National Institutes of Health
Office of Data Science Strategy

Toward a modern and FAIR biomedical data ecosystem

Laura Biven, PhD
Data Science Technical Lead
Office of Data Science Strategy
National Institutes of Health
Laura.Biven@nih.gov

1 April, 2021

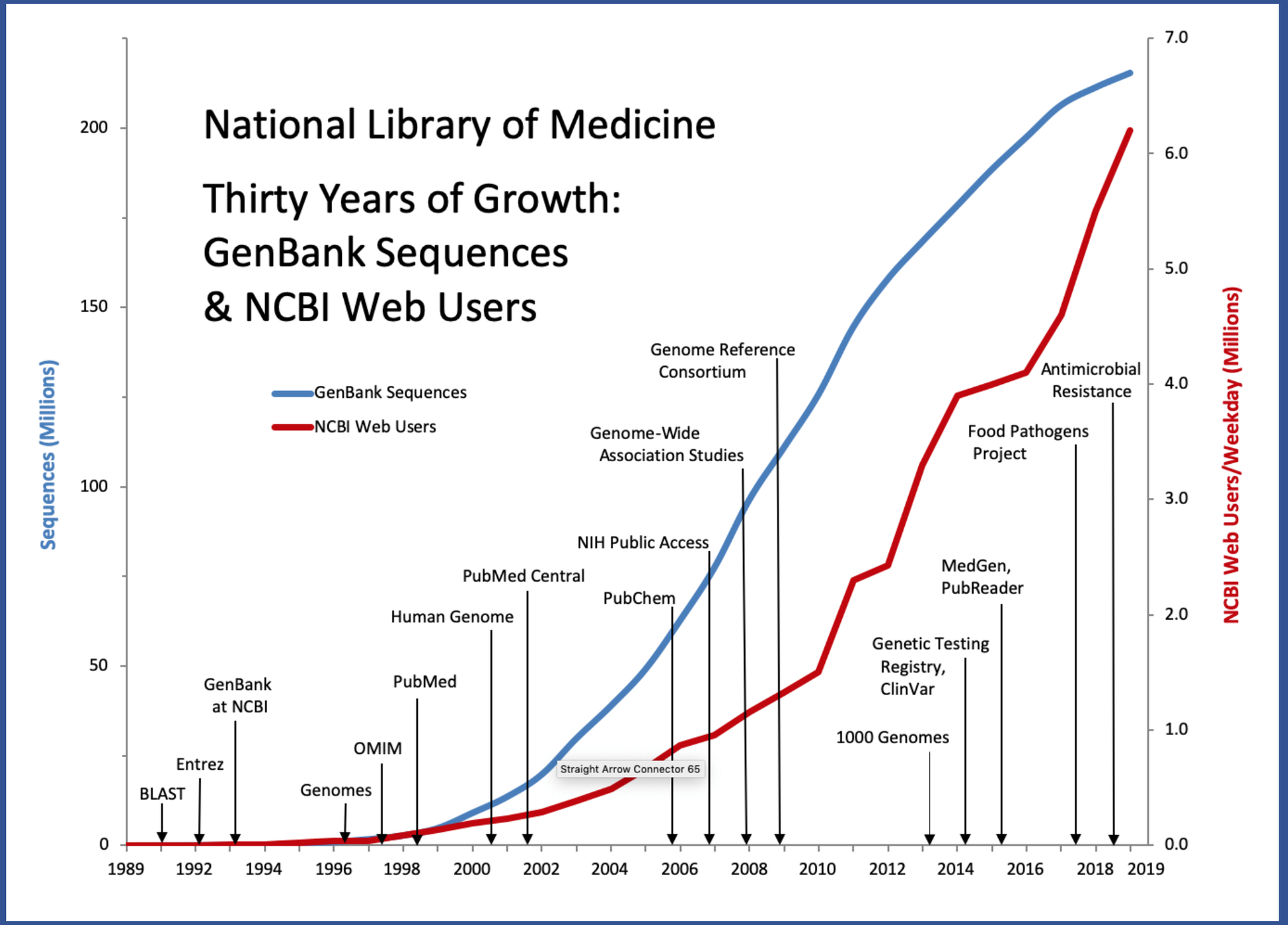
The National Institutes of Health

The Nation's Steward of Medical & Behavioral Research

"Science in pursuit of fundamental knowledge about the nature and behavior of living systems… and the application of that knowledge to extend healthy life and reduce illness and disability."

NIH History

- 1887: One-room “Laboratory of Hygiene” established by Dr. Joseph Kinyoun
  - Becomes lab of U.S. Public Health Service
- 1930: Ransdell Act: Hygienic Laboratory became National Institute (singular) of Health
- 1937: National Cancer Institute established with sponsorship by every U.S. Senator
- 1940: President Franklin D. Roosevelt dedicated buildings and grounds of Bethesda campus
- 1944: President Roosevelt created Federal Funding Law for NIH



