CIESIN
Columbia University

The Earth Institute
AT COLUMBIA UNIVERSITY

WDC
World Data Center for Human Interactions in the Environment

Implementing a Digital Repository for the Preservation of Interdisciplinary Data

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

Prepared for Presentation to the
International Association for Social Science Information Services & Technology (IASSIST) 2008 Conference
Technology of Data: Collection, Communication, Access and Preservation

Stanford University, Palo Alto, California
May 30, 2008

CIESIN
Columbia University

The Earth Institute
AT COLUMBIA UNIVERSITY

World Data Center for Human Interactions in the Environment

Implementing a Digital Repository for the Preservation of Interdisciplinary Data

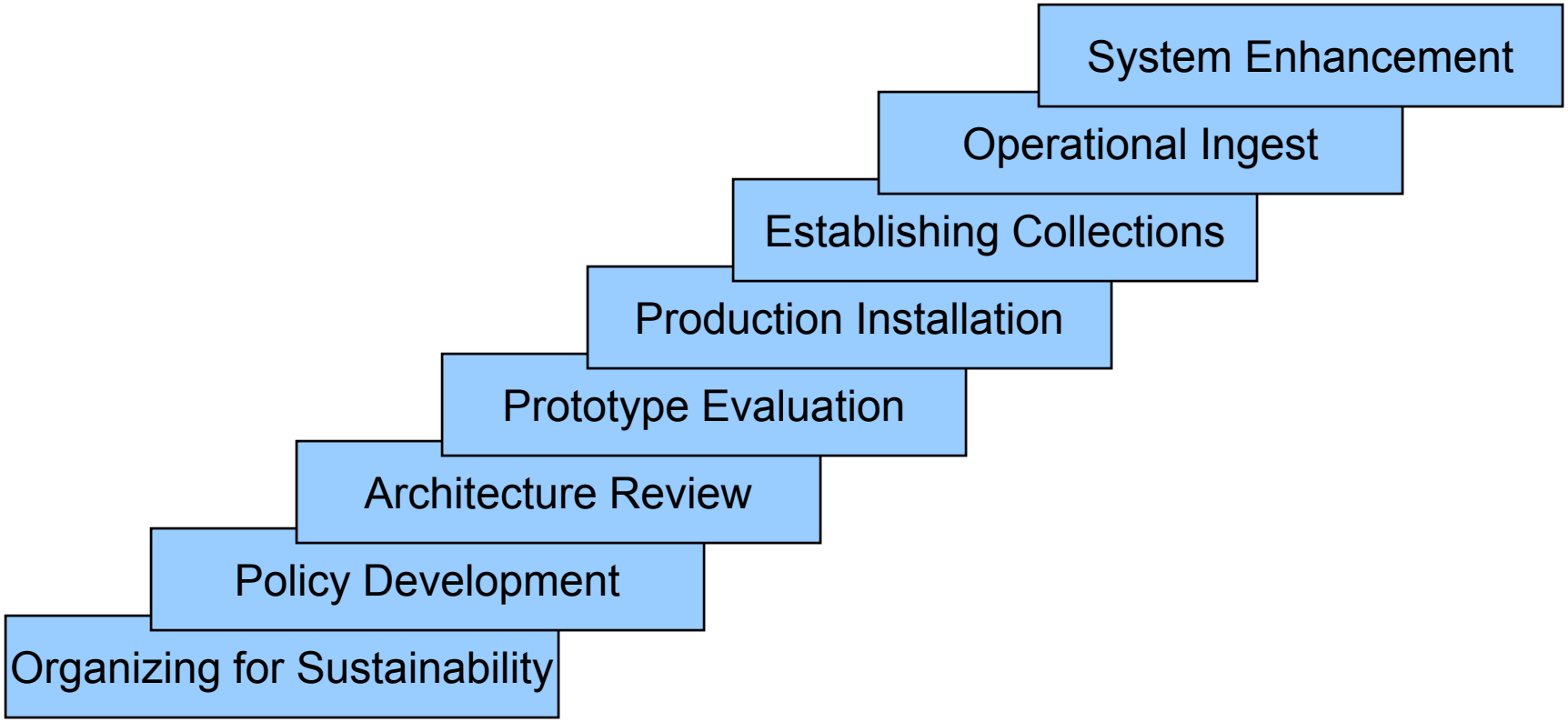
Robert R. Downs and Robert S. Chen

Digital scientific data created during the last few decades offer potential for analysis by future users and for integration with other data from different disciplines to support interdisciplinary analysis, discovery, decision-making, and education. However, significant barriers remain in managing and documenting such data sufficiently to meet the needs of future and interdisciplinary users. One possible approach to overcoming these barriers is to develop and implement digital repository systems within an appropriate institutional context. We report here on progress in implementing a digital repository using the Fedora open source software, working with the Columbia University Libraries. After discussing platform selection, feasibility testing, and collection development policy issues, we describe our experience with data migration and parallel ingest of data. We then discuss current system enhancements, challenges, and plans to improve capabilities for ingesting data and for enabling dissemination that supports future applications and use.

Challenges for Enabling Future and Interdisciplinary Use of Today's Data

- Provide sustainable long-term preservation of interdisciplinary data
- Facilitate acquisition of interdisciplinary data and descriptive information
- Ensure review and preparation of data for preservation and use
- Afford integration of data with other data to foster new analyses
- Foster discovery by current and future user communities
- Support interoperable access and use with new tools and services

Digital Repository Development



Organizing for Sustainability

- Experiment in Organizational Sustainability for Digital Preservation
- SEDAC Long-Term Archive Board Established with
  - Columbia University Libraries and Information Technology
  - The Earth Institute of Columbia University
  - SEDAC Project and Archives Management
- Contingency plans for Board representation and archive management in the event of a lapse in project funding

Policy Development

- Policies Pertaining to Digital Repository
  - CIESIN Policy for Preservation of Digital Resources
  - CIESIN Data and Information Management Policy
  - CIESIN Data Policy
  - CIESIN Digital Repository Collections Development and Use (Draft)
  - CIESIN Statement on the Responsible Use of Data and Information Resources (Draft)
- Collection-Level Policies Pertaining to Digital Repository
  - SEDAC Long-Term Archive Mission Statement (Draft)
  - SEDAC Long-Term Archive Management Structure (Draft)
  - SEDAC Operational Enhancements for Submission of Data to the Long-Term Archive (Draft)
  - SEDAC Long-Term Archive Management and Operations (Draft)

CIESIN Policy for Preservation of Digital Resource

CIESIN Columbia University
Center for International Earth Science Information Network
- Columbia University

CIESIN Policy for Preservation of Digital Resources

August 2004

CIESIN recognizes that the vulnerability of digital resources and evolving information technology pose considerable risks for facilitating persistent access to and use of digital resources and has developed the CIESIN Policy for Preservation of Digital Resources to manage this risk. This policy is also designed to establish practices for data stewardship that ensure the quality, integrity, confidentiality, security of digital resources over time; and to manage the intellectual property rights associated with digital resources archived at CIESIN indefinitely.

Guidelines:

In a proactive and ongoing effort to employ practices that preserve its digital resources, CIESIN Staff will:

1. Work with creators, owners, developers, and users to identify candidate data for archiving and to appraise, archive, obtain rights, describe, and document the preservation of specified digital resources and monitor their preservation status during their entire life cycle.
2. Work to identify and employ recognized standards and maintain currency of hardware, software, metadata and data formats to eliminate potential loss of digital resources resulting from storage media deterioration or technological obsolescence.
3. Identify and employ relevant, current practices and procedures for creating, acquiring, archiving, appraising, recovering, securing, and preserving digital resources.
4. Where relevant, obtain training on current practices for archiving, managing, and preserving digital resources to improve policies, plans, and procedures that provide enduring support for their discovery, access and use.
5. Work to identify and employ pertinent digital preservation resources, both external to and within Columbia University, the Columbia University Libraries, the Earth Institute at Columbia University, and CIESIN.
6. Routinely review, identify and improve the use of hardware, software, data formats, and standards to reduce the risk of storage media deterioration or technological obsolescence.
7. Routinely review, identify and improve policies, plans, and procedures to ensure completeness and coverage for contingencies such as changes in technology, standards and/or project requirements and capabilities, as well as to proactively engage in risk management.
8. Provide assurance for the quality of each digital resource by routine reporting on the status of archived data sets.
9. Identify research initiatives and results that would improve digital preservation planning and practices at CIESIN and at Columbia University as appropriate.
10. Work with creators, owners, developers, and users to identify and appraise candidate data for long term archiving.
11. CIESIN digital preservation policies and procedures will conform to other CIESIN policies such as CIESIN data policy and CIESIN data and information management policies as appropriate.

Architecture Review

- Reviewed commercial and open source systems to facilitate ingest, preservation, and access
  - Digital asset management systems
  - Electronic records management systems
  - Document management systems
  - Digital repository systems
- Decided to focus on open source approaches to avoid proprietary dependencies
  - Dspace
  - Eprints
  - Fedora
  - Greenstone
- Selected the Flexible Extensible Digital Object Repository Architecture (Fedora)
  - Developed by Cornell University and the University of Virginia
  - Modular approach to facilitate enhancement
  - Active user community of developers and implementers

Prototype Evaluation

- Installed Fedora on a development server as a prototype implementation for evaluation
- Ingested SEDAC datasets being reviewed for the SEDAC Long-Term Archive (LTA)
- Demonstrated ingest and access capabilities
- Evaluated operational prototype for a year prior to implementing Fedora digital repository in production

CIESIN Columbia University

The Earth Institute AT COLUMBIA UNIVERSITY

WDC

World Data Center for Human Interactions in the Environment

Searching the Fedora Prototype Implementation

Search Repository - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://webdev0:8080/fedora/search

Fedora Repository

Find Objects

Fields to display:

☑ pid	☐ bDef	☐ type
☐ label	☐ bMech	☐ format
☐ fType	☑ title	☐ identifier
☐ cModel	☐ creator	☐ source
☐ state	☐ subject	☐ language
☐ ownerId	☐ description	☐ relation
☐ cDate	☐ publisher	☐ coverage
☐ mDate	☐ contributor	☐ rights
☐ dcmDate	☐ date	

Search all fields for phrase: data *help*

Or search specific field(s): *help*

Maximum Results: 20

Search

pid	title
test:27	Environmental Treaties and Resource Indicators (ENTRI) - The Update of the Treaty Status Data
test:1	SEDAC Archives
test:3	SEDAC Long-Term Archive
test:30	HALOPH: A Data Base of Salt Tolerant Plants of the World
test:29	World Resources 1998-99: A Guide to the Global Environment: Environmental Change and Human Health (Data Tables)
test:31	Gridded Population of the World (GPW) Version 1
test:2	SEDAC Active Archive
test:102	Confidentiality Issues and Policies Related to the Utilization and Dissemination of Geospatial Data for Public Health Applications
test:25	Freedom in the World (1995-1996)
test:33	Gridded Population of the World (GPW) Version 2
test:34	Gridded Population of the World, Version 2: Ancillary Data

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

World Data Center for Human Interactions in the Environment

Searching the CIESIN Digital Repository

List of Titles | VTLS VITAL 3.1 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://vtlprod0.ciesin.columbia.edu:8080/vital/access/manager/Repository?query=information&public=true&start=16

CIESIN Digital Repository, Columbia University
Powered by VITAL

SIGN OUT

Home

Contributions
Contribute a New Item

Repository
Show All 43
Show Quick Collection 0

Search Advanced Search
data Search
Show Public Objects
Show Hidden Objects

Browse
Communities & Collections
By Title
By Creator
By Subject
By Date
Additional Resources

Highlights
Most Accessed Papers
Most Accessed Items
Most Accessed Authors
Recent Additions
Author Highlights

Work of The Day
CIESIN Data and Information Management Policy
Downs, Robert R. 2004

Resources
Help
About CIESIN

Home ▷ List of Titles
Your search on information produced these results.

