World Data Center for Human Interactions in the Environment

Creating a Trustworthy Digital Repository for a Long-Term Archive of Interdisciplinary Data: A Case Study

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

21st International CODATA Conference

5-8 October, 2008

Kyiv, Ukraine

Need for a Long-Term Archive of Scientific Data

- Current and recent scientific observations are at risk of being lost
- Future tools can be used to analyze today's scientific data
- Future communities can discover and use today's scientific data
- Longitudinal analysis can compare today's observations with future observations

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

United Nations Educational, Scientific and Cultural Organization

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."

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

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

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

CIESIN Columbia University

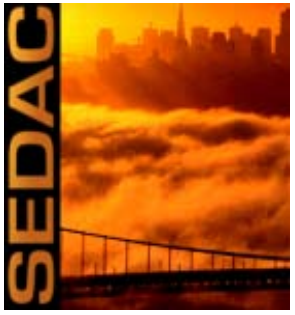
The Earth Institute AT COLUMBIA UNIVERSITY

WDC

World Data Center for Human Interactions in the Environment

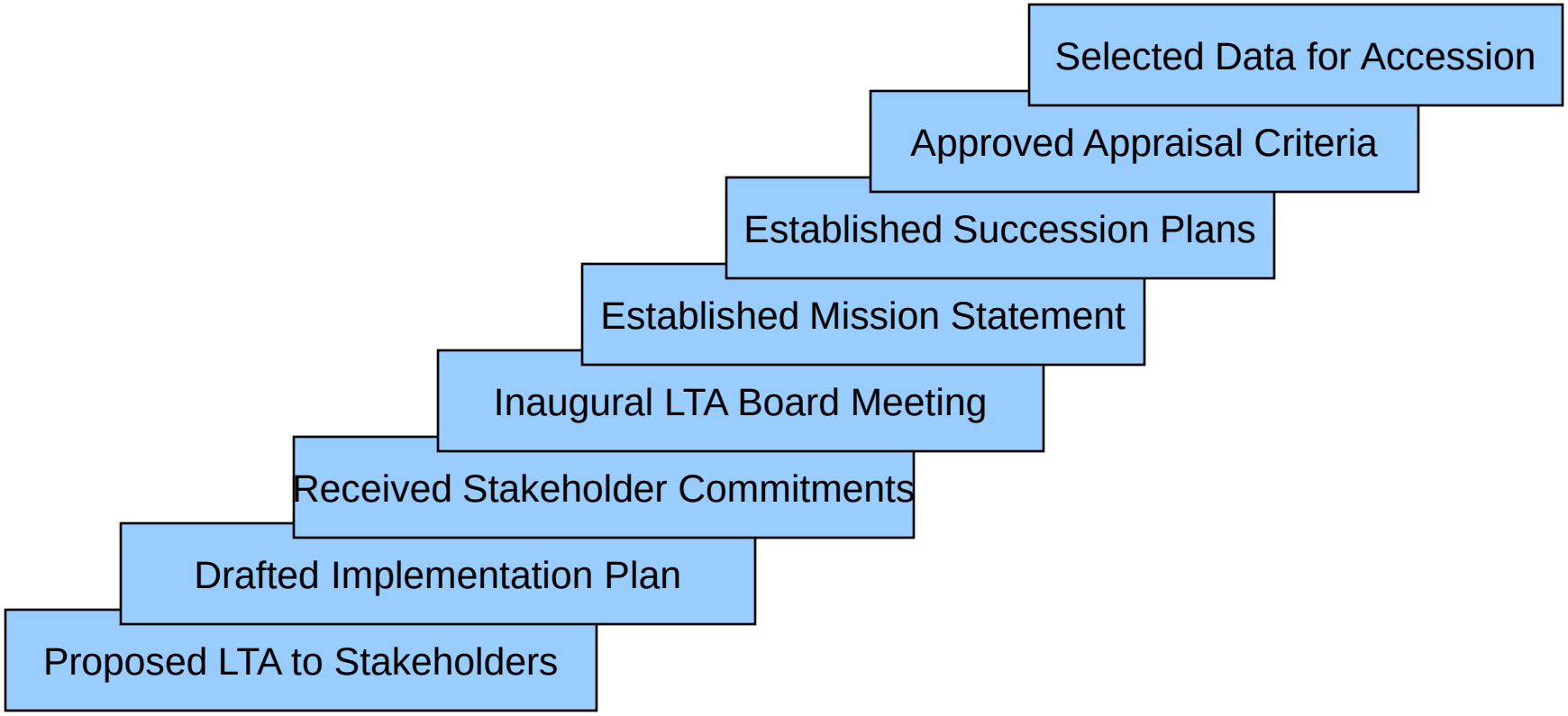
The SEDAC Long-Term Archive: An Experiment in Sustainable Governance for Stewardship of Interdisciplinary Scientific Data

