Data Conservancy
Collection Policy Mockup 0.55

Palmer, Cragin, Weber, Duerr, Khalsa,
Choudhury, Dilauro, Evans, & Rippin

The Data Conservancy embraces a shared vision: scientific data curation is a means to collect, organize, validate, and preserve data so that scientists can find new ways to address the grand research challenges that face society. The Data Conservancy will research, design, implement, deploy, and sustain data curation infrastructure for cross-disciplinary discovery with an emphasis on observational data

In support of this mission, the Data Conservancy seeks to collect and curate data that meet the needs of its target scientific communities and advance the Data Conservancy's research and development program.

1. Introduction

The Data Conservancy (DC) does not follow a universal method for collection development, but instead emphasizes the collection of data that support its mission to advance science and the state of the art in data curation. This is accomplished by:

- Providing systematic access to digital data of value for scientific research.
- Archiving and preserving acquired data and sustaining links to distributed data.
- Providing a data registry for systematic representation of data, and search and browse functions across data holdings.
- Supporting open access to data.

Data may be contributed through submission processes, acquired through standing partnerships, as part of a regularly scheduled ingest and/or solicited by the DC. A DC curatorial team will assign a level of service to submitted data based on the scope and criteria described below.

2. Communities Served

The DC communities represent a broad spectrum of scientific domains that generate and use data. Data contributors include individuals and groups in astronomy, life sciences, earth sciences, and social sciences. Researchers in these domains make up the primary user communities as well, but data in the DCS will also be valuable to educators, students, citizen scientists, policy makers, and others who can benefit from access to it.

3. Scope of collections

The scope of data of interest to the DC is aligned with the following four domain areas:

3.1 Astronomy

Astronomical data originate from telescopes on the ground and on orbiting spacecraft, and comprise both pointed observations (where a telescope is trained on a

particular object or field of view) and survey observations (where a telescope systematically scans a large region of sky). Data may be spatial (images), spectral, or time series, or hybrid combinations (such as time-resolved spectroscopy or "data cubes"—image stacks with spectral or velocity planes). Of particular interest in the context of the DC is enabling the discovery, intercomparison, and synthesis of observations from multiple telescopes and instruments, a goal shared with the Virtual Astronomical Observatory.

Current Areas of Collection Emphasis:
The Sloan Digital Sky Survey is serving as an initial testbed for data ingest. This five-spectral band survey is one of the most widely used data sets in astronomy owing to the uniformity and quality of the data processing and calibration. It consists of the five-band imaging data, a source catalog, and spectra for ~one million quasars.

Another primary focus is on the data associated with publications in peer-reviewed journals. Even in this age of digital journals, it is most often the case that data are presented only in tabular or graphical form, and the underlying digital representations of the data are not part of the official scientific record. In collaboration with the VAO, DC will build a data capture and preservation process for the digital data sets represented and referenced in the literature. The resulting archive of highly processed datasets will be accessible through DC and VAO data discovery and delivery servicers, with links provided to the journal publishers for incorporation into their citations.

3.2 Life Sciences

Life science holdings include non-experimental and experimental observational data and derived (computed) data. The data relate to phenomena that occur across a time scale that ranges from less than a millisecond to over 3 billion years, and on spatial scales that range from subatomic to 20,000 kms. These products span across biological subdisciplines including molecular biology and biochemistry, genetics, cell biology, anatomy physiology, ecology, taxonomy and evolution. The needs of the agricultural and medical sciences, including biomedical or bioengineering data, are not included within the scope of Data Conservancy. Our interests overlap with DataOne. Socio-biological data will be considered on a case-by-case basis (Thessen, 2010).

Future Areas of Collection Emphasis:
Broadly, occurrence records and data related to the distribution of organisms, ecological interactions, components of a taxonomic ontology for Life Sciences, including data associated with data management plans.

3.3 Earth Sciences

Earth science data supports a number of research communities such as geoinformatics and fluid-earth researchers, this includes domains such as geology,

volcanology, soil science, oceanography, cryospheric science, and atmospheric science.

Future areas of Collection Emphasis:
Sea ice data, Earth deformation data, climate simulations.

