World Data Center for Human
Interactions in the Environment

Exploring Collaborative Models for Sustainable Governance of Digital Collections of Scientific Data

Robert R. Downs and Robert S. Chen
Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

Prepared for presentation to the
Workshop on Digital Archive Preservation and Sustainability (DAPS'08)
held in conjunction with the
IEEE Symposium on Massive Storage Systems and Technologies
Baltimore, Maryland

September 22, 2008

Sustainable Scientific Data Stewardship: Critical to the Future of Science

- **Enabled** by sustainable scientific data stewardship:
  - Data discovery and use by future communities of users
  - Expanded opportunities for data-driven science
  - Longitudinal studies of digital and digitized observations
- **At risk** without sustainable scientific data stewardship:
  - Records of non-replicable observations
  - Legacy data that underpin the cumulative scientific knowledge base
  - The ability to replicate past and current scientific analyses

Note: “Stewardship” encompasses both preservation and curation

Sustainable Cyberinfrastructure for Preserving Today's Scientific Data and Research-Related Information

- **Technical Infrastructure**
  - Information and communication technologies and skills enabling continuing access and interoperability
- **Standards**
  - Classifications, persistent identifiers, intellectual property rights, specifications, and ontological frameworks enabling discovery and use
- **Sustainable Governance**
  - Institutional governance and resource commitments enabling continuing stewardship

Responsibilities for Preserving Our Digital Heritage

- "Measures should be taken to … encourage universities and other research organizations, both public and private, to ensure preservation of research data"
- "Preservation of the digital heritage requires sustained efforts on the part of governments, creators, publishers, relevant industries and heritage institutions"

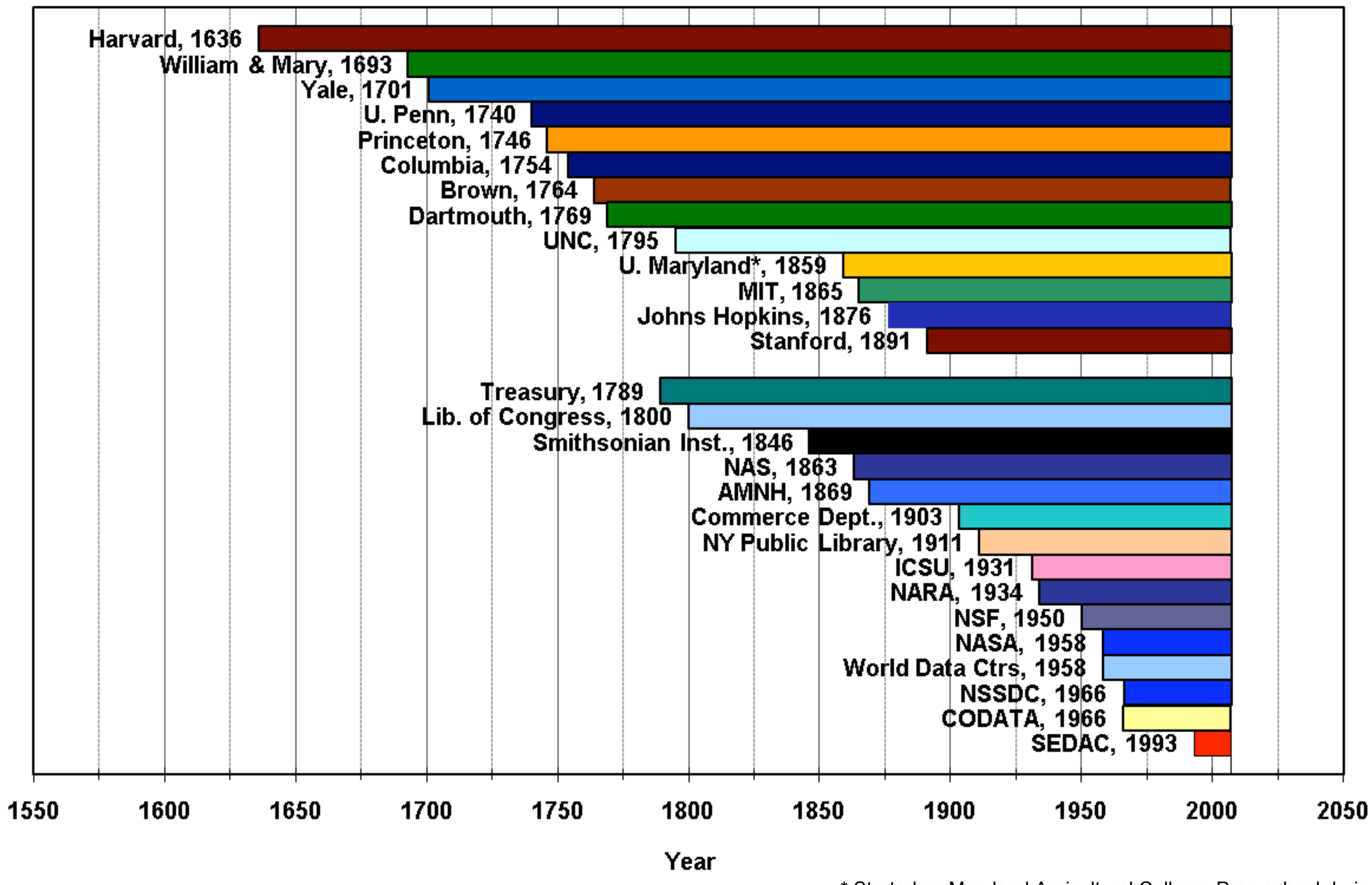
Source: United Nations Educational, Scientific and Cultural Organization. Charter on the Preservation of the Digital Heritage. (2003) http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf

UNESCO
United Nations Educational,
Scientific and Cultural Organization

Why Should University Libraries and Scientific Data Centers Collaborate on Scientific Data Stewardship?

- **Mutual interest** in the long-term stewardship of digital scientific data
  - University Libraries are realizing that the diverse and rapidly growing amounts of digital data generated by their faculty, staff, and students are at significant risk of loss
  - Scientific data centers are realizing that longevity in data preservation and access will depend on continuous, long-term institutional support with resources, staff, and infrastructure less vulnerable to short-term fluctuations

Longevity of Selected Universities, Government Agencies, and Other Institutions



* Started as Maryland Agricultural College; Prep school during 1864-66

What are Some of the Potential Benefits of Collaboration Between University Libraries and Scientific Data Centers?

- Complementary expertise and experience
  - University Libraries offer a tradition of sustainable information services to support various learning and research activities of the university community
  - Scientific data centers offer progressive data services to support research and related activities within specific scientific disciplines
- Broader engagement with the relevant science communities
  - University Libraries have long-term links with campus departments, faculty, staff, and students who can contribute needed data and expertise
  - Scientific data centers have strong collaborations with national and international data networks, scientific societies, and disciplinary and interdisciplinary science communities

Columbia Libraries Have Recognized Need for Collaboration on Long Term Digital Archiving

- “Work with campus partners such as CIESEN, the NASA Socioeconomic Data and Applications Center (SEDAC), and Columbia University IT on strategic and implementation planning for the creation of a Columbia Long-Term Digital Archiving Service.”

CU Home > Libraries Home
Search | Site Index | FAQ | Help
Columbia University LIBRARIES
Libraries & Collections | Catalogs | E-Resources | Request It | Ask Us | Using the Libraries
Spotlight
crossSearch
learn more
News
Quick SEARCH
CLIO | Databases | E-Journals | Website
Search Library Catalog: Keyword(s) GO
Go To CLIO >>
INFORMATION FOR...
Alumni & Friends
Jobseekers
Visitors
FEATURED RESOURCE
Text-only version of this page
© Columbia University Libraries
My Library Account | Hours | Contacts | Suggestions

Search CLIO, Proquest, Web of Science simultaneously

Columbia University's Rare Book & Manuscript Library acquires manuscript collection from novelist Oscar Hijuelos ... >>