- 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



Summary of Current Selection Criteria for Accession to SEDAC Long-Term Archive

- **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

Title, Publication Date, Preservation and Dissemination Services for LTA Data

Data Set Title	Pub. Date	Preservation Service – Preserve and Maintain Content In:	Dissemination Service
Environmental Subset of Collection of Multilateral Conventions at the Fletcher School of Law and Diplomacy	1992	Supported Formats	Public
Environmental Treaties and Resource Indicators (ENTRI) - The Update of the Treaty Status Data	1998	Original Formats	Public
Freedom in the World (1995-1996)	1995, 1996	Original Formats	Public
HALOPH: A Data Base of Salt Tolerant Plants of the World	1989	Original Formats	Public
Gridded Population of the World (GPW) Version 1	1995	Supported Formats	Public
Gridded Population of the World (GPW) Version 2	2000	Supported Formats	Public
Gridded Population of the World, Version 2: Ancillary Data	2000	Supported Formats	Public
World Resources 1996-97	1996	Original Formats	Public
World Resources 1998-99: A Guide to the Global Environment: Environmental Change and Human Health (Data Tables)	1998	Original Formats	Public

Conducting a Self-Assessment of the Long-Term Archive as a Trustworthy Digital Repository

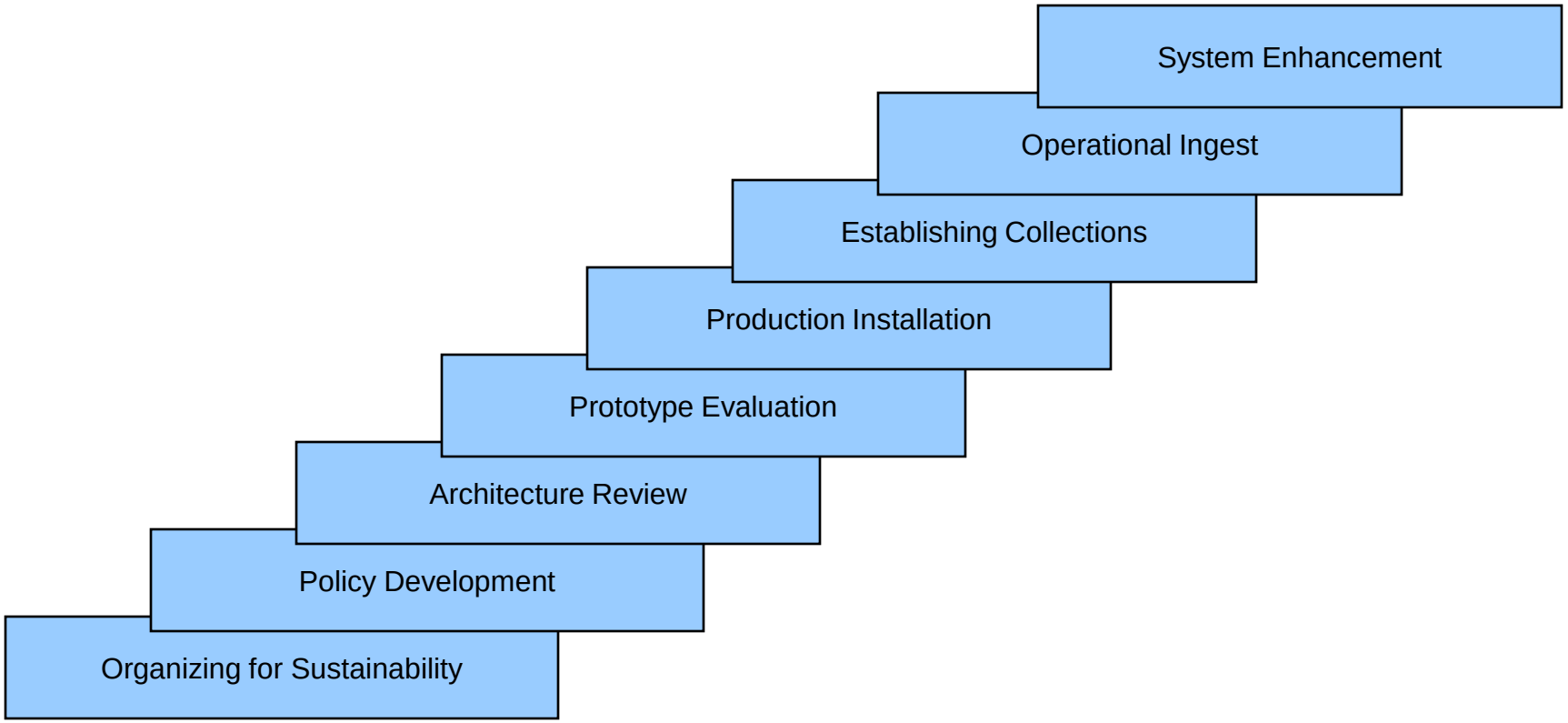
- Reviewing requirements for a trusted repository
- Examining documents and describing evidence verified for requirements met by the LTA
- Analyzing and categorizing requirements where need for improvements have been identified
- Proposing revisions for plans, policies, and procedures to meet requirements pertaining to management and practices
- Designing enhancements to meet system capabilities identified for improvement

