Content transfer: Getting data moved around the network

The NC Geospatial Data Archiving Project experience

Steve Morris
NCSU Libraries

DLF Fall Forum
November 12, 2008

NCGDAP Content transfer: Overview

- Content type: State and Local geospatial data from the NC Geospatial Data Archiving Project
- Three content transfers:
  - External drive transfer to LC (2 TB)
  - Network transfer to LC (remaining 3.2 TB)
  - Network transfer to San Diego Supercomputer Center (5.2 TB) as part of Chronopolis project
- Technical issues
  - Errors in initial drive manifests
  - Checksumming times for large TIFFs
  - Illegal characters in uncurated data

Importance of data transfer in project

- Very large data collections originating from small local agencies
  - 100's of GB for a single county orthophoto flight year
  - Limited network capacity at agency level -- need for simple network transfer solutions that can be implemented at local or state agency level
- Reality of external drive transfer
  - e.g. "orthophoto sneakernet" among agencies
  - Need better data integrity management in disk transfers
- New local-state-federal collaborations on data sharing infrastructure
  - Project focus: capture data "in motion"

Content transfer as aid to preservation

- Preparation of data for transfer spurs introspection and can catalyze improvements
  - Documentation and metadata
  - Clarification of rights
  - Extraction and encapsulation of data
- Requires and helps cultivate a common understanding of data
  - Data naming
  - Content packaging (e.g., .zip file practices)
  - How and when the data can be used

Content transfer in GeoMAPP

- Geospatial Multi-state Archival and Preservation Partnership (2007-2009 NDIIPP project)
  - State Archives and state geospatial organizations of North Carolina, Kentucky, and Utah
- Intrastate data transfer scenarios
  - G → G: Local-to-state, between state agencies
  - G → A: State/local geo agency-to-Archives
  - A → G: Retrieval of historical data from Archives
- Inter-state data transfer scenarios
  - G → G: Reciprocal continuity of operations support (?)
  - A → A: Distributed preservation networks