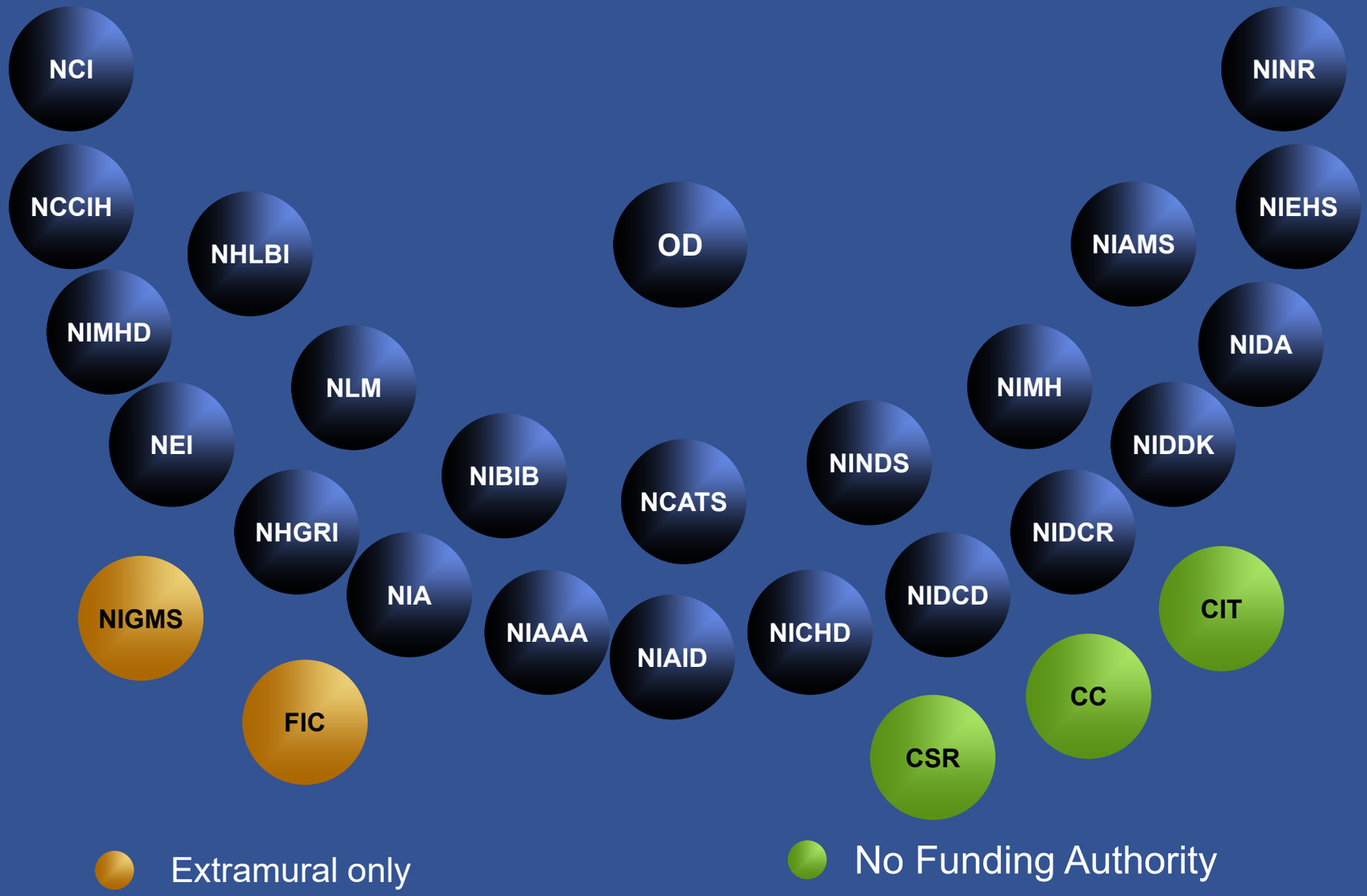
FY21 Congressional Justification NLM
https://www.nlm.nih.gov/about/2021CJ.html

NIH Today

- Conducts research in its own laboratories
- Supports research of non-Federal scientists
  - In universities, medical schools, hospitals, and research institutions throughout United States and overseas
- Helps train research investigators
- Fosters communication of medical information
- 153 NIH-supported researchers have become Nobel Laureates*

* As of 10/04/2017

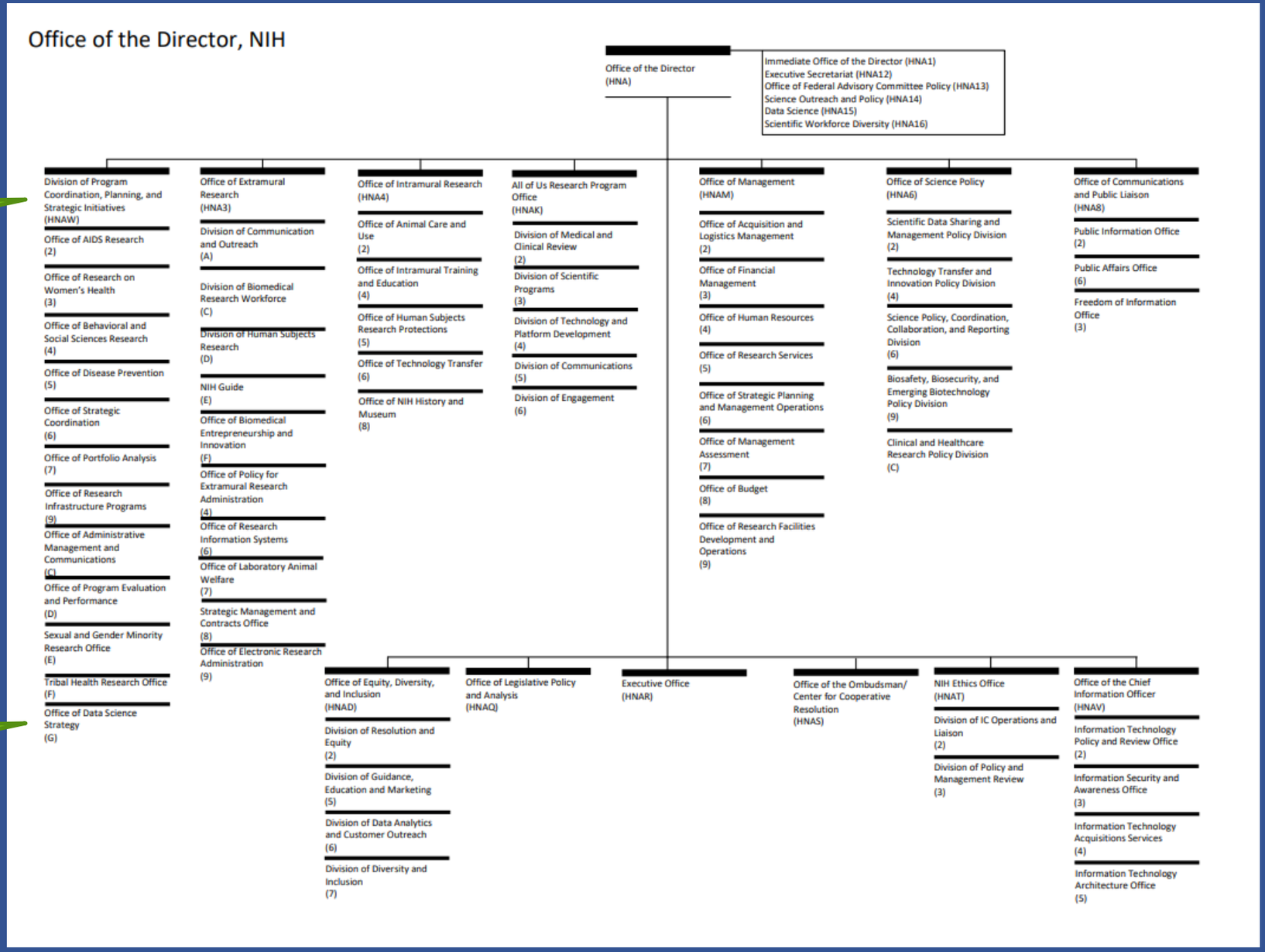
The National Institutes of Health



NIH Organization – Office of the Director

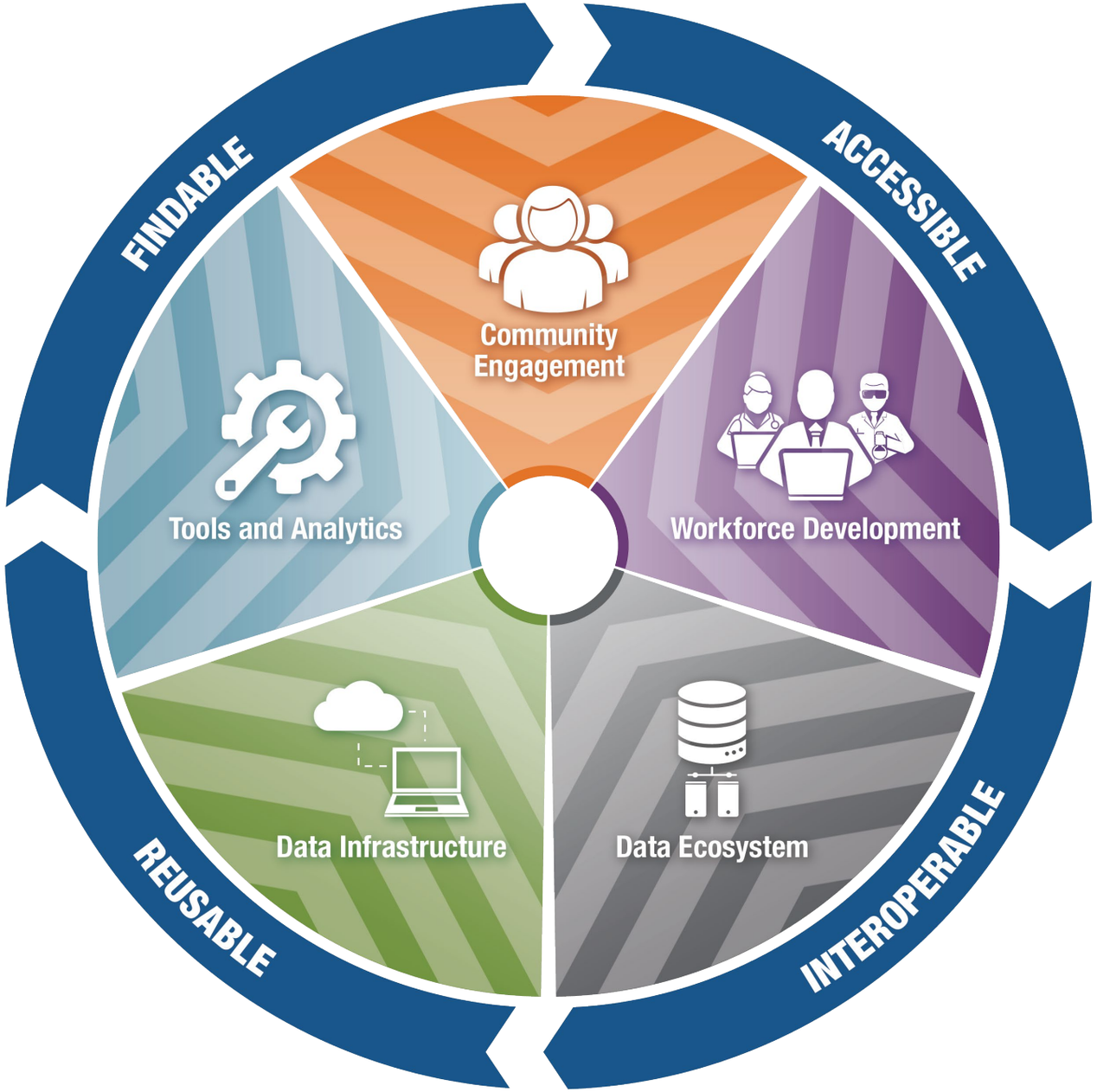
Division of Program Coordination, Planning, and Strategic Initiatives

Office of Data Science Strategy



VISION

a modernized, integrated, FAIR biomedical data ecosystem



https://datascience.nih.gov/nih-strategic-plan-data-science

Strategic Plan for Data Science: Goals and Objectives

1. Data Infrastructure

- Optimize data storage and security
- Connect NIH data systems

2. Modernized Data Ecosystem

- Modernize data repository ecosystems
- Support storage and sharing of individual datasets
- Better integrate clinical and observational data into biomedical data science

3. Data Management, Analytics, and Tools

- Support useful, generalizable, and accessible tools
- Broaden utility of, and access to, specialized tools
- Improve discovery and cataloging resources

4. Workforce Development

- Enhance the NIH data science workforce
- Expand the national research workforce
- Engage a broader community

5. Stewardship and Sustainability

- Develop policies for a FAIR data ecosystem
- Enhance stewardship

Making Data FAIR

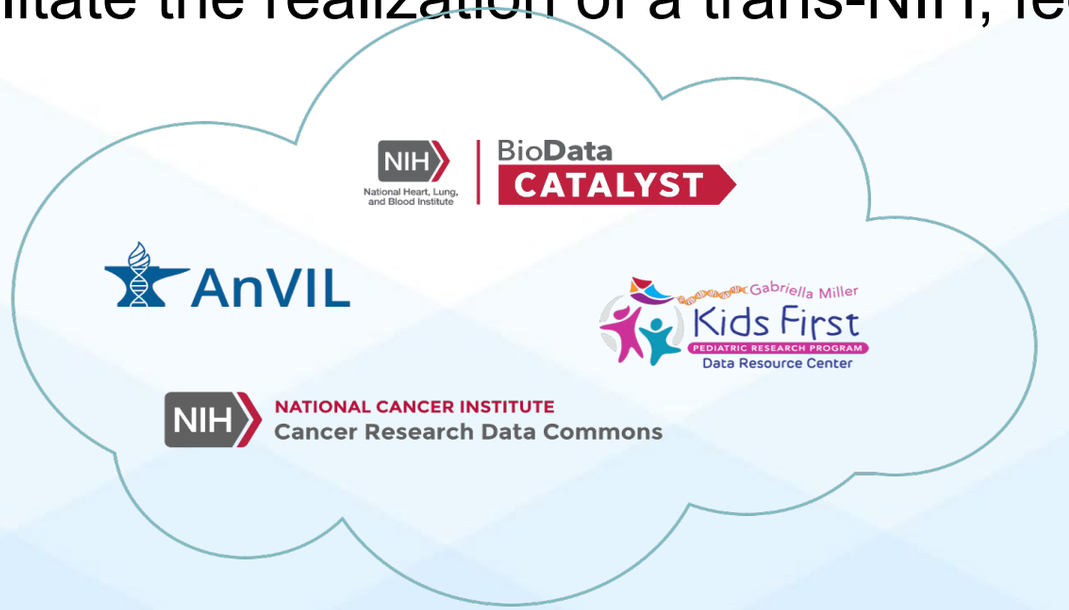
Findable	• must have unique identifiers, effectively labeling it within searchable resources.
Accessible	• must be easily retrievable via open systems and effective and secure authentication and authorization procedures.
Interoperable	• should “use and speak the same language” via use of standardized vocabularies.
Reusable	• must be adequately described to a new user, have clear information about data-usage licenses, and have a traceable “owner’s manual,” or provenance.

Two example initiatives

- NIH Cloud Platforms Interoperability (NCPI)
- NIH Researcher Auth Service (RAS)

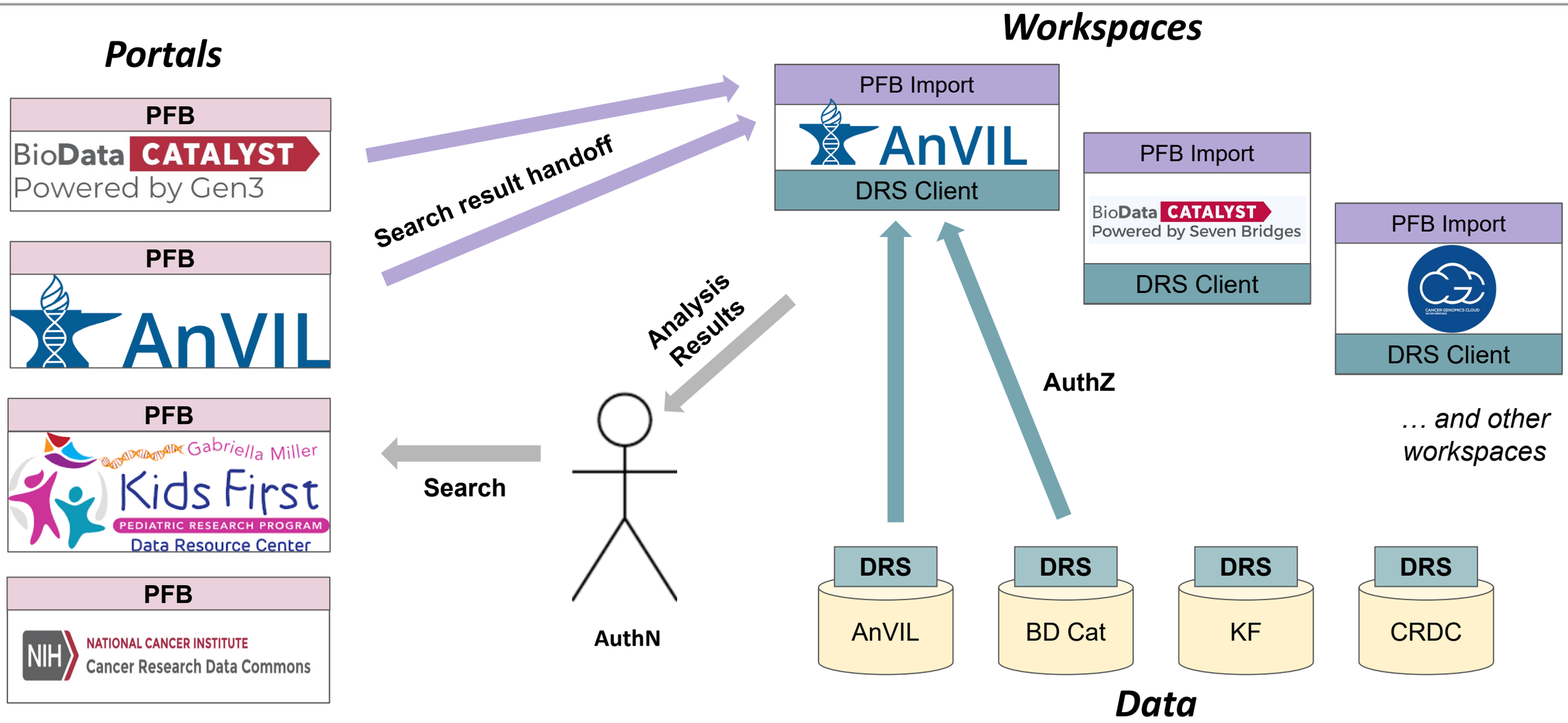
What is NCPI?

The NIH Cloud Platform Interoperability (NCPI) effort aims to establish and implement guidelines and technical standards to empower end-user analyses across participating NIH cloud platforms, to facilitate the realization of a trans-NIH, federated data ecosystem.



https://anvilproject.org/ncpi

Systems Interoperation WG - Technical 1st Year Vision



Tackling Multiple Layers of Interoperability

Challenge	Working Group	NCPI Activities
Operational barriers to trans-platform data sharing	**Community Governance**	Establish principles for promoting interoperability across multiple platforms; evaluate operational barriers
Inability to search & access data across platforms	**Systems Interoperation**	Test & implement technical standards for auth (RAS) & data exchange (e.g. GA4GH DRS) based on key use cases
Transitioning researchers to use the cloud	**Outreach & Training**	Create public “knowledge base”; create training materials
Lack of standards for clinical data exchange	**FHIR**	Pilot and assess FHIR resources to model and share complex clinical and phenotypic data

Systems Interoperation WG - 2020 Accomplishments

*Collectively, we have achieved improved interoperability in 2020 across multiple systems through **PFB**, **GA4GH DRS**, and **GA4GH Passports**.*

2020 Results

- **Search Result Handoff:** PFB

  2 portals
  ~417K subjects accessible

- **Data Access:** DRS 1.1

  4 DRS Servers
  ~6PB of data

- **Auth**: RAS for AuthN

  RAS

Supported Platforms

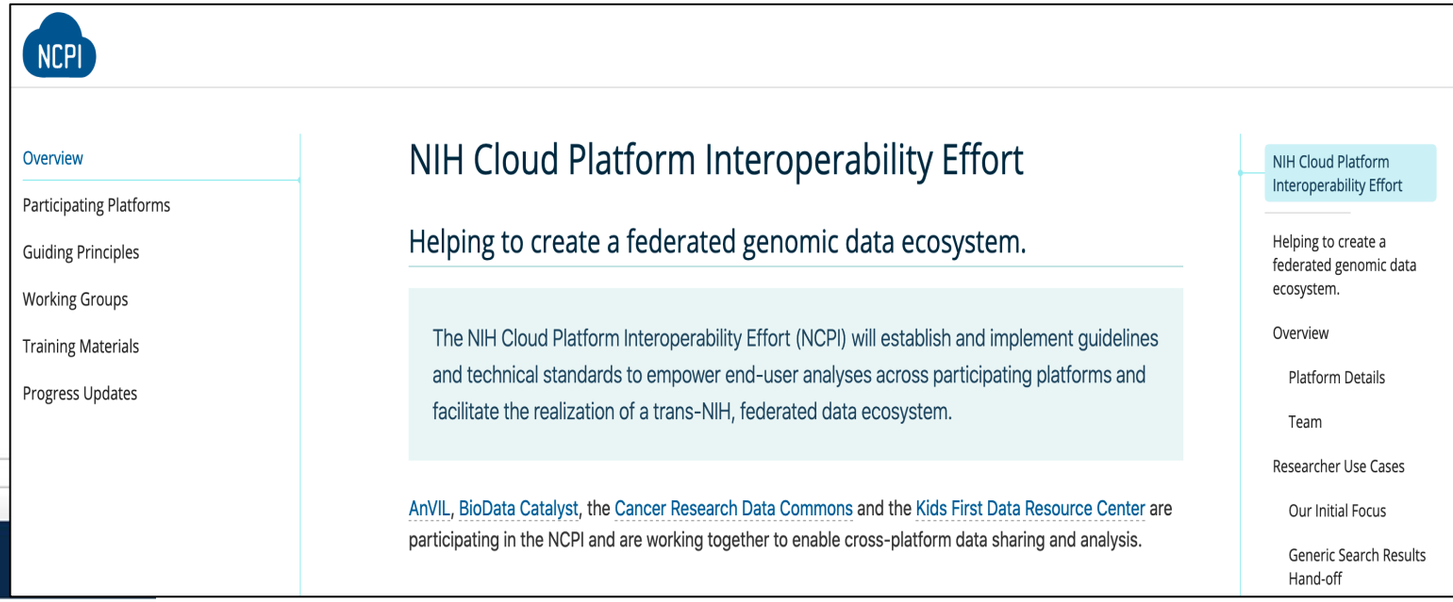
- The **NHGRI AnVIL** and **NHLBI BioData Catalyst** portals both support handoff of search results to **workspaces** (Terra, Gen3, SBG)
- We have data accessible on **AnVIL, BDCat, CRDC,** and **Kids First** via **DRS 1.1** support
- **GA4GH Passports** are in use by **RAS** and support visas from dbGaP made accessible by Gen3.

