April 2024

The Future of Decentralized Data

Karl Gutwin, Ph.D.
Principal Consultant
Data Strategy and Ecosystems

bioteam.net

BioTeam
Accelerate Science

The Grand Vision: Interconnecting Biomedical Data

Human-driven Barriers to Data Access

1

The **data accessibility gap** between the "haves" and the "have nots" is a *systemic equity problem*

- The "haves" - those who have the time, skill, and resources to wrangle data

2

FAIR *can only be achieved* by a diverse set of technologies, speaking the same language

- This cannot be one vendor, one approach - science is too diverse

Limitations of typical Data Platforms

Rigid Schema

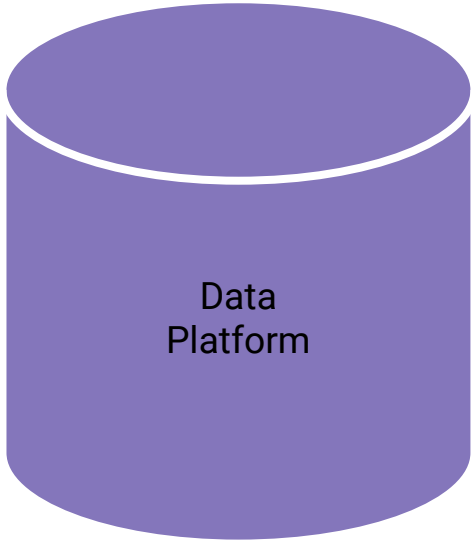
✅ predictable data model
!! inflexible; ETL data loss

Experts Load Data

✅ enforce standards
!! gatekeeping

Code-Heavy Data Operations

✅ use best technologies
!! excludes low-code users



SILO?

Purpose-Built

✅ designed to meet use cases
!! reinventing the wheel

Limited Data Ingest Formats

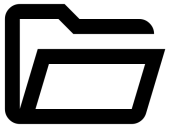
✅ work with known formats
!! constant ETL maintenance

Data Catalog

✅ centralized external docs
!! separate meaning from schema

Common Alternatives

File/Object Store



✅ holds everything

!! structured data handling needs extra work

NoSQL Database

✅ highly dynamic

!! schema mutation over time is painful

SaaS Data Workbench

✅ lots of slick tooling

!! walled garden, only works as long as you keep paying the bill

Open-Source Code Workbench

✅ do anything with code

!! high skill barrier of entry

What if we could build a Data Platform…

… more flexible than a traditional data platform:

- Dynamic schema, supports any data, with integrated tools for users to manipulate and create derivative datasets

… more interactive than a shared filesystem or cloud drive:

- Data, documentation, and visualizations all in one place

… more open than a commercial vendor's product:

- Use existing standards to foster interoperability via federation

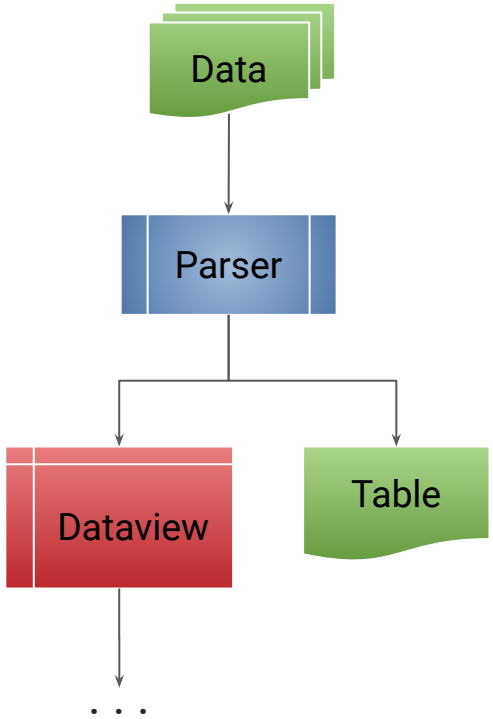
The Path to an Interoperable Ecosystem

Current State		Ideal
 Harmonizing scientific data from multiple sources is time consuming, generally *ad hoc*, and the results are difficult to share with others.		 **Facilitated Curation:** A platform for people to find, browse, share, combine and remix datasets on computable community-derived data dictionaries and CDEs.
 Exploring and merging datasets often requires writing code or performing complex operations.		 **Dynamic Transformation.** A "no-code" interface for users to transform data based on specified rules, allowing for datasets from multiple sources to be analyzed without creating unnecessary artifacts.
 Data commons systems are purpose-built for specific use cases and have to "own" all the data.		 **Federation By Design:** A federated system model enabling interaction with distributed data managed by others and providing a fabric on which existing data commons will be able to interoperate.

Four Core Principles

Principle 1: Accept all data, transform it live

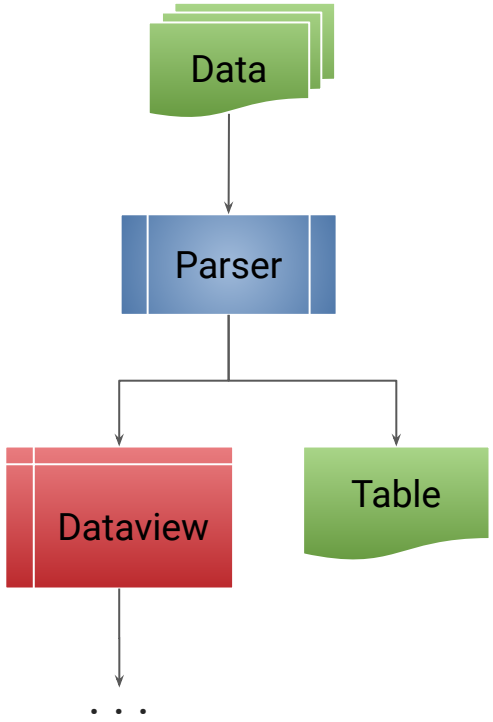
- Also known as ELT (extract, load, transform)
- Too many problems with "define a schema, transform all data before loading"
  - There is no "one true model" of biomedical data
  - Data models change over time
- Support data from all sources using a plugin-driven interface
- Incorporate a high-performance transformation engine based on SQL
- Allow **users** to easily create their own transformations and compose them using building blocks developed by the community



Principle 1: Accept all data, transform it live

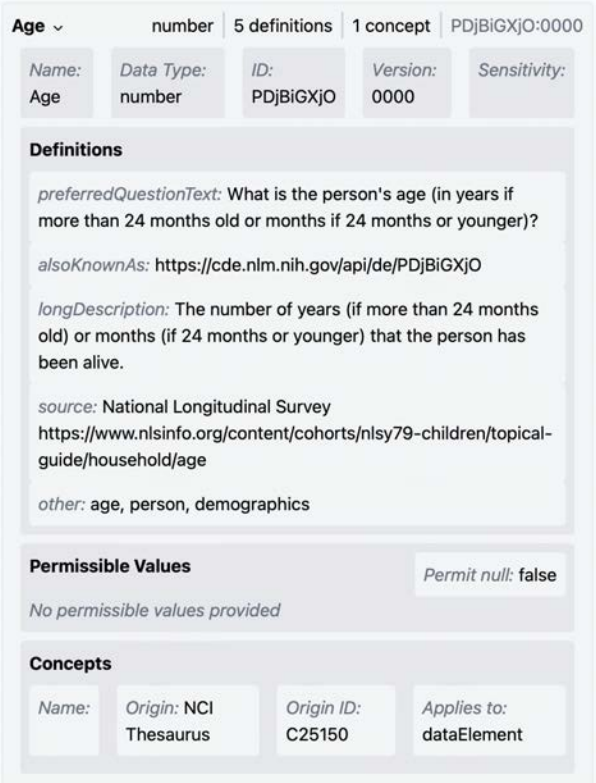
Approach

- `Python` *flexibility, data science heritage*
- `Polars` *high-performance dataframes*
- `DuckDB` *in-memory SQL OLAP engine*
- `PRQL` *pipeline-oriented transformation DSL*
- `Parquet` *data caching*



Principle 2: Data Elements as metadata

- Current practice is to store column-level metadata in a data catalog, data dictionary, or external documentation
  - Distant from the data, disconnected from transformations
- Data Elements answer the question "what does the data mean?"
  - How it was collected, what values are allowed, what concepts are represented
- Transformations should affect both the data and the data elements
- Common Data Elements - reuse as much as possible



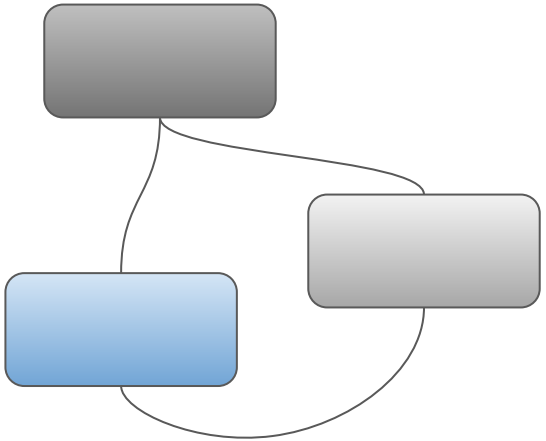
Principle 2: Data Elements as metadata

Approach

- Data Elements integrated from plugins, such as the NIH CDE Repository
- Track column/data element associations, propagate through data transformations
- Automatic data validation based on data element permissible values

Principle 3: Federation over centralization

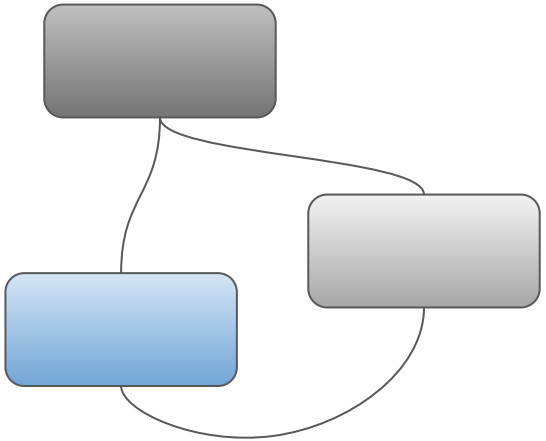
- Science is distributed, and the technology should reflect and support that model
- Allow user communities to maintain their own data, and share it securely with others
- The platform should abstract away the technical details of data movement
  - Links to remote data ought to behave just like local files, from the user's perspective
- Reuse existing open protocols
  - Other implementations of these concepts should be welcomed
- Future vision: Global federated search engine



Principle 3: Federation over centralization

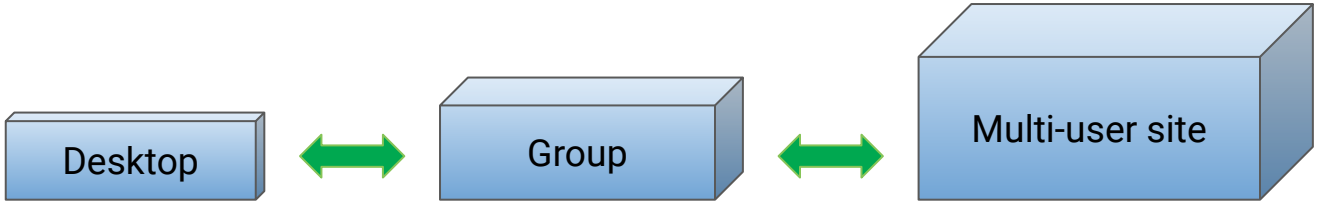
Approach

- Automated data transfer and caching
- Implementation of existing open standard protocols
  - GA4GH Data Connect for tabular data
  - ActivityPub for site-to-site messaging and access requests



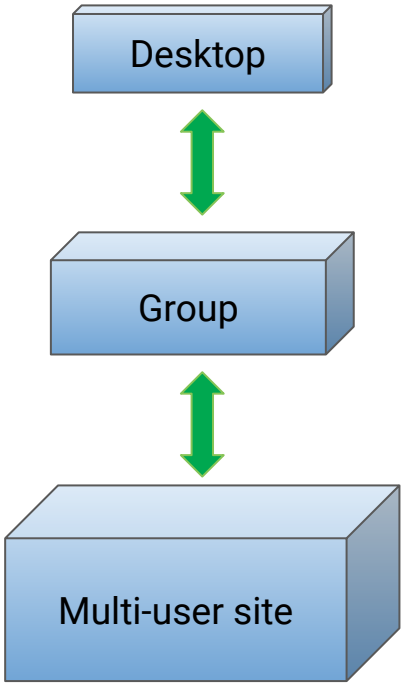
Principle 4: One platform, at any scale

- All of the previous principles apply at all scales, whether the user is working on their desktop, using a group server, or a large multi-user deployment



- It *should be easier* for a curious user to try the platform themselves than a sysadmin should be able to deploy it to the cloud

Principle 4: One platform, at any scale



Approach

- Single Docker container has all necessary dependencies
- Designed for horizontal scalability using load balancing, with multiple supported storage backends
- Can be deployed at scale using cloud-native or Kubernetes orchestration

Empowering Decentralized Data

A New Model for Data Infrastructure

A platform for *any* data

- The universality of tabular data
- Computable metadata
- Structured data is validated live

Expanding the definition of data items

- “Links”, to provide access to remote data
- “Dataviews", to enable on-the-fly transformations

Federated, not siloed

- Independently operated instances agreeing to a lightweight, standards-based set of communications protocols
- One core platform at any scale - from desktop to multi-center institution
- Enables search and sharing across labs/institutions/organizations

Framework for collaboration

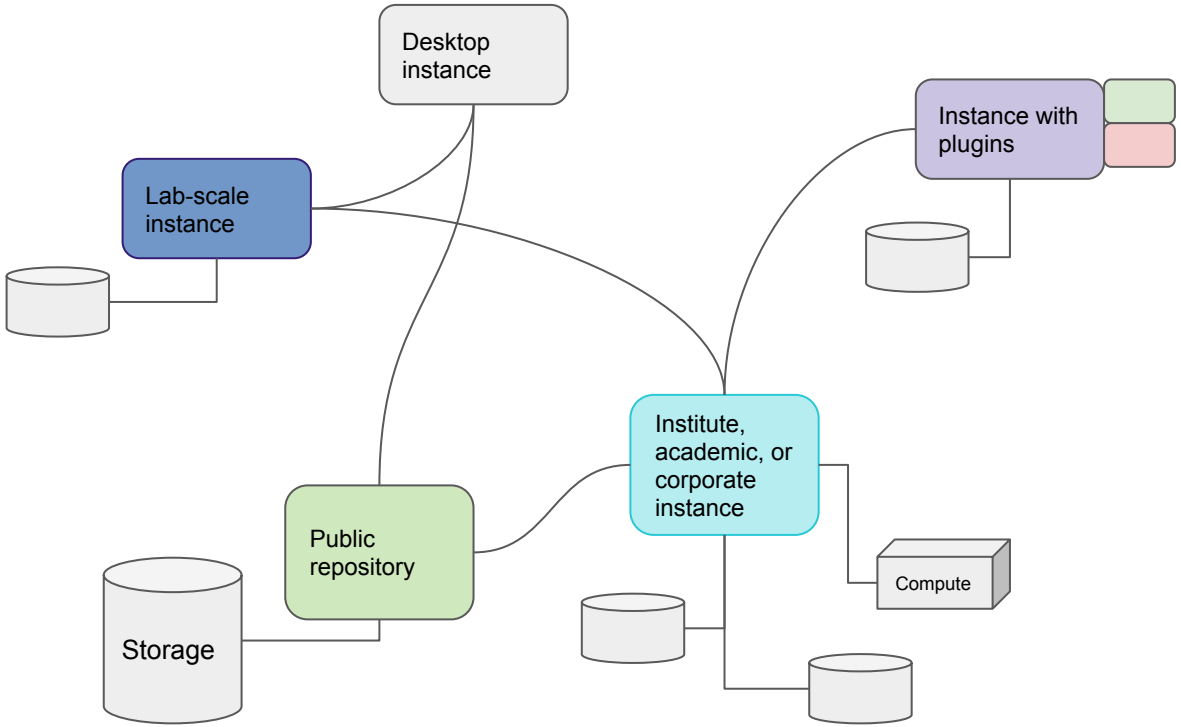
- Data dictionaries developed as open-source resources
- Share the data, and the *transformations* that make the data, to make data more reusable

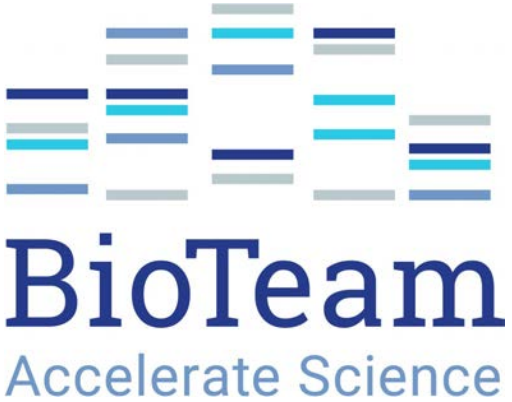
The Grand Vision: Interconnecting Biomedical Data

The Big Idea:

We break down barriers to sharing and interoperability by envisioning an open-source solution that anyone can run, customize to their needs, publish and remix data, with appropriate access controls.

In this way, we contemplate not just a platform, but a true *global ecosystem* of users and use cases, across any kind of biomedical data.





Thank You!

bioteam.net