3.4 Social Sciences

Generally, social science data consists of quantitative and qualitative data originating from research or administrative records from which statistics are produced, as well as from surveys, questionnaires and interviews. In addition to numeric and textual files, data may also include audio, video, encoded text files, arrays, URIs, etc. For the Data Conservancy, social science data is collected to support interdisciplinary analysis and is focused on material that complements other data in the DC.

Future areas of Collection Emphasis:
An initial primary focus for the collection will be on data related to urban vulnerability to climate change and weather hazards. These data come from both environmental and social science domains, including such information about urban areas as temperature, carbon emissions, mortality rate, population size and density, age structure, gender composition, education attainment, income level, and gross domestic product. Existing sociodemographic, economic, and environmental data on urban areas involve different spatial and temporal scales. Datasets generated from individual studies are usually not included in the academic publication process. It is of great importance to develop an integrated data system to collect, organize, synthesize, and preserve data from individual projects in this interdisciplinary field.

Initially, the collection will include data sets represented in and resulting from a meta-analysis of peer-reviewed articles on urban vulnerability. It will also include some data sets from specific projects (e.g., ADAPTE). These data will be prioritized for collection when not included in other social science or environmental data repositories.

4. Data Types

The types of data needed to support the communities served by DC are varied and will continue to evolve over time. DC will accept numerous types and formats of primary and analyzed data, including digitized content such as photographs or logbooks. However, some information objects such as dissertations, e-journals, article preprints, and physical specimens are not appropriate for submission to the DC. In such cases an effort will be made to identify an appropriate alternative repository (please see ‘levels of service’, below).

5. Criteria for Inclusion

Data that fit into the scope of the DC’s research domains will be evaluated based on the criteria below. While there are no strict metrics used, DC seeks data that meet most or all of the criteria below. In addition, the costs associated with acquisition and long-term service needs will be weighed along with contribution to the DC mission.

Value: Data that are useful to a wide constituency across multiple disciplines, or that are a significant asset for a single research area are considered highly import for DC. Emphasis is placed on data that extend or add value to existing DC data, have long-term value, and/or uniquely fill unmet research needs.

Uniqueness: Data that represent a time period, subject, or phenomena not currently accounted for or underrepresented in public data repositories will be a high priority for DC. Unique data that expand currently available holdings or represent new dimensions of inquiry are also prioritized.

Risk of Loss: The DC is interested in preserving data that may be lost without proper digital curation; this includes data that is in danger of obsolescence, currently stored on aging hard drives or media, orphaned from inactive labs or long-term research projects, or data for which there are no other repositories.

Preservation Readiness: Data are 'preservation ready' when the producer has provided DC all of the documentation needed for accurate long-term preservation. This documentation includes workflow, provenance, and contextual information used in the creation or interpretation of the submitted data. DC will curate data without complete metadata or documentation, but such submissions will require ample coordination between the repository staff and producers. Depositors must also understand that the level of service offered by DC will be dependent on the completeness of the documentation they provide.

DC will follow OAIS guidelines, collecting the following types of documentation about each submission:

Element	Collection Information	Examples
Representation Information	This should at minimum include a description of the significant properties of a data object, both Structure (format) and Semantic(human language) components	ASCII, HDF, TIFF, OWL, etc.
Provenance	Describes the source of the deposit, who has had custody since its origination, and its processing history.	PI, Project Affiliation, Institutional affiliation, NASA data level.
Context	Context might answer questions like why the content was produced, and it may include a description of how it relates to another object that is available (either internally or externally.)	Abstract of proposal for project that produced the data. Calibration information, related data sets,

		collection metadata, etc.
Reference	Provides one or more identifiers by which the content may be uniquely identified.	title, file names, formats, creators, URI or DOI.
Fixity	Provides for the stability of the content, for instance this may involve a check sum over the of a digital Information Package	Check-sum, wrapper.

OAIS, Section 2 Preservation Description Information

Cost: For data that require extensive investment of resources in processing, preservation, or other curatorial activities the potential value of the data should strongly outweigh the cost of acquisition. An assessment of the level of service (see below) will be performed on all submissions.

