Harvard Data Commons

Mercè Crosas, Ph.D. @ mercecrosas
University Research Data Management Officer
Chief Data Science and Technology Officer, IQSS

With Bill Barnett, Paul DiBello, Piotr Sliz, Stu Snydman, Scott Yockel
Harvard University

2020 Open Cloud Workshop
What is a Data Commons?
“... a data commons brings together (or co-locates) data with cloud computing infrastructure and commonly used software services, tools & applications for managing, analyzing and sharing data to create an interoperable resource for a research community”

https://medium.com/@rgrossman1/a-proposed-end-to-end-principle-for-data-commons-5872f2fa8a47
DATA COMMONS COMPONENTS

Active Research:
Collection, Cleaning, Process, Analysis, Exploration, Visualization

Research Tools

Data Repository

Cloud computation and storage

Computing Resources

Storage (security/data enclaves)

Researcher Facing

Data Management & long-term access:
Global Persistent IDs, Metadata, Data Dictionaries, Provenance, Versions, Access controls, Data curation
EXAMPLE: AUSTRALIAN RESEARCH DATA COMMONS

A Data Commons at the National Level

Nectar Research Cloud
Our Nectar Research Cloud is Australia’s first federated research cloud. This service provides Australia’s research community with computing infrastructure and software. Researchers can store, access, and run data, remotely, rapidly and autonomously.

Research Data Australia
Research Data Australia (RDA) is an online portal for finding research data and associated projects, researchers, and data services. You can find, access, and reuse data for research from over one hundred Australian research organisations, government agencies, and cultural institutions.

Identifier Services
We provide services to create and manage persistent identifiers for research data, research samples, files, documents or other digital objects. Identifiers connect objects to important context surrounding the objects and adds value to them.

Research Vocabularies Australia
Research Vocabularies Australia (RVA) makes it easy to find and use controlled vocabularies used in research. It also makes it possible for Australian research organisations to publish, re-purpose, create, and manage their own controlled vocabularies.
EXAMPLE: GENOMIC DATA COMMONS

A Data Commons for a scientific domain

The Next Generation Cancer Knowledge Network

Cases by Major Primary Site

The NCI's Genomic Data Commons (GDC) provides the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs at the NCI Center for Cancer Genomics (CCG), including The Cancer Genome Atlas (TCGA) and...
What is the Harvard Data Commons?
HARVARD DATA COMMONS: A PROPOSAL

At its early stage:
- Pilots in 2020
- Defining architecture and use cases
A collaboration across Harvard units and schools:

- IT/Research Computing
- Library
- Harvard Dataverse
- Schools (initially Faculty of Arts & Sciences, Medical School, Business School)
Research and Data Management Tools

Interoperability Middleware

Extract and generate:
- Metadata
- Workflows
- Provenance
- Research Objects
- Containers

from researcher's tools and computing environments

Leverage and expand computational and reproducibility platforms (e.g., WholeTale, Renku)

Published Data

DRS Collections Preservation at Harvard Library

HARVARD Dataverse

Storage, Computing

Notary Service

NERC NESE MGHPC

Trusted Remote Storage Agents

Leverage ImPACT

protocols.io

DMP Tool

OPEN SCIENCE FRAMEWORK

globus

jupyter Studio

RENU K 連句

qualtrics REDCap
HARVARD DATA COMMONS: NEW FEATURES

- Dashboard to find and access unpublished and published datasets from Harvard researchers
- An interoperability middleware to add metadata and provenance to outputs of research tools (e.g., create a container with notebook + data + metadata + provenance) and deposit to the repository
- Multiple Trusted Remote Storages to host sensitive and large data while (only) metadata is in the repository