USGS Archive Technology and Future Trends

Tom Sohre, USGS

USGS: 2022.10.04_14.15

WGISS-54

Tokyo, Japan (JAXA)

3-7 October 2022

CEOS
Committee on
Earth Observation Satellites

USGS POLICIES

Fundamental Science Practices (FSP)

Scientific Data Management Foundation *(requires Data Management Plans)*

Metadata for Scientific Data, Software, and Other Information Products

Review and Approval of Scientific Data for Release

Preservation Requirements for Digital Scientific Data

USGS Public Access Plan

The USGS Public Access Plan, which is reflective of broad government requirements for scientific public access to scientific data, requires USGS scientists to release the data upon which their scientific publications are based.

- Foundations for Evidence-Based Policymaking Act Of 2018
- Increasing Public Access to the Results of Federally Funded Scientific Research Feb. 22, 2013
- Making Open and Machine Readable: the New Default for Government Information May 9, 2013
- Managing Information as an Asset

USGS
science for a changing world

USGS APPLICATIONS and TOOLS

- Data Management Website
- ScienceBase
- Digital Object Identifier (DOI) Tool
- Metadata Editors
- Science Data Catalog
- Model Catalog
- Publications Warehouse
- Information Product Data System (IPDS)/ORCHID
- Code.usgs.gov
- CHS Cloud Services (Dremio, Tableau)
- USGS Web
- Dashboards
- Checklists for data/metadata review
- ISO Metadata Tool (with FWS, BOEM)

Model Catalog
data.usgs.gov/modelcatalog

ScienceBase
sciencebase.gov

DOI Creation Tool
www1.usgs.gov/csas/doi

Online Metadata Editor
www1.usgs.gov/csas/ome

Metadata Wizard
https://code.usgs.gov/usgs/fort-pymdwizard

Science Data Catalog
data.usgs.gov

USGS
science for a changing world

ASSESSMENTS and ROADMAPS

USGS State of the Data

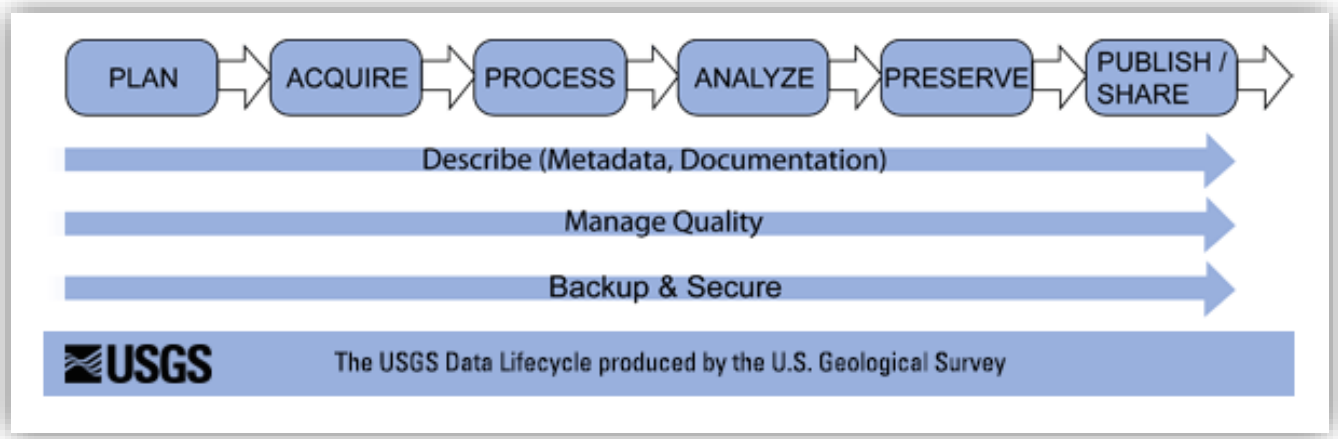
Trusted Digital Repositories Certification

Review and Approval for Data Release

USGS FAIR Roadmap

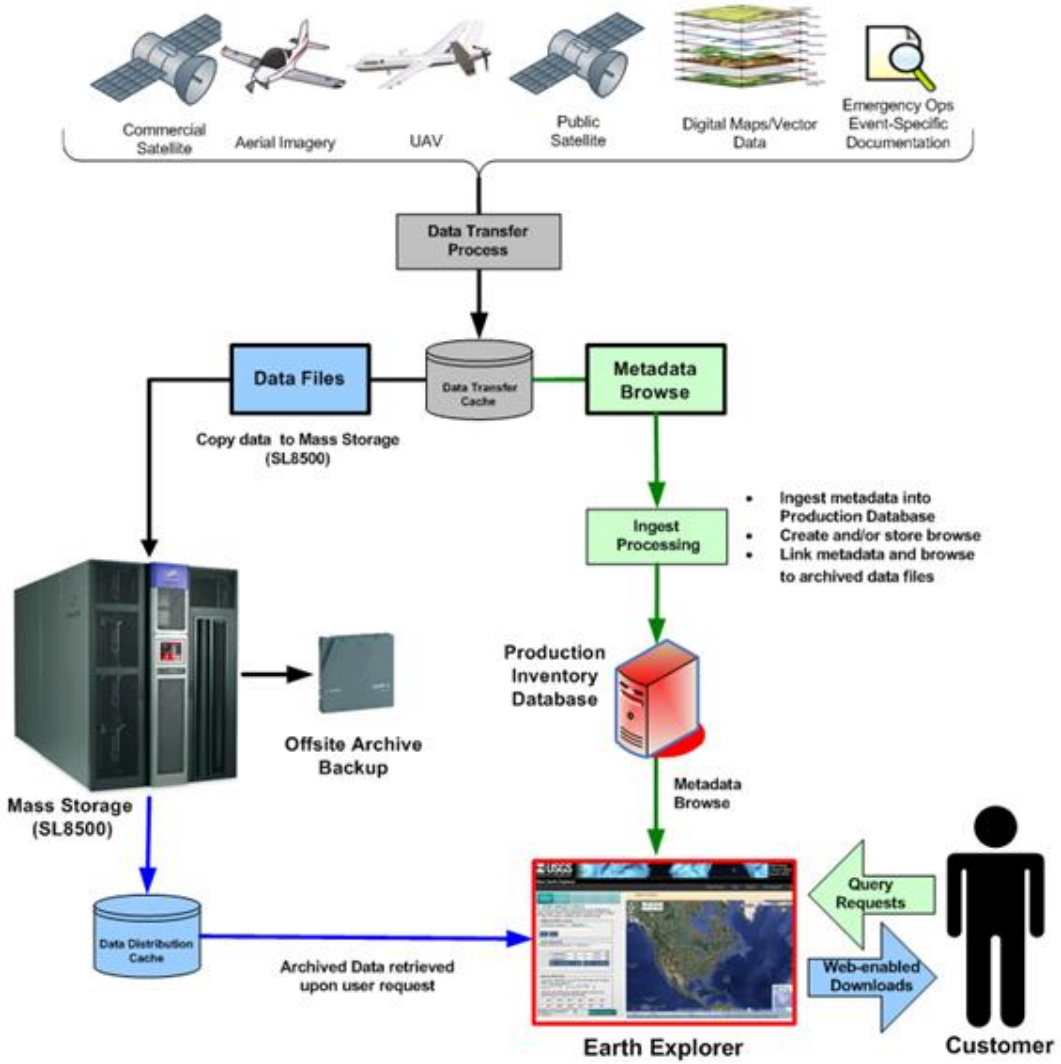
USGS EROS Data Management Goals

- Provide data management, access, archive, and distribution for all data sets within the USGS Historical Archives that have long term relevance to science and support the USGS mission
- Improve access to the land archive
- Utilize consistent data management approaches across all data sets.
- Develop and maintain data access, preservation, and distribution infrastructure to support the mission and other projects.



USGS EROS Data Management Process

- **Utilize repeatable functions and processes**
- **Develop efficient automated data ingest routines**
  - Data receipt, transfer from internet or media, QA/QC
- **Utilize consistent data management approach**
  - FGDC metadata, browse and image data linked and data base driven
  - Key to web-enabling diverse collections
- **Maintain effective long-term records management**
  - Inventory controls, data base catalog management



USGS EROS Data Archiving

- Data is managed and preserved using 3 copies: nearline, offline, and offsite
  - Offsite copies are stored at NARA's Kansas City facility
- Multiple media types are utilized to ensure long term readability
  - Currently using LTO-8 and IBM TS1160 tapes
  - Media are periodically updated as tape technology advances.
- Records management schedules are worked with National Archives and Records Administration (NARA)

USGS EROS Tape Library System

- Utilizes a Spectralogic Tfinity tape library systems
- Currently holds approximately 30 PB with 150 PB capacity
- Offsite copies are stored NARA's Kansas City facility
- Media are periodically updated as new technology becomes available

- Implementation of hierarchical store management enables distribution directly from the archived
  - Utilizes varying "tiers" of disk (SATA/SAS/Solid State) that are used as a front-end cache to the tape

USGS Cloud Methodology

- Currently using “Hybrid” approach for processing and distribution
  - Collection level processing utilizes cloud infrastructure, forward processing relies on EROS hardware
    - Collection processing takes weeks instead of months (Scalable)
    - Forward processing sits next to the archive (Cost control)
  - New acquisitions (<=90 days) are delivered on-prem, legacy collections are distributed via the cloud copy
    - Reduces on-prem IT footprint, while still controlling egress costs
      - ~70% of distribution is of acquisitions < 90 days old
    - Allows the entire collection to be cloud accessible and available for Direct Access
      - Process next to the data, not through data replication
    - Bundle and band files available
      - Reduces data transfer by allowing users to download only the data desired

Future Archive Trends

- Continue to evolve Cloud capabilities
- Data volumes growing rapidly … new products, new sensors
- Cloud egress costs large … prefer “algorithms next to archive”
- Archive replication becomes challenging … prefer data services

Data Authenticity

- Several end-users (both academia and commercial) continue to replicate the USGS archive for utilization and redistribution
- Challenges exist regarding community awareness on legitimacy and differences contained within these remote collections
- Q: Investigations are underway to provide redistributors authenticated or certified status
- Q: Currently looking at services to compare checksum hash values back against the USGS holdings for validation

Data Safeguarding and Preservation

- Archive is managed and preserved using 3 copies: nearline, offline, and offsite
  - Offsite copies are stored at NARA's Kansas City facility
  - Multiple media types are utilized to ensure long term readability (Currently using LTO-9 and IBM TS1160 tapes)

- Currently, only USGS Trusted Digital Repositories (TDR) are realized as authoritative data holders (https://www.usgs.gov/office-of-science-quality-and-integrity/acceptable-digital-repositories-usgs-scientific)
- Q: Given today's IT services, are there new ways to maintain long-term preservation while still meeting programmatic requirements?

Managing Interim Data Versions

- Advancements in sensor models and calibration/validation continue to drive product improvements
- Q: How should these versions between major processing campaigns be handled?
- Q: What retention and long-term preservation requirements exist?
- Q: How are other organizations handling this challenge?