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

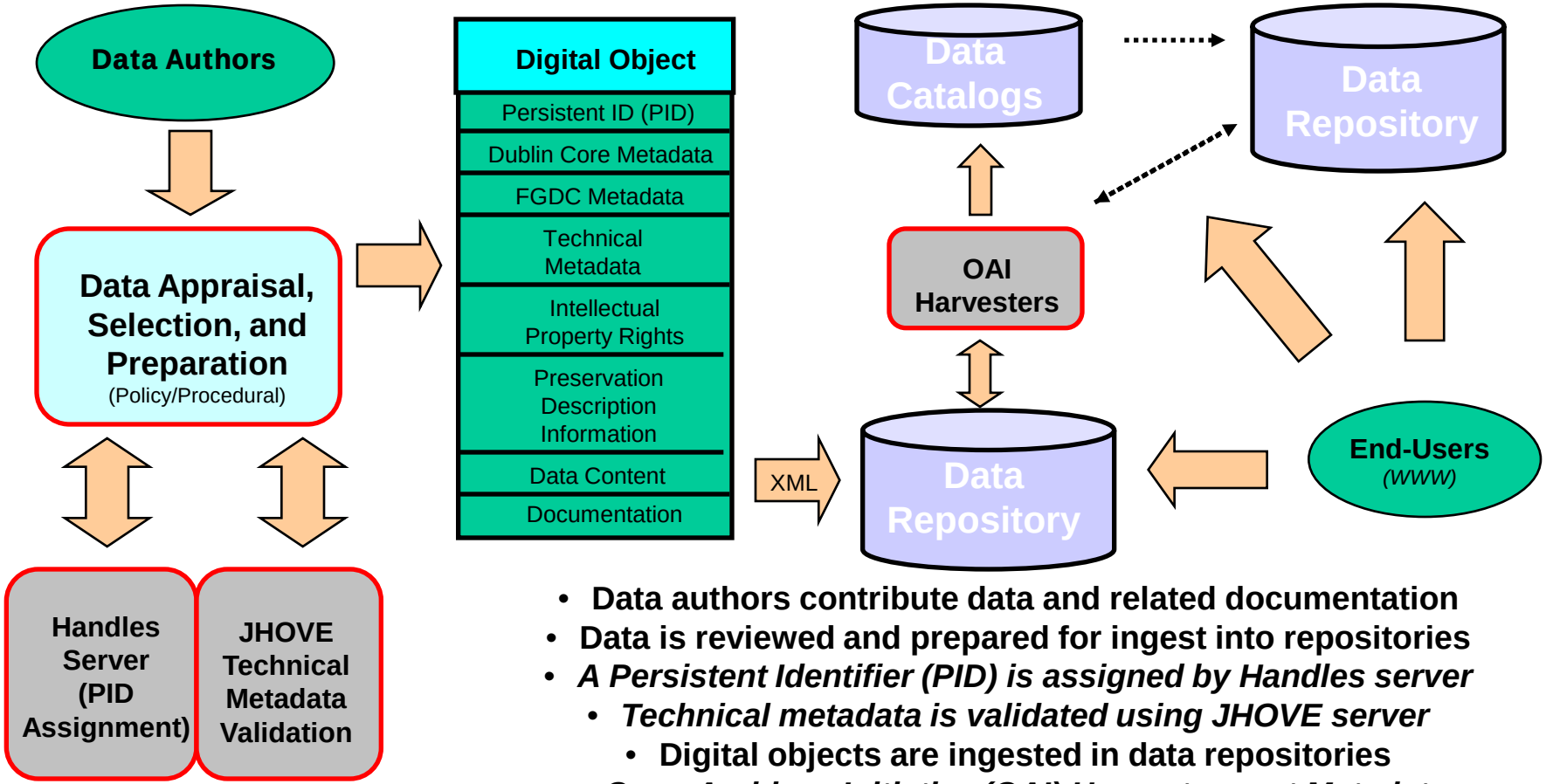
Digital Repository Development



CIESIN
Columbia University

The Earth Institute
AT COLUMBIA UNIVERSITY

WDC
World Data Center for Human Interactions in the Environment



- **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

fedora

VITAL

VTLS

Production Implementation

- Decision to implement Fedora for production digital repository
- Purchased VITAL with Fedora from VTLS
- Installed VITAL 3.0, including Fedora 2.1 on production and failover server
- Trained system and administrative staff on VITAL/Fedora
- Developed and tested procedures for ingesting and updating objects
- Purged data ingested during test period
- Successive upgrades to VITAL 3.1.1 and Fedora 2.2

