April 2014

USGS Guidelines for the Preservation of Digital Scientific Data

Introduction

This document provides guidelines for use by USGS scientists, management, and IT staff in technical evaluation of systems for preserving digital scientific data. These guidelines will assist in selecting, specifying, building, operating, or enhancing data repositories. The USGS Fundamental Science Practices Advisory Committee – Data Preservation Subcommittee developed these guidelines based on material from the Library of Congress-sponsored National Digital Stewardship Alliance (National Digital Stewardship Alliance, 2013). This document does not cover additional non-technical issues such as preservation policies, funding, or organizational competency and longevity, which are critical for data preservation, but beyond the scope of this document. More information about this topic can be found in *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (OCLC-NARA, 2007)[1].

When considering how to preserve digital data you should address these questions:

- Where are the data stored?
- Do you store a copy of the data off-site?
- How do you ensure the integrity of the data over time?
- What IT security features does USGS require for storing and accessing the data?
- What additional IT security features do you need?
- What metadata standard should be used to document the data?
- What sustainable file formats should be used for long-term storage?
- What are the applicable USGS rules regarding record retention periods for the data?[2]
- Who can you ask for assistance?

Table Element Definitions

Each row in the table below represents a technical element of digital data preservation.

- Storage & Geographic Location – Storage systems, locations, and multiple copies to prevent loss of data.
- Data Integrity – Procedures to prevent, detect, and recover from unexpected or deliberate changes to data.
- Information Security – Procedures to prevent human-caused corruption of data, deletion, and unauthorized access.
- Metadata – Documentation of the data to enable contextual understanding and long-term usability.
- File Formats – File types, data structures, and naming conventions to aid long-term preservation and reuse.
- Physical Media – Basic recommendations to reduce obsolescence risks that can threaten the readability of physical media.

"Long-term" and "Sustainable Format" Definitions

Long-term: A period of time long enough for there to be concern about the loss of integrity of digital information held in a repository, including deterioration of storage media, changing technologies, support for old and new media and data formats, and a changing user community. This period extends into the indefinite future.

Sustainable format: The ability to access an electronic record throughout its lifecycle, regardless of the technology used when it was originally created. A sustainable format is one that increases the likelihood of a record being accessible in the future.

Levels of Digital Preservation

For each element, the columns in the following table describe four levels of increasing assurance that digital data will be preserved. The table is based upon a left-to-right progression.

- Each level adds requirements to the previous levels.
- To enhance an existing data repository, upgrade all elements to the same level.
- For highest assurance of data preservation, specify all elements at Level Four.

The table can be used to evaluate a repository's compliance with the desired levels. For each cell in the table, assign a grade to the repository of Pass, Incomplete, or Fail. Mark each cell (e.g. green, yellow, red backgrounds), to create a quick visual assessment. Pay attention to the lower levels marked Incomplete or Fail. If you are proposing improvements to an existing repository, you could create *Before* and *After* tables, to highlight proposed or actual progress.

Level One: The minimum criteria and activities needed to maintain data through the life of a research project.
Level Two: Better. Implement Level Two elements after all Level One elements are in place.
Level Three: Even better. Implement Level Three elements after all Level Two elements are in place..
Level Four: Best. USGS should plan to provide repositories that meet these criteria for all long-term USGS records.

ELEMENT	LEVEL ONE	LEVEL TWO	LEVEL THREE	LEVEL FOUR
Storage and Geographic Location	• Two complete copies stored physically separate from each other • Transfer the digital content from temporary media into an established storage system • Managed storage system in place	• Three complete copies • At least one copy in a different geographic location (offsite locations must follow NARA 1571 guidelines[3]) • Document the storage system and storage media	• At least one copy in a geographic location with a different disaster threat (e.g. hurricane area versus an earthquake area) • Maintain an obsolescence monitoring process for the storage system and media	• At least 3 copies in geographic locations with different disaster threats • Implement a comprehensive plan that keeps files and metadata on currently accessible systems and media
Data Integrity	• Verify checksums on ingest, if provided	• Verify checksums on all data ingest	• Verify checksums at fixed intervals	• Verify checksums of all content in response to

	• Create checksums if not provided • Virus check all content	• Use read only procedures when working with original media	• Maintain logs of checksums and supply audit information on demand • Maintain procedures to detect corrupt data • Virus check all content	specific events or activities • Maintain procedures to replace or repair corrupted data • Ensure no one person has write access to all copies • Create, store, and verify a second, different checksum for all content
Information Security	• Identify who has authorization to read, write, move, and delete individual files • Limit authorizations to individual files	• Document access restrictions for content	• Maintain logs of who performed what actions on files, including deletions and preservation actions	• Perform audit of logs
Metadata	• Inventory of content and its storage location • Ensure backup and physical separation of inventory information • Adhere to current USGS metadata standards	• Store all relevant database management information • Store information describing changes to the structure or format of the data, including time of occurrence • Provide access to all forms of the metadata	• Preserve standard technical, descriptive, and preservation metadata	• Same as Level 3
File Formats	• Encourage the use of a limited set of documented and open file formats, codecs, compression schemes, and encapsulation schemes	• Inventory the file formats in use	• Monitor file format obsolescence issues	Perform format migrations, emulations (a virtual instance of a previous operating system or procedure) and similar activities
Physical Media	• Inventory all physical media utilized including hard disks.	• Develop a plan to utilize trade studies to evaluate medias	• All non-recommended media have been properly disposed of	• Base all media choices on trade studies. • All information is

		suitable for USGS purposes. • Begin to transition away from all media utilized that are 10 years or more in age.	following transition activities.	migrated from an older media to a newer media every 3 to 5 years including hard disks.

Derived from: Library of Congress, National Digital Stewardship Alliance, NDSA Levels of Digital Preservation: Version 1, February 2013.

Roles and Responsibilities

The repository manager or project chief ensures that all of the table elements are addressed, though others may be responsible for implementation and operation (e.g. data managers, IT specialists).
Scientists and research staff can use these criteria to judge the suitability of a data repository for preserving data.
Management can use these criteria for specifying or selecting data repositories.
IT personnel can use these criteria for building, buying, enhancing, or operating data repositories.

Checksums

A checksum is a short mathematical digest of a file, which changes if any bit in the file changes. You use checksums to detect unexpected changes in file content. Federal agencies, including USGS, should use NIST-approved checksums for new systems.
Currently, those are: SHA-224, SHA-256, SHA-384, and SHA-512.
MD5 and SHA-1 checksums are widely used, but not approved for new systems.
For more information, see http://csrc.nist.gov/groups/ST/toolkit/secure_hashing.html.

Assistance

- Storage & Geographic Location – John Faundeen (faundeen@usgs.gov, 605-594-6092)
- Data Integrity – Rex Sanders (rsanders@usgs.gov, 831-460-7555)
- Information Security – See your local IT security point of contact.
- Metadata – Vivian Hutchison (vhutchison@usgs.gov, 303-202-4227)

Additional USGS information and contacts can be found at http://www.usgs.gov/datamanagement/.

Other Recommendations

Individual scientists who want to improve the viable longevity of their personal archives could use the Library of Congress recommendations for Personal Archiving at http://www.digitalpreservation.gov/personalarchiving/.

Footnotes

[1] http://www.oclc.org/research/news/2007/03-12.html
[2] http://internal.usgs.gov/gio/irm/fmref2.html
[3] http://www.archives.gov/foia/directives/nara1571.pdf