Shakespeare Collection
An extensive collection of authoritative materials and the complete Arden editions

crossSearch for books and articles

Source: Columbia University Libraries Strategic Plan 2006-2009 (October 2006). http://www.columbia.edu/cu/lweb/img/assets/6675/strategicplan_2002-2009.pdf

Key Data Stewardship Challenges for the Columbia Libraries

- Increasing number, size, and complexity of digital collections derived from print and analog sources
  - E.g., digital copies of fragile audio tapes, manuscripts, photos
- Increasing number, size, and complexity of "Born digital" data
  - Digital data, databases, documents, images, etc. generated by Columbia faculty, staff and students
  - Collections of digital materials obtained by the Libraries for research and preservation, e.g., architectural drawings in CAD format, Geographic Information System (GIS) files
  - Community data collections and databases developed and maintained by campus organizations

Key Data Stewardship Challenges for SEDAC

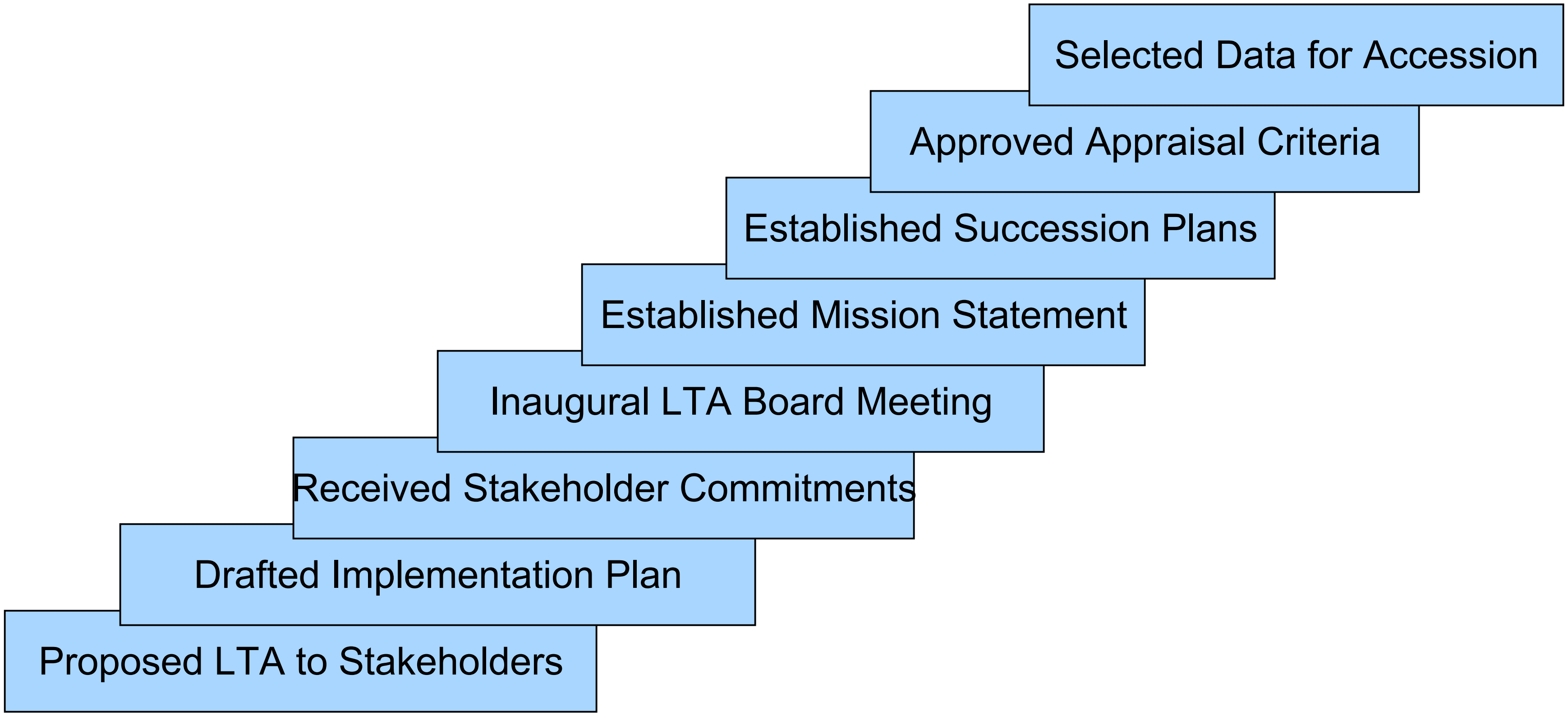
- SEDAC has limited funding
  - Priority is on supporting users with "latest and greatest" data
  - Older data still valuable, but maintenance of continually growing amounts of older data should not eat into new data activities
- SEDAC funding could end at any time
  - SEDAC operates on a five-year contract from NASA, but funding allocations are annual and sensitive to NASA's budget situation and programmatic priorities
- CIESIN, which operates SEDAC, and even the Earth Institute do not (yet) have long-term institutional homes within Columbia

