

# 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 Snyderman , Scott Yockel  
Harvard University

2020 Open Cloud Workshop

What is a Data Commons?

# DATA COMMONS DEFINITION

"... 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

Researcher Facing

Research Tools

Data Repository

## Data Management & long-term access:

Global Persistent IDs  
Metadata  
Data Dictionaries  
Provenance  
Versions  
Access controls  
Data curation

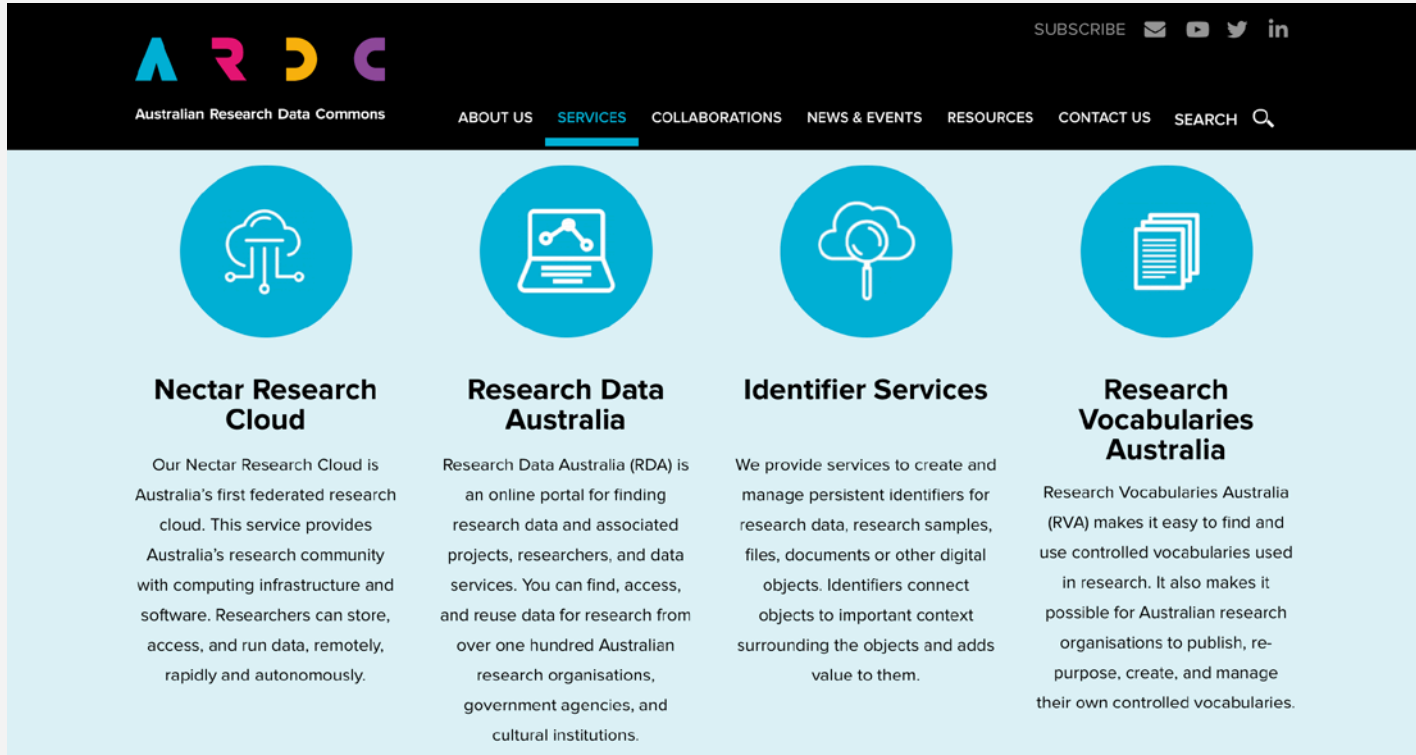
Cloud  
computation  
and storage

Computing Resources

Storage (security/data enclaves)





# EXAMPLE: AUSTRALIAN RESEARCH DATA COMMONS


## A Data Commons at the National Level



The image shows a screenshot of the Australian Research Data Commons website. At the top, there is a black navigation bar with the ARDC logo on the left and social media icons (SUBSCRIBE, email, YouTube, Twitter, LinkedIn) on the right. Below the navigation bar, the text "Australian Research Data Commons" is displayed. The main content area features four service cards, each with a blue circular icon and a title. The cards are: 1. Nectar Research Cloud (cloud icon), 2. Research Data Australia (laptop icon), 3. Identifier Services (cloud with magnifying glass icon), and 4. Research Vocabularies Australia (stack of papers icon). Each card contains a brief description of the service.

**ARDC**  
Australian Research Data Commons

SUBSCRIBE    

ABOUT US **SERVICES** COLLABORATIONS NEWS & EVENTS RESOURCES CONTACT US SEARCH 

**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

NATIONAL CANCER INSTITUTE - CANCER.GOV

NIH NATIONAL CANCER INSTITUTE  
Genomic Data Commons

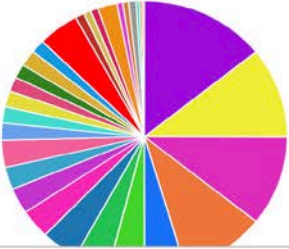
CCG Web Site | Contact Us | [Launch Data Portal](#) | GDC Apps

Search this website

[About the GDC](#) | [About the Data](#) | [Analyze Data](#) | [Access Data](#) | [Submit Data](#) | [For Developers](#) | [Support](#) | [News](#)

### 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


### Analyze Data



The **GDC Data Analysis, Visualization, and Exploration (DAVE) Tools** allow users to interact intuitively with the GDC data and promote the development of a true cancer genomics knowledge base.

[→ More about Analyzing Data](#)

### Access Data



The **GDC Data Portal** provides a platform for efficiently

What is the Harvard Data Commons?

# HARVARD DATA COMMONS: A PROPOSAL

At its early stage:

- Pilots in 2020
- Defining architecture and use cases



# HARVARD DATA COMMONS: A COLLABORATION

A collaboration across Harvard units and schools:

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

## Active Data

## Published Data

### 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)

A  
P  
I



### HARVARD Dataverse

DRS  
Collections  
Preservation at  
Harvard Library

Storage,  
Computing

Notary  
Service

Trusted  
Remote  
Storage  
Agents

Leverage  
ImPACT

NERC  
NESE  
MGHPCC



# 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