Harvard Data Commons

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

European Dataverse Workshop 2020, Tromso, Norway

THIS TALK

- What is a Data Commons
- Data Commons Types
- Harvard Data Commons
- Impact on Dataverse

What is a Data Commons?

DATA COMMONS DEFINITION 1

"The NIH Data Commons will accelerate biomedical discovery by providing a cloud-based platform where investigators can store, share, access, and compute on digital objects including data, software, workflows, and more."

https://nihdatacommons.us/

DATA COMMONS DEFINITION 2

"More formally, 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

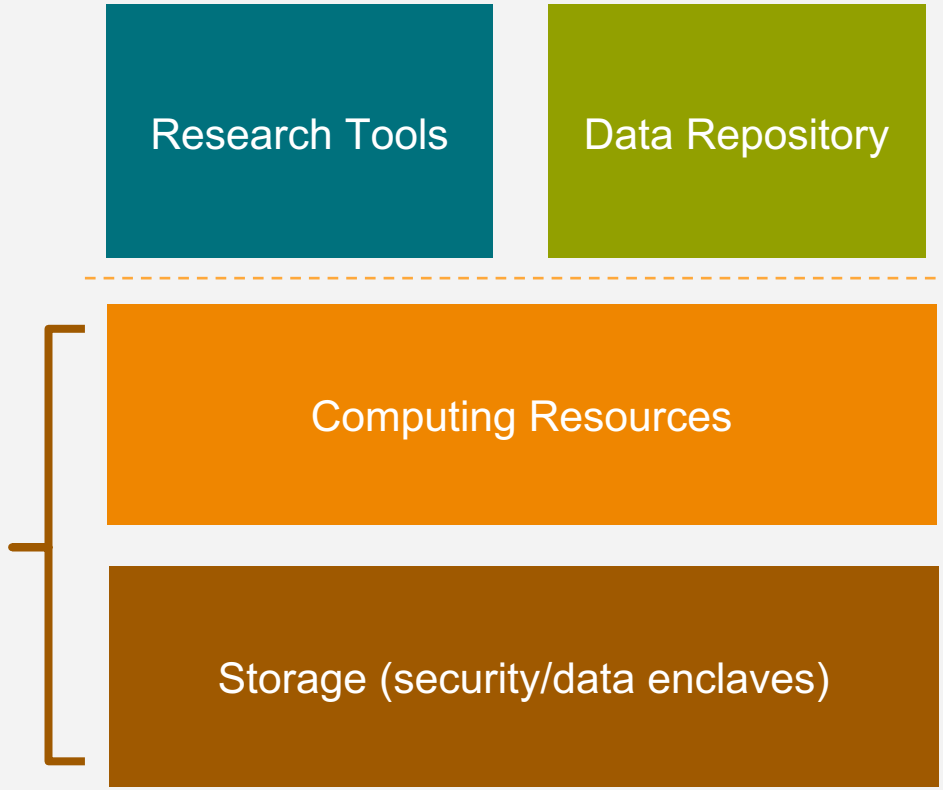
A Data Commons integrates active data-centric research with data management and archival best practices.

- Tools and services for active research
- Data Repositories for archival
- Cloud computation and storage for scalability

DATA COMMONS COMPONENTS

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

Researcher Interfacing



Data Management & Archival:
Global Persistent IDs
Metadata
Data Dictionaries
Provenance
Versions
Access controls

Cloud computation and storage

Data Commons Types

- National
- For a Research Community
- Institutional

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.

A Research Data Commons concept by the German Data Infrastructure

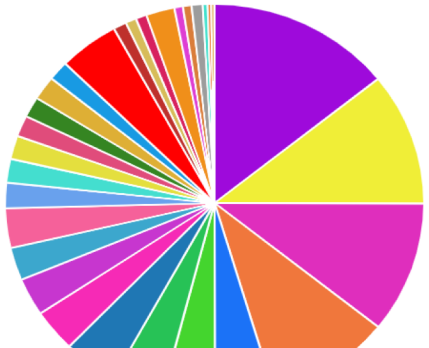
"... the view of a Research Data Commons (RDC) as an overarching virtual expandable infrastructure to leverage user involvement and collaborative data driven research. This includes for example joint cloud services, access to computing power and collaborative workspaces, and a common authentication and authorisation infrastructure (AAI). The RDC calls for a common strategy for interacting with the existing large-scale compute and data infrastructures in Germany and the need for harmonisation among these centers."

Berlin Declaration on NFDI Cross-Cutting Topics, September 22, 2019

- National
- For a Research Community
- Institutional

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

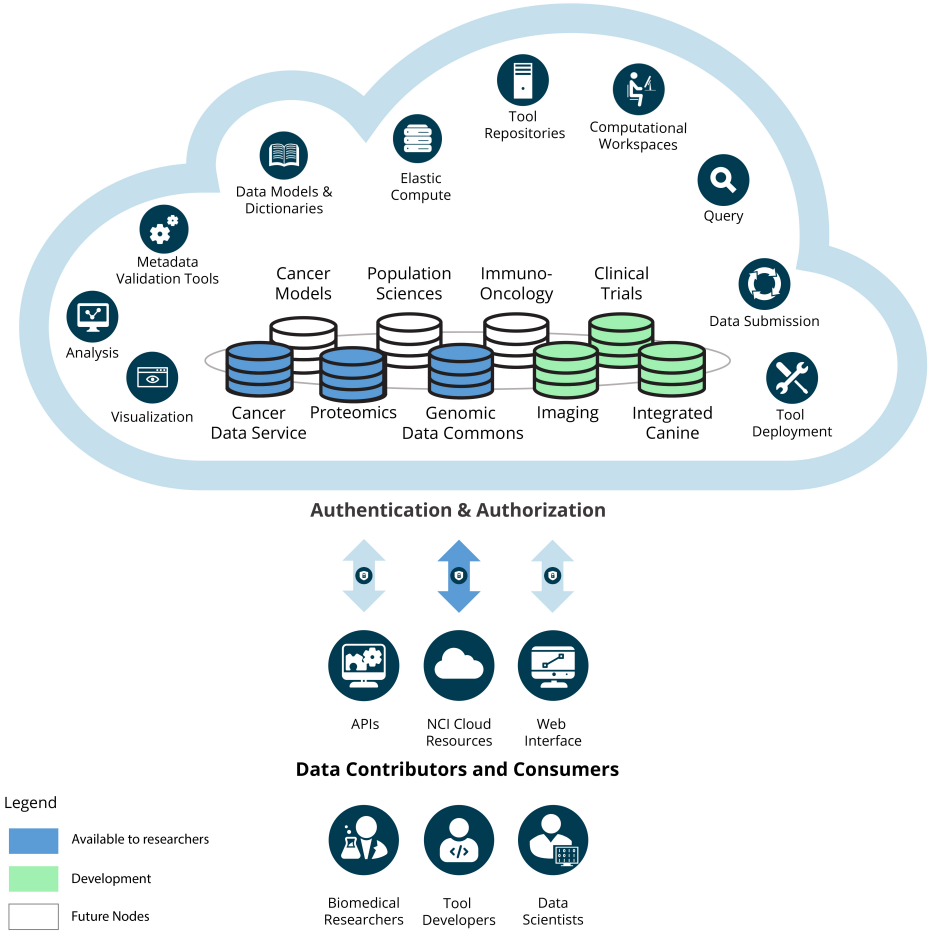
Access Data

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

NCI Cancer Research Data Commons

The vision for the Cancer Research Data Commons (CRDC) is a virtual, expandable infrastructure that provides secure access to many different data types across scientific domains, allowing users to analyze, share,

NCI Cancer Research Data Commons (CRDC)



Brings together:

- *Research tools*
- *Metadata*
- *Data models*
- *Relevant datasets*
- *Cloud Resources*
- *Auth/Authz*

- National
- For a Research Community
- Institutional

HARVARD DATA COMMONS: A PROPOSAL

At its early stage:

- Pilots in 2020
- Defining architecture and use cases

HARVARD DATA COMMONS: A COLLABORATION

Across:

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

HARVARD DATA COMMONS: WHY

To address current Research Data Management challenges:

- Incomplete knowledge of research data use
- Difficulty sharing active data across groups
- Duplication of datasets resulting in high costs
- Lack of documentation and provenance
- Risk of compliance with data policies

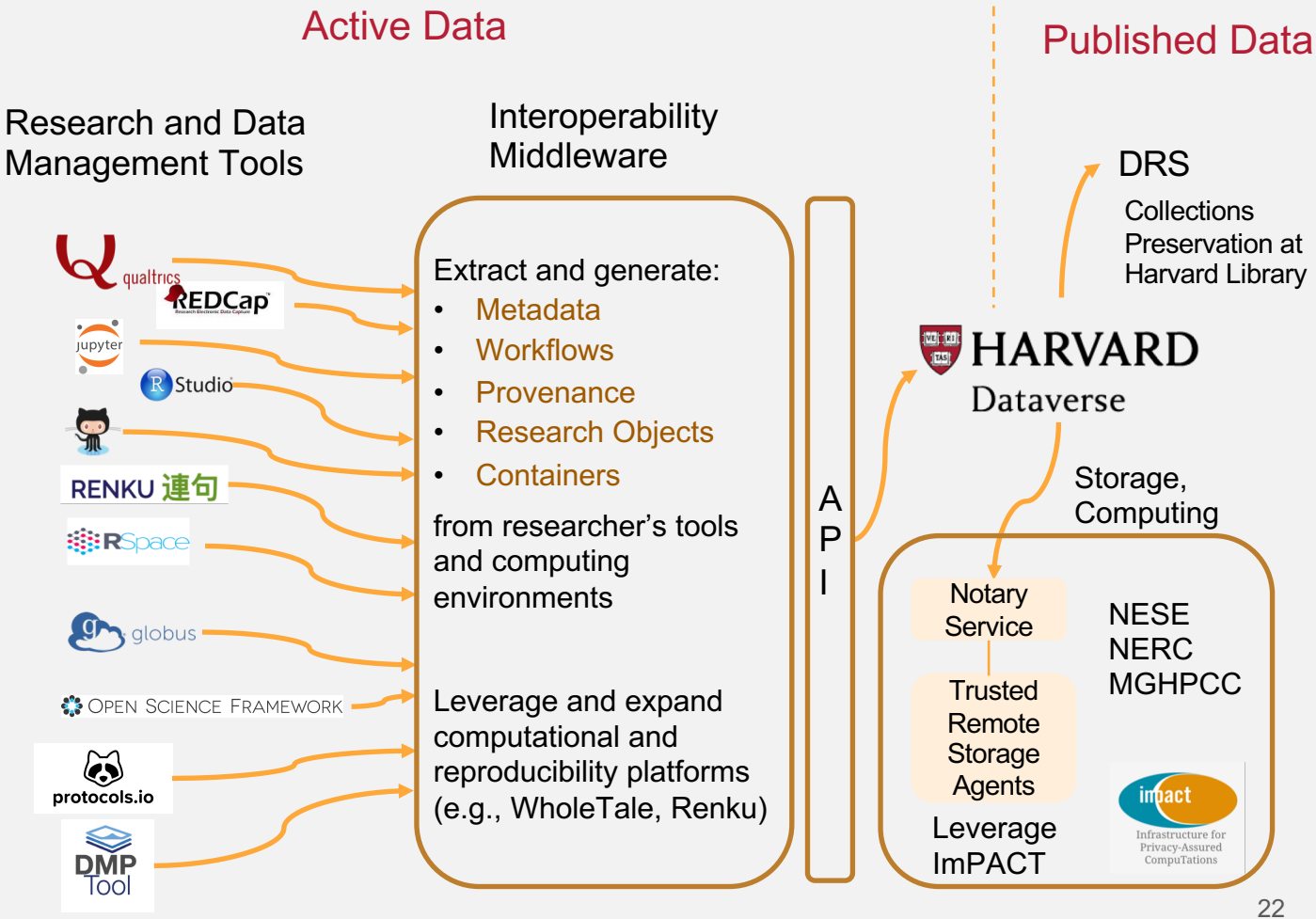
HARVARD DATA COMMONS: AIMS

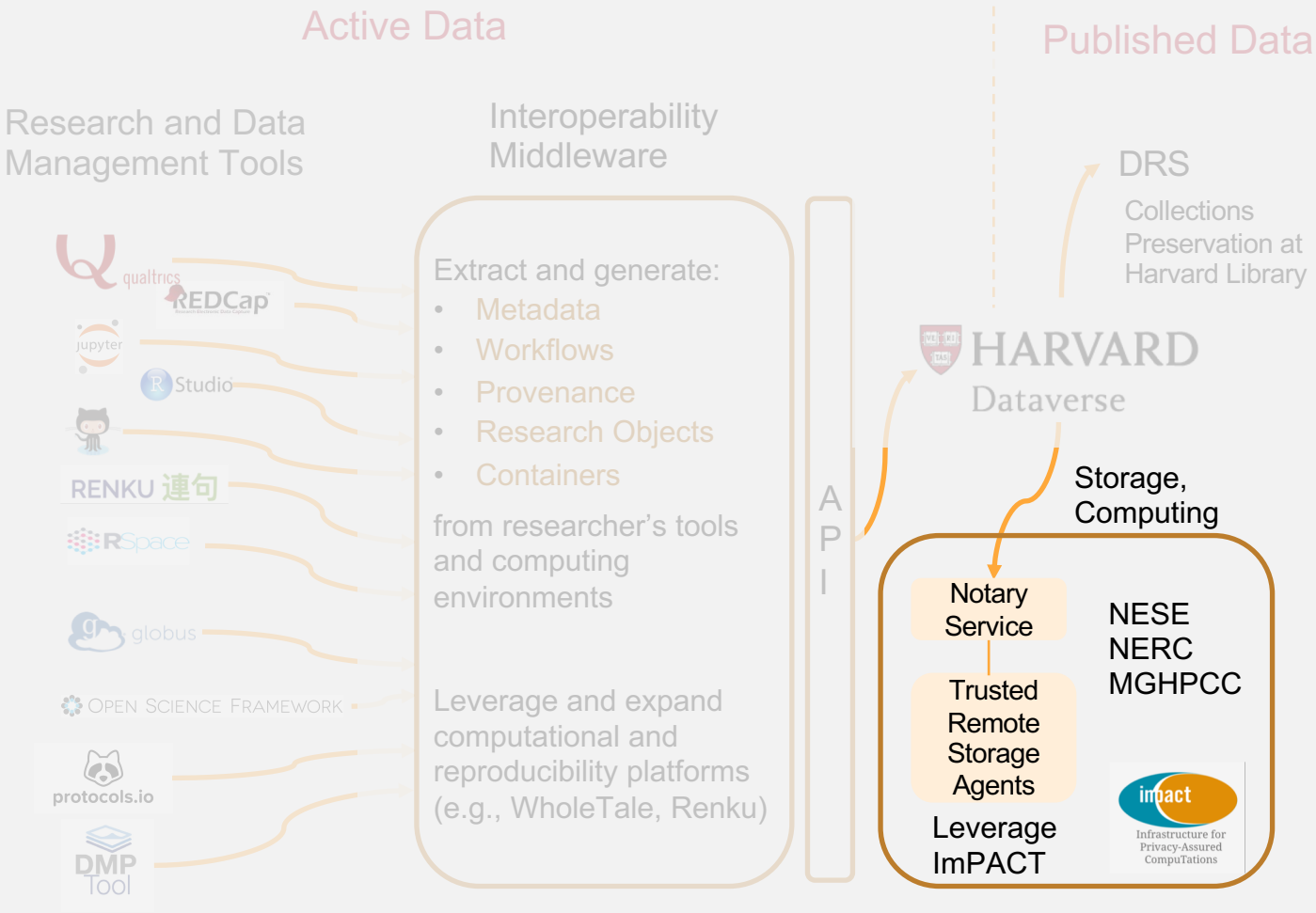
Improve data management and collaboration through:

- Discovery and access of unpublished data
- Metadata registry for tracking active datasets
- Entry point for collaboration on private datasets
- Support for publishing large and sensitive data

HARVARD DATA COMMONS: IMPLEMENTATION

An interoperability framework between research tools, research cloud computing, and Dataverse

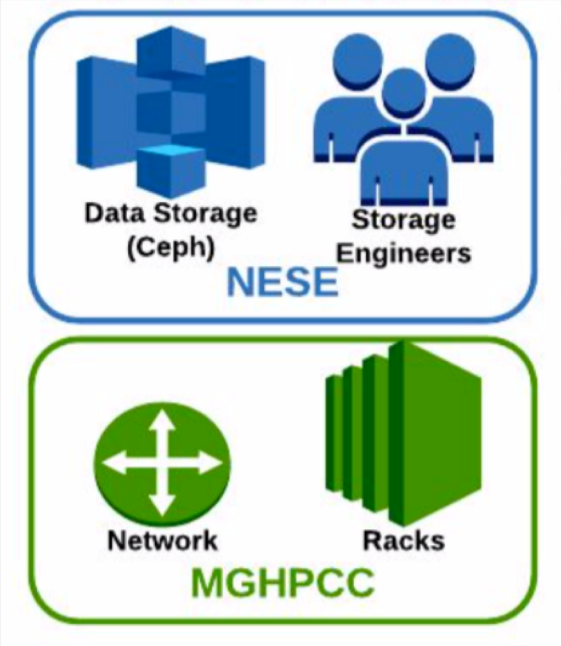

Active Data
Published Data
Research and Data Management Tools
Interoperability Middleware
DRS
Collections Preservation at Harvard Library
qualtrics
REDCap
jupyter
RStudio
RENKU 連句
RSpace
globus
OPEN SCIENCE FRAMEWORK
protocols.io
DMP Tool
Extract and generate:
• Metadata
• Workflows
• Provenance
• Research 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
NESE
NERC
MGHPCC
Leverage ImPACT
impact
Infrastructure for Privacy-Assured CompuTations


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
NESE
NERC
MGHPCC
Leverage ImPACT
impact
Infrastructure for Privacy-Assured CompuTations

THE DATA CENTER: MGHPCC

- **Mass. Green High Performance Computing Center (MGHPCC)**
- **Used by Harvard, BU, MIT, NE, UMass**
- **$95 Million**
- **90,000 square foot**
- **> 750 racks of computing equipment**

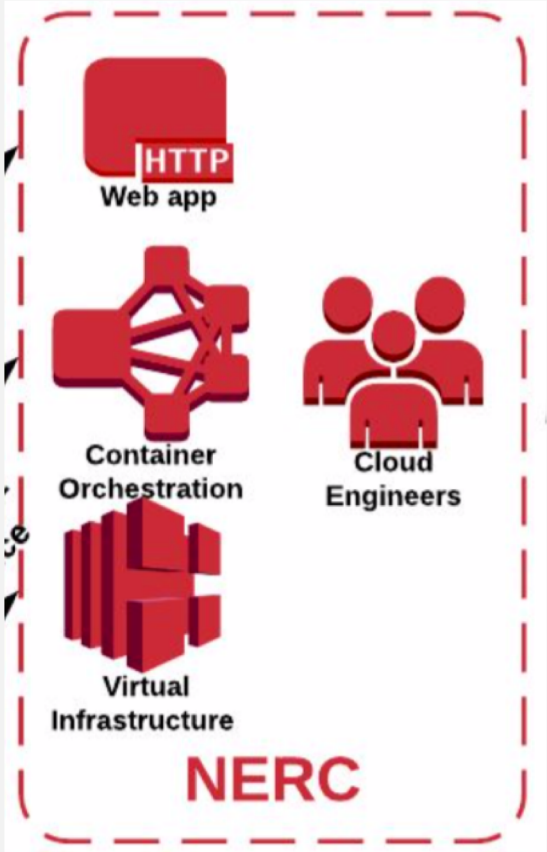
DATA STORAGE: NESE



Northeast Storage Exchange (NESE):

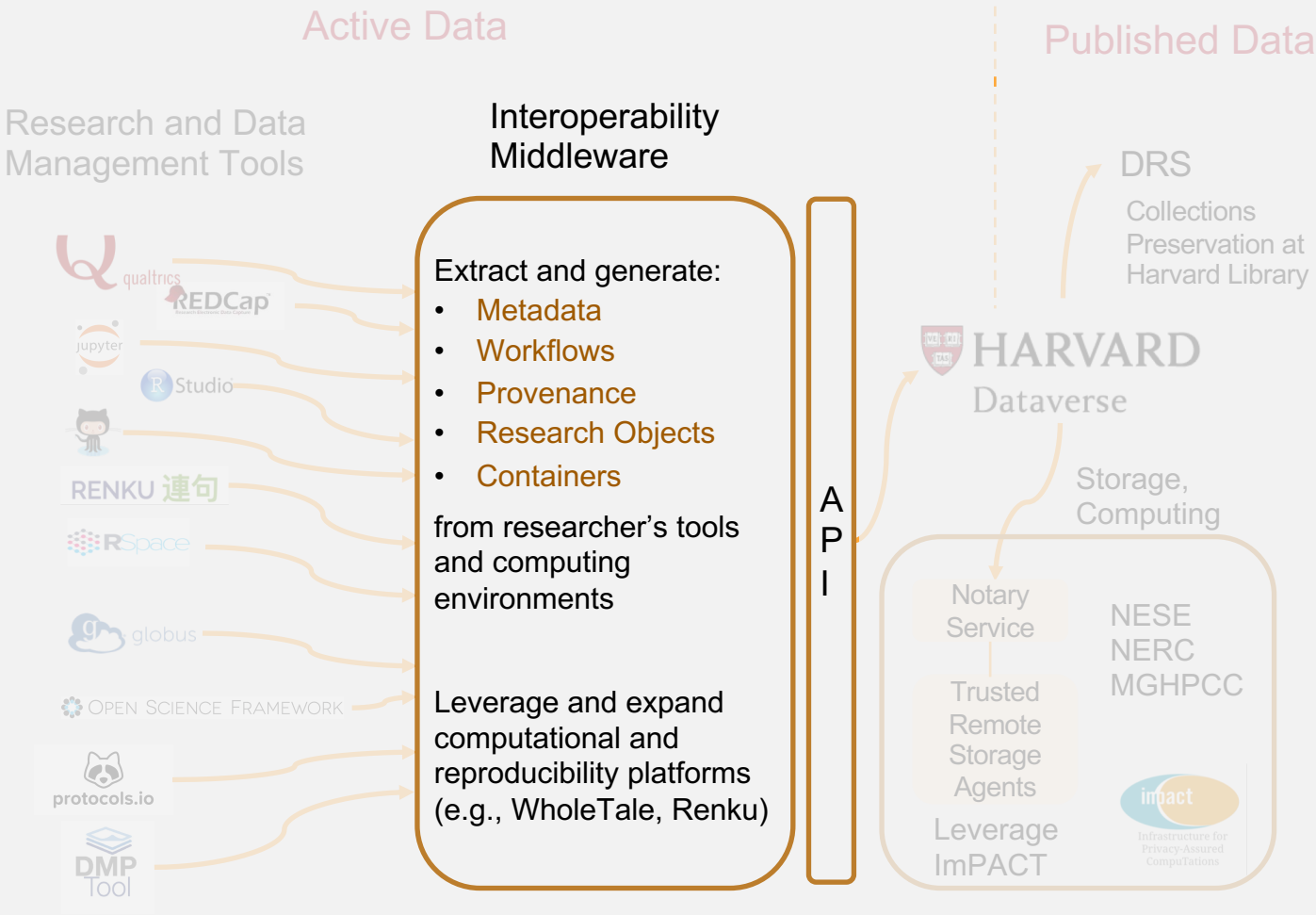
- Part of MGHPCC
- $4 Million, NSF funded
- A shared "data lake"
- Phase 1: 20 Petabytes
- 25% of the cost of equivalent AWS S3 storage

STORAGE: NERC

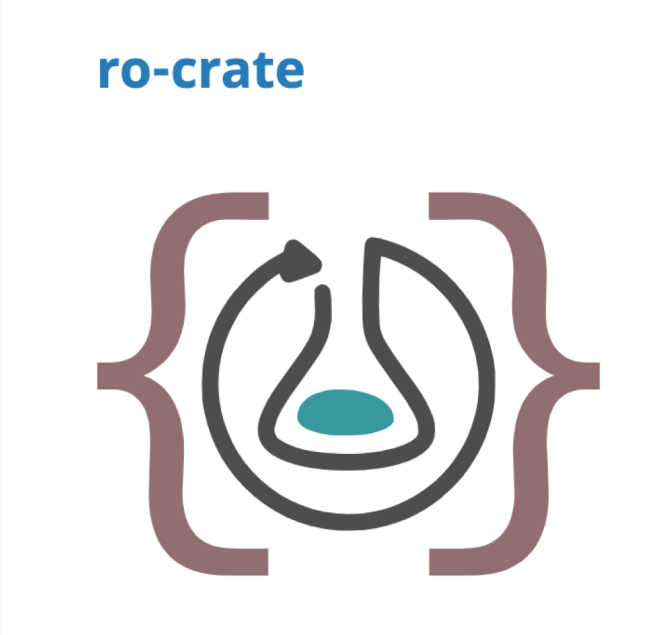


New England Research Computing (NERC):

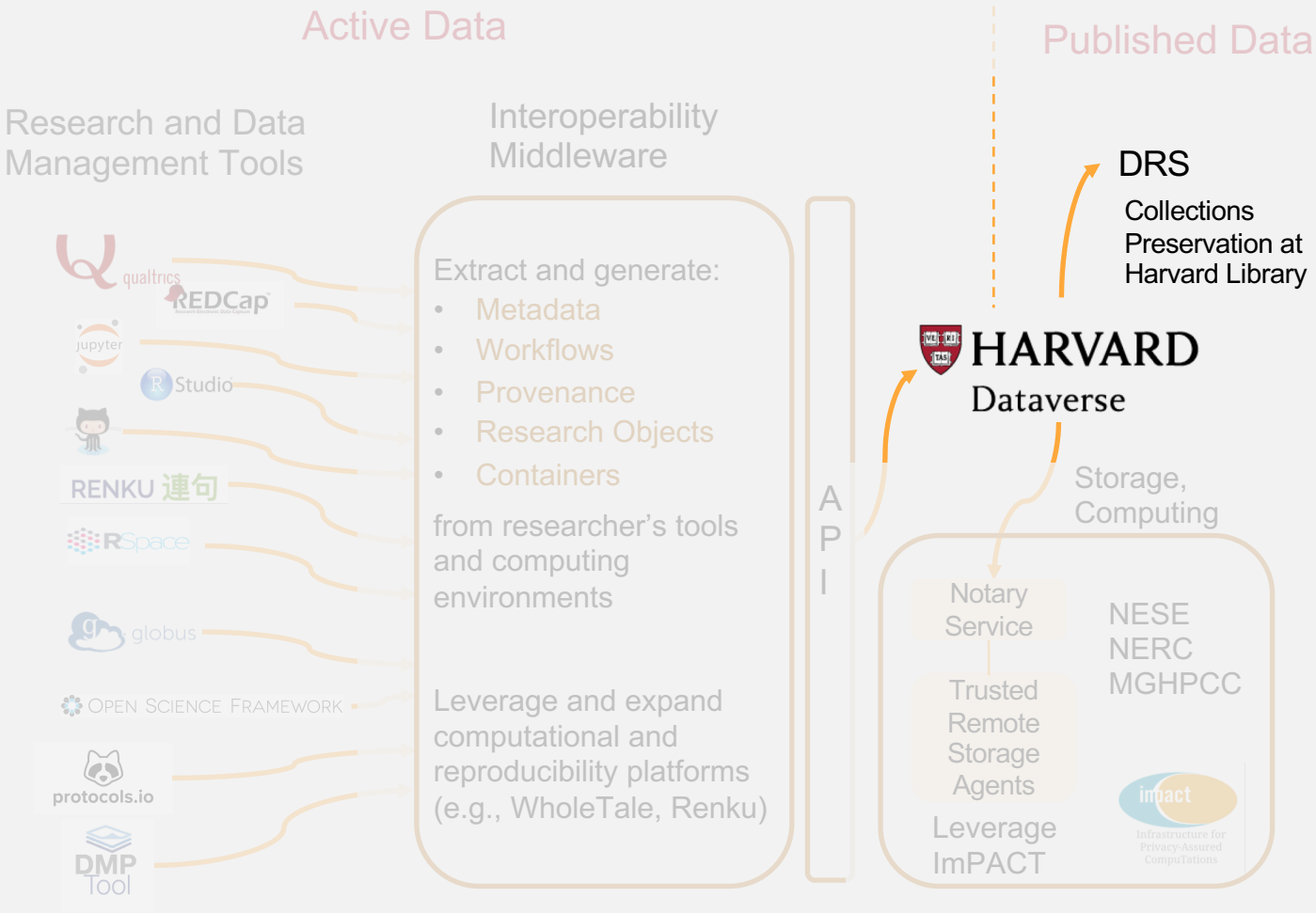
- Part of MGHPCC
- Aims to provide a common computing cloud platform
- Cost-effective
- Production, professional services
- SaaS, PaaS, IaaS
- OpenStack open-source community


Active Data
Published Data
Research and Data Management Tools
Interoperability Middleware
qualtrics
REDCap
jupyter
Studio
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
DRS
Collections Preservation at Harvard Library
HARVARD Dataverse
Storage, Computing
Notary Service
Trusted Remote Storage Agents
NESE
NERC
MGHPCC
Leverage ImPACT
impact
Infrastructure for Privacy-Assured CompuTations

INTEROPERABILITY MIDDLEWARE



- Create Research Objects/containers from research tools outputs
- Consider using RO-Crate metadata file: descriptor, root data entity, (data entities, contextual entities)
- Allows to extend metadata about objects
- Dashboard for searching and managing tools, ROs, data


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
NESE
NERC
MGHPCC
Leverage ImPACT
impact
Infrastructure for Privacy-Assured CompuTations

DATAVERSE IMPROVEMENTS

- Better support for unpublished datasets (versions, search)
- Flexible architecture to support multiple remote storage for large, sensitive data, tiers
- Support for Research Objects and RO-Crate metadata
- Option to archive collections to long-term preservation solution

PILOT with HARVARD MEDICAL SCHOOL

The pilot aims to develop standards for the generation, management, and sharing of research provenance and documentation, utilizing the Research Object Crate (RO-Crate) or other relevant technology.

The pilot aims to offer electronic laboratory notebooks (ELNs), as a service to the research community, to support research data generation, management, and analytics, integrated with sharing and archiving through Dataverse.

SUMMARY

- A Data Commons brings together active research with data management and archival
- The main components of a Data Commons are research tools, data repositories, and infrastructure for cloud research computing and storage
- Data Commons can be for a domain-specific community, a national infrastructure, or for one or more institutions
- Harvard Data Commons is an example of an institutional Data Commons which extends and integrates Dataverse with tools and cloud computing

Thanks