The SEDAC Long-Term Archive: An Experiment in Sustainable Governance for Stewardship of Interdisciplinary Scientific Data

- Initiated in 2004 to preserve scientific data and research-related information disseminated by the NASA-supported Socioeconomic Data and Applications Center (SEDAC) for future access and use
- Managed collaboratively by SEDAC and the Columbia University Libraries

Low Library, Columbia University

Steps in the Establishment of the SEDAC Long-Term Archive



Selected Data for Accession

Approved Appraisal Criteria

Established Succession Plans

Established Mission Statement

Inaugural LTA Board Meeting

Received Stakeholder Commitments

Drafted Implementation Plan

Proposed LTA to Stakeholders

TRAC Requirements for Governance & Organizational Structure

- A1. Governance & organizational viability
  - A1.1. Repository has a **mission statement** that reflects a commitment to the long-term retention of, management of, and access to digital information.
  - A1.2. Repository has an appropriate, **formal succession plan, contingency plans, and/or escrow arrangements** in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope.
- A2. Organizational structure & staffing
  - A2.1. Repository has identified and established the duties that it needs to perform and has appointed **staff with adequate skills and experience** to fulfill these duties.
  - A2.2. Repository has the appropriate number of staff to support all functions and services.
  - A2.3. Repository has an **active professional development program** in place that provides staff with skills and expertise development opportunities.

Source: Trustworthy Repositories Audit & Certification: Criteria and Checklist. OCLC and CRL. (2007). http://bibpurl.oclc.org/web/16712

傳 auditing and certification of Digital Archives

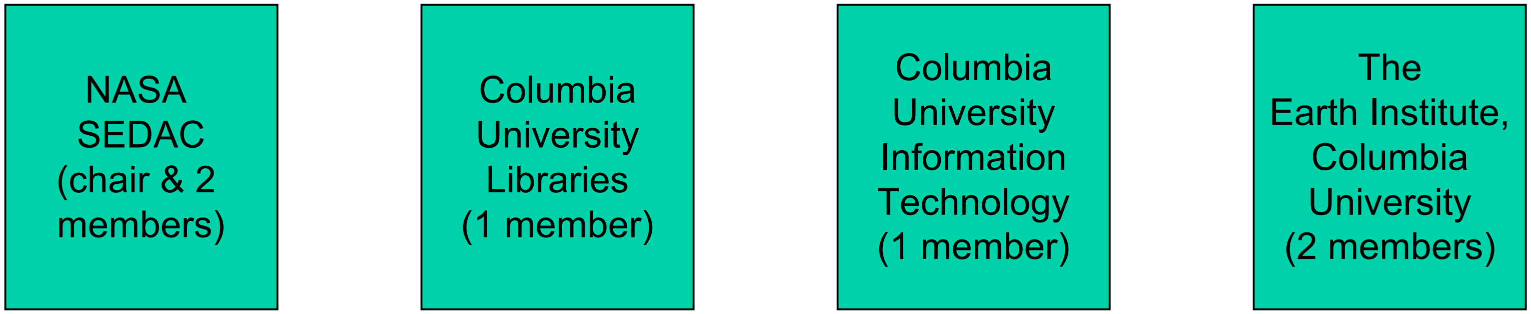
The Chinese character *chuan*, meaning to pass on over time and space or to hand down from generation to generation.

SEDAC Long-Term Archive Mission Statement

"The SEDAC Long-Term Archive acquires, preserves, and maintains the content of selected high-quality data, data products, documentation, and services relevant to human dimensions of global change in a digital form to support the discovery, access, and use of archived resources by scientific, educational, and decision-making communities for at least the next 50 years."

Source: SEDAC Long-Term Archive Implementation Plan (Draft revised 2008)

Organizational Representation on the SEDAC Long-Term Archive Board



NASA SEDAC (chair & 2 members)

Columbia University Libraries (1 member)

Columbia University Information Technology (1 member)

The Earth Institute, Columbia University (2 members)

- In the event of a lapse in SEDAC funding:
  - Libraries will replace chair and one of the SEDAC members
  - CIESIN will name the other SEDAC member

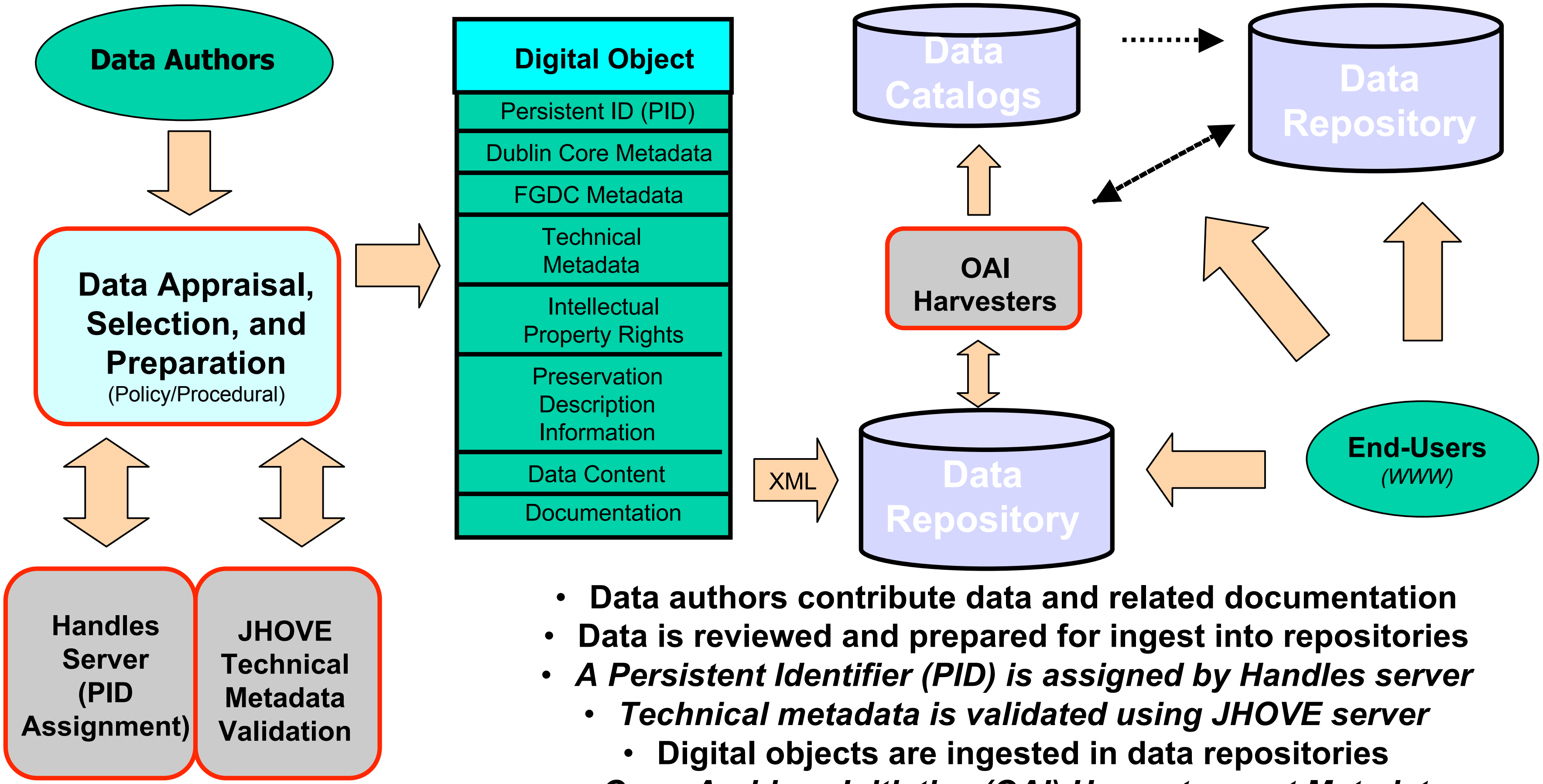
  => Libraries and CUIT will have majority of members
  - Columbia University will appoint the Long-Term Archive Manager and other staff as needed

Source: SEDAC Long-Term Archive Implementation Plan (Draft revised 2008)

CIESIN
Columbia University

The Earth Institute
AT COLUMBIA UNIVERSITY

WDC
World Data Center for Human Interactions in the Environment



Data Authors

Data Appraisal, Selection, and Preparation
(Policy/Procedural)

Handles Server (PID Assignment)

JHOVE Technical Metadata Validation

Digital Object

Persistent ID (PID)

Dublin Core Metadata

FGDC Metadata

Technical Metadata

Intellectual Property Rights

Preservation Description Information

Data Content

Documentation

XML

Data Repository

OAI Harvesters

Data Catalogs

Data Repository

End-Users
(WWW)

fedora

VITAL

VTLS

- Data authors contribute data and related documentation
- Data is reviewed and prepared for ingest into repositories
- *A Persistent Identifier (PID) is assigned by Handles server*
- *Technical metadata is validated using JHOVE server*
- Digital objects are ingested in data repositories
- *Open Archives Initiative (OAI) Harvesters get Metadata*
- *OAI Harvesters deposit metadata in data catalogs*
- End-users discover data in data catalogs
- End-users access data from data repositories

N.B.: Italics indicates machine-to-machine, automated or semi-automated

Columbia Libraries Have Recently Initiated Efforts to Build a Digital Archiving Infrastructure

- Plans for 4 copies of all digital data holdings
  - 3 online (disk) copies in 2 locations (NYC and upstate NY)
  - 1 tape copy in Iron Mountain facility
  - Working with vendors such as Sun on storage and retrieval technologies (some hardware already purchased)
- Planning to use Fedora as platform for digital asset management
- Developing migration and "exit" strategies for all technologies

Opportunities to Explore Integration of SEDAC LTA with the Libraries' Long Term Digital Archives

- Test case for coordinating a community data collection with the archive
  - Precursor to a distributed network of community data holdings across the University linked to the "main" digital archive?
- Test case for transfer of Archival Information Packages (AIPs) generated by SEDAC LTA into the archive
  - SEDAC LTA could become a "virtual collection" managed entirely by Columbia's infrastructure?
- Model for management of other digital data collections
  - Tools and interfaces developed for SEDAC LTA could facilitate stewardship of other types of data and data collections in both natural and social sciences

Current and Near-Term Collaborative Efforts

- LTA Governance and Management
  - Completing self-assessment as a trustworthy repository
  - Improving data selection and appraisal process
  - Improving preservation and dissemination services offered
- Information Technology Infrastructure
  - Testing transfer of digital objects and adequacy of current standards
  - Access control and public access
  - Capturing additional provenance metadata
  - Submission interface and workflow system
  - Developing interfaces between digital repositories
    - catalog interoperability and/or metadata harvesting
    - data migration
    - backup and recovery

