THE EARTH INSTITUTE
COLUMBIA UNIVERSITY

Geospatial Data Stewardship at an Interdisciplinary Data Center

Robert R. Downs, PhD
Senior Digital Archivist and Senior Staff Associate Officer or Research
Acting Head of Cyberinfrastructure and Informatics Research and Development

Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

Prepared for presentation to the
Geospatial Summit on
Framing a National Preservation and Access Strategy for Geospatial Data
November 12-13, 2009
Library of Congress

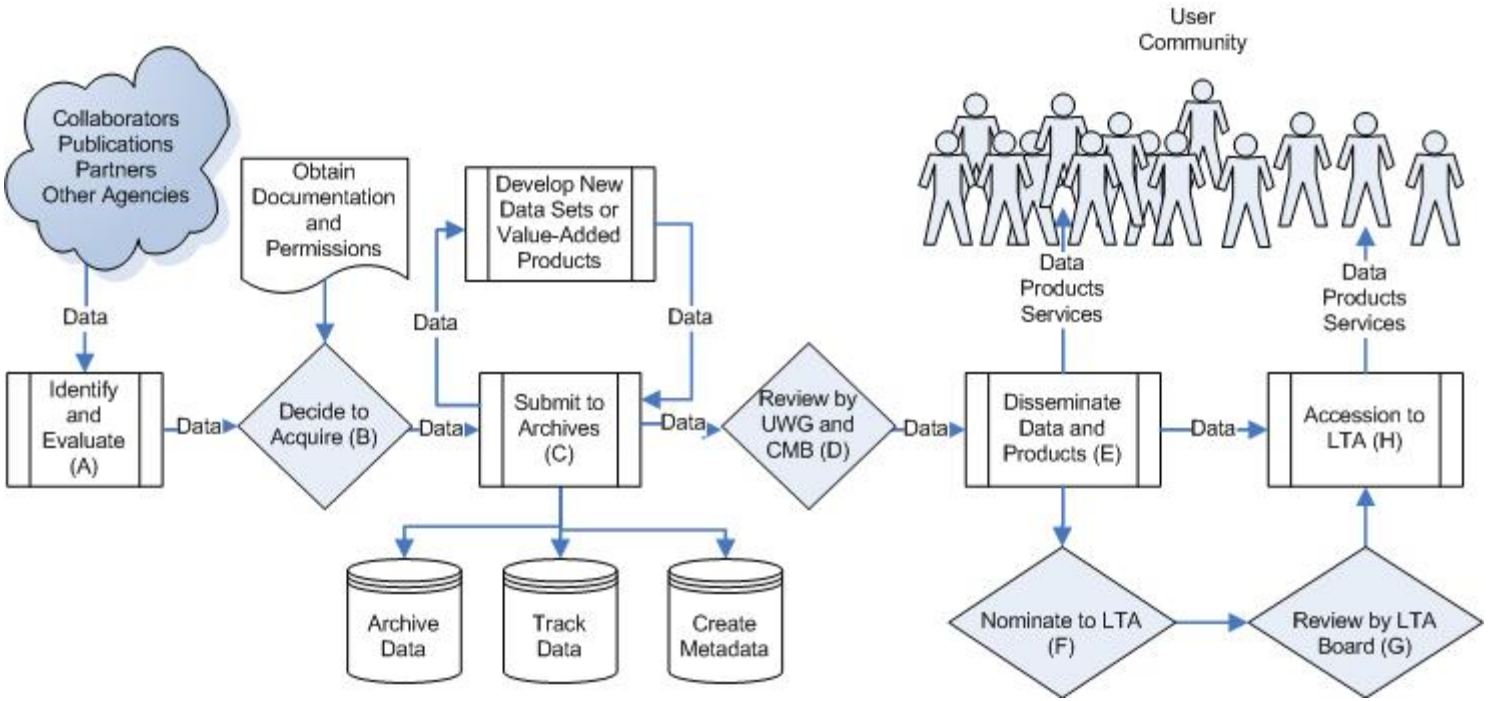
Geospatial Activities at CIESIN

- Examples of Current Projects
  - NASA Socioeconomic Data and Applications Center
  - World Data Center for Human Interactions in the Environment
  - Northeast Information Node for the National Biological Information Infrastructure (NBII)
  - Cyberinfrastructure for the Earth Institute (CI4EI)
  - Africa Soil Information Service (AfSIS) Digital Soils Map Project
  - Millennium Villages
  - Polar Information Commons
- Examples of Previous Projects
  - Climate Change Information Resources for the New York Metropolitan Region (CCIR-NYC)
  - Global Natural Disaster Hotspots
  - Managing and Preserving Geospatial Electronic Records

CIESIN Geospatial Data Lifecycle Activities

- Development of Data Products and Services
  - Data creation, enhancement, and description; map creation; facilitating data analysis
- Archiving of Data and Research-Related Information
  - Data, data products, scientific reports, maps, guides, methodological descriptions, documentation
- Data Dissemination
  - Enabling and supporting use of data by interdisciplinary scientific, educational, and decision-making communities
  - Long-Term Archiving
  - Establishing a Long-Term Archive for the NASA Socioeconomic Data and Applications Center (SEDAC) in collaboration with the Columbia University Libraries and the Earth Institute of Columbia University

SEDAC Data Workflow



Selection Criteria for LTA Data Appraisal

Scientific or Historical Value

- citation, research, and educational use as published in refereed scientific publications/reports from recognized committee of scientists

Potential Usability and Use

- evidence of usability, usefulness, and sufficient usage by the community interested in human dimensions of the environment. Adequate evidence indicate potential for future use justifies costs of long-term archiving

Uniqueness of Data (non-redundant stewardship)

- not being preserved in any form in another archive and is at risk of loss if not accessioned into the Long-Term Archive

Relevance to LTA Mission

- currently endorsed or approved by community interested in human interactions in the environment. For the short-term, relevance includes content germane to SEDAC mission and SEDAC strategic plan

Documented for Accessibility

- completeness and correctness of documentation to facilitate future discovery, access, and use

Technological Accessibility (feasibility)

- received in format meeting technical criteria for the Service Level designated for the resource

Legality and Confidentiality

- unrestricted permissions for preservation and future dissemination. No information that is confidential or prohibited from dissemination

Non-Replicability

- data replication not feasible, excessively costly or prohibitive

Interdisciplinary Data and Use

- Data Holdings Represent Various Kinds of Data
  - GIS, Maps, and Remote Sensing Data
  - Model data and Programs
  - Demographic and Survey Data
  - Examples of Diverse Publications Citing Data Holdings
  - Agriculture, Ecosystems & Environment; American Journal of Public Health; Atmospheric Research; Biodiversity and Conservation; Biological Conservation; Bioscience; Climatic Change; Conservation Biology; Ecological Economics; Ecological Indicators; Ecology Letters; Ecosystems; Energy Economics; Energy Policy; Environment, Global Environmental Change; Hydrological Processes; Journal of Risk Research; Marine Policy; National Geographic; Nature; Science; Sustainability Science; The Bulletin of the Atomic Scientists; The Journal of Environment Development; The Journal of Geology; The Lancet Oncology; Trends in Parasitology; Water Policy; Water Resources Management

Examples of Formats for Archived Data

- GIS, Remote Sensing, and Map Imagery:
  - ESRI ArcGIS, ArcInfo Workspace, American Standard Code for Information Interchange (ASCII), GRID, ArcIMS XML (axl), ArcView Theme Legend (avl), Shapefile Attribute Format (dbf), Shapefile Projection (prj), Shapefile Spatial Index (sbn and sbx), Shapefile Shape (shp), Header (hdr), Band Interleaved by Line (bil), Tagged Image File and GeoTiff (tif), Graphics Interchange Format (gif), personal Geodatabase (mdb),Comma-Separated Values (csv), Joint Photographic Experts Group (jpg), Microsoft Excel (xls), Adobe Portable Document Format (pdf), Microsoft Word (doc), Powerpoint (ppt), Hypertext Markup Language (html and html), JavaScript Source Code (js), Extensible Markup Language (xml)
- Modeling:
  - ASCII Byte (byt), ASCII Output (out), ASCII Data (dat), ASCII Text (txt) Btreive Unformatted (unf), Unix Shell Script (sh), Scene (scen), Initialization (in, ini & init), Model Parameter (par), Configuration (con), Control (ctl), Workbook (wk1), Fortran (for), Executable (run), Gemstone (gs), Word (doc)
- Tabular:
  - Comma-Separated Values (csv), Excel spreadsheet (xls), Word (doc),
- Unstructured :
  - Word (doc), Text (txt), Adobe Portable Document Format (pdf), Hypertext Markup Language (html and html),

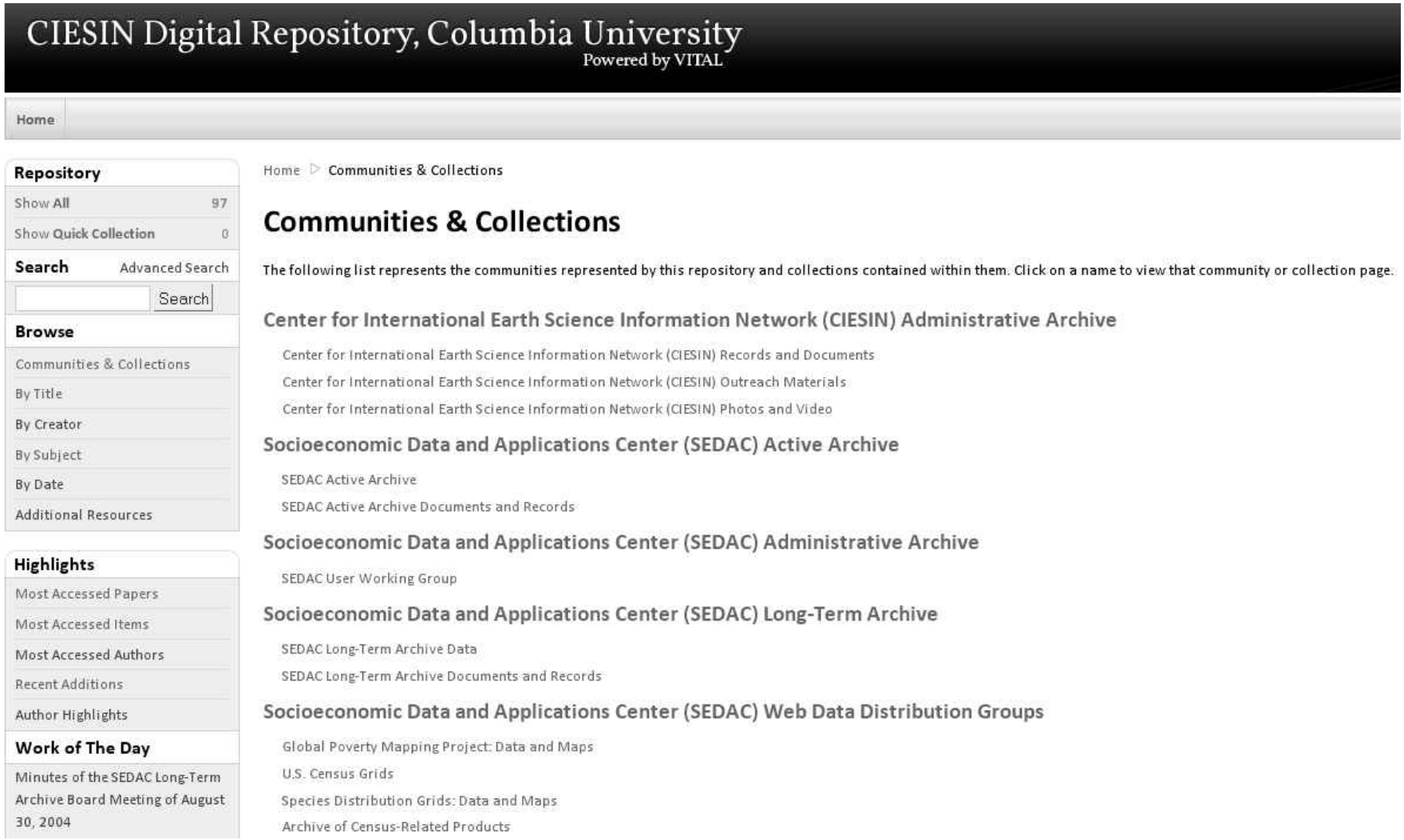
Example Formats of Additional Materials Archived

- Documentation:
  - Word (doc), Portable Document Format (pdf), Text (txt), Word doc, (pdf)
- Metadata
  - Extensible Markup Language (xml), Hypertext Markup Language (html)
- Permissions
  - Portable Document Format (pdf), Text (txt),
  - Also received as email, fax, or signed original
- Fixity Information:
  - Text (txt), Extensible Markup Language (xml)

Archival Evolution

- Implemented VITAL / Fedora digital repository
  - Currently operating VITAL 3.1.1, Fedora 2.2, VALET 1.1.3
  - Duplicated in failover system and nightly backup copies
  - Managing Multiple Collections
- Migrating from traditional digital data archiving procedures to managing collections in digital repository
  - Parallel ingest for selected data sets
  - Migration of selected archived data sets to digital repository
  - Continuing traditional archival operations onsite and offsite

Digital Repository Collections



Digital Repository Development

- Conducting a Self-Assessment of LTA Collection
  - Based on the Trusted Repositories Audit and Certification: Criteria and Checklist (TRAC)
- Developing and testing online submission services
  - Enabling self-submission by data producers
- Assessing schemas for provenance of pre-ingest activities
  - (PREMIS Event, DDI Archive)
- Customizing pre-ingest workflow
  - Establishing processes for online review and approval
- Enabling access control
  - Defining privileges for collections, objects, and datastreams
- Developing capabilities to test transfer between repositories
  - Collaborating with Columbia University Libraries digital repository