Additional Challenges for Potential NCPI Roadmap

Challenge	Potential NCPI Activities?
Users don't want to use the cloud if their favorite tools and workflows are not there	Potential new WG to port workflows to the cloud?
New programs, platforms, and databases want to play in the sandbox	How do we onboard new programs or development teams to NCPI?
How to estimate cloud costs for researcher analyses	Benchmark pipelines? Create public cloud cost guide?
Complex clinical and phenotypic data (that don't map to CDMs/CDEs)	FHIR as a flexible structure for clinical data interoperability (even if not derived from EHRs)

Web Presence

https://datascience.nih.gov/nih-cloud-platform-interoperability



https://anvilproject.org/ncpi

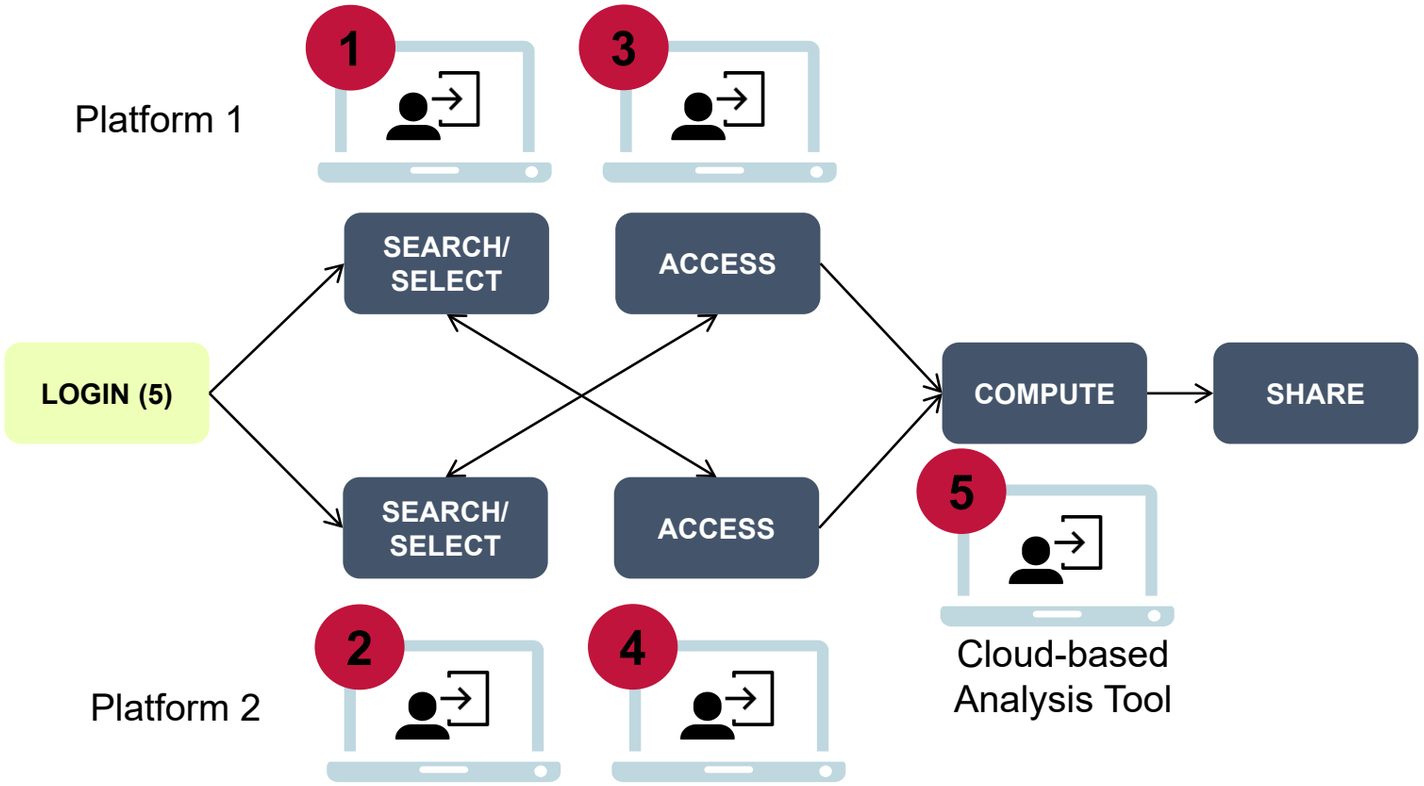
Why an NIH Researcher Auth Service?

"There is a strong need to simplify interoperability between resources by **enabling a researcher to log in once** and then have their appropriate authentication and authorization information travel with them as they move between platforms.

Without a **standard protocol for describing authorization information** or passing it between platforms, the burden shifts to the user to configure each system - currently a user may have to authenticate multiple times using the same username/password in order to access the same dataset across different platforms."

Researcher Workflows Before NIH RAS

Problem statement: Researchers login and/or give consent **at least 5 times** to search for and analyze controlled access data in 2 NIH data repositories

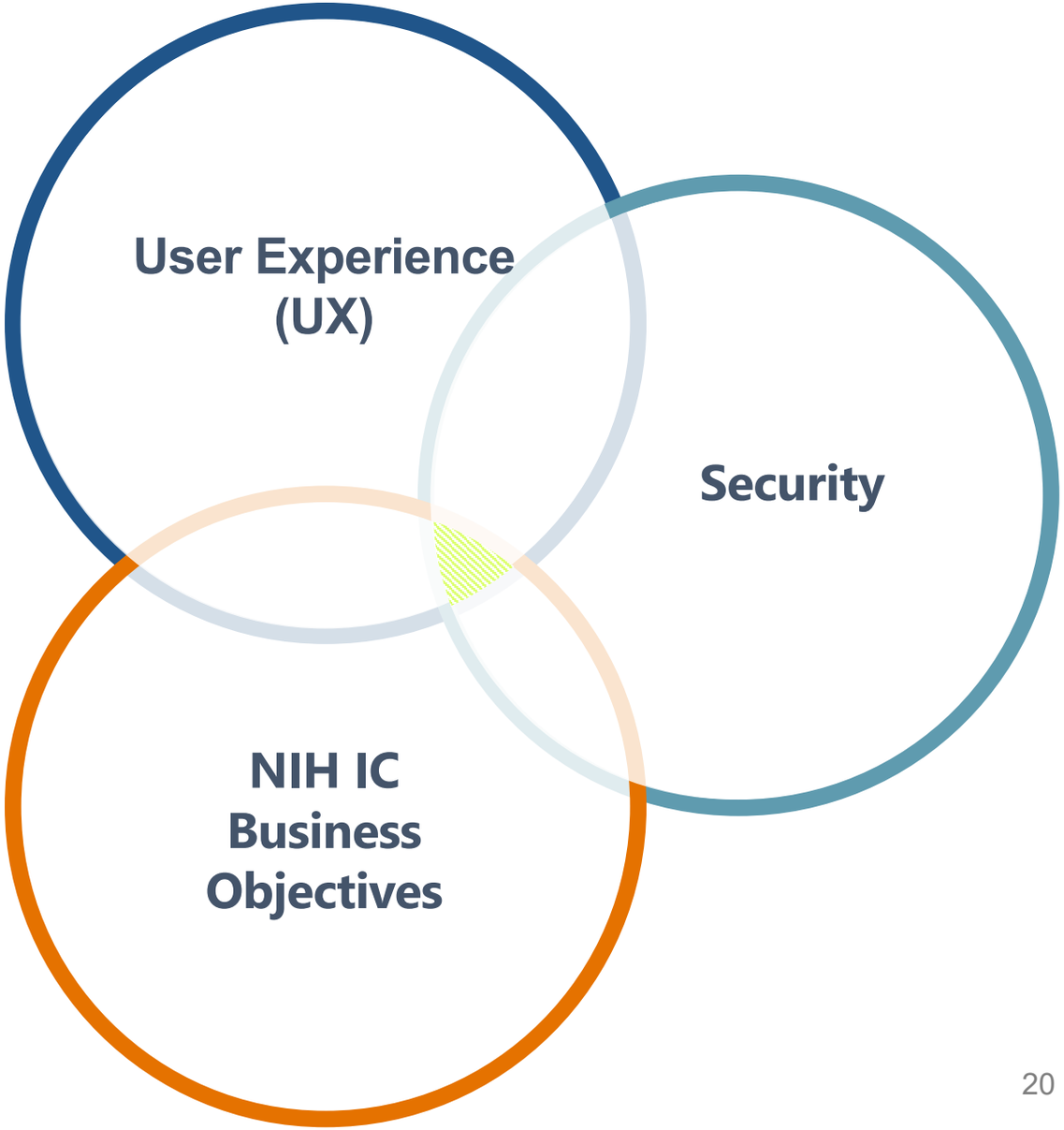


NIH RAS Scope & Agile Strategy

AUTHENTICATION
No matter what preferred credentials researchers enter, their account identity is recognized

AUTHORIZATION
Researchers have access to the datasets they need and expect in ways they understand; web of trust

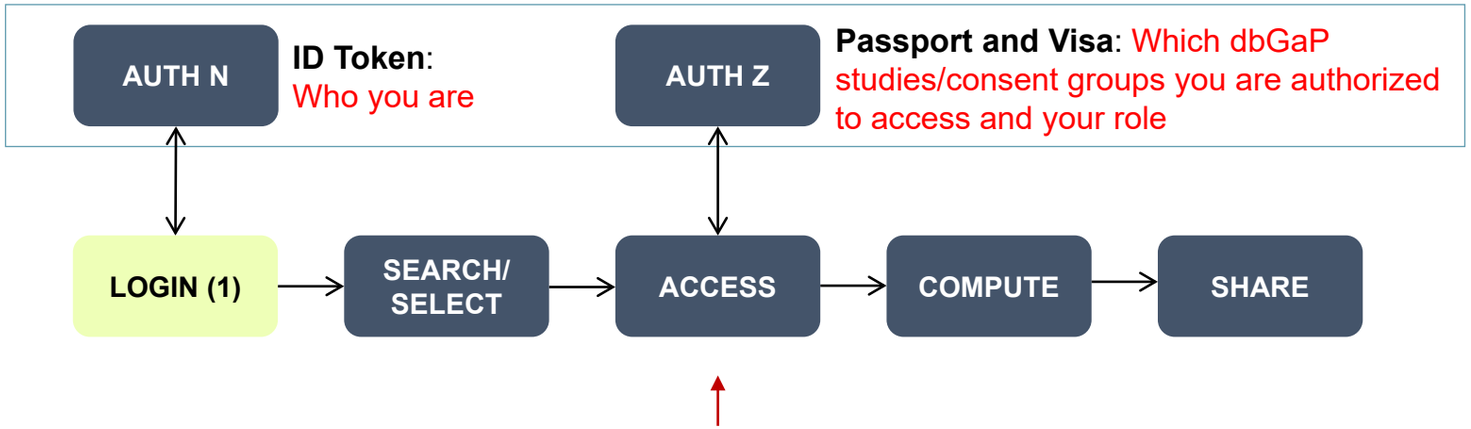
AUDITING
Trace and log data in a standard way to protect staff, intellectual property, and human data



Researcher Workflows After RAS Phase 1 Release

RAS V1: Authentication and Authorization provided by a central NIH service.

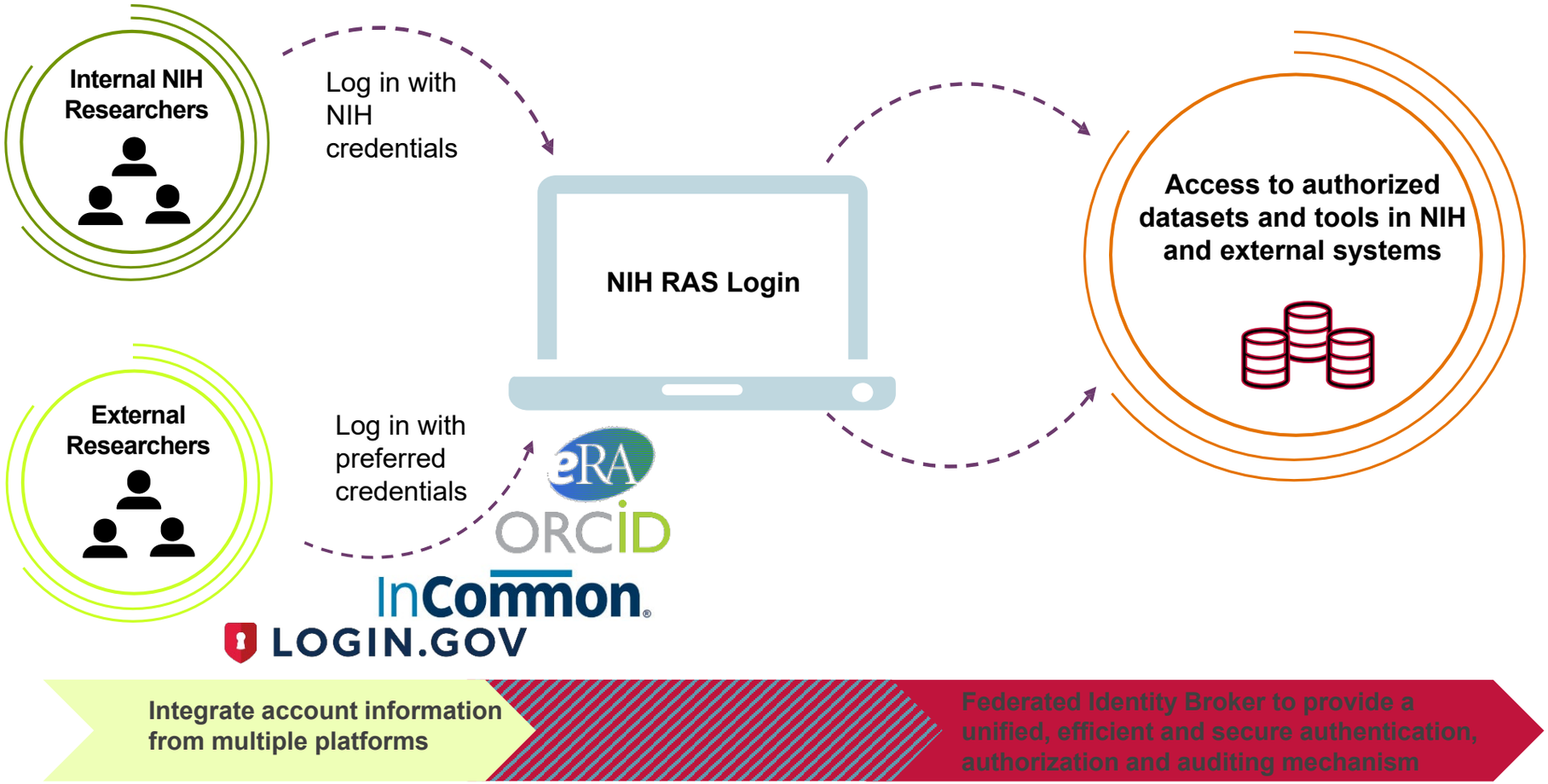
Auth tokens move with the user as they navigate to any of the four Phase 1 Data Platforms so that the researcher only logs in **one time to RAS**



Before provisioning data, the platform validates the passport/visa by calling RAS, so access information is always up to date within the last hour

NIH RAS Vision

NIH Researcher Auth Service (RAS) is a unified, efficient, and secure authentication and authorization service that enables streamlined access by researchers to NIH-funded data resources across multiple systems, and provides standardized methods of logging and auditing such access.



Design Informed by Early Adopter IC Systems

Phase 1		Phase 2	
Gabriella Miller Kids First Pediatric Research Program Data Resource Center	Common Fund Kids First Data Resource Center (KFDRC)	dbGaP Genotypes and Phenotypes	NCBI Database of Genotypes and Phenotypes (dbGaP)
BioData Catalyst	NHLBI Biodata Catalyst (formerly DataSTAGE)	Common Fund Data Ecosystem	Common Fund Data Ecosystem (CFDE)
National Cancer Institute GDC Data Portal	NCI Cancer Research Data Commons (CRDC)	All of Us Research Program	All of Us (AoU)
AnVIL	NHGRI Analysis, Visualization and Informatics Lab-space (AnVIL)	NIMH National Institute of Mental Health	NIMH Data Archive (NDA)

NIH National Institutes of Health
Turning Discovery Into Health

https://auth.nih.gov/docs/RAS/serviceofferings.html

Researcher Auth Service (RAS) Project Service Offerings

Department of Health & Human Services (HHS)
National Institutes of Health (NIH)
Center for Information Technology (CIT)
Identity & Access Management (IAM) Services

Getting Started

This document explains the authentication, authorization, and logging services available to NIH Institutes and Centers and extramural systems desiring information about users requesting to access NIH's open and controlled data assets and repositories through the NIH Researcher Auth Service (RAS).
The intended audience for this document is technical developers and leaders responsible for design and implementation of the integrations with the RAS APIs.

Prerequisites

The following are the recommended prerequisites to using this document:

- Familiarize with the **OAuth 2.0 Specification, RFC 6749**.
- Familiarize with the **OpenID Connect 1.0 Specifications**.
- Familiarize with the **JSON Web Token (JWT) specification, RFC 7519** and understanding of claims and tokens.
- Familiarize with **Global Alliance for Genomics & Health (GA4GH) Passport Specifications**.

Office of Data Science Strategy

www.datascience.nih.gov

A modernized, integrated, FAIR biomedical data ecosystem