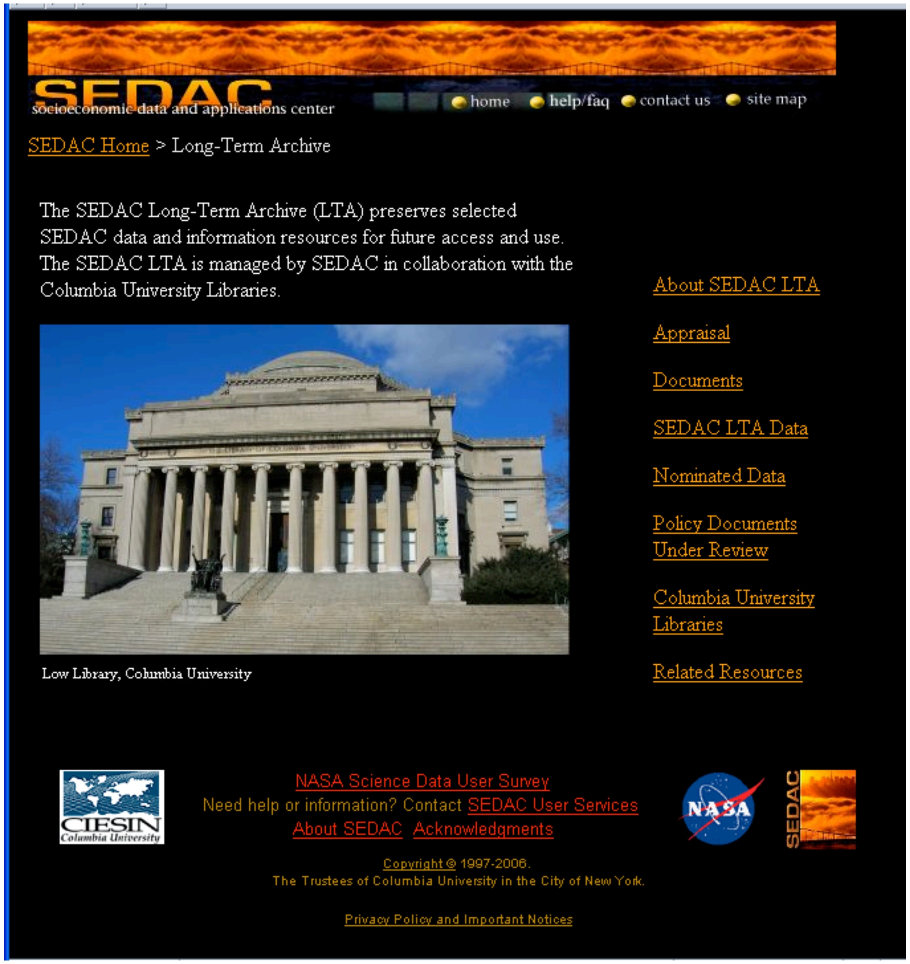
Summary: Benefits of Collaborative Governance

- The Columbia University community has >250 years of experience in preserving knowledge for future generations
- The Columbia University Libraries have a long-term role in digital data stewardship and in ensuring access by future faculty, staff, and students
- SEDAC has experience in managing specific data types using state-of-the-art tools, and resources and skills to develop, test, and implement archival systems for effective and efficient data curation
- Jointly developing the SEDAC LTA has facilitated:
  - Learning about LTA needs from both data center and library viewpoints
  - Collaborative activities to improve LTA implementation and governance
  - Increased awareness of current and future challenges in data stewardship
  - Establishment of a University-wide E-Science Task Force led by the Libraries!

World Data Center for Human
Interactions in the Environment

SEDAC LTA

http://sedac.ciesin.columbia.edu/lta/



SEDAC Home > Long-Term Archive

The SEDAC Long-Term Archive (LTA) preserves selected SEDAC data and information resources for future access and use. The SEDAC LTA is managed by SEDAC in collaboration with the Columbia University Libraries.

Low Library, Columbia University

- About SEDAC LTA
- Appraisal
- Documents
- SEDAC LTA Data
- Nominated Data
- Policy Documents Under Review
- Columbia University Libraries
- Related Resources

NASA Science Data User Survey
Need help or information? Contact SEDAC User Services
About SEDAC Acknowledgments

Copyright © 1997-2006.
The Trustees of Columbia University in the City of New York.

Privacy Policy and Important Notices

The authors gratefully acknowledge support received from NASA for the operation of SEDAC under contract NNG08-HZ11C.