNGDC.pdf

Category	Content
Value	of scientific discovery)? • Does the data involve changes in processing methods, new standards, or precedents? • Does the data support a trend or current project in science? • Is there potential for more research in the relevant scientific field? • Is it likely to meet the future needs/directions of the scientific community? • Is the data contributing to a broader collection? • Is the data likely to be reused? • Is the data cited in publications?
Original	• Is the data unique? • Does the data remain unchanged and maintain its existing integrity? • Would it be cost-prohibitive to reproduce or re-collect the data? • Is this believed to be the primary copy of this data? • Are copies of this data at risk?
Conditions	• Are the data accompanied by relevant metadata? • Are there more scientific value data than non-scientific value data? • Can the data be ingested without additional processing (e.g., differentiation, format conversion, etc.)? • Are the data in good condition to be added to the collection (i.e., readable, intact, and robust enough to be handled)?
Storage and Preservation	• Can the data be stored without special requirements (digital or hard copy)? • Can the data be preserved without special requirements (digital or hard copy)?
Access/Use	• Can the data be deposited without intellectual property or copyright restrictions? • Can the data be deposited without conditions imposed by external sources or existing terms and conditions? • Can the data be deposited without any temporal restrictions on its the use ?
Format/ technical limitations	• Is the deposit in an acceptable data format? (examples[9]) • Are the data accessible without specialized (and generally unavailable) software? • Is specialized software readily available from the Geoscience Data Center? • If the data is not in an acceptable format, can it be transferred to an appropriate storage/archiving system or converted into a commonly used format?

4 Data Repository

【 Definition of Data Repository 】

☐ A data repository is an online database service, an archive that manages the long-term storage and preservation of digital data resources and provides a catalog for navigation and access.[10)]

【 Considerations for Choosing a Data Repository 】[11)]

☐ Provide a persistent identifier for the submitted dataset.

☐ After exploring the dataset, metadata that supports checking and using the contents of the dataset are provided as a landing page for the dataset.

☐ Support tracking of data use.

☐ Respond to community needs or be recognized as a “trusted data repository.”

☐ Meet legal requirements, such as data protection, and allow for data reuse without unnecessary licensing requirements.

【 Examples of Data Repository 】

☐ General-purpose repositories:

- FigShare(http://figshare.com)
- Dryad(https://datadryad.org)
- Zenodo(http://zenodo.org/)
- DataHub(http://datahub.io)
- DANS(http://www.dans.knaw.nl/)

9) NGDC. (2023.). NGDC Depositing Data - Preferred Digital Formats. Retrieved May 19, 2023 Available: https://www.bgs.ac.uk/download/ngdc-depositing-data-preferred-digital-formats/

10) The University of Sheffield. (n.d.). Research data repositories. Retrieved November 15, 2019. Available: https://www.sheffield.ac.uk/library/rdm/repositories

11) DCC (2014). 'Five steps to decide what data to keep: a checklist for appraising research data v.1'. Edinburgh: Digital Curation Centre. Available: http://www.dcc.ac.uk/resources/how-guides

□ <Table 3> shows repositories in the field of geoscience.

<Table 3> Repository for geoscience datasets

Repository	Explanation
National Geoscience Data Centre (NGDC)	• The National Geoscience Data Centre (NGDC) is a repository that manages datasets from the British Geological Survey (BGS) in the UK, collecting and preserving geoscience data and information for long-term use by the community. • http://www.bgs.ac.uk/services/ngdc/
Centre for Environmental Data Analysis (CEDA)	• CEDA operates the Atmospheric and Earth Observation Data Center function on behalf of the Natural Environmental Research Council (NERC) for the UK atmospheric science and Earth observation community. • https://www.ceda.ac.uk/
UK Polar Data Centre (UK PDC)	• The UK Polar Data Centre (UK PDC) is the center for Arctic and Antarctic environmental data management in the UK and is part of the Natural Environmental Research Council's (NERC) Environmental Data Network. • https://www.bas.ac.uk/data/uk-pdc/
PANGAEA	• PANGAEA has a 30-year history as an open access library for archiving, publishing, and distributing georeferenced data in the earth, environmental, and biodiversity sciences. • https://www.pangaea.de/
TOAR Surface Observation Database	• The Tropospheric Ozone Assessment Report (TOAR) database is the world's most extensive database of surface ozone measurements. • https://toar-data.fz-juelich.de/
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)	• The Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) is one of the Earth Observing System Data and Information System (EOSDIS) data centers managed by the National Aeronautics and Space Administration (NASA) Earth Science Data and Information System. • https://daac.ornl.gov/
Norwegian Marine Data Center (NMD)	• The Norwegian Marine Data Center (NMD) is a national data center dedicated to the specialized processing and long-term storage of marine environmental and fisheries data and the production of data products. • https://www.hi.no/en/hi/forskning/research-groups-1/the-norwe

Repository	Explanation
	gian-marine-data-centre-nmd
International Council for the Exploration of the Sea (ICES)	• ICES is an intergovernmental marine science organization that advises on conservation, management, and sustainability. It aims to increase and share scientific knowledge of the marine environment and its living resources and to utilize it. • https://ecosystemdata.ices.dk/
ICTS SOCIB Data Repository	• Balearic Islands Coastal Observing and Forecasting System (ICTS SOCIB) is a multi-platform distributed and integrated system that provides oceanographic data product streams and modeling services. • https://www.socib.es/data/

【 Storing Data 】

☐ Identify how many copies of data to store and how to synchronize them.

☐ Provide storage for data.

☐ Retain a backup site for data transfer systems stored in cloud-based services.

☐ Provide a data download service from the backup site in the event of a power outage.

☐ Provide criteria for comparing storage (storage space) solutions.

☐ Ensure integrity and accessibility when backing up data.

【 Backup and Recovery 】

☐ To prevent and protect data loss and damage, researchers are responsible for regularly and automatically backing up their data to multiple locations.

☐ The backup system of the Geo Big data Open Platform consists of a double backup with InnoStor Appliance (ISA-2000) and Quantum Scalar i500, and backup is performed by storing the data periodically backed up from the service storage in the backup system.

- ○ Backup target: Performs backups for data, databases, and user data files of the Geo Big Data Open Platform.
- ○ Backup cycle

- Backup of user data files, databases, and system data: Daily
- Full backup of user data, database, and research data (files): Saturday

☐ Recovery Policy and Guidelines:

- ○ System and application software are recovered from local GIT repositories.
- ○ Recovery for research data, database, and user data files is performed from data stored on the backup device.
- ○ Perform recovery from tape backups at the point of origin if the backed-up data fails.

【 Archiving and Preserving Data 】

☐ Periodically archive (magnetic tape) research data to preserve research data.

☐ Vaulting and archiving of backup tapes to a remote location through tape backup (yearly).

☐ Archiving tapes are retained for a minimum of five years.

【 Preserving Strategies for Descriptional and Procedural Stability 】[12)]

☐ Migration: Convert file formats from less common or deprecated file formats to current file formats.

☐ Emulation: Emulation, which involves mimicking the functionality of an older or obsolete computer, allows a computer to read an older file format and then save it in a current file format (a combination of emulation and migration) or a technique for reading and using older, obsolete files in the future.

☐ Normalization: Restrict data formats to common formats for preservation (e.g., limiting text files to open document formats or Word format) or converted software-dependent file formats to software-independent file formats (e.g., SPSS system files) or software-dependent file formats (e.g., ASCII or XML-based formats).

12) Inter-university Consortium for Political and Social Research (ICPSR). 2009. "Principles and Good Practice for Preserving Data", International Household Survey Network, IHSN Working Paper No 003, December 2009.

Appendix Digital Assets Preservation Framework

【 Overview of Digital Assets Preservation Framework 】[13)]

□ The Digital Asset Preservation Framework was presented by the National Digital Stewardship Alliance in 2013. This framework can be used to assess the level of digital preservation using <Table 4>.

□ The appendix is a guide for assessing the level of preservation of digital assets, which can be used to evaluate the state of preservation in a repository and provide a year-by-year indication of where the level of preservation should be increased in the future.

<Table 4> Digital Asset Preservation Framework by Level

Content	Level 1 (Data Protection)	Level 2 (Data Recognition)	Level 3 (Data Monitoring)	Level 4 (Data Recovery)
Storage and Geolocation	• Two complete copies stored physically separate from each other • For data on heterogeneous media (optical disks, hard drives, etc.), transferring content from that media to the storage system.	• At least three complete copies • At least one copy in a different geographic location • Document the storage system, storage media, and what you need to use the storage.	• One or more copies in geographic locations with different disaster threats (e.g., hurricane zone vs. earthquake zone) • Maintain an obsolescence monitoring process for storage systems and media	• At least three copies in geographic locations with different disaster threats. • Have a comprehensive plan for archiving files and metadata on systems and media that are currently accessible.
File fixity and data integrity	• Verify file integrity on ingest (if provided) • Generate checksums, if not provided	• Integrity checks on all data collection • Read-only when working with source media	• Integrity checks at regular intervals • Maintain integrity logs; provide audit information as needed	• Check the integrity of all content in response to specific events or activities

13) Phillips, M., Bailey, J., Goethals, A., & Owens, T. (2013, January). The NDSA levels of digital preservation: Explanation and uses. In Archiving Conference (Vol. 2013, No. 1, pp. 216-222). Society for Imaging Science and Technology.

Content	Level 1 (Data Protection)	Level 2 (Data Recognition)	Level 3 (Data Monitoring)	Level 4 (Data Recovery)
	• Virus scanning of all content	• Virus scanning for high-risk content	• Maintain procedures to detect compromised data • Virus scanning of all content	• Maintain procedures for replacing or repairing corrupted data • Ensuring that no one person has write access to all copies of a file
Information security	• Identify users with permissions to read, write, move, and delete individual files • Restrict permissions on individual files	• Restricting document access to content	• Maintain a log of users who have taken actions on files, including delete and retention actions.	• Perform a log audit
Metadata	• Inventory and storage locations of content • Ensure backup and physical separation of inventory information	• Storing administrative metadata • Store transformative metadata and log events	• Preserving standard technical and descriptive metadata •	• Store standard retention metadata
File Format	• Encourage limited use of known open formats and codecs if they can be used to create digital files.	• Inventory of file types in use	• Monitoring file types that are no longer supported	• Perform format migration, emulation, and similar tasks.

Version No.	Date	Contents
0.1	2023. 03. 20.	Create document outline
0.6	2023. 04. 28.	Create draft
0.8	2023. 05. 08.	Guideline review
1.0	2023. 05. 19.	Accept review comments