Funding Requirements: Research projects that require a data management plan and have funding allocated for archiving and sharing data will receive priority over those that do not, but these submissions should align with the general scope of DC research domains and meet other basic criteria for inclusion. Pre-award or proposal agreements are possible through coordination with DC.

Ruth Duerr 4/7/11 10:34 PM

Comment [1]: From SJSK: The importance of having a data management plan might warrant calling out in a separate item?

Copyright and Right of Deposit: Data free of licensing restrictions are preferred. Embargo periods for ongoing or recently completed research projects can be negotiated. Agreement of Deposit statements are required for all submissions.

Resource Allocation Assessment
In addition to available funding, data submissions will also be assessed for the level of processing that will be required to prepare the data for re-use. For example this assessment will include consideration of the authentication requirements, resulting in preference being given to data that is unencumbered, i.e. immediately available for unrestricted and public use. These types of data are expected to require lower levels of systems management services, and thus lower maintenance costs over time.

6. Reappraisal

All data hosted by DC will be reappraised on a regular schedule. Reappraisal assessments are based on use, current collection emphases, relation to other holdings, and the general quality of the information package.

Reappraisal may result in a downgrade of the level of support for data, but when appropriate, may also result in more resources being applied to raise the level of service. Data that are

expected to be of high value in the future will be identified early in the curation process to ensure their retention. These data may receive limited support or varied support over the long term. Data that no longer comply with policies or legal requirements may be removed from the archive after appropriate assessment. Disposal does not necessarily mean that data will be destroyed; they may be removed from the archive and transferred to another site. DC will alert depositors and community stakeholders well in advance of any change in service-level.

7. Levels of Service [DC service levels are under development in another document]

Data that are ingested, linked to, or aggregated by the DC will receive varying degrees of service. Some lower levels will include minimum preservation and access; higher levels will guarantee services such as format migration and normalization.

DC is committed to ensuring long-term archiving options for all submissions. If submitted data are outside the scope of the DC collection policy, curatorial team members will work with contributors to find an appropriate repository or suggest alternative services available for preserving digital data.

8. Collection Board

Acquisition of data that are determined to meet the DC scope and criteria for inclusion is subject to approval by a curatorial team that broadly acts in consultation with the Collection Board. The Collection Board includes members of DC and the DC service communities that do not have conflicts of interest in maintaining the collection policy. Board members come from a variety of professional backgrounds and include scientific domain experts, repository technical staff, and representatives of other DC research and education constituencies.

Ruth Duerr 4/7/11 10:40 PM
Comment [2]: SJSK: I would think DC members would very much have an interest in maintaining a collection policy. Do you mean DC members who do not have a conflict of interest when applying the collections policy to a particular submission?

9. Maintenance of this Policy

This collection policy is a 'living document' driven by the mission of the DC. Maintenance of this policy will be coordinated with the development of the repository and related DC policies, such as user agreements and service levels. The collection policy is subject to annual review by the Collection Board. The DC welcomes public review and comments on this document.

Sources

Caplan, Priscilla. 2009. Understanding PREMIS. Washington, DC: Library of Congress. Available at www. loc. gov/standards/premis/ understanding-premis. Pdf

Data Conservancy Data Model Y1: https://wiki.library.jhu.edu/display/DATACON/Design+Document+-+Data+Model

Hanisch, Robert; Choudhury, Sayeed (2009). The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation.

ISO. Space data and information transfer systems - Open archival information system - Reference model (ISO 14721:2003), 2003.

Thessen, Anne. (2010). Life Sciences White Paper. Available internally at: https://wiki.library.jhu.edu/display/DATACON/Life+Sciences

Policies Consulted:

ICPSR Collection Policy: icpsr.umich.edu/icpsrweb/ICPSR/org/policies/colldev.jsp

ADS Collection Policy: ads.ahds.ac.uk/project/collpol.html

IMLS Digital Collections & Content, Opening History Collection Policy: http://imlsdcc.grainger.uiuc.edu/docs/CollectionDevelopmentPolicy.pdf

NSDL Collection Policy: *onramp.**nsdl**.org/eserv/.../**NSDL_Collection**_Development_**Policy**.pdf*

NSIDC Service Model, BADC collection levels (documents shared privately)