List View | Icon View

Showing results 16 - 30 of 36.

First Previous | 1 2 3 | Next Last

	Title	Creator	Date	Full Text
	Environmental Vulnerability Index (EVI) 2004			
	Pilot 2006 Environmental Performance Index (EPI): Summary for Policymakers Brochure			
	Pilot 2006 Environmental Performance Index (EPI): Press Release			
	Summary For Policymakers: 2005 Environmental Sustainability Index: Benchmarking National Environmental Stewardship			
	Pilot Environmental Performance Index			
	2005 Environmental Sustainability Index: Benchmarking National Environmental Stewardship			
	Pilot Environmental Sustainability Index (ESI)			
	2008 Environmental Performance Index (2008 EPI)			
	Environmental Subset of Collection of Multilateral Conventions at the Fletcher School of Law and Diplomacy			
	Pilot 2006 Environmental Performance Index (EPI): Full 2006 EPI Report			
	Status of the CIESIN Digital Repository of Columbia University: Report to the SEDAC Long-Term Archive Board	Downs, Robert R.	2007	
	Minutes of the SEDAC Long-Term Archive Board Meeting of July 6, 2006	Downs, Robert R.; Center for International Earth Science Information Network (CIESIN), Columbia University	2006	–
	Minutes of the SEDAC Long-Term Archive Board Meeting of September 8, 2005	Downs, Robert R.; Center for International Earth Science Information Network (CIESIN), Columbia University	2005	–
	CIESIN Data and Information Management Policy	Downs, Robert R.	2004	
	CIESIN Data Policy	Downs, Robert R.; Lenhardt, W. Christopher	2004	–

First Previous | 1 2 3 | Next Last

Creator
Downs, Robert R. (6)
Center For International Earth Science Information Network (CIESIN), Columbia University (6)
Lenhardt, W. Christopher (1)
Center For International Earth Science Information Network (CIESIN), Columbia University, (1)
Gitelman, Yitzhak (1)

Subject
Permissions Form (4)
Intellectual Property Rights (4)
Restrictions (4)
Digital Asset Management (3)
License Agreement (3)
Data Management (2)
Information Management (2)
Data License (2)
Data Agreement (2)
Resource License (2)

Format Type
Adobe Acrobat PDF (22)
XML Document (18)
Text Document (16)
Microsoft Word Document (14)
(12)
Application/x-msword (9)
Microsoft Excel Document (7)
TIFF Image (1)
Microsoft Powerpoint Presentation (1)
HTML Document (1)

Resource Type
Working Paper (5)
Form (5)
Report (1)

Disclaimer | Copyright | Contact | Back To Top

English (United States) Change Locale

Customize Page Content

Edit Page Content
Annote this page with simple markup.

Reporting

View Statistics
View statistics and usage patterns for visitors to the repository.

Configure Statistics
Configure the displaying and recording of statistics and usage patterns for visitors to the repository.

View Current Activity
View a snapshot of recent user activity.

View Administrative Reports
Create and run reports on content, usage statistics and metadata stored in the repository.

Validate Repository Links
Find external URLs that may have moved or changed since the content was originally added.

Repository

Repository Indexing
Control the status of the background indexing service.

Edit Indexing Configuration
Edit the search indices and XPath definitions used to index the repository.

Establishing Collections

- Center for International Earth Science Information Network (CIESIN) Administrative Archive
  - Center for International Earth Science Information Network (CIESIN) Records and Documents
- Socioeconomic Data and Applications Center (SEDAC) Active Archive
  - SEDAC Active Archive
  - SEDAC Active Archive Documents and Records
- Socioeconomic Data and Applications Center (SEDAC) Administrative Archive
  - SEDAC User Working Group
- Socioeconomic Data and Applications Center (SEDAC) Long-Term Archive
  - SEDAC Long-Term Archive Data
  - SEDAC Long-Term Archive Documents and Records

CIESIN Columbia University

The Earth Institute AT COLUMBIA UNIVERSITY

World Data Center for Human Interactions in the Environment

CIESIN Digital Repository Communities and Collections Screen

CIESIN Digital Repository, Columbia University
Powered by VITAL

SIGN IN

Home

Repository
Show All 44
Show Quick Collection 0

Search Advanced Search
Search

Browse
Communities & Collections
By Title
By Creator
By Subject
By Date
Additional Resources

Highlights
Most Accessed Papers
Most Accessed Items
Most Accessed Authors
Recent Additions
Author Highlights

Work of The Day
CIESIN Data and Information Management Policy
Downs, Robert R. 2004

Resources
Help
About CIESIN

Home ▷ Communities & Collections

Communities & Collections

The following list represents the communities represented by this repository and collections contained within them. Click on a name to view that community or collection page.

Center for International Earth Science Information Network (CIESIN) Administrative Archive

Center for International Earth Science Information Network (CIESIN) Records and Documents

Socioeconomic Data and Applications Center (SEDAC) Active Archive

SEDAC Active Archive
SEDAC Active Archive Documents and Records

Socioeconomic Data and Applications Center (SEDAC) Administrative Archive

SEDAC User Working Group

Socioeconomic Data and Applications Center (SEDAC) Long-Term Archive

SEDAC Long-Term Archive Data
SEDAC Long-Term Archive Documents and Records

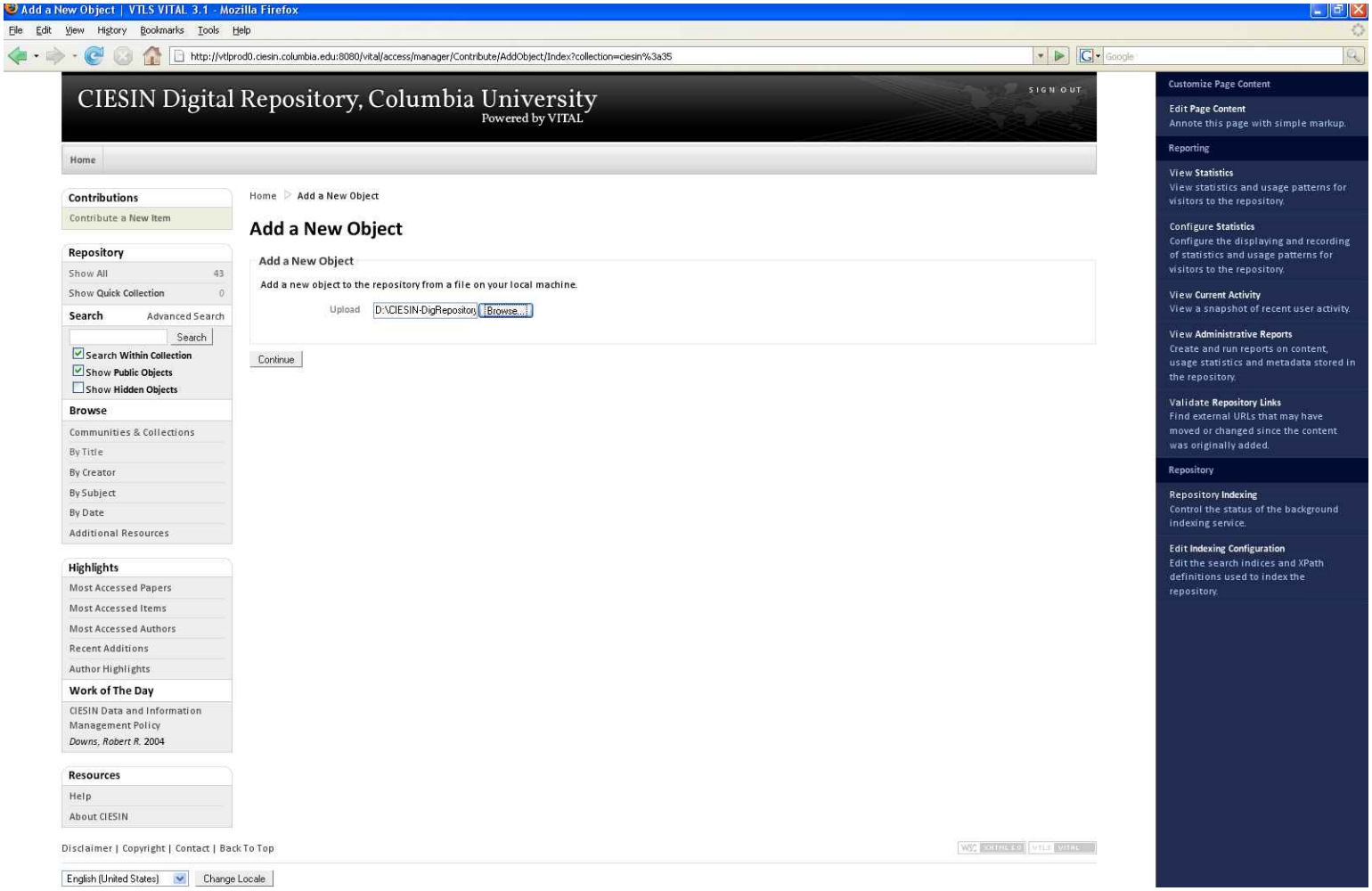
Disclaimer | Copyright | Contact | Back To Top

English (United States) Change Locale

Operational Ingest

- Data Migration
  - Migration of data previously archived on portable media
- Parallel Ingest
  - Ingest of data during accession in parallel with traditional archiving
- Self-Submission Workflow
  - Submission by data producers and their representatives

Adding a New Object Using the Administrative Interface



CIESIN Columbia University

The Earth Institute AT COLUMBIA UNIVERSITY

World Data Center for Human Interactions in the Environment

Describing Object Using the Administrative Interface

Object Properties | VTLS VITAL 3.1 - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://vtlprod0.ciesin.columbia.edu:8080/vital/access/manager/administration/ObjectProperties/ciesin:48?expert=title%3a%222008+Environmental+Performance+Index+(2008+EPI)%22

CIESIN Digital Repository, Columbia University
Powered by VITAL

SIGN OUT

Home

Contributions
Contribute a New Item

Repository
Show All 43
Show Quick Collection 0

Search Advanced Search
Search
Show Public Objects
Show Hidden Objects

Browse
Communities & Collections
By Title
By Creator
By Subject
By Date
Additional Resources

Highlights
Most Accessed Papers
Most Accessed Items
Most Accessed Authors
Recent Additions
Author Highlights

Work of The Day
CIESIN Data and Information Management Policy
Downs, Robert R. 2004

Resources
Help
About CIESIN

Home ▷ Object Properties

Object Properties

Object Properties

State: Active
Label: 2008 Environmental Performance Index (2008 EPI)
Content Model:
Created: May 9, 2008 12:58:55 PM EDT
Last Modified: May 9, 2008 1:19:37 PM EDT
Owner:

Update Object

Datastreams

The table below lists all the datastreams available in this object.

Version	Label	MIME Type	Created	State
SOURCE1	2008 Environmental Performance Index Main Report (23 Jan. 2008)	Adobe Acrobat PDF (application/pdf)	May 9, 2008 1:01:29 PM EDT	A
SOURCE10	2008EPI_rankingsandscores_23Jan08.pdf	Adobe Acrobat PDF (application/pdf)	May 9, 2008 1:16:04 PM EDT	I
SOURCE11	2008EPIPolicymakerSummary.pdf	Adobe Acrobat PDF (application/pdf)	May 9, 2008 1:16:35 PM EDT	I
SOURCE12	EPI2008IndicatorsMetadata.pdf	Adobe Acrobat PDF (application/pdf)	May 9, 2008 1:18:36 PM EDT	I
SOURCE2	E-Mail: 2008 EPI archive ingest form.eml	application/octet-stream	May 9, 2008 1:17:03 PM EDT	I
SOURCE3	2008 Environmental Performance Index (2008 EPI)	Microsoft Excel Document (application/msexcel)	May 9, 2008 1:05:14 PM EDT	A
SOURCE4	CIESIN Ingest Form	Microsoft Word Document (application/msword)	May 9, 2008 1:17:39 PM EDT	I
SOURCE5	E-Mail: Formal UWG Approval of 2008 Environmental Performance Index data set.eml	application/octet-stream	May 9, 2008 1:18:08 PM EDT	I
SOURCE6	CIESIN Permission to Use Data, Signed April 7, 2008 by Daniel Esty, Yale Center for Environmental Law and Policy	Adobe Acrobat PDF (application/pdf)	May 9, 2008 1:15:07 PM EDT	I
SOURCE7	2008EPI_Data.xls	Microsoft Excel Document (application/msexcel)	May 9, 2008 1:19:37 PM EDT	I

Object Properties

Cancel
Return to the Item Display page.

Download FOXML
Download a full copy of the object's exported FOXML metadata.

Update Dublin Core
Update the content or properties of this object's Dublin Core datastream.

Add Datastream
Add a datastream to the current object from a file on your local machine or from a predefined datastream model.

Delete Object
Delete this object permanently.

Customize Page Content

Edit Page Content
Annote this page with simple markup.

Reporting

View Statistics
View statistics and usage patterns for visitors to the repository.

Configure Statistics
Configure the displaying and recording of statistics and usage patterns for visitors to the repository.

View Current Activity
View a snapshot of recent user activity.

View Administrative Reports
Create and run reports on content, usage statistics and metadata stored in the repository.

Validate Repository Links
Find external URLs that may have moved or changed since the content was originally added.

Repository

Repository Indexing
Control the status of the background indexing service.

Edit Indexing Configuration
Edit the search indices and XPath definitions used to index the repository.

CIESIN Columbia University

The Earth Institute AT COLUMBIA UNIVERSITY

World Data Center for Human Interactions in the Environment

Self-Submission and Review Workflow Interface

VALET Resource Submission (Deposit Form) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://vtlprod0.ciesin.columbia.edu:8888/cgi-bin/valet/submit.cgi

CIESIN Digital Repository, Columbia University (Deposit Form)

Instructions

You are about to begin the process for depositing resources into the digital repository at the Center for International Earth Science Information Network (CIESIN) of Columbia University. You will be asked to attach the file(s) you wish to deposit and then provide some information to describe them.

You will be presented with the following steps:

1. Information collection
   - About the files
   - Select a resource type
   - General resource description
   - About the creator
   - About the association with a faculty/school, etc.
   - About the rights
2. Review/Change the information you entered.
3. Confirm submission.

Mandatory fields are marked with an *.

About the files

help?

Attach file(s) Browse... Description: upload

		Original Filename	File size	Description
Download	Remove	104131.afs.lis	140	Readme File
Download	Remove	104131.afs.lis	140	afs directory listing of files to be disseminated

General resource description:

help?

* Title: lerant Plants of the World

Description/Abstract: exploited land, preferably on land degraded by man's profligate behavior, in the event that appropriates halophytes can be identified and developed to make use of these water sources.

* Year: 1987

Language: English

Coverage: Global

Keyword(s):

+ Add additional keywords

Subject(s): Farming, Biota, Environment

Choose all subjects that apply

Inland Waters
Location
Oceans
Planning Cadastre
Society

About the creator

Digital Repository System Enhancement

- Conduct self-assessment for compliance with OAIS framework as a trustworthy digital repository
- Improve capabilities for self-submission of data
- Customize workflow processes for review and approval for ingest
- Explore opportunities to record provenance events
- Establish capabilities for batch ingest of objects
- Enable access control to collections, objects, and datastreams
- Experiment with access to datastreams from applications and services
- Test the system's ability to retrieve different combinations of objects in support of different user needs for retrieval and access