Childhood Cancer Data Initiative Webinar Series

CCDI Federated Data: Enhancing Data Discoverability

*Geoff Lyle, Clay McLeod, Martin Ferguson,
and Allison Heath*

NIH NATIONAL CANCER INSTITUTE

August 13, 2024

Agenda

1. *CCDI Data Federation Resource: Background*
2. *Data Harmonization: Aligning the Data to NCI Standards*
3. *Data Federation Resource API: Designing the API*
4. *Using the Federation Data Demo: Leveraging Jupyter Notebook*
5. *Future Applications*
6. *Q&A*

Today’s Speakers

Geoff Lyle
Technical Project Manager, Treehouse Childhood Cancer Initiative

Clay McLeod
Director, Product Development and Engineering, St. Jude Children’s Research Hospital

Martin Ferguson
External Consultant, National Cancer Institute

Allison Heath
Director of Data Technology and Innovation, Children's Hospital of Philadelphia

CCDI Data Federation Resource: Background

Geoff Lyle

Why a Federated Childhood Cancer Data Ecosystem?

- Pediatric cancer data are currently siloed
  - Reduced and delayed access to data
  - Missed therapeutic opportunities

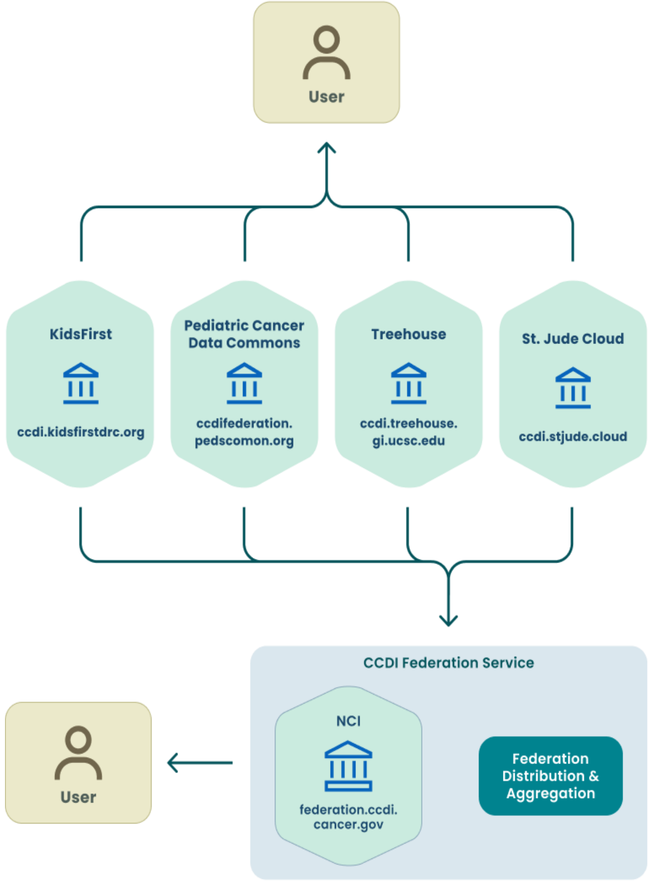
Data Federation Objectives

Facilitate large-scale biomedical research via a federated, real-time data search API

- Develop and implement a common application programming interface (API) specification where deidentified, participant-level data from each member can be queried.
- Results from queries return responses, leveraging an ever-growing, harmonized set of metadata values.
- Rich, faceted search across the supported information.
- Data will not be moved or centrally warehoused; instead, users can access the data where it resides.

Data Federation Status

- Four current federation members:
  - Kids First Data Resource Center
  - Pediatric Cancer Data Commons
  - Treehouse Childhood Cancer Initiative
  - St. Jude Cloud
- Version 1.0 API implementation and demos
  - If you're interested in joining, please email NCIChildhoodCancerDataInitiative@mail.nih.gov


User
KidsFirst
ccdi.kidsfirstdrc.org
Pediatric Cancer Data Commons
ccdifederation.pedscomon.org
Treehouse
ccdi.treehouse.gi.ucsc.edu
St. Jude Cloud
ccdi.stjude.cloud
CCDI Federation Service
NCI
federation.ccdi.cancer.gov
Federation Distribution & Aggregation
User

Driver Scientific Use Cases

Scientific Use Case	Description
1. Disease and Genomic Variant Querying	Search for diseases or genomic variants to gather data on alterations, uncertain variants, or mutations
2. Participant Cross-reference System for Data Retrieval	Retrieve comprehensive clinical and genomic information across institutions
3. RNA Sequencing Map t-SNE	Generate a global gene expression map using t-distributed stochastic neighbor embedding (t-SNE) analysis, providing visual insights into RNA expression patterns worldwide
4. Flexible, Tiered User Query Handling	Allow users to submit general queries and explore detailed matches to simplify data exploration and retrieval
5. Harmonizing Multiple Batches of Samples	Utilize internal patterns within datasets to align multiple batches of samples for a comprehensive analysis
6. Biospecimen Metadata Integration	Incorporate biospecimen information as an additional layer, expanding beyond clinical features and data types for users

Implementation Approach

Established two working groups:

1. **Data harmonization** – Ensures federation member data are harmonized to NCI standards
2. **API development** – Determines the best methods for delivering data that is accessible and useful to users querying information
   - Documents implementation guidelines for federation members to share data via the open API

Data Harmonization: Aligning the Data to NCI Standards

Geoff Lyle

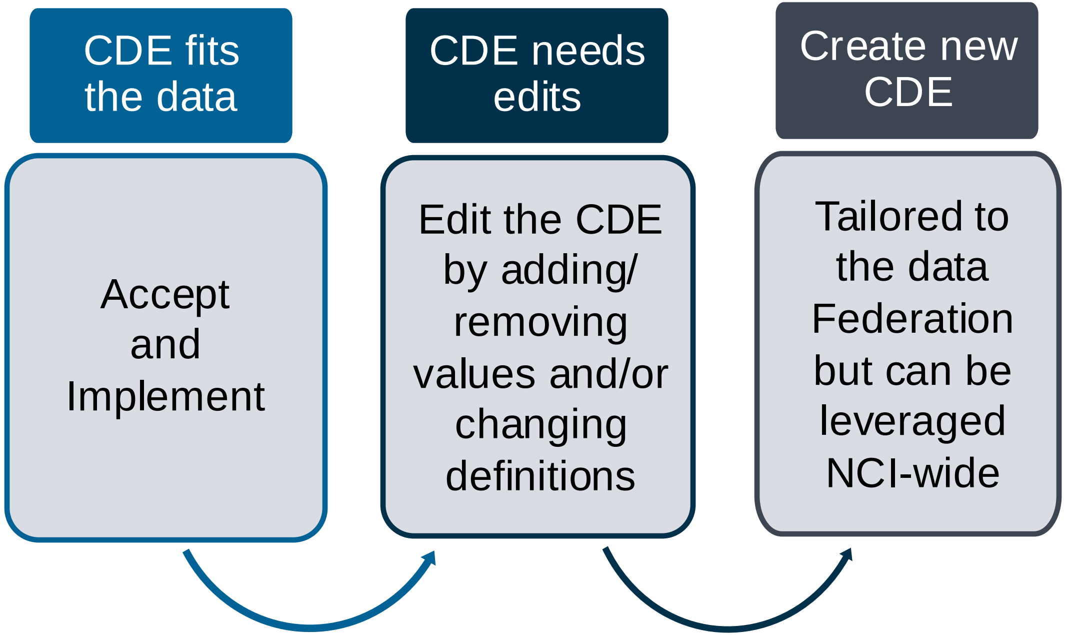
V1 Federation Resource Data Summary

Through the API, the federation delivers **metadata** that will help users create a synthetic cohort across multiple institutions/data types.

Source	Data Types	Data Level	Subjects (Participants)	Samples	Files
St. Jude Cloud	Genomic data, gene expression, imaging	Participant	13,956	19,866	133,579
UCSC - Treehouse	Genomic data, gene expression	Participant	12,483	12,770	6
Kids First - CHOP	Genomic data, gene expression, clinical data, imaging	Participant	34,066	162,549	327,893
PCDC - UChicago	Clinical data, imaging	Aggregate	22,667		

Data Harmonization Approach

- Leverage existing CCDI standards
- Utilize caDSR Common Data Elements (CDEs) to map attributes and allowable values (cadsr.cancer.gov)
- Develop harmonization guidelines when no NCI standard exists
- All discussions are tracked on GitHub



CDE fits the data

Accept and Implement

CDE needs edits

Edit the CDE by adding/ removing values and/or changing definitions

Create new CDE

Tailored to the data Federation but can be leveraged NCI-wide

V1 Federation Resource Scope – Harmonized Fields

Common Data Elements (CDEs) from Cancer Data Standards Registry and Repository (caDSR) (https://cadsr.cancer.gov)

Subject (Participant)	Sample	Study and File
Sex (6343385)	Sample tumor status (5432687)	Study short title (11459812)
Race (2192199)	Tumor classification (12922545)	Study name (11459810)
Ethnicity (2192217)	Age at diagnosis (3225640)	dbGaP phs accession (11524544)
Vital status (2847330)	Age at collection (14473376)	Institution (12662779)
Age at vital status (14480965)	Library strategy (6273393)	File location (Link/Gateway) (11556141)
Subject ID (6867052)	Preservation method (8028962)	File description (11280338)
	Disease diagnosis (ICD-O; WHO CNS5)	File size (11479876)
	Disease phase (12217251)	md5sum (11556150)
	ICD-O morphology code & term (11326261)	File type (11416926)

Data Federation Resource API: Designing the API

Clay McLeod

API Strategy

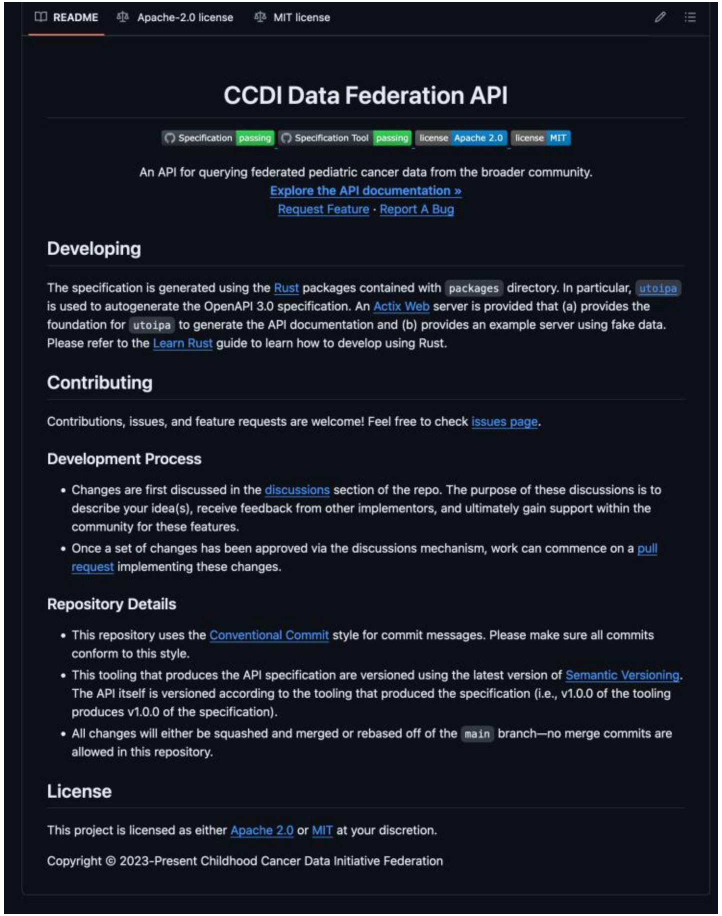
- **Select a standard that is purpose-built** for indexing the specific types of data we wanted to share (no enforced metadata standards).
- **Chose a scalable foundation** that will work for hundreds of thousands of samples and millions of files from day one.
- **Enable relaying information provided by source servers** with as little on-the-fly transformation as possible to enable high-performance aggregator services (e.g., the NCI aggregation server).
- **Ensure that joining the federation is as accessible as possible.**
  - The specification should not be more complicated than necessary.
  - The specification can be readily implemented using multiple open-source frameworks.

API Strategy (cont.)

- **Considered three strategies/standards for creating an API:**
  - FHIR API
  - Beacon V2 API (GA4GH standard)
  - Bespoke API
- After careful review of the existing standards, discussion with the specification designers (for Beacon only), and internal discussion, we jointly decided that, today, **neither FHIR nor Beacon V2 met all our criteria**.
- Given this, we decided to start by **creating a simple, bespoke API** that was purpose-built for indexing federation data and metadata.

Development Approach

- The specification itself was designed to be robust and stand the test of time.
  - The OpenAPI specification itself, and all other source code, was written in Rust.
  - An example server with fake data that implementors can refer to.
  - A toolset for implementations to test compliance with the spec.
  - Testing to ensure the spec remains well-formed on every proposed change.
- All discussions regarding both the design of the API are open and searchable in the federation GitHub repository.



CCDI Data Federation API

An API for querying federated pediatric cancer data from the broader community.
Explore the API documentation »
Request Feature · Report A Bug

Developing

The specification is generated using the Rust packages contained with `packages` directory. In particular, `utoipa` is used to autogenerate the OpenAPI 3.0 specification. An Actix Web server is provided that (a) provides the foundation for `utoipa` to generate the API documentation and (b) provides an example server using fake data. Please refer to the Learn Rust guide to learn how to develop using Rust.

Contributing

Contributions, issues, and feature requests are welcome! Feel free to check issues page.

Development Process

- Changes are first discussed in the discussions section of the repo. The purpose of these discussions is to describe your idea(s), receive feedback from other implementors, and ultimately gain support within the community for these features.
- Once a set of changes has been approved via the discussions mechanism, work can commence on a pull request implementing these changes.

Repository Details

- This repository uses the Conventional Commit style for commit messages. Please make sure all commits conform to this style.
- This tooling that produces the API specification are versioned using the latest version of Semantic Versioning. The API itself is versioned according to the tooling that produced the specification (i.e., v1.0.0 of the tooling produces v1.0.0 of the specification).
- All changes will either be squashed and merged or rebased off of the `main` branch—no merge commits are allowed in this repository.

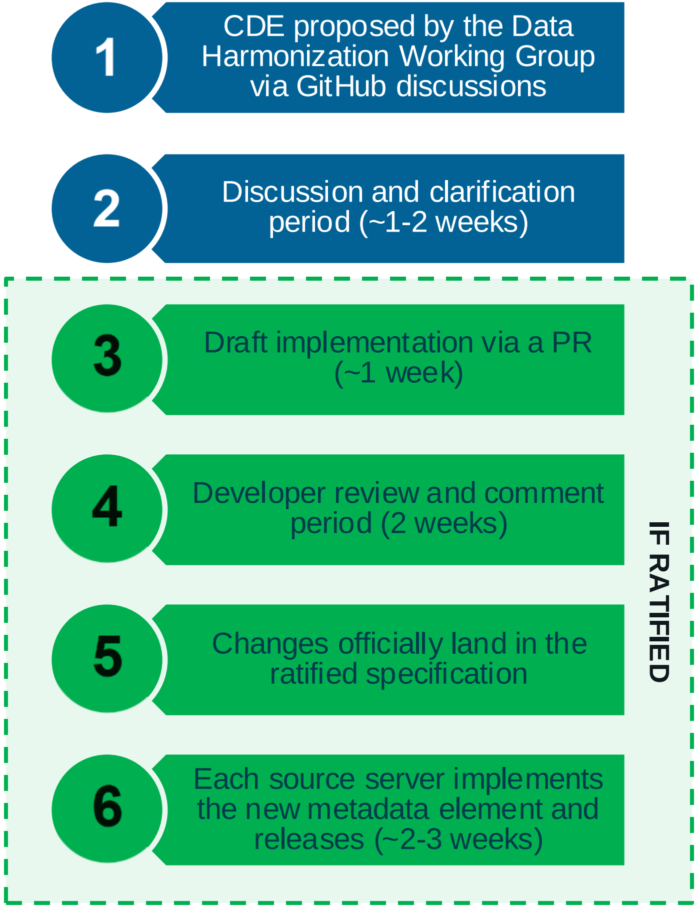
License

This project is licensed as either Apache 2.0 or MIT at your discretion.

Copyright © 2023-Present Childhood Cancer Data Initiative Federation

Development Approach (cont.)

- GitHub was leveraged during metadata ratification, specification design, and development.
  - All CDEs followed a proposal to implementation lifecycle that was about ~1 month end-to-end.
- Slack was used for real-time, informal discussions amongst stakeholders as well as implementors.
- Each federation member, as well as the NCI aggregator, implemented the specification using their own framework/infrastructure.



1 CDE proposed by the Data Harmonization Working Group via GitHub discussions

2 Discussion and clarification period (~1-2 weeks)

3 Draft implementation via a PR (~1 week)

4 Developer review and comment period (2 weeks)

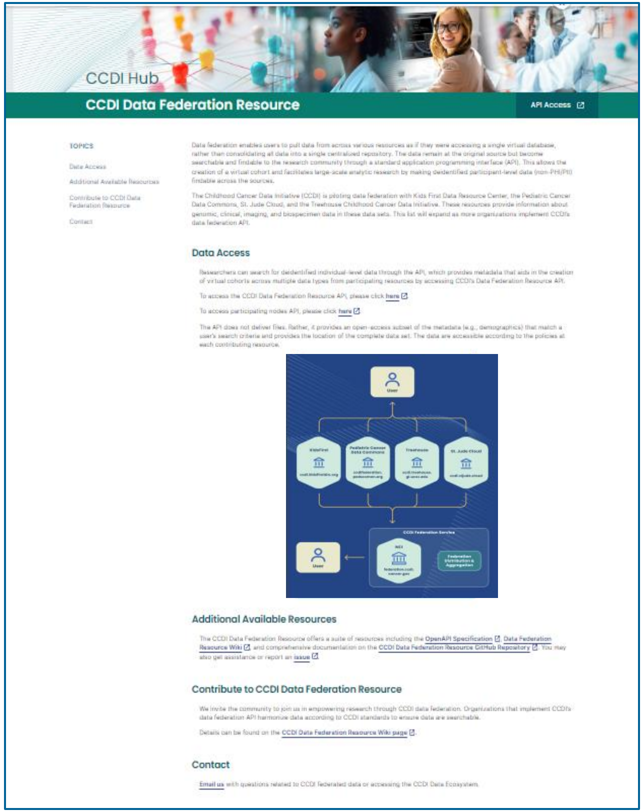
5 Changes officially land in the ratified specification

6 Each source server implements the new metadata element and releases (~2-3 weeks)

IF RATIFIED

Information About the API

- Future blog post from API authors diving into the specification in-depth
- CCDI Hub (https://ccdi.cancer.gov/explore)
  - Information about the project
  - Link to GitHub
- CCDI Data Federation Resource GitHub (https://cbiit.github.io/ccdi-federation-api-aggregation/)
  - OpenAPI Specification
  - GitHub Wiki for metadata descriptions
  - GitHub issues/discussions for questions
  - Links to participating nodes API spec



CCDI Hub

CCDI Data Federation Resource

API Access

TOPICS

Data Access

Additional Available Resources

Contribute to CCDI Data Federation Resource

Contact

Data federation enables users to pull data from across various resources as if they were accessing a single virtual database, rather than consolidating all data into a single centralized repository. The data remain at the original source but become searchable and findable to the research community through a standard application programming interface (API). This allows the creation of a virtual cohort and facilitates large-scale analytic research by making deidentified participant-level data (non-PHI/PII) findable across the sources.

The Childhood Cancer Data Initiative (CCDI) is piloting data federation with Kids First Data Resource Center, the Pediatric Cancer Data Commons, St. Jude Cloud, and the Treehouse Childhood Cancer Data Initiative. These resources provide information about genomic, clinical, imaging, and biospecimen data in these data sets. This list will expand as more organizations implement CCDI's data federation API.

Data Access

Researchers can search for deidentified individual-level data through the API, which provides metadata that aids in the creation of virtual cohorts across multiple data types from participating resources by accessing CCDI's Data Federation Resource API.

To access the CCDI Data Federation Resource API, please click here.

To access participating nodes API, please click here.

The API does not deliver files. Rather, it provides an open-access subset of the metadata (e.g., demographics) that match a user's search criteria and provides the location of the complete data set. The data are accessible according to the policies at each contributing resource.

User
KidsFirst
ccdi.kidsfirstdrc.org
Pediatric Cancer Data Commons
ccdifederation.pedscommons.org
Treehouse
ccdi.treehouse.gi.ucsc.edu
St. Jude Cloud
ccdi.stjude.cloud
CCDI Federation Service
User
NCI
federation.ccdi.cancer.gov
Federation Distribution & Aggregation

Additional Available Resources

The CCDI Data Federation Resource offers a suite of resources including the OpenAPI Specification, Data Federation Resource Wiki, and comprehensive documentation on the CCDI Data Federation Resource GitHub Repository. You may also get assistance or report an issue.

Contribute to CCDI Data Federation Resource

We invite the community to join us in empowering research through CCDI data federation. Organizations that implement CCDI's data federation API harmonize data according to CCDI standards to ensure data are searchable.

Details can be found on the CCDI Data Federation Resource Wiki page.

Contact

Email us with questions related to CCDI federated data or accessing the CCDI Data Ecosystem.

https://ccdi.cancer.gov/data-federation-resource

Using the Federation Data Demo: Leveraging Jupyter Notebook

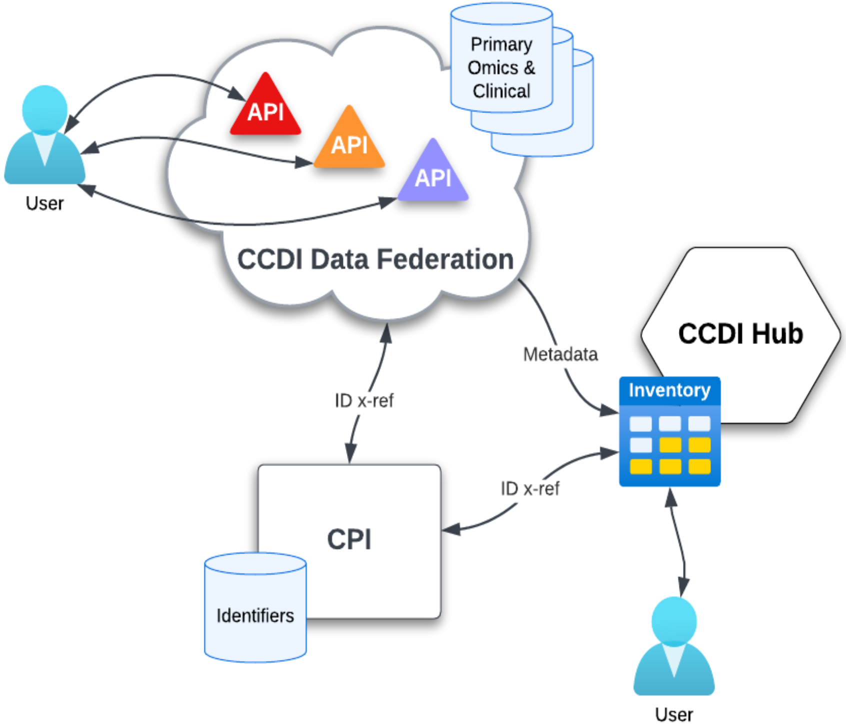
Martin Ferguson

Future Applications

Allison Heath

Further Enabling Scientific Use Cases: Childhood Cancer Data Initiative Participant Index (CPI)

- Link participants’ data from the Federation and CCDI across:
  - Time (longitudinal)
  - Space (institutions, studies, trials)
  - Modalities (clinical, molecular, imaging)
- Minimize double counting
- Create a “Cohort of One”
  - Integrate data for a patient across federation
  - Live updates of new data for patients of interest



