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?

DATA COMMONS DEFINITION

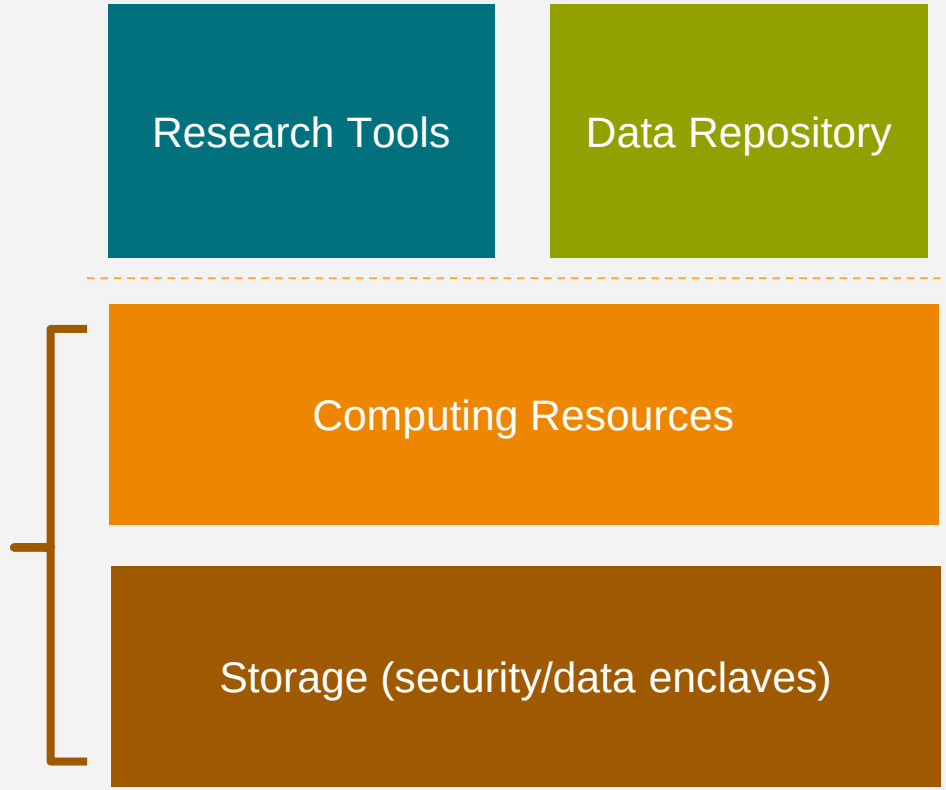
"… a data commons brings together (or co-locates) data with cloud computing infrastructure and commonly used software services, tools & applications for managing, analyzing and sharing data to create an interoperable resource for a research community"

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

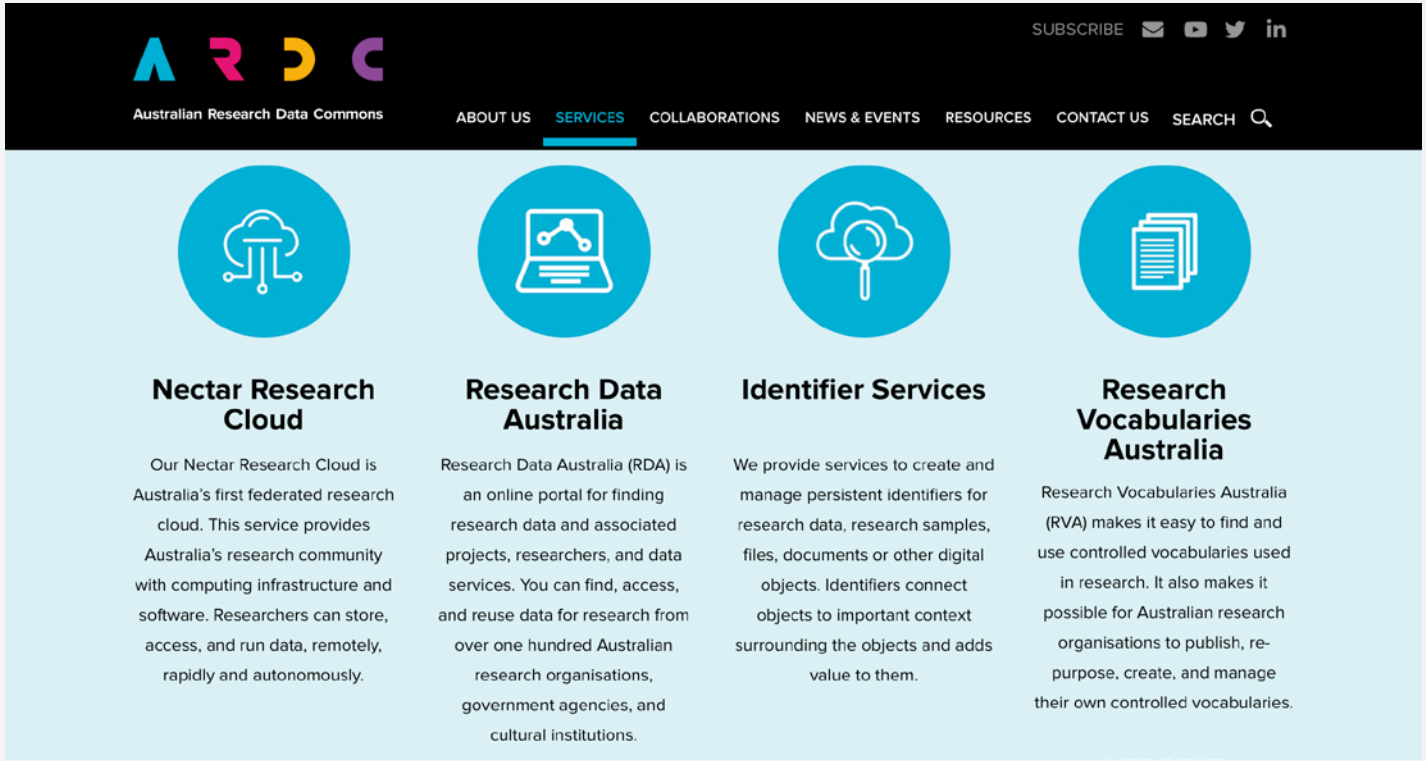


Cloud computation and storage

Data Management & long-term access:
Global Persistent IDs
Metadata
Data Dictionaries
Provenance
Versions
Access controls
Data curation

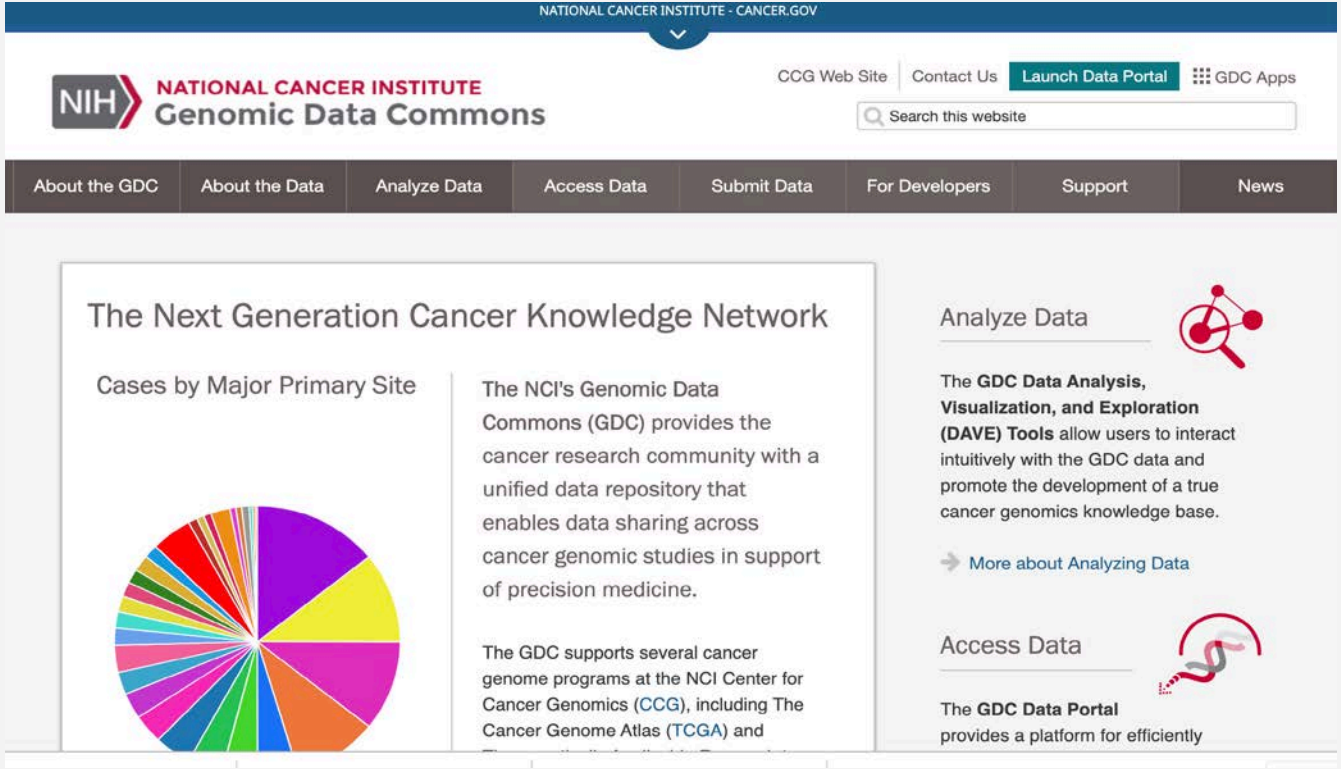
EXAMPLE: AUSTRALIAN RESEARCH DATA COMMONS

A Data Commons at the National Level



EXAMPLE: GENOMIC DATA COMMONS

A Data Commons for a scientific domain



What is the Harvard Data Commons?

HARVARD DATA COMMONS: A PROPOSAL

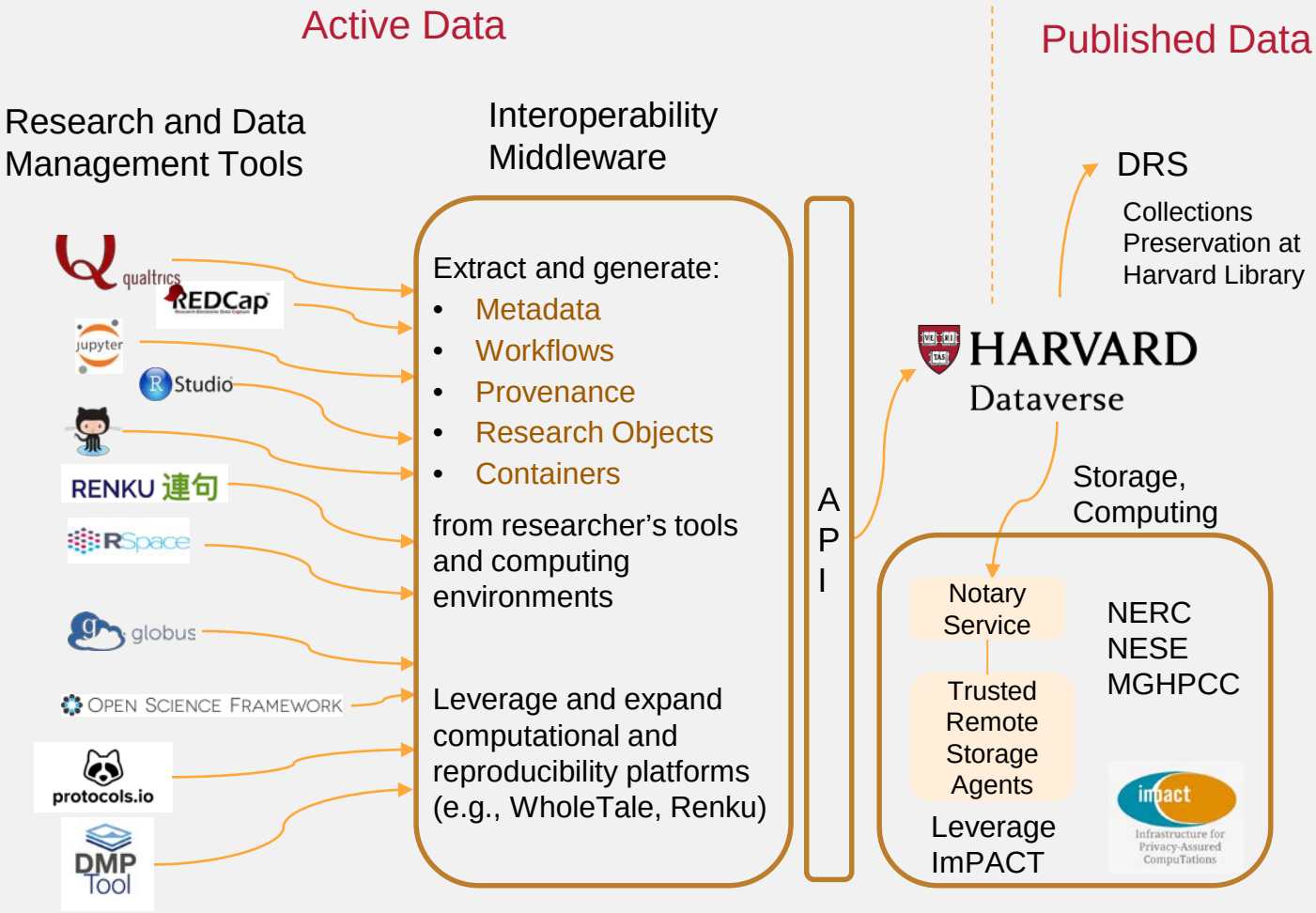
At its early stage:

- 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
DRS
Collections Preservation at Harvard Library
qualtrics
REDCap
jupyter
RStudio
RENKU 連句
RSpace
globus
OPEN SCIENCE FRAMEWORK
protocols.io
DMP Tool
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)
API
HARVARD Dataverse
Storage, Computing
Notary Service
Trusted Remote Storage Agents
NERC
NESE
MGHPCC
Leverage ImPACT
impact
Infrastructure for Privacy-Assured CompuTations

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