The Road from Data Commons to Data Meshes: Challenges, Opportunities, and Emerging Best Practices

Robert L. Grossman

Center for Translational Data Science

University of Chicago

September 12, 2022

1. What is a Data Commons?

What is a Commons?

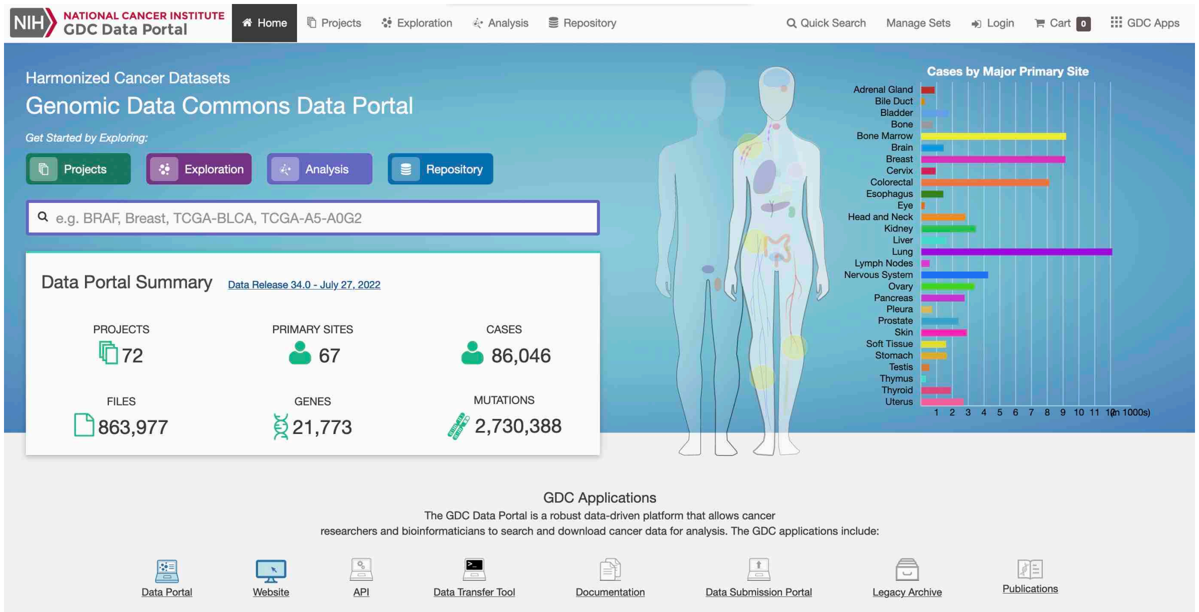
Commons are **resources** that are held in common (and not owned privately) that a group or **community** manage for individual and collective benefit.*

A **data commons** is a software platform along with a governance framework that together allow a community to manage, analyze and share its data.**

*Ostrom, Elinor. Governing the commons: The evolution of institutions for collective action. Cambridge university press, 1990.

**Grossman, Robert L. "Data lakes, clouds, and commons: A review of platforms for analyzing and sharing genomic data." Trends in Genetics 35, no. 3 (2019): 223-234.

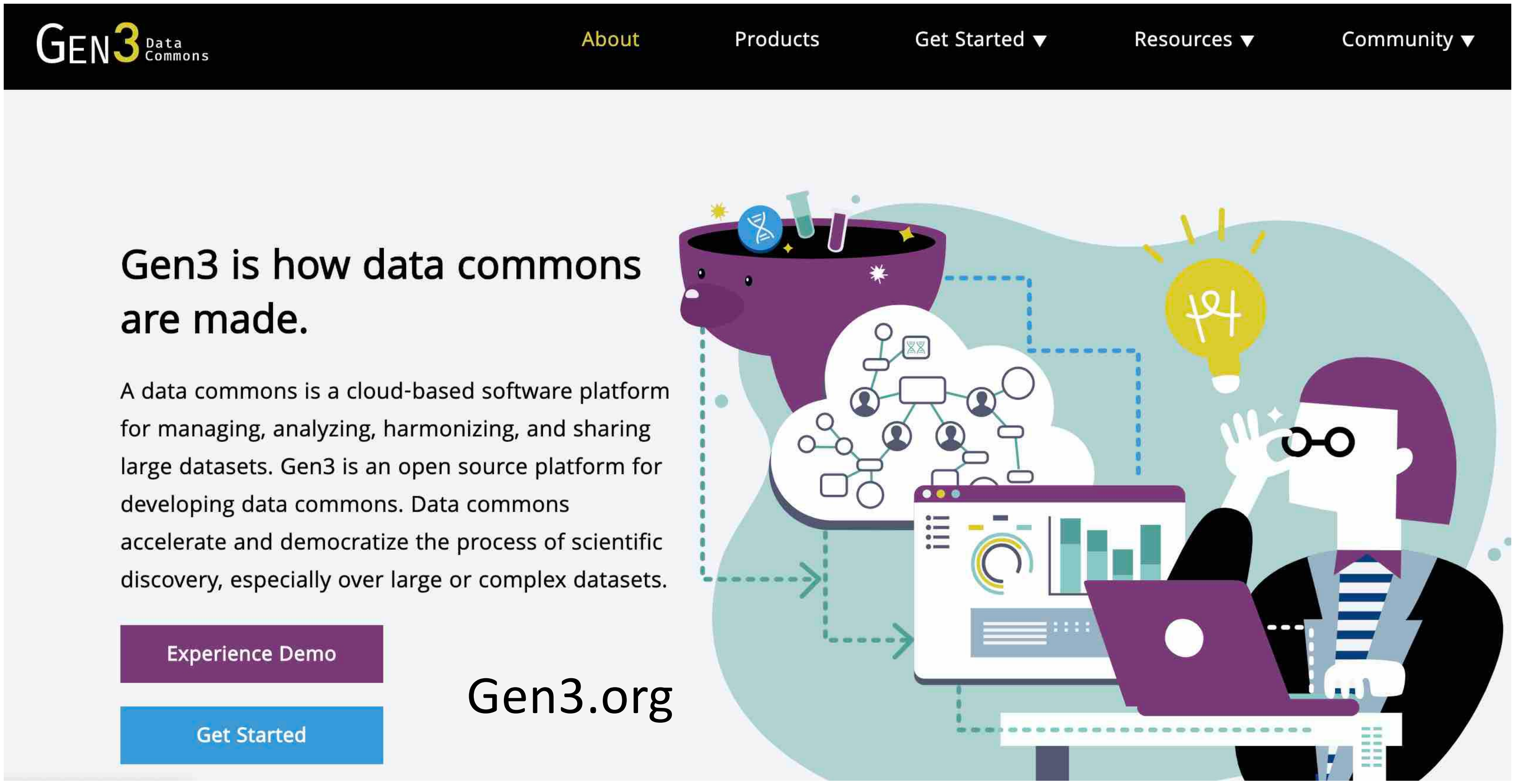
NCI Genomic Data Commons*



The GDC is a system of systems, including 1) data exploration & visualization portal; 2) data submission portal; 3) data analysis and harmonization system system (GPAS); 4) an API so third party can build applications.

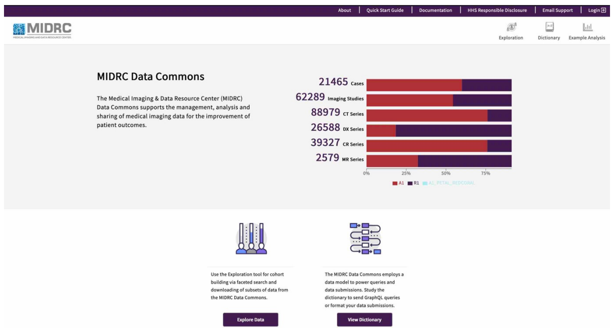
- The GDC makes over 4.3 PB of data available for access via an **API**, analysis by cloud resources on public clouds, and downloading.
- In an average month, the GDC is used by over 60,000 users, over 2 PB of data is accessed, and over 25,000 container based bioinformatics pipelines are run.
- The GDC is based upon an open source software stack that can be used to build other data commons.

*Heath AP, Ferretti V, ... and Grossman RL, The NCI Genomic Data Commons, Nature Genetics 2021 Mar;53(3):257-262. PMID: 33619384 doi: 10.1038/s41588-021-00791-5. PMID: 33619384

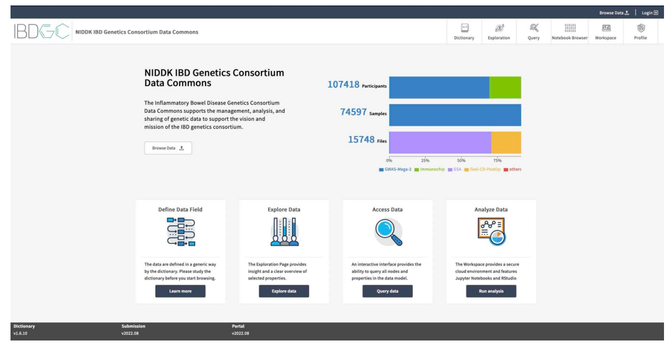


Gen3 is an open-source software platform to build and operate data commons and data meshes.

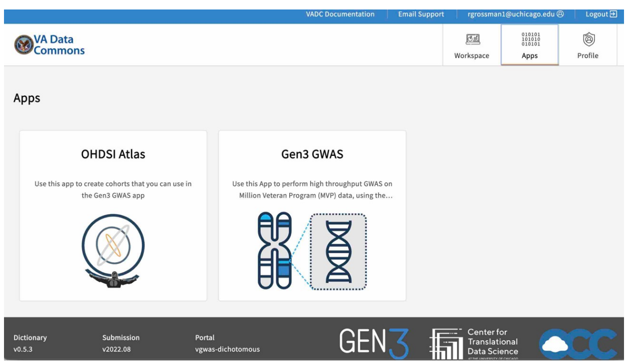
Selected Gen3 Data Commons



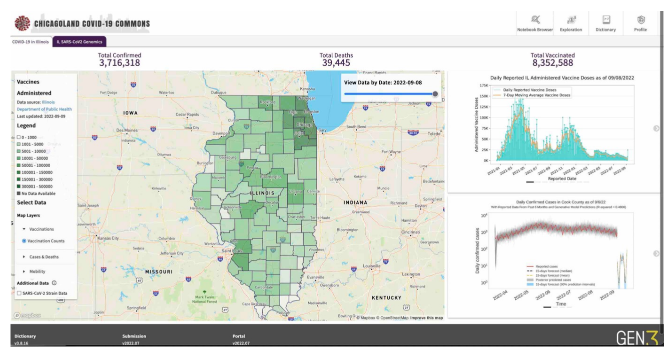
NIBIB MIDRC Data Commons



NIDDK IBDGC Data Commons



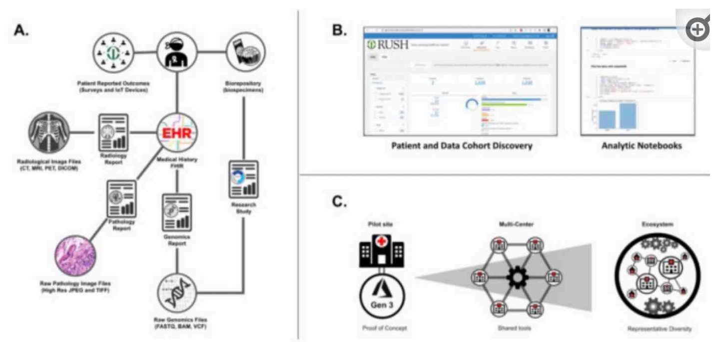
VA Data Commons



OCC Pandemic Response Commons (Chicago region)

Selected Gen3 Data Commons (continued)

Figure 2

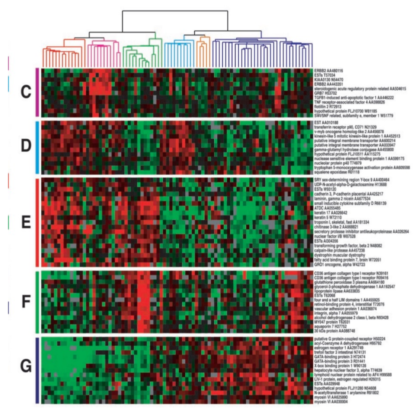


A) The underlying data model captures raw and processed multimodal data for each patient journey including medical history, imaging, and genomic data used in care practice as well as patient reported outcomes and research use only specimens; B) The interface for cohort discovery and data analyses are graphical and interactive; C) The ecosystem of reusable analytic and data engineering tools will expand and diversify as more users and use cases engage.

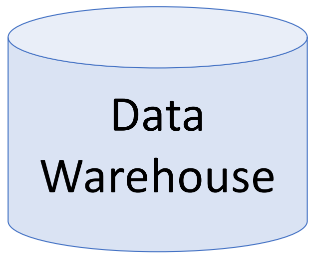
Data Model, Data Discovery, and Data Analysis

- Rush University Medical Center built a Gen3 data commons on Microsoft Azure within their security and compliance boundary to integrate, manage, analyze and share their internal research data.
- They built a number of Azure applications over the commons.
- They also created export mechanism to share data with third party systems, including the Gen3-based Pandemic Response Commons.

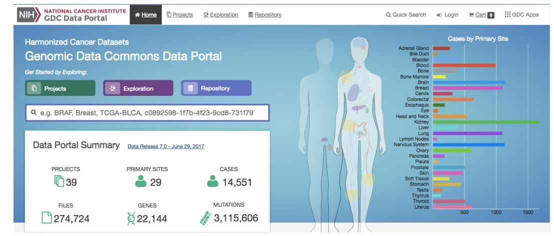
Source: O'Hara, Thomas, Anil Saldanha, Matthew Trunnell, Robert L. Grossman, Bala Hota, and Casey Frankenberger. "Economical Utilization of Health Information with Learning Healthcare System Data Commons." Perspectives in Health Information Management 19, no. Spring (2022).



Databases organize data around a **project** (1980's)



Data warehouses organize the data for an **organization** (1990's)



Data commons organize the data for a scientific **discipline** or field (2010's)

After a decade, cloud computing is now ubiquitous in data driven research and provides a good foundation for data science.

Data Clouds

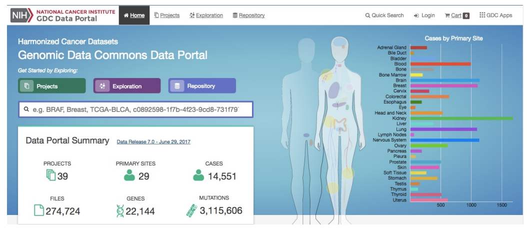
- Emerged around 2010
- Persistent Identifiers for **data objects in clouds**
- Researchers can use **cloud computing** to analyze data so it does not have to downloaded
- **Workflow languages, container repositories, workflow execution services** for large scale computation

Data Cloud Architecture

- Data lake model
- Some standards have emerged for the data objects
- No standards yet for the data object's metadata
- Data is pulled into a computing environment for analysis
- Slow consensus on workflow languages
- No real consensus on workflow execution orchestration

Source: Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403

Data Clouds



Data Commons

- Data objects in clouds
- Data workspaces in clouds
- **Common data models**
- **Harmonized data**
- **Core data services w APIs**
- **Data & Commons Governance**
- **Data sharing**
- **Reproducible research**

Data Commons Architecture

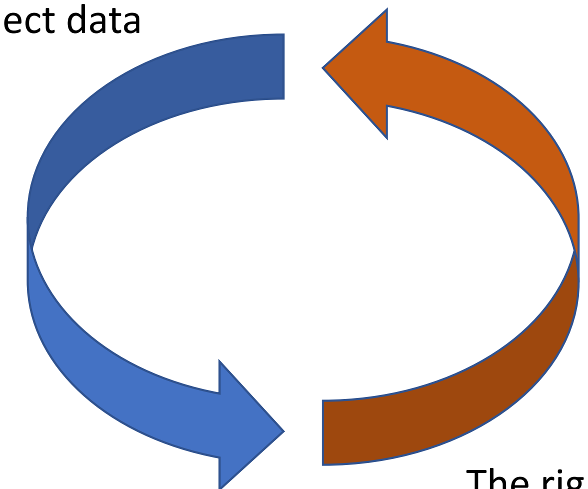
- Data lake model for data objects
- Graph (or other model) for clinical, biospecimen and other structured data
- Container based workflows to uniformly process submitted data (data harmonization)
- Open APIs to support portals, workspaces and third party applications

Source: Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403

Data commons balance protecting human subject data with open research that benefits patients:

Research ethics committees (RECs) review the ethical acceptability of research involving human participants. Historically, the principal emphases of RECs have been to protect participants from physical harms and to provide assurance as to participants’ interests and welfare.*

Protect human subject data



The right of human subjects to **benefit** from research.

[The Framework] is guided by, Article 27 of the 1948 Universal Declaration of Human Rights. Article 27 guarantees the rights of every individual in the world "to share in scientific advancement and its benefits" (including to freely engage in responsible scientific inquiry)…*

Data sharing with **protections** provides the evidence so patients can **benefit** from advances in research.

*GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data, https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/

2. What is a data mesh?

There are now over a dozen data commons, all sponsored by separate organizations, and the number is growing. How do researchers find datasets of interest, explore them, and analyze them?

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

BLOODPAC

RUSH

CHICAGOLAND COVID-19 COMMONS

NIH National Heart, Lung, and Blood Institute BioData CATALYST

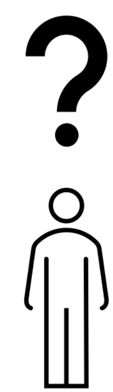
NIH NIDDK BDGC

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

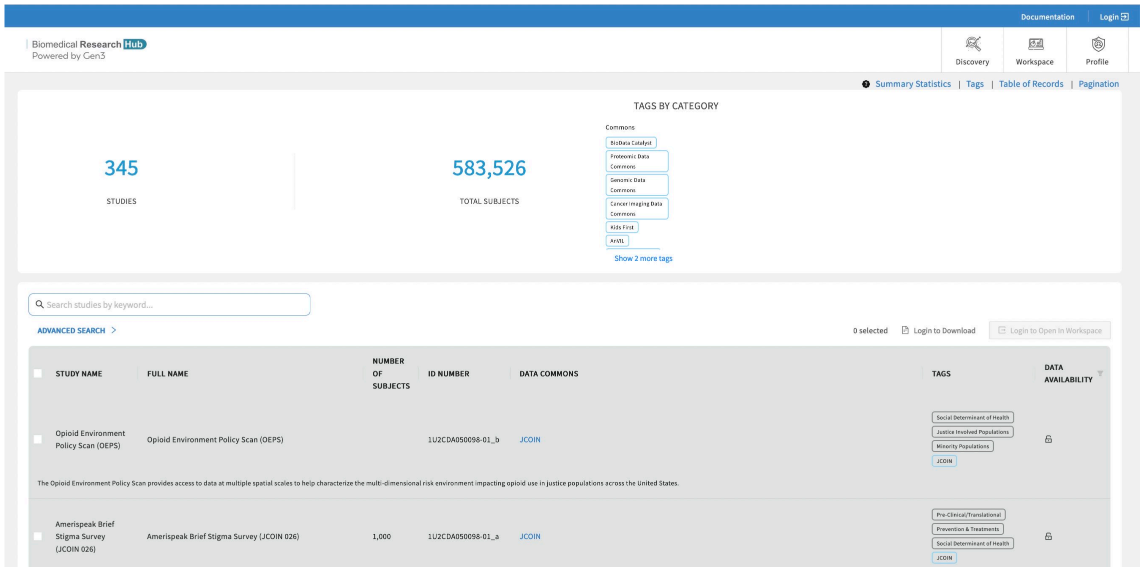
Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center

NIH NIBIB MIDRC EDICAL IMAGING AND DATA RESOURCE CENTER.

NIH NHGRI AnVIL



The Biomedical Research Hub (BRH)



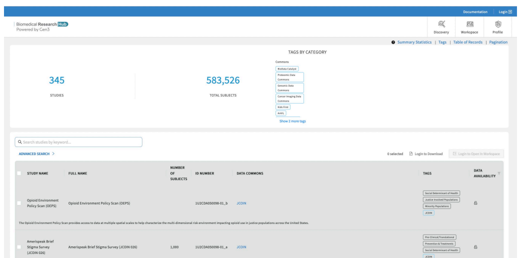
BRH Discovery Portal. Each data commons or data resource in the BRH data mesh exposes metadata about its datasets through FAIR APIs. The Gen3 Discovery Portal uses the metadata to power search. Data can then be explored and analyzed in workspaces. BRH is a joint project between the Center for Translational Data Science at the University of Chicago, OCC and AWS.

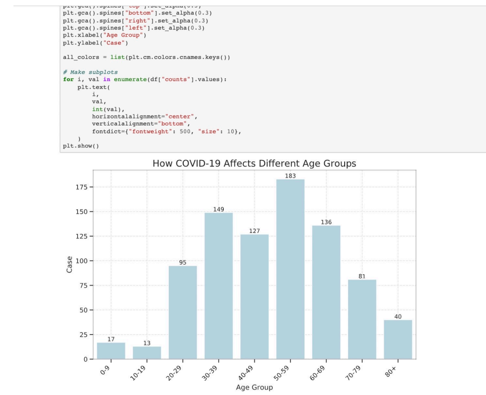
THE UNIVERSITY OF CHICAGO | Center for Translational Data Science

Open Commons Consortium

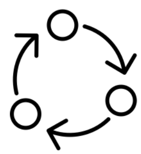
aws

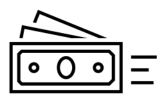
The Biomedical Research Hub (BRH) User Flow





1. Researcher uses BRH Discovery Portal to find one or more datasets of interest.



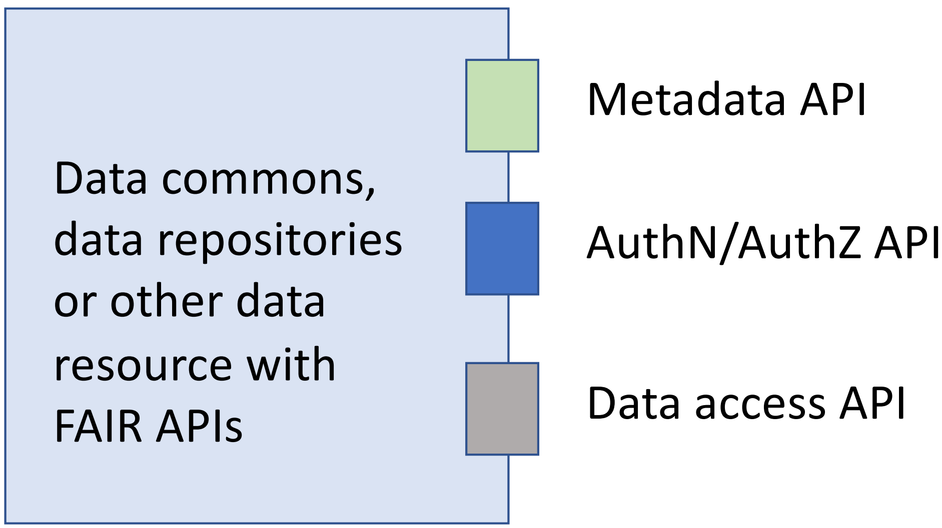


NIH STRIDES

2. Users register for BRH workspaces and a) uses a cloud-based workspace to analyze data after providing payment; or b) uses cloud platform data commons hosting the data to analyze the data; c) downloads / transfers the data to their own computing infrastructure.

3. User launches a BRH Workspace and accesses, explores and analyzes data of interest. A workspace can access data from multiple data commons and data resources.

Biomedical Research Hub Assumptions – Connecting to the BRH



To be part of the BRH data ecosystem a data commons or data resources must expose three APIs:

1. AuthN/AuthZ API
2. (FAIR) Metadata API
3. (FAIR) Data API

BRH enables interactive data discovery and data exploration over all data commons, data repositories, and other cloud-based resources that expose FAIR APIs.

Findable, Accessible, Interoperable & Reusable (FAIR) Data

Box 2 | The FAIR Guiding Principles

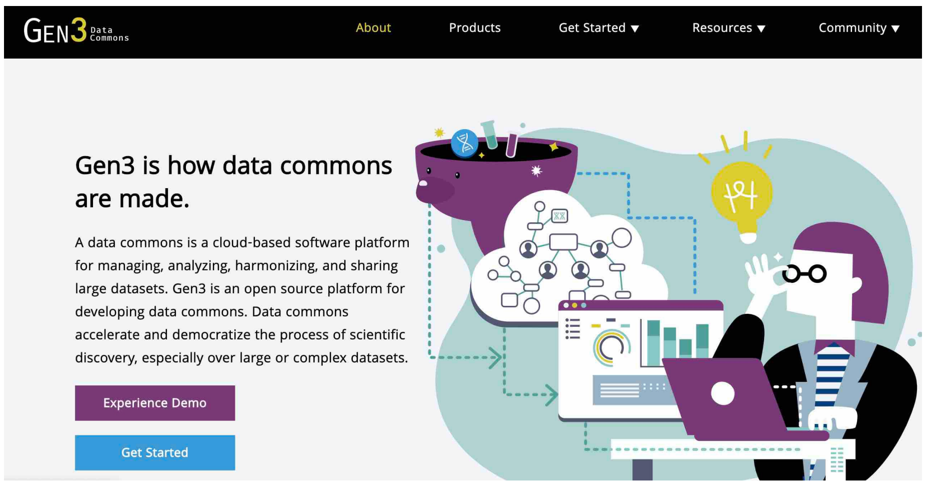
To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

The Gen3 Indexd Service and the Gen3 Metadata Service provides the services required to make data FAIR.

Figure from: Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1 (2016): 1-9.

Adding a Gen3 Data Commons to the Biomedical Research Hub



Gen3.org

1. Define a data model using Gen3.
2. Use the Gen3 platform to *auto-generate* the data commons and associated API (based upon your data model).
3. **This also creates the 3 APIs required to connect to the Biomedical Research Hub data ecosystem.**
4. Import data into the commons using Gen3 data submission portal or Gen3 data submission API.
5. Use the Gen3 Data Exploration portal to explore your data and create virtual (synthetic) cohorts.
6. Use Gen3 workspaces, notebooks (Jupyter and Rstudio) to analyze the data.

End to End Design Principle

End-To-End Arguments in System Design

J. H. SALTZER, D. P. REED, and D. D. CLARK
Massachusetts Institute of Technology Laboratory for Computer Science

This paper presents a design principle that helps guide placement of functions among the modules of a distributed computer system. The principle, called the end-to-end argument, suggests that functions placed at low levels of a system may be redundant or of little value when compared with the cost of providing them at that low level. Examples discussed in the paper include bit-error recovery, security using encryption, duplicate message suppression, recovery from system crashes, and delivery acknowledgment. Low-level mechanisms to support these functions are justified only as performance enhancements.

CR Categories and Subject Descriptors: C.0 [**General**] Computer System Organization—*system architectures*; C.2.2 [**Computer-Communication Networks**]: Network Protocols—*protocol architecture*; C.2.4 [**Computer-Communication Networks**]: Distributed Systems; D.4.7 [**Operating Systems**]: Organization and Design—*distributed systems*

General Terms: Design

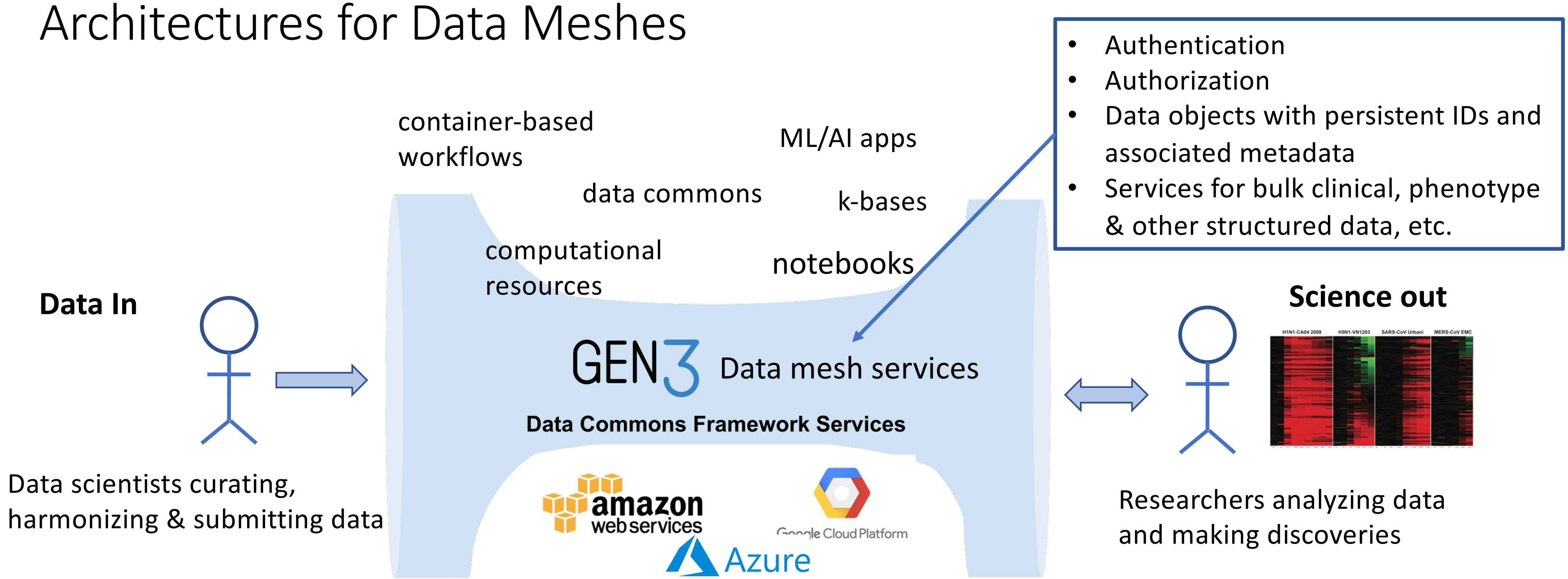
Additional Key Words and Phrases: Data communication, protocol design, design principles

1. INTRODUCTION

Choosing the proper boundaries between functions is perhaps the primary activity of the computer system designer. Design principles that provide guidance in this choice of function placement are among the most important tools of a system designer. This paper discusses one class of function placement argument that

Source: ACM Transactions on Computer Systems (TOCS), Volume 2 Issue 4, Nov. 1984, Pages 277-288

Architectures for Data Meshes



- Data meshes (aka data ecosystems) arise when multiple data commons and computational resources interoperate and support a collection of third party applications using a common set of core services (called data mesh or framework services)*
- This architecture using data mesh services is an example of the end-to-end design principle (aka “narrow middle” architecture)**

Sources: *Grossman, Robert L., Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles and Practice of Oncology, May/June 2018, Vol 24 (3), pg 122-126
**Saltzer, Reed and Clark, End-to-End Arguments in Systems Design, ACM Transactions on Computer Systems (TOCS), Vol 2 (4), Nov. 1984, pg 277-288

Gen3 Data Mesh Services for BRH

Gen3 Data Mesh Services

- Authentication & Authorization Infrastructure (AAI) – Gen3 Fence and Arborist
- Services for FAIR Data (Gen3 Indexd & Metadata Services)
- Service for ingesting data into the platform / mesh (Gen3 DIIRM)
- Services for Interop (Gen3 Crosswalk services)
- Services for bulk structured data (Avro-based formats for importing, exporting, versioning, and updating data)

Standards

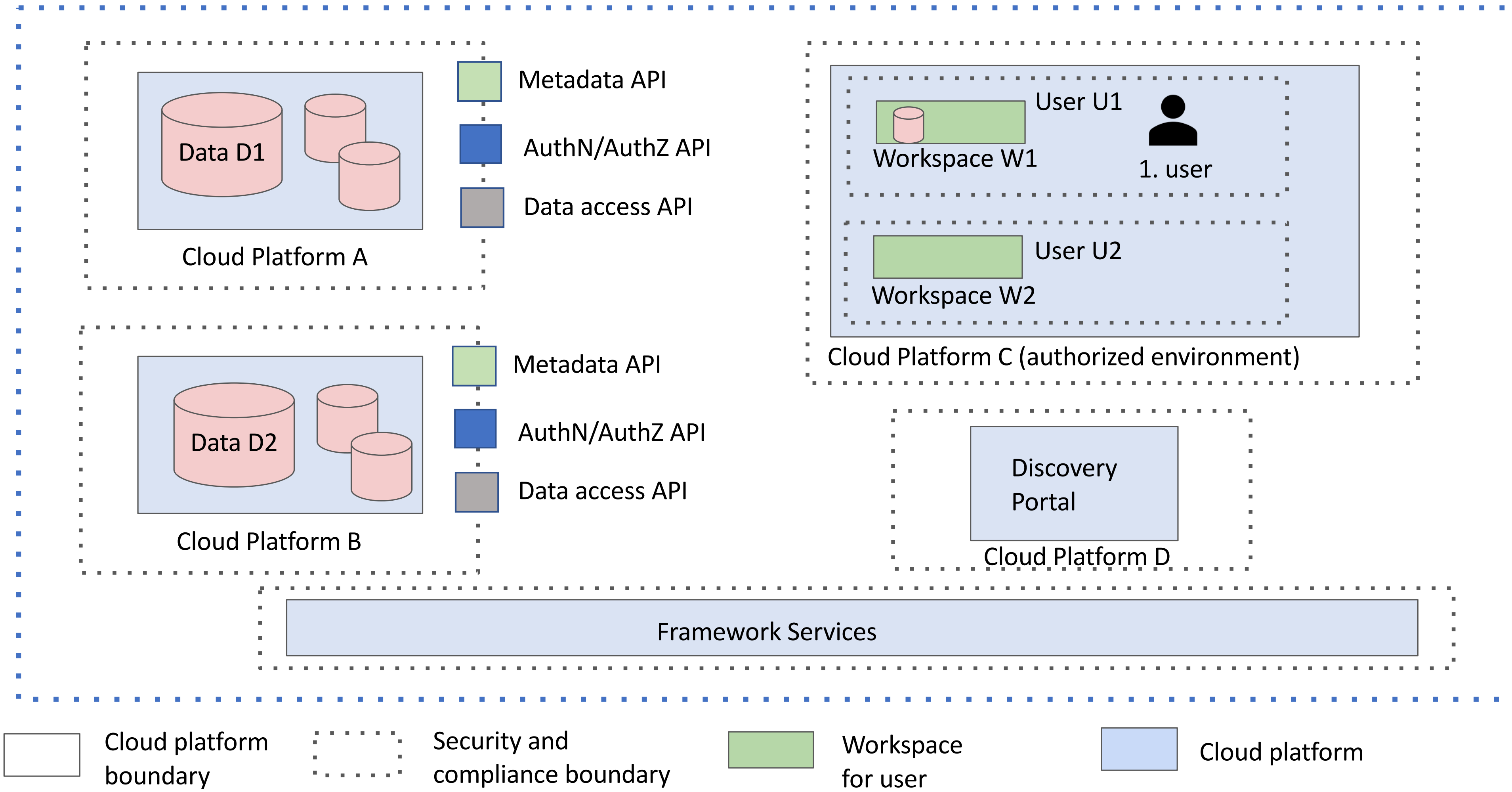
- GA4GH standards
  - GA4GH DRS
  - GA4GH Visas & Passports
- NIH RAS Service
- W3C & Research Data Alliance (RDA)

Security & Compliance

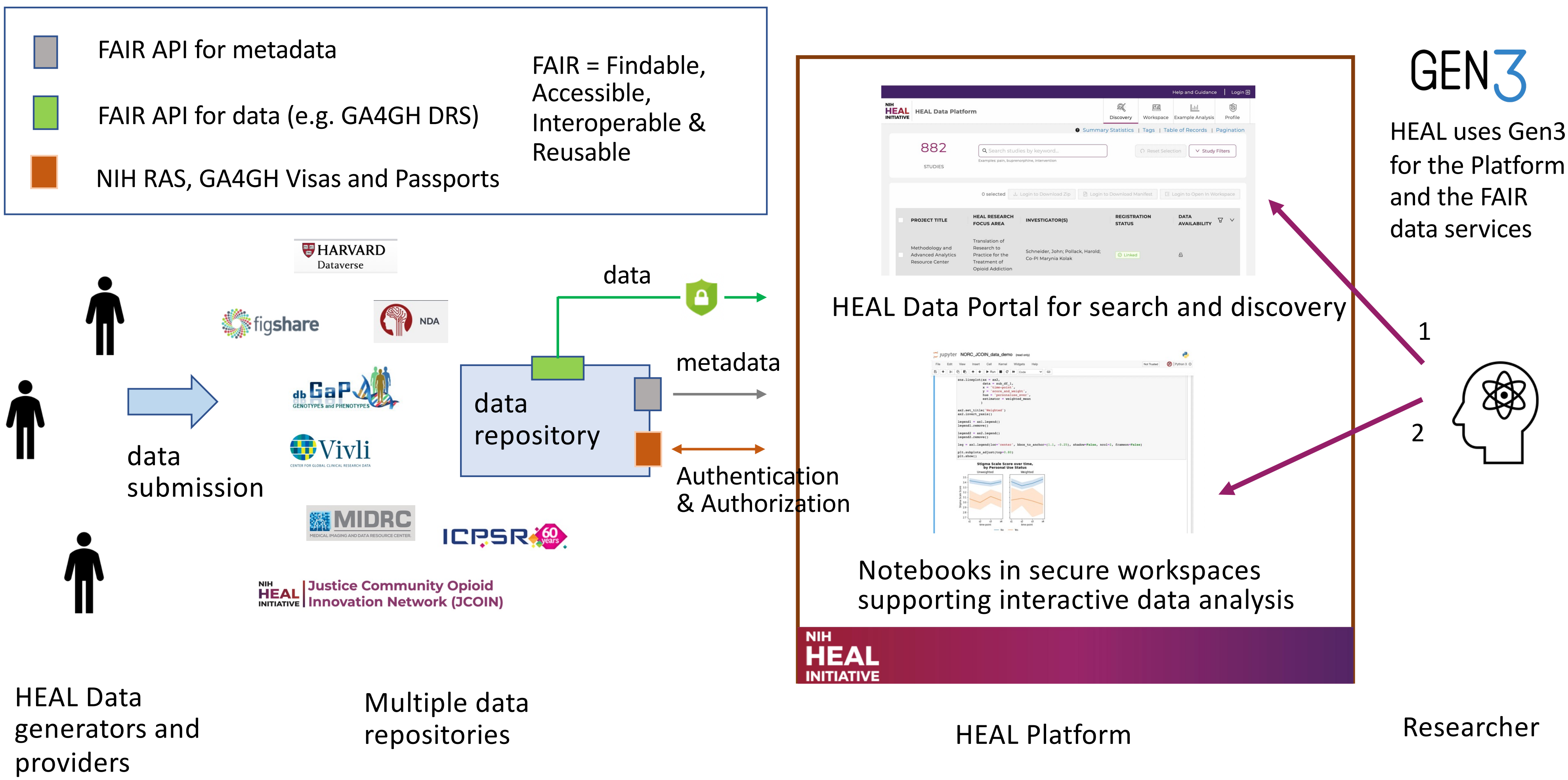
- NIST 800-53 (Moderate) ATO

Source: Craig Barnes, Binam Bajracharya, …, and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.

BRH Security & Compliance Follows NIST SP 800-53 Moderate



The HEAL Data Platform is a Gen3 Data Mesh for the NIH HEAL Initiative

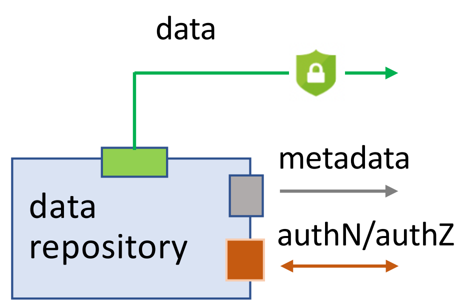


Hybrid Governance Model

Data Repository Governance

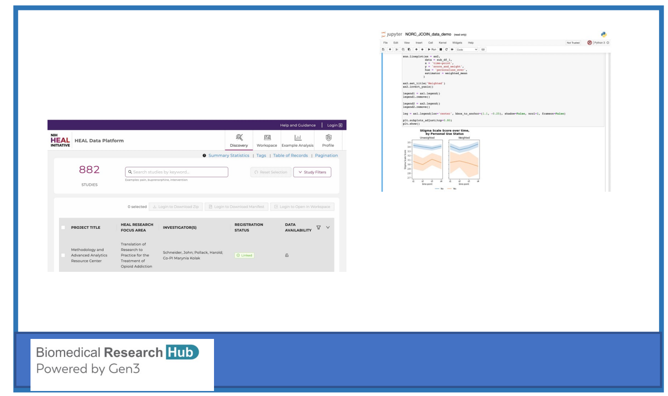
- DUA agreements between data submitters & repositories
- Required metadata, CDE, etc.
- Any data curation, etc.

Shared Governance (between repositories & mesh platform)



- Data Mesh Services
- FAIR APIs
- Interoperating AuthN/AuthZ
- System "Interoperability Agreement"

Data Mesh Governance (mesh platform)



- Which data repositories to connect to
- Governance rules for workspaces

Source: Craig Barnes, Binam Bajracharya, …, and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.

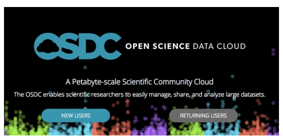
Summary

- **Data commons** software platforms that co-locate: 1) curated data, 2) cloud-based computing infrastructure, and 3) commonly used software applications, tools and services to create a governed resource for managing, analyzing and sharing data with a research community.

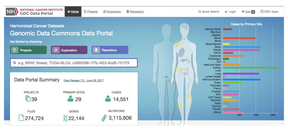
- **Data meshes** (aka data ecosystems) integrate multiple data commons, computational platforms, and other cloud-based resources operated by different organizations, along with a hybrid governance framework, and enable the management, discovery, analysis and sharing of data.

- **Data Mesh Services** (aka Data Commons Framework Services) are a set of services to to develop and operate data commons and data meshes.

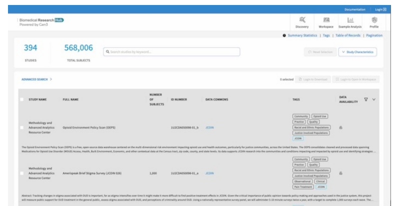
Source: - Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403
- Craig Barnes, Binam Bajracharya, …, and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.



Data Clouds
(2010)



Data Commons
(2016)



Data Meshes
(2020)

- Interoperates **multiple data commons, databases, knowledge bases**, and other resources
- Supports **mesh / ecosystem of commons, portals, notebooks, applications & simulations** across multiple disciplines
- Data meshes support multiple data models.

Data Meshes Architecture

- Data lake model for data objects
- Framework services with AuthN/AuthZ, data objects and services for clinical/phenotype data
- Open APIs to support other commons, portals, workspaces and third-party applications
- Container based workflows to uniformly process submitted data (data harmonization)
- Governance model that supports data sharing

Source: - Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403
- Craig Barnes, Binam Bajracharya, …, and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.

3. Building Data Commons and Data Meshes with the Open Source Gen3 Data Platform

GEN3

About ▼ Products ▼ Get Started ▼ Resources ▼ Community ▼

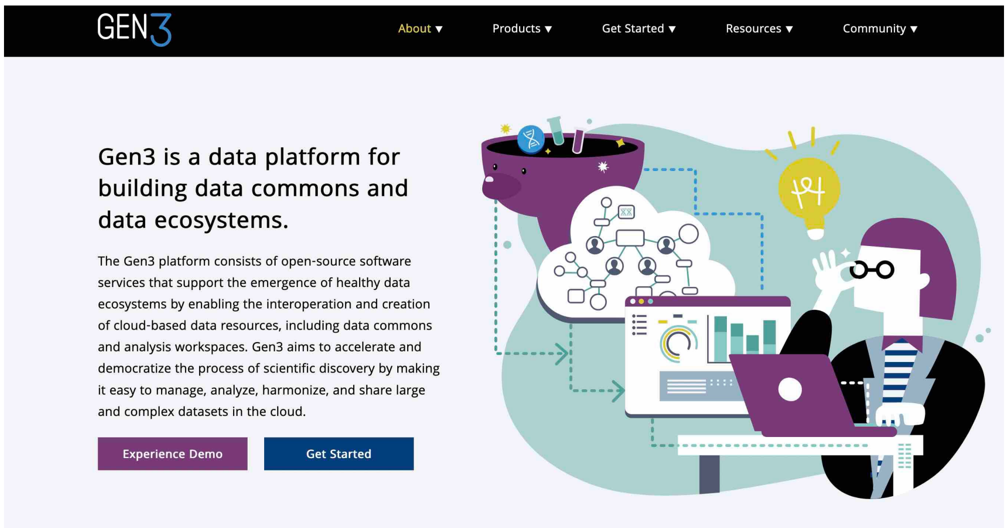
Gen3 is a data platform for building data commons and data ecosystems.

The Gen3 platform consists of open-source software services that support the emergence of healthy data ecosystems by enabling the interoperation and creation of cloud-based data resources, including data commons and analysis workspaces. Gen3 aims to accelerate and democratize the process of scientific discovery by making it easy to manage, analyze, harmonize, and share large and complex datasets in the cloud.

Experience Demo Get Started

Gen3.org

Five Steps to Build a Gen3 Data Commons



Gen3.org

1. Define a data model using Gen3.
2. Use the Gen3 platform to *auto-generate* the data commons and associated API (based upon your data model).
3. Import data into the commons using Gen3 data submission portal or Gen3 data submission API.
4. Use Gen3 data exploration portal to explore your data and create synthetic cohorts.
5. Use existing workspaces, (Jupyter, RStudio, Stata) notebooks and applications to analyze the data or develop your own.

GEN3

1,445,041 Total Subjects | **56,332,082** Total Files | **15.52 PB** Total File Size

NIH HEAL INITIATIVE
153 Attributes
85 Files
Total Size 14.88 GB

Veterans Affairs Data Commons
658,278 Subjects
1,606 Attributes
223 Files
Total Size 1.21 TB

BloodPAC (BLOOD PROFILING ATLAS IN CANCER)
4,839 Subjects
888 Attributes
35,549 Files
Total Size 34.57 TB

NIH National Institute of Allergy and Infectious Diseases – AccessClinicalData@NIAID
2,096 Subjects
151 Attributes
10 Files
Total Size 3.88 MB

CANINE Data Commons
1,499 Subjects
1,048 Attributes
3,820 Files
Total Size 1.88 TB

NIH National Heart, Lung, and Blood Institute – BioData CATALYST
240,460 Subjects
770 Attributes
667,328 Files
Total Size 3.74 PB

GenoMEL – the Melanoma Genetics Consortium
1,390 Subjects
387 Attributes
6,555 Files
Total Size 31.6 TB

The AnVIL
26,636 Subjects
551 Attributes
187,134 Files
Total Size 502.27 TB

VPO – Veterans Precision Oncology Data Commons
163,695 Subjects
1,606 Attributes
352,786 Files
Total Size 2.18 TB

GEN3 – Open Access Data Commons
1,366 Subjects
1,452 Attributes
1,598 Files
Total Size 13.77 TB

NIH NATIONAL CANCER INSTITUTE – Cancer Research Data Commons
83,709 Subjects
622 Attributes
17,166,700 Files
Total Size 4.32 PB

CHICAGOLAND COVID-19 COMMONS
53,728 Subjects
1,464 Attributes
285,653 Files
Total Size 117.64 TB

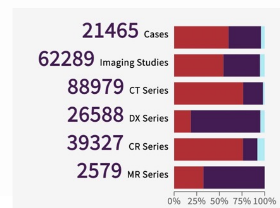
NIH HEAL INITIATIVE – Justice Community Opioid Innovation Network (JCOIN)
237 Subjects
509 Attributes
369 Files
Total Size 1.27 GB

BDGC
107,418 Subjects
786 Attributes
15,977 Files
Total Size 4.2 TB

ACCOUNT
1,516 Subjects
985 Attributes
5,661 Files
Total Size 7.77 TB

Gabriella Miller Kids First – PEDIATRIC RESEARCH PROGRAM – Data Resource Center
21,833 Subjects
776 Attributes
740,249 Files
Total Size 6.64 PB

MIDRC – MEDICAL IMAGING AND DATA RESOURCE CENTER



Environmental Data Commons
265 Attributes
33,441,289 Files
Total Size 99.2 TB

Gen3 Provides the Data Mesh Services to Make Data Findable, Accessible, Interoperable & Reusable (FAIR)

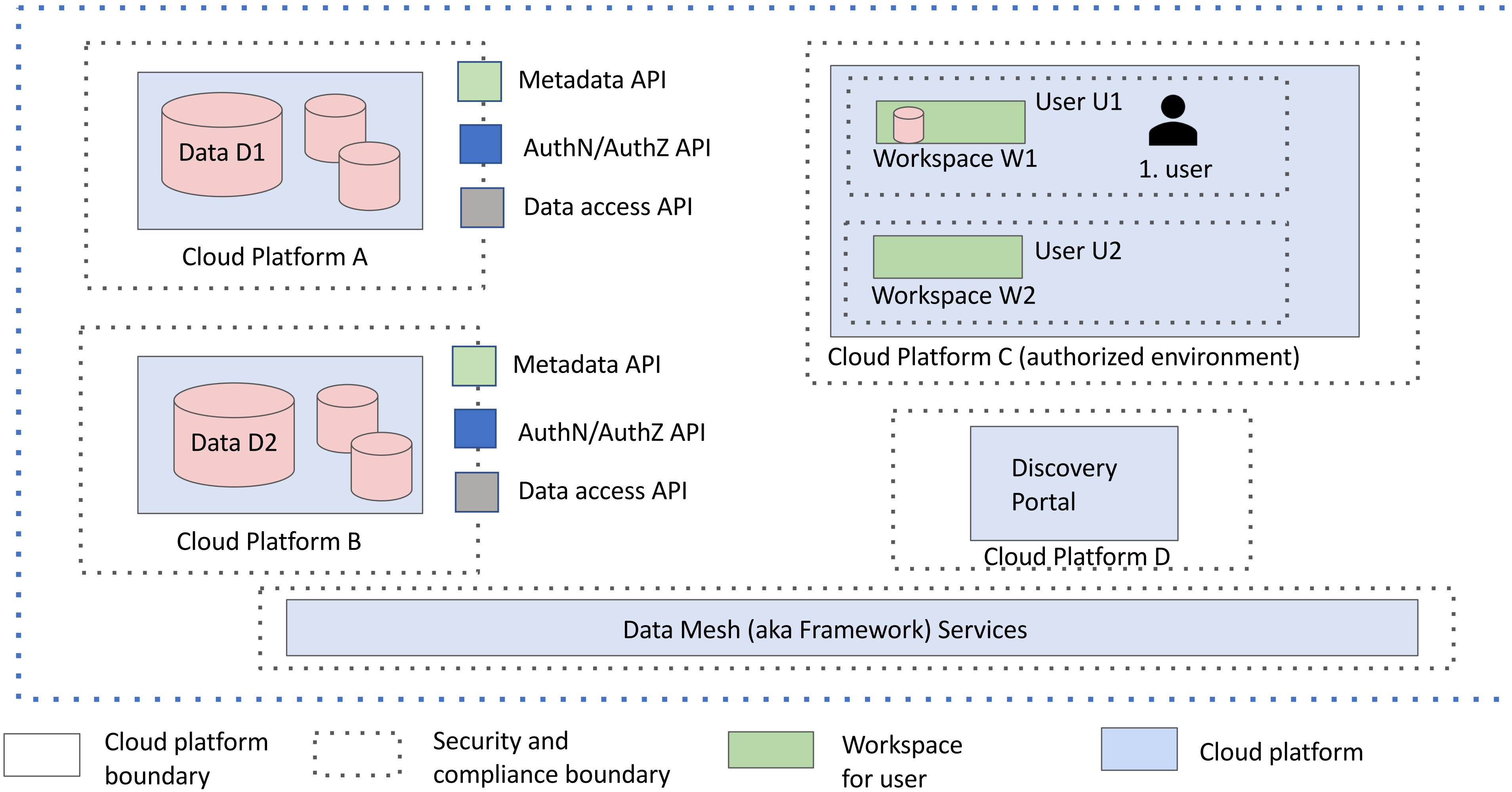
Box 2 | The FAIR Guiding Principles

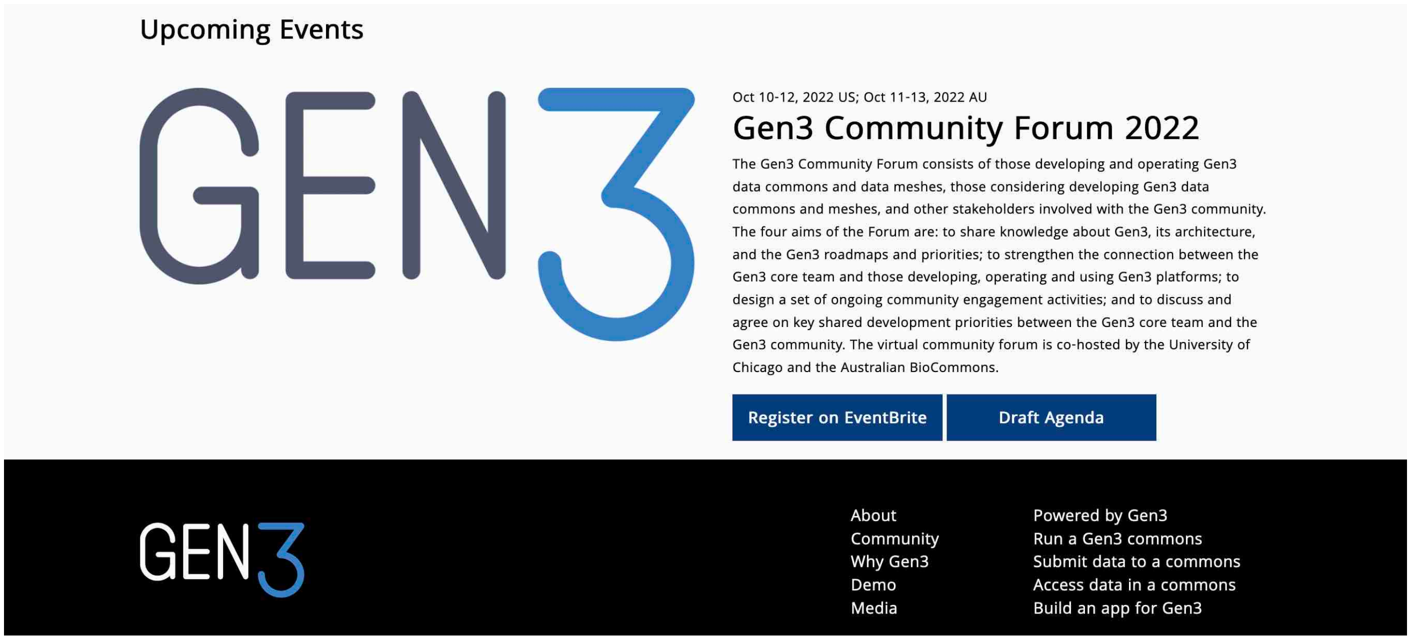
To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

Figure from: Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3.1 (2016): 1-9.

Gen3 Security & Compliance Follows NIST SP 800-53 Moderate





- First Gen3 Community Forum will take place on Oct 10-12, 2022.
- It’s virtual and free.
- Gen3.org/community/events/

To build a data commons

Gen3.org

Gen3 Data Commons

- Open source
- Define data model
- Import and curate data
- Create and export synthetic cohorts
- Analyze data, share data

Governance, legal agreements & best practices for building data commons & ecosystems

OPEN COMMONS CONSORTIUM

OCC-data.org

Open Commons Consortium

- Not-for-profit
- Data commons governance
- Data ecosystems governance
- Security & compliance services
- Legal templates
- Outsource operating data commons & ecosystems

To build a data mesh

DCF DataCommons Framework

Data mesh services

DCF.Gen3.org

Gen3 Data Commons Framework Services (DCFS)

- AuthN/AuthZ
- Digital ID and metadata services for data objects
- Emerging services for clinical, phenotype & other structured data

4. Lessons Learned

Six Reasons for Building Data Commons

Briefly, a data commons is a cloud-based software platform with a governance structure that allows a community to manage, analyze and share its data.

1. The functionality is compelling.
2. To speed the pace of research discoveries.
3. To create network effects.
4. To host data that is too large to be managed easily by research groups.
5. To reduce cost.
6. To protect sensitive data.

Source: Grossman, Robert L. "Ten Lessons for Data Sharing With a Data Commons." arXiv preprint arXiv:2207.11167 (2022).

Ten Lessons

1. Build a commons for a specific community with a specific set of research challenges.
2. Successful commons curate and harmonize the data.
3. It’s ultimately about the data and its value to generate new research discoveries.

Source: Grossman, Robert L. "Ten Lessons for Data Sharing With a Data Commons." arXiv preprint arXiv:2207.11167 (2022).

Ten Lessons (continued)

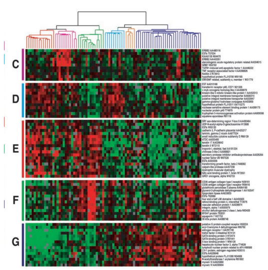
4. It is very important to reduce barriers to access to increase usage.
5. Data curation and developing interactive user interfaces is expensive.
6. Support an ecosystem of applications, not just a single system.
7. Security and compliance for data commons are expensive.

Ten Lessons (continued)

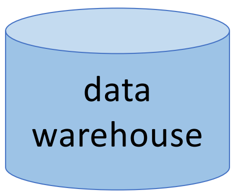
8. It’s not easy to predict what archived data will lead to great science.
9. Over time, the value of data commons will grow if it is part of a data mesh.
10. Resist the temptation to build a cloud-based walled garden.

Source: Grossman, Robert L. "Ten Lessons for Data Sharing With a Data Commons." arXiv preprint arXiv:2207.11167 (2022).

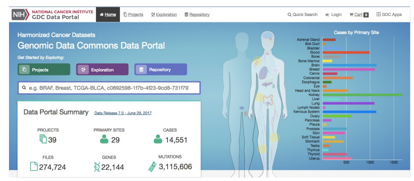
5. Conclusion and Summary



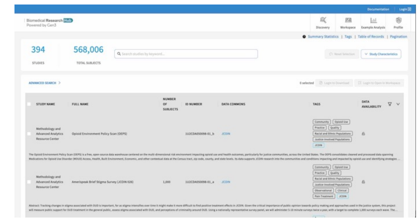
Databases organize data around a **project** (1970's)



Data warehouses organize the data for an **organization (**1990's)



Data commons organize the data for a scientific **discipline** or field (2010's)



Data meshes (aka data ecosystems) enable discoveries **across multiple commons operated by different organizations.** (2020's)

Data commons are software platforms that co-locate: 1) **well-curated data**, 2) cloud-based computing infrastructure, and 3) commonly used software applications, tools and services to create a governed resource for managing, analyzing, integrating and sharing data with a community.

Data meshes integrate multiple data commons and other data and computational resources.

Adapted from Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234

Proposed Open Commons Principles* (2017)

1. Agencies and foundations that fund biomedical research should require that researchers share the data generated.

2. Agencies and foundations that fund biomedical research should provide the computing infrastructure (“commons”) and bioinformatics resources that are required to support data sharing.

3. The data commons developed by agencies and foundations should themselves share data and interoperate with other data commons to create a data ecosystem (aka data mesh).

*Source: Robert L. Grossman, Supporting Open Data and Open Science With Data Commons: Some Suggested Guidelines for Funding Organizations, March 23, 2017, https://www.healthra.org/download-resource/?resource-url=/wp-content/uploads/2017/08/Data-Commons-Guidelines_Grossman_8_2017.pdf

Benefits of Data Commons and Data Sharing

1. Move the research **field forward faster**.
2. Support **repeatable, reproducible and open** research.
3. We have the statistical power to study **weaker effects.**
4. Researchers can work with **large datasets at much lower cost** and make discoveries of phenomena that are not evident at smaller scale.
5. Data commons can **interoperate** with each other to create a data mesh so that over time data sharing can benefit from a “network effect”

Source: Grossman, Robert L. "Ten Lessons for Data Sharing With a Data Commons." arXiv preprint arXiv:2207.11167 (2022).

Questions?

rgrossman.com
@bobgrossman

We are hiring and also looking for volunteers that want to impact biology, medicine, healthcare and the environment using data science and cloud computing. Please contact us at the CTDS or the OCC.

For more information (1 of 5)

Reviews and lessons learned about data lakes, data commons and data meshes for biomedical data

Grossman, R.L., 2022. Ten Lessons for Data Sharing With a Data Commons. arXiv preprint arXiv:2207.11167.

Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. PMID: 30691868 PMCID: PMC6474403

Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles and Practice of Oncology, May/June 2018, Volume 24 Number 3, pages 122-126 doi: 10.1097/PPO.0000000000000318. PMID: 29794537

For more information (2 of 5)

Genomic Data Commons (GDC)

An overview of the GDC:

Heath AP, Ferretti V, ... and Grossman RL, The NCI Genomic Data Commons, Nature Genetics 2021 Mar;53(3):257-262. PMID: 33619384 doi: 10.1038/s41588-021-00791-5. PMID: 33619384

Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt, Toward a Shared Vision for Cancer Genomic Data, New England Journal of Medicine, September 22, 2016, Volume 375, Number 12, pages 1109--12

An overview of the data processing for the GDC:

Zhenyu Zhang, Kyle Hernandez, Jeremiah Savage, Shenglai Li, Dan Miller, Stuti Agrawal, Francisco Ortuno, Louis M. Staudt, Allison Heath, and Robert L. Grossman, Uniform genomic data analysis in the NCI Genomic Data Commons, Nature communications 12, no. 1 (2021), pages 1-11. PMID: 33619257 doi: 10.1038/s41467-021-21254-9.

GDC API:

Shane Wilson, Michael Fitzsimons, Martin Ferguson, …, Robert L. Grossman, Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API, Cancer Research, volume 77, number 21, 2017, pages e15-e18. PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5683428/

For more information (3 of 5)

Gen3 Data Commons

The BloodPAC Data Commons:

Robert L. Grossman, Jonathan R. Dry, Sean E. Hanlon, Donald J. Johann, Anand Kolatkar, Jerry SH Lee, Christopher Meyer, Lea Salvatore, Walt Wells, and Lauren Leiman, BloodPAC Data Commons for liquid biopsy data, JCO Clinical Cancer Informatics Volume 5, 2021, pages 479-486. PMID: 33929890 DOI: 10.1200/CCI.20.00179

VA Data Commons:

Danne C. Elbers, Nathanael R. Fillmore, Feng-Chi Sung, Spyridon S. Ganas, Andrew Prokhorenkov, Christopher Meyer, Robert B. Hall, Samuel J. Ajjarapu, Daniel C. Chen, Frank Meng, Robert L. Grossman, Mary T. Brophy, and Nhan V. Do, The Veterans Affairs Precision Oncology Data Repository, a Clinical, Genomic, and Imaging Research Database, Patterns Volume 1 (2020) 100083. DOI: 10.1016/j.patter.2020.100083

Pandemic Response Commons:

Trunnell, Matthew, Casey Frankenberger, Bala Hota, Troy Hughes, Plamen Martinov, Urmila Ravichandran, Nirav S. Shah, Robert L. Grossman, and Pandemic Response Commons Consortium. "The Pandemic Response Commons." medRxiv (2022). doi: https://doi.org/10.1101/2022.06.20.22276542

For more information (4 of 5)

Gen3 Data Commons (cont'd)

Rush University Medical Center Gen3 Data Commons

O'Hara, Thomas, Anil Saldanha, Matthew Trunnell, Robert L. Grossman, Bala Hota, and Casey Frankenberger. "Economical Utilization of Health Information with Learning Healthcare System Data Commons." Perspectives in Health Information Management 19, no. Spring (2022).

Data Meshes

Biomedical Research Hub:

Craig Barnes, Binam Bajracharya, …, and Robert L. Grossman, The Biomedical Research Hub: A Federated Platform for Patient Research Data, Journal of the American Medical Informatics Association, 2021, doi:10.1093/jamia/ocab247.

For more information (5 of 5)

More about data commons:

Robert L. Grossman, et. al. A Case for Data Commons: Toward Data Science as a Service, Computing in Science & Engineering 18.5 (2016): 10-20. Also https://arxiv.org/abs/1604.02608

Data clouds for biomedical data:

Heath, Allison P., Matthew Greenway, Raymond Powell , …, Robert L. Grossman, Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. Journal of the American Medical Informatics Association 21.6 (2014): 969-975. DOI: 10.1136/amiajnl-2013-002155.

Interoperability of data commons:

Robert L. Grossman, Some Proposed Principles for Interoperating Data Commons, Medium, October 1, 2019., https://medium.com/@rgrossman1/some-proposed-principles-for-interoperating-data-commons-3668c6cf48df

Grossman, Robert L. "Supporting Open Data and Open Science With Data Commons: Some Suggested Guidelines for Funding Organizations." (2017), https://www.healthra.org/wp-content/uploads/2018/08/Data-Commons-Guidelines_Grossman_8_2017.pdf

Contact Information

Robert L. Grossman
rgrossman.com

@BobGrossman
robert.grossman@uchicago.edu

THE UNIVERSITY OF CHICAGO | Center for Translational Data Science

ctds.uchicago.edu

OPEN COMMONS CONSORTIUM

occ-data.org