API-based Integration for Discoverability

● = CCDI federated data of interest

St. Jude Cloud

DATA FOR THE COMMON GOOD

Pediatric Cancer Data Commons

Resources interested in CCDI data

Gabriella Miller Kids First

PEDIATRIC RESEARCH PROGRAM

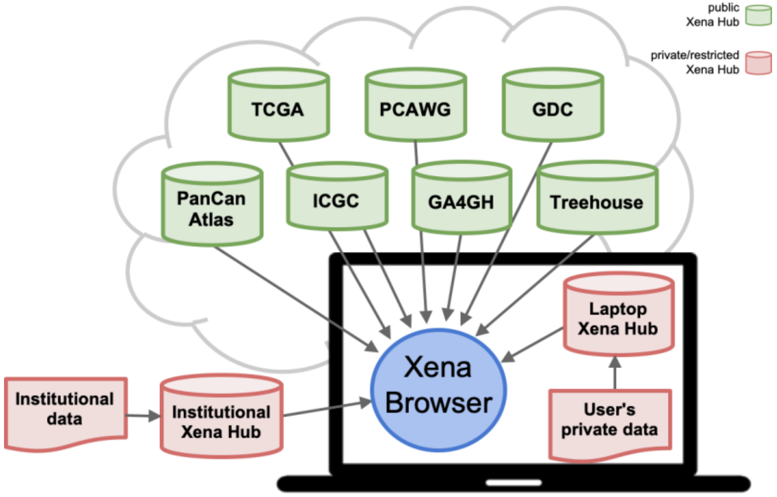
Data Resource Center

Treehouse

CHILDHOOD CANCER INITIATIVE

API-Based Integration with Local Data for Analysis

- Empower individuals to analyze their tumor data using the full scope of the CCDI federation data
- Data analysis tools utilizing CCDI API for rapid analysis by the research community
- Enable users to upload data and compare it to federation data (similar to matchmaker exchange

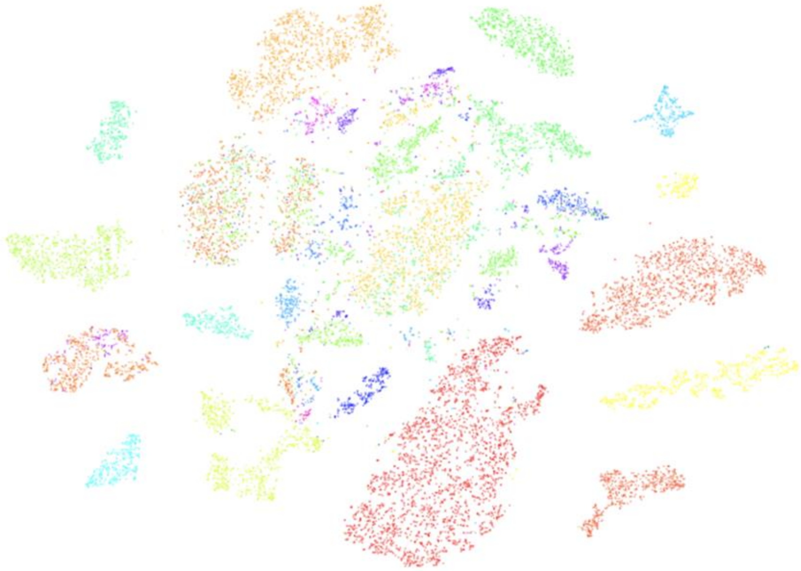


PEDIATRIC CANCER KNOWLEDGE BASE
PeCan

PedcBioPortal
for Integrated
Childhood Cancer Genomics

RNA-Seq “World Map”: Foundational Layer for Scientific Use Cases

- Generate a gene expression “world map” using RNA-Seq data from all partners
  - Enables finding tumors that are “acting similarly” from an expression perspective
  - Individual institutions have already found these tools useful in diagnosing rare diseases as well as checking for high-quality data space (institutions, studies, trials)
- API-based mechanism would enable ongoing map updates and refinement as new data is generated



Q&A

Contacts

- Leveraging the federation resource
  - Email: NCIChildhoodCancerDataInitiative@mail.nih.gov with questions related to CCDI federated data or accessing the CCDI Data Ecosystem
- Questions related to individual APIs
  - St. Jude Cloud: support@stjude.cloud
  - Treehouse: treehousegenomics@ucsc.edu
  - Kids First – CHOP: nemarichc@chop.edu
  - PCDC – UChicago: lgraglia@bsd.uchicago.edu

How You Can Engage with CCDI

Learn about CCDI and subscribe to our monthly newsletter:
cancer.gov/CCDI

Access CCDI data and resources:
ccdi.cancer.gov

Questions? Email us at:
NCIChildhoodCancerDataInitiative@mail.nih.gov

Thank you for attending!

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

NIH NATIONAL CANCER INSTITUTE

cancer.gov

cancer.gov/espanol