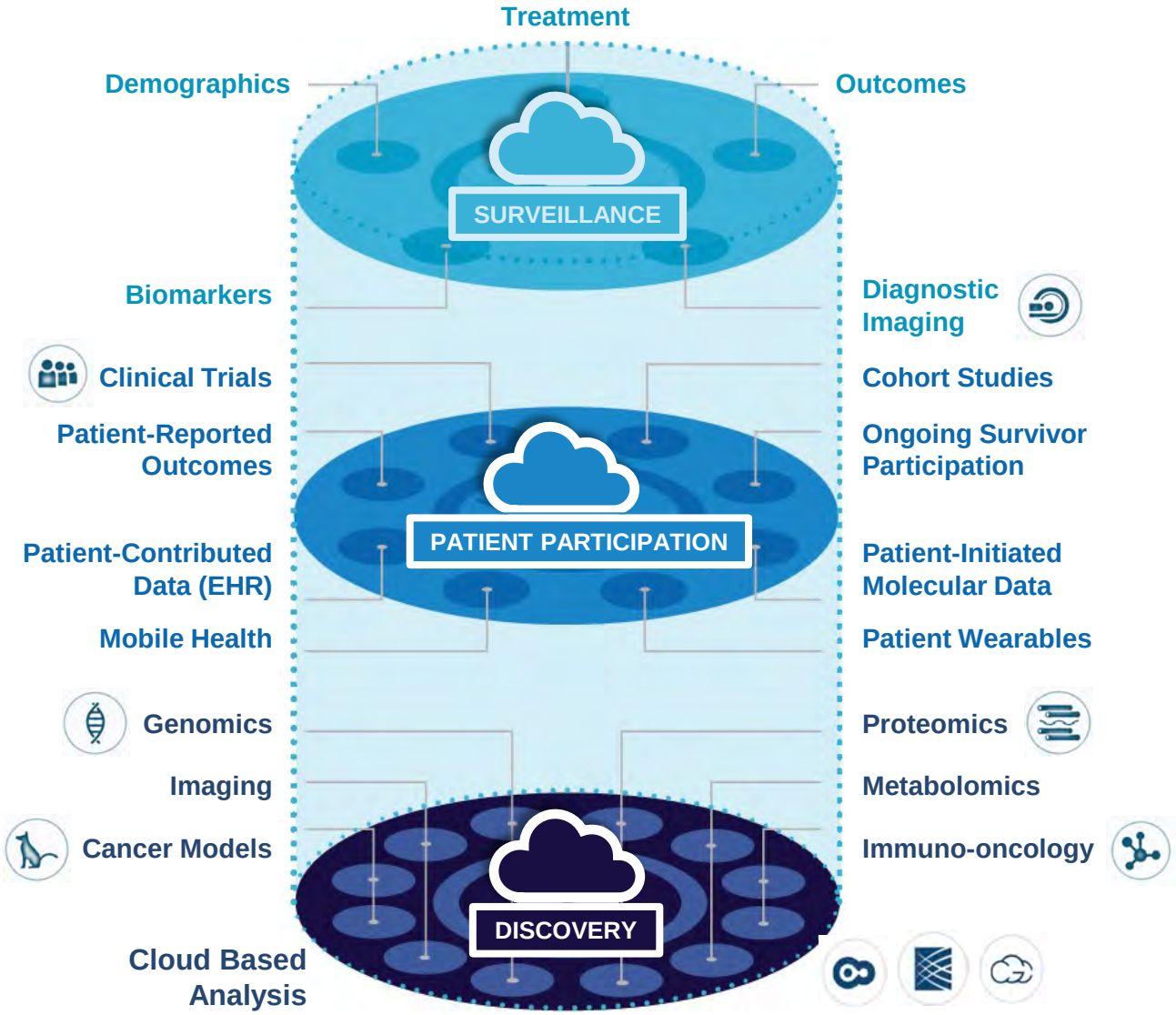
Introduction to the Cancer Research Data Commons (CRDC)

Tanja Davidsen, Ph.D.