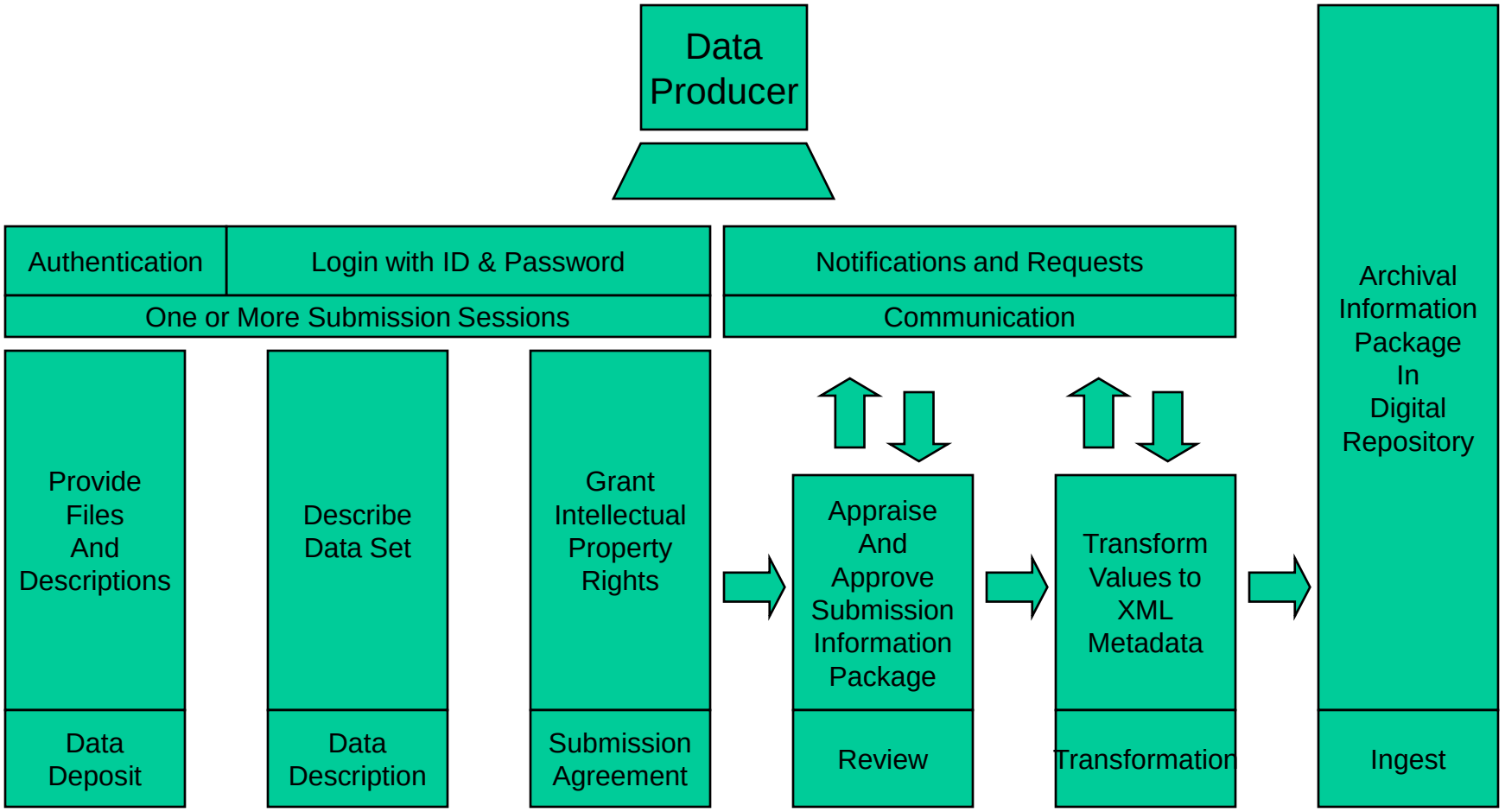
CIESIN Columbia University

The Earth Institute AT COLUMBIA UNIVERSITY

WDC

World Data Center for Human Interactions in the Environment

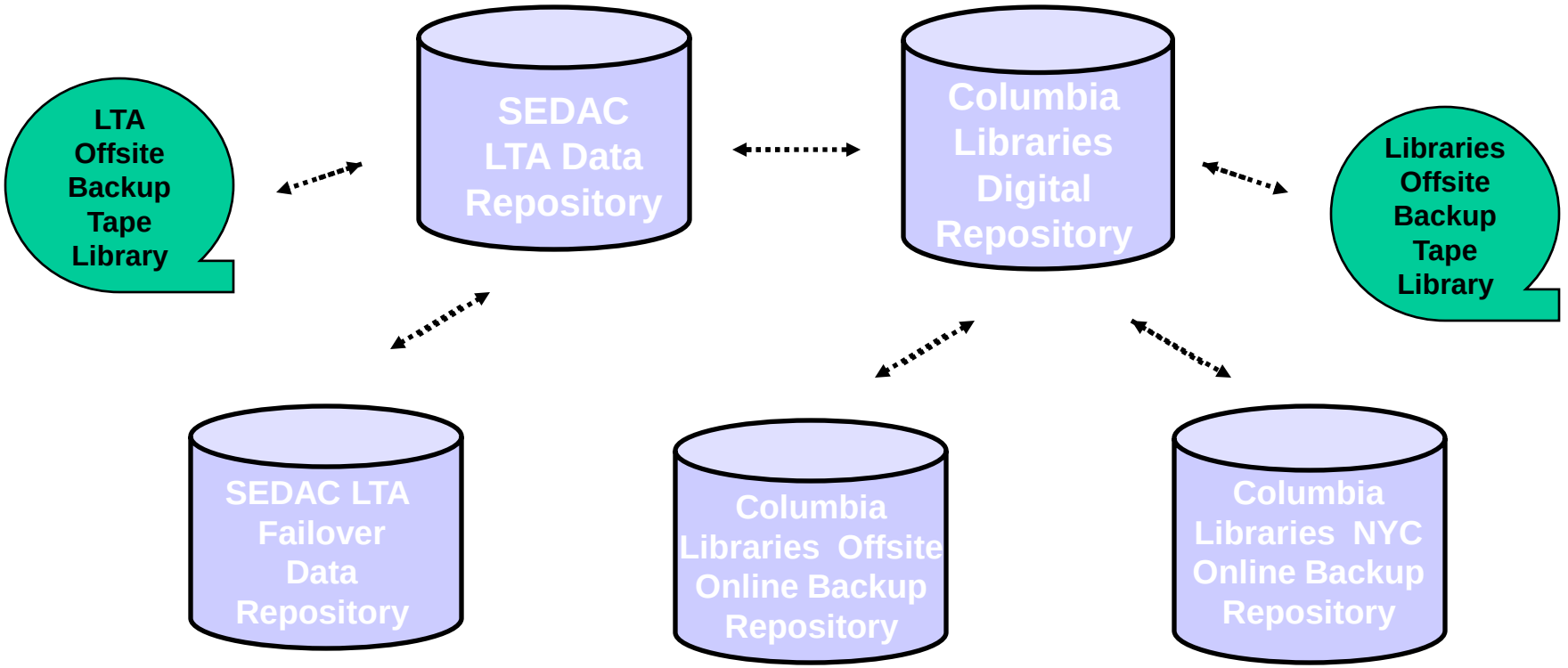
Model for Web-Based Data Submission and Workflow



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

Planned Tests for Transfer, Backup, and Recovery of Digital Objects



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/

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

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)

Sources Being Consulted for Self-Assessment of the SEDAC Long-Term Archive as a Trustworthy Digital Repository

- **Reference Model for an Open Archival Information System (OAIS).** Consultative Committee for Space Data Systems. January 2002. Adopted as: Space data and information transfer systems - Open archival information system - Reference model (ISO 14721:2003). http://public.ccsds.org/publications/archive/650x0b1.pdf

- **Producer-Archive Interface Methodology Abstract Standard.** Consultative Committee for Space Data Systems (CCSDS 651.0-B-1). May 2004. http://public.ccsds.org/publications/archive/651x0b1.pdf

- **Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), Version 1.0.** OCLC & CRL. February 2007 http://www.crl.edu/PDF/trac.pdf

- **Digital Repository Audit Method Based on Risk Assessment (DRAMBORA). The Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE).** 2007. http://www.repositoryaudit.eu/

- **Catalogue of Criteria for Trusted Digital Repositories, Version 1.** Nestor Working Group, Trusted Repositories - Certification. December 2006. http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf