National Geospatial Digital Archive

Greg Janée

University of California at Santa Barbara
Overview

• One of 8 NDIIPP projects funded by Library of Congress
  – joint project with Stanford University

• Goal: long-term, wide-scale preservation of geospatial data

• Preservation architecture & prototype archive
  – single-digit terabytes
  – CaSIL: GIS datasets, remote-sensing imagery, aerial photography
  – Rumsey collection: scanned maps
Common starting hypothesis

recent content

now

take action

now

+ 100 years
NGDA starting hypothesis

“mid-century perspective”

content  \[\text{now - 50}\]  \(\rightarrow\)  old content  \[\text{now}\]  \(\rightarrow\)  ancient content  \[\text{now + 50}\]

\(\uparrow\)  take action
Mid-century perspective

• Repeated migrations across storage media and storage systems
  – past and future

• Repeated migrations across archive management systems
  – each possibly necessitating transformation and reorganization of archived content

• Repeated handoffs between institutions
  – each implementing different policies
Mid-century perspective

- Migrations/handoffs may occur asynchronously
  - different evolution rates, pressures

- Ability to interpret archived data may change and deteriorate

- Information value, resource levels change over time
  - need an ultra-low cost, “fallback” preservation mode
NGDA architecture goals

• Facilitate migration at all levels
  – separate levels to accommodate asynchronicity

• Provide fallback mode
  – for individual objects and entire archives

• Capture semantics

• Cheap & easy
  – or preservation can’t be large-scale
Semantics

• *Def*: knowledge needed to interpret and use information that is not shared by the target user community

• Simple documents
  – descriptive metadata, format specification sufficient

• Remote sensing imagery
  – need data interpretation, usage, processing, calibration
  – in practice, such semantics are packaged separately

• Climate data records
  – require periodic reprocessing
Ozone reprocessing requirements

- xDRs
- Delivered IPs
- Engineering data (incl. C3S data if not in RDRs)
- Upload files
- Databases
- Software (source code)
- Calibration artifacts
  - data
  - analysis tools
  - tables
  - logs
  - notebooks
  - instrument design
- All project documentation
- All scientific papers
- All reports

*Courtesy of Mike Linda, NASA GSFC; from 2006 NOAA CLASS workshop*
NGDA architecture

1. format registry
   - maintains directory of formats; stores specification documents; models inter-format relationships

2. registry wiki
   - supports collaborative management of format registry

3. ingest crawler
   - crawls provider content; maps content to archival objects; maintains identifier associations

4. NGDA archive data model
   - defines uniform, self-contained representation of archival objects, object semantics, and inter-object relationships

5. storage API
   - abstracts storage subsystem

6. reliable storage subsystem
   - Archivas cluster

7. archive server
   - builds and validates archival objects; associates objects with semantics

8. webview
   - crawlable, HTML view of archive

9. ADL
   - provides spatiotemporal, other types of search; integrated OAI server

10. ADL mapper
    - maps archival objects to ADL items

11. ingest
    - “single item ingest”; archive management

12. export

Greg Janée • Geospatial data preservation workshop • 2006-10-27
Federation interaction points

1. Format registry...
   • provides a central place for data providers to describe file semantics, and for archives and end users to reference those semantics.

2. Ingest services and tools...
   • allow data providers to transfer content into an archive.

3. Access services...
   • allow end users to search for and use content across the entire federation, and allow third parties to provide value-added access services.

4. Archive data model...
   • defines a uniform representation of archive content; archives that implement or map to the data model can employ NGDA tools to provide access and export services.

5. Export function...
   • transfers archive content in bulk to other archives for replication and migration purposes; ancillary object semantics are automatically included.
Storage system requirements

• Req’s:
  – associate UUIDs/RIDs with bitstreams
  – retrieve global/local bitstream by UUID/RID
  – determine (parent) UUID of any bitstream
  – list all UUIDs

• Satisfied by:
  – any filesystem
  – any kind of UUIDs
    • tag:library.ucsb.edu,2005:identifier
Data model

• Physical implementation of OAIS logical model
  – filesystem
  – files and directories identified by UUIDs
  – XML manifests

• Organizing principle: archival object
  – one individually reusable unit of information
  – groups metadata, data, derivatives, etc.

• Inter-object relationships
  – semantic definitions
  – lineage
  – collections and other aggregations
Archival objects

manifest

component
Towards a more layered architecture
Towards a more layered architecture

**archive**
- asserts control
- defines policy

**archive object layer**
- defines standard structuring of content
- maintains persistent associations to semantics

**storage virtualization layer**
- provides structure-neutral storage
- interoperability between archival, working storage
- implements storage policies
Questions?