October 16, 2024

NIH NATIONAL CANCER INSTITUTE

National Data Ecosystem: Integrating Cancer Research



The Cancer Research Data Commons (CRDC)

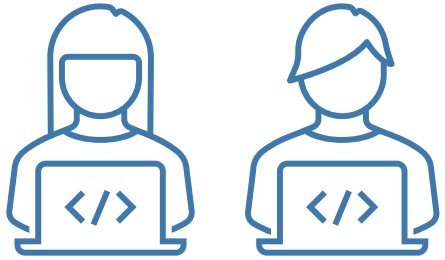
Limitations for Data Driven Research

Data Types

- NCI funds research that generates valuable data
  - Basic research, Clinical trials, Population studies
- Difficult to find and analyze multiple data types from multiple data sources

User Skills and Tools

- Most researchers are not data scientists or informaticians
- Skill levels for data handling and data analysis varies
- Availability of analysis tools varies on platforms

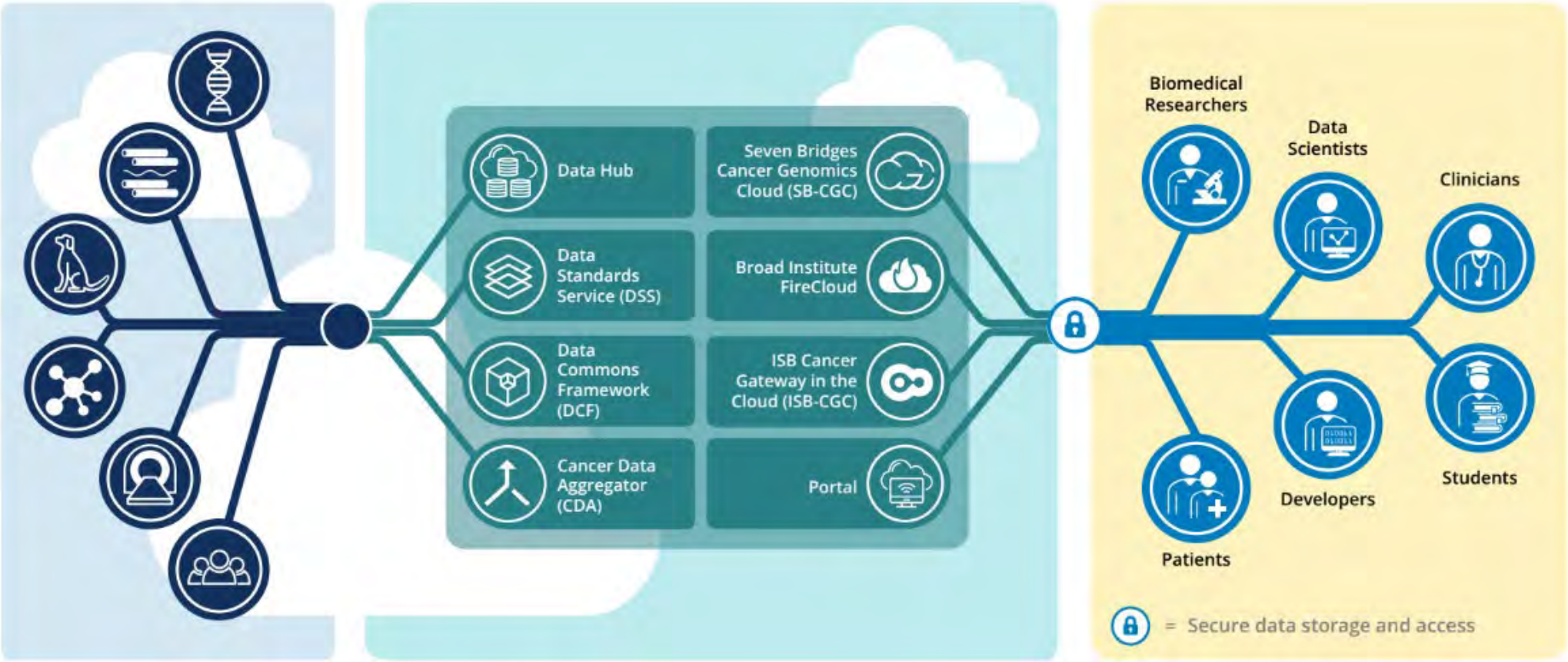


Data Storage and Usage

- Data often stored in separate data commons or repositories for download
- Data usage & combining datasets may require multiple downloads or moving data

Cancer Research Data Commons (CRDC)

NCI's primary data science platform for cancer research



Data Commons **Infrastructure & Cloud Resources** **User Community**

Cancer Research Data Commons (CRDC)

NCI's primary data science platform for cancer research

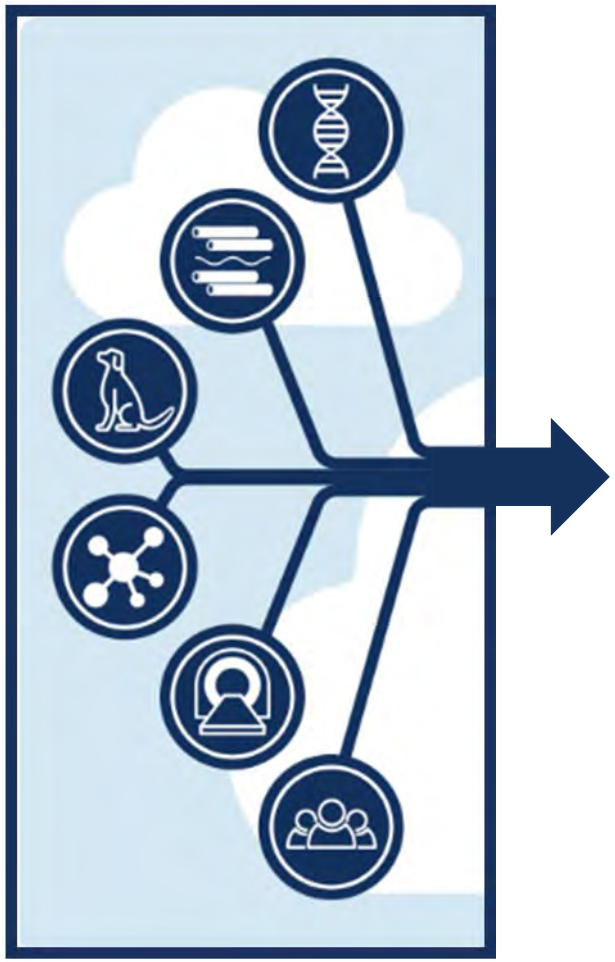
Mission

- To empower researchers by providing a cancer data ecosystem with state-of-the-art **visualization, analysis, and interoperability tools in a flexible, cloud-based** computational environment

Goals

- Preserve **long-term value** of NCI-funded data
- Improve **data submission, access, and interoperability**
- Accelerate cancer research through integrative analysis of **multi-modal data**

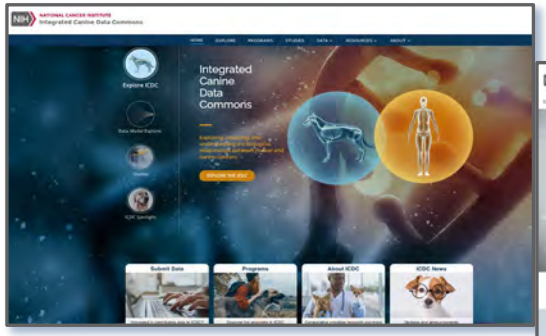
CRDC Ecosystem: Data Commons



Genomic Data Commons (GDC)

Proteomic Data Commons (PDC)

Imaging Data Commons (IDC)



Integrated Canine DC (ICDC)

Clinical & Translational DC (CTDC)

Cancer Data Service (CDS)

CRDC: NCI Cloud Resources

Democratizing access to cancer research data

- Access to **large cancer data sets** without need to download or move data
- Access to **workspaces, analysis tools, and workflows/pipelines**
- **Bring your own data and tools**: collaborative pre-publication workspaces

ISB's Cancer Gateway in the Cloud

Great for command-line, BigQuery, Specialty DBs

Broad's FireCloud

Great for running production pipelines

Seven Bridges' Cancer Genomics Cloud

Great for non-technical user Interface, visual displays

AACR Cancer Research Series

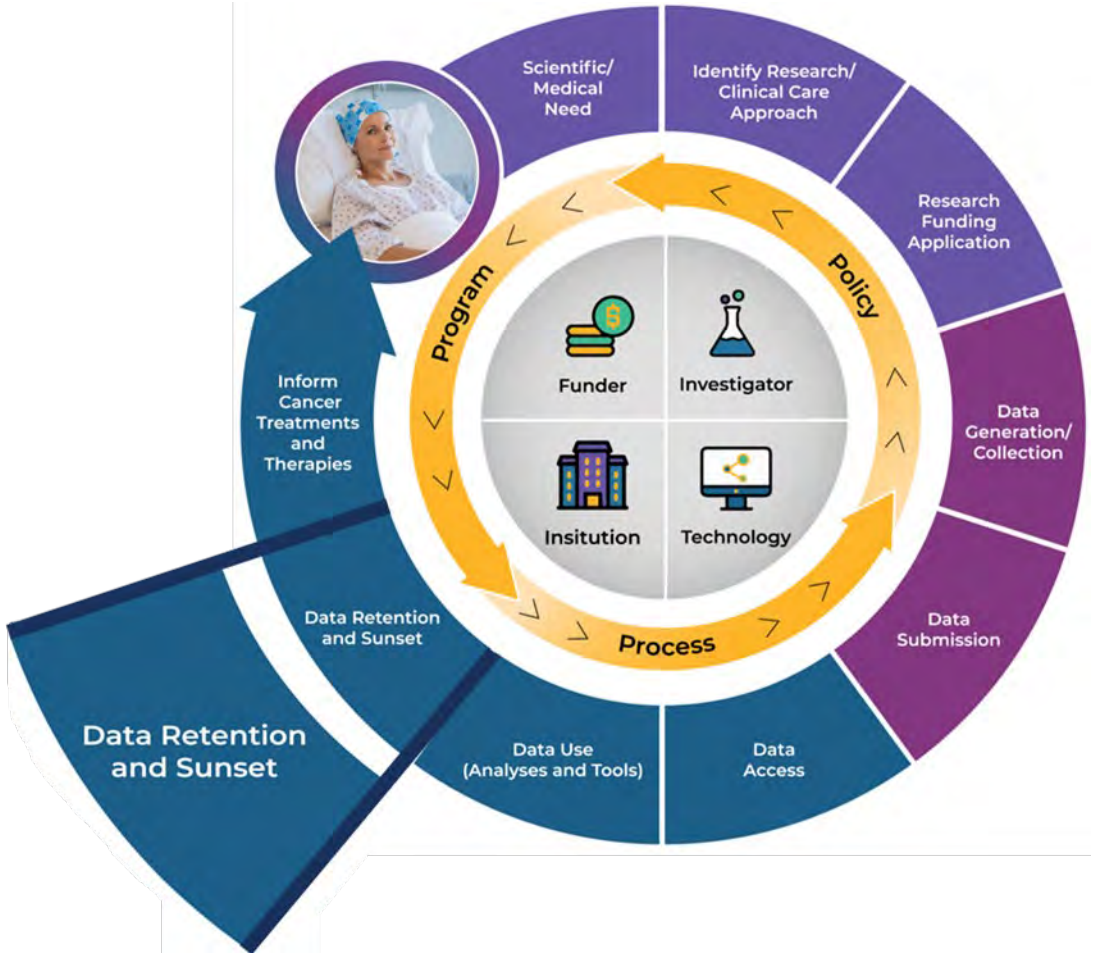
A four-part invited series published online in March 2024 highlighting the CRDC's accomplishments from the past 10 years.

- LESSONS LEARNED AND FUTURE STATE
- RESOURCES TO SHARE KEY CANCER DATA
- CLOUD-BASED ANALYTICAL RESOURCE
- CORE STANDARDS AND SERVICES

Learn more about the series on the CRDC Website

NCI Data Lifecycle

CRDC as an Exemplar



Structured Data Management for FAIR Sharing: CRDC Data Governance and Submission

Ina Felau & Durga Addepalli, Ph.D.

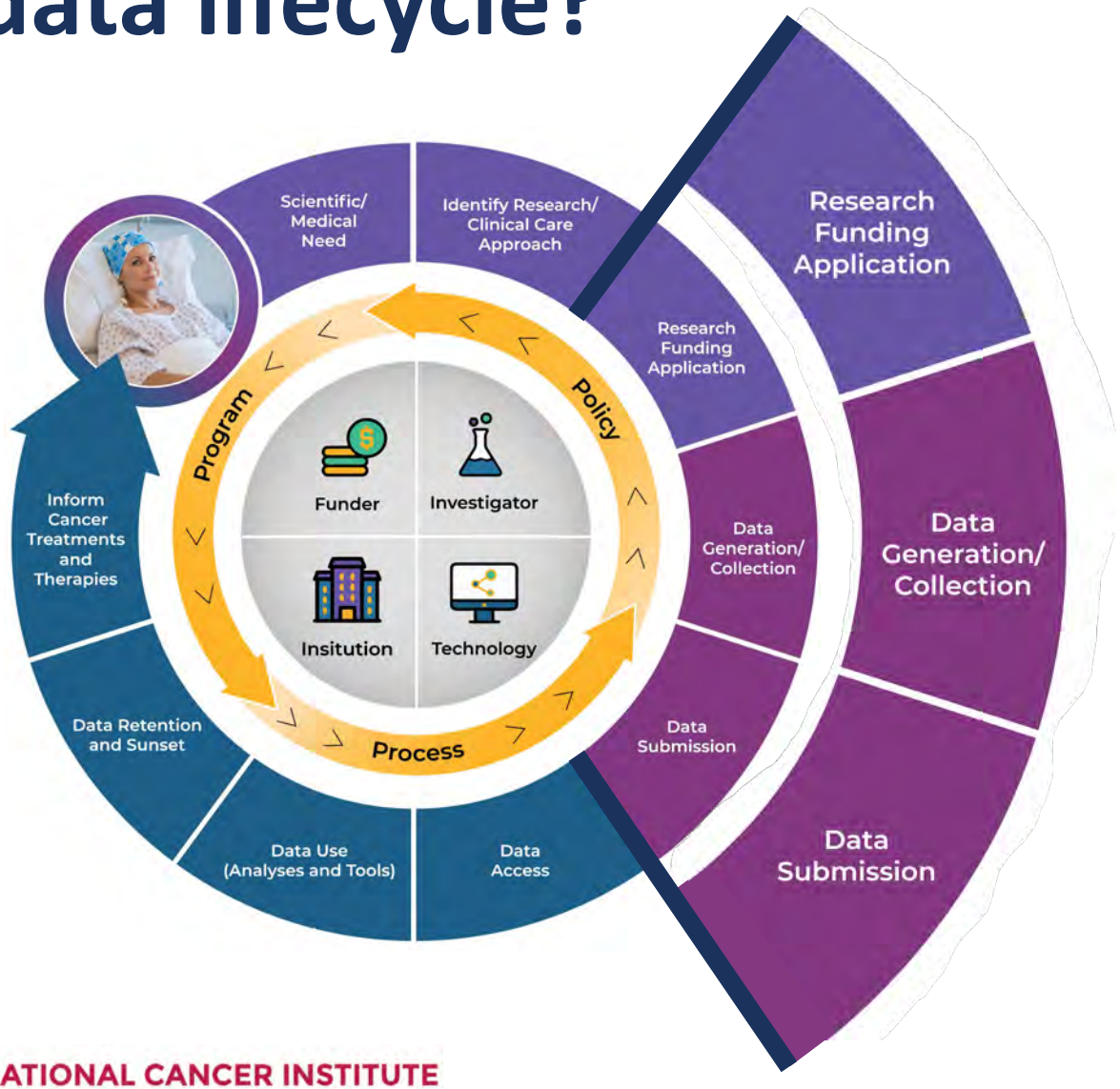
October 16, 2024

NIH NATIONAL CANCER INSTITUTE

Part 1 - Agenda

- How can CRDC Support You?
- CRDC Vision for Data Submission
- CRDC Submission Portal: Submission Request Process

How can CRDC support you through the scientific data lifecycle?



Research Funding Application

Consider CRDC when preparing your DMS Plan

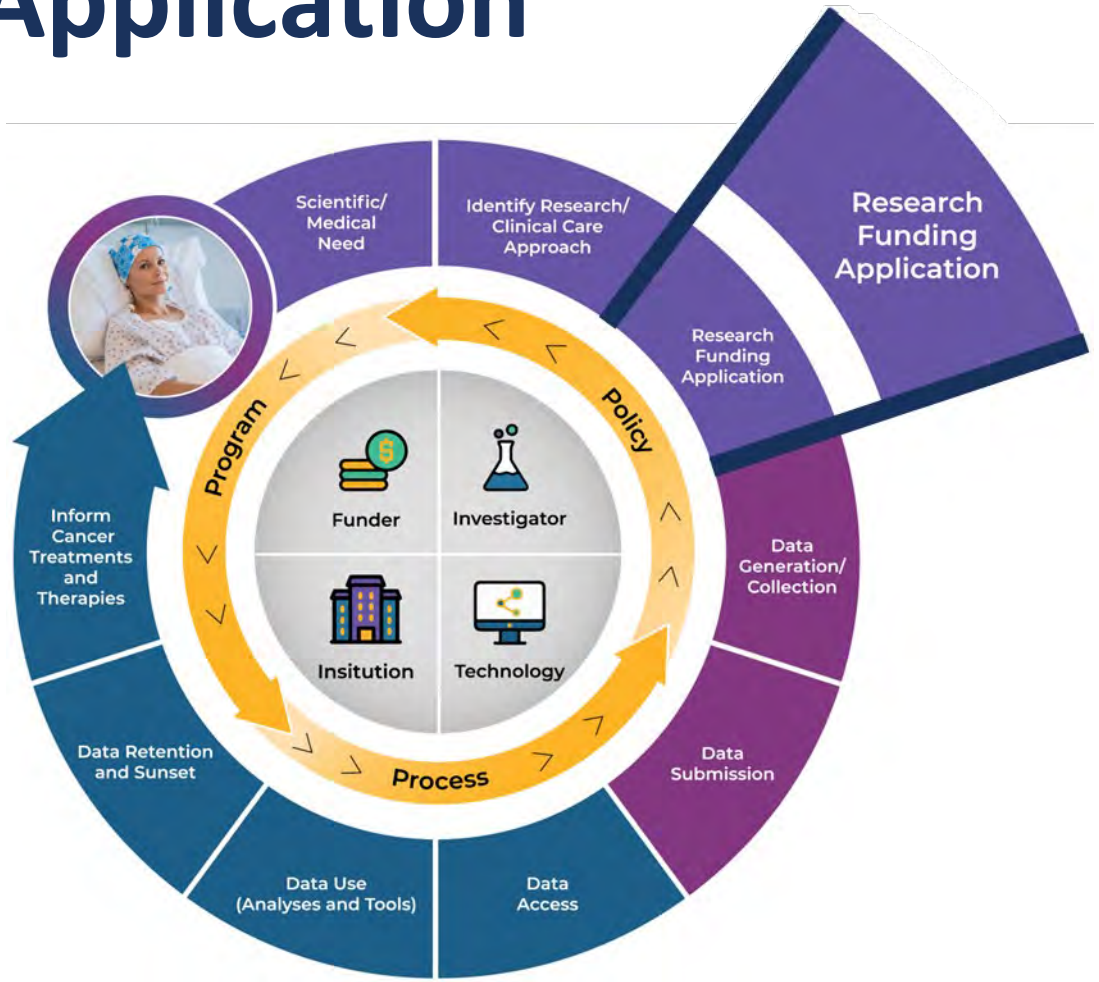
Data Generation/Collection

Consider CRDC metadata & data standards

Data Submission

The CRDC Submission Portal

NCI Data Lifecycle: Research Funding Application

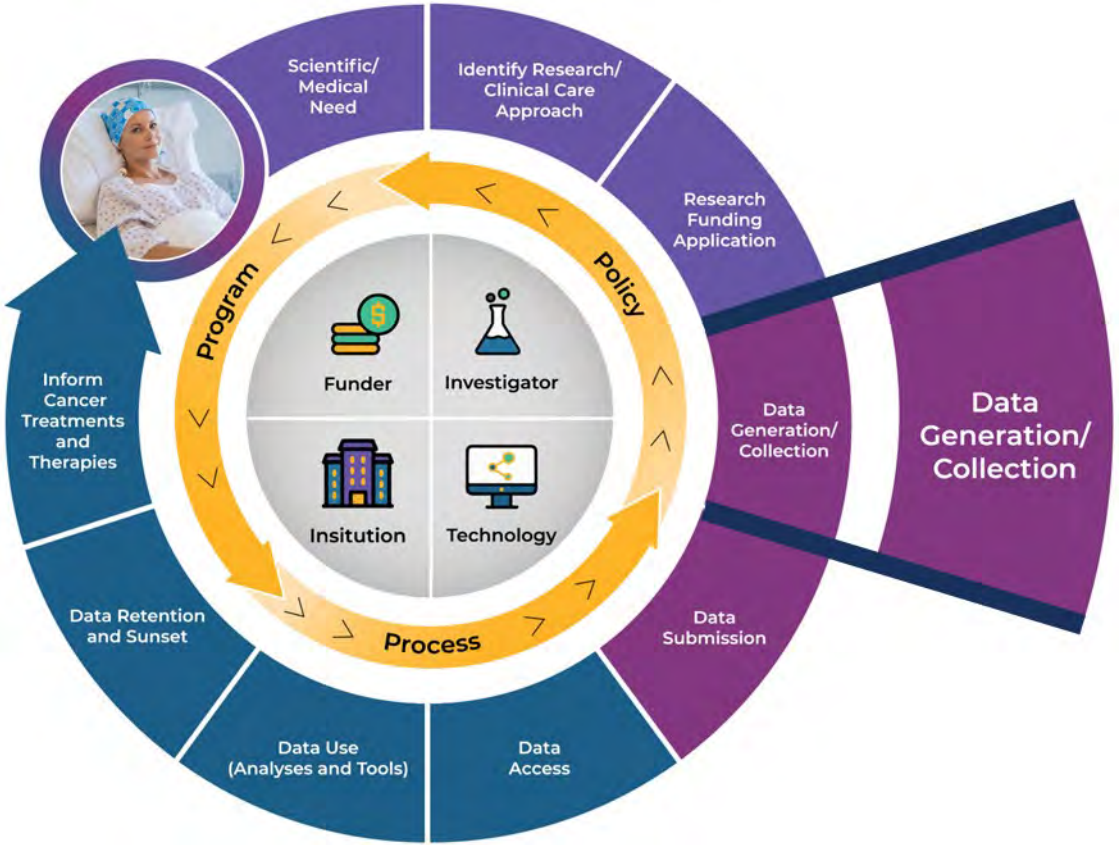


Research Funding Application

Consider CRDC when preparing your DMS Plan:

- Genomic, proteomic, imaging, and other cancer research data
- Types of metadata
- Data standards

NCI Data Lifecycle: Data Generation & Collection

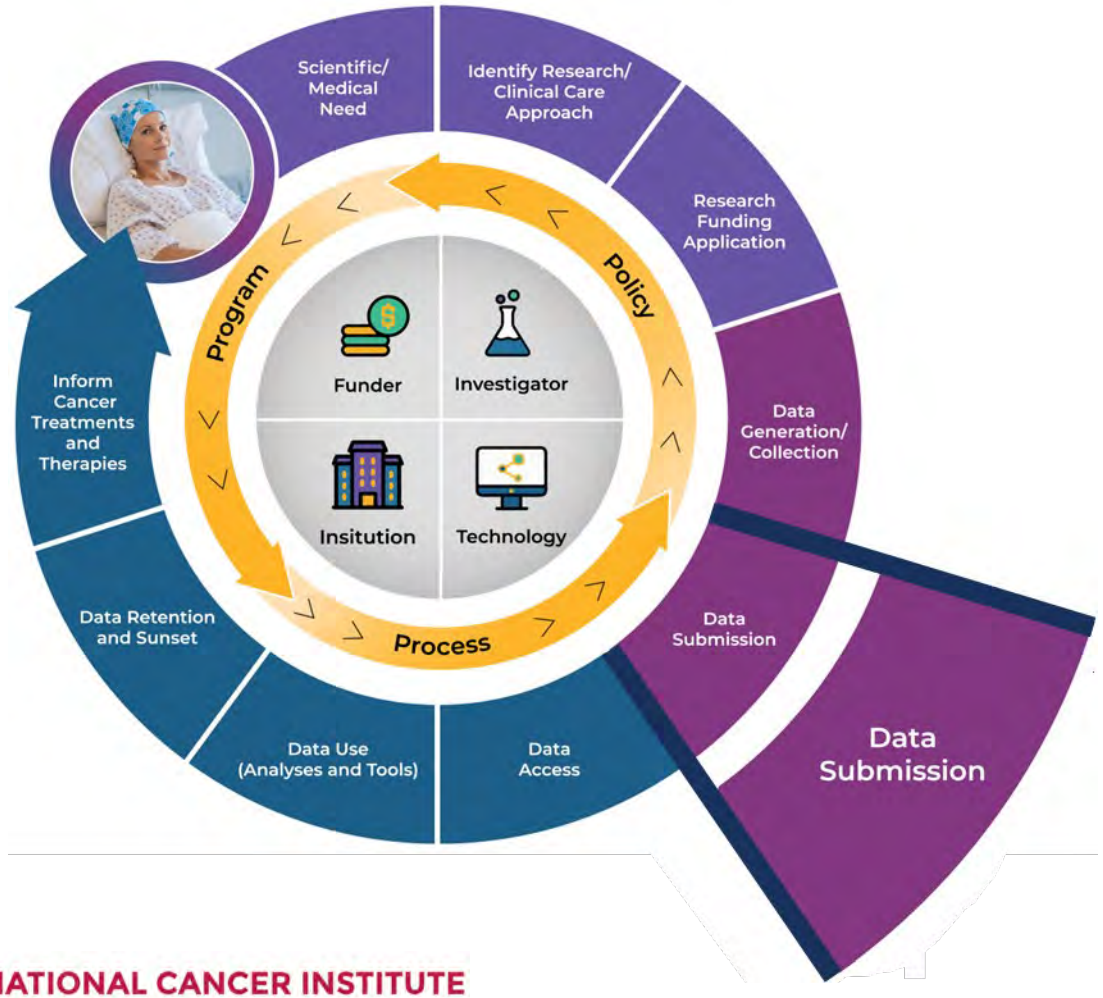


Data Generation/Collection

Consider CRDC metadata & data standards:

- CRDC data dictionary
- CRDC standard Common Data Elements (CDEs)

NCI Data Lifecycle: Data Submission Parts 1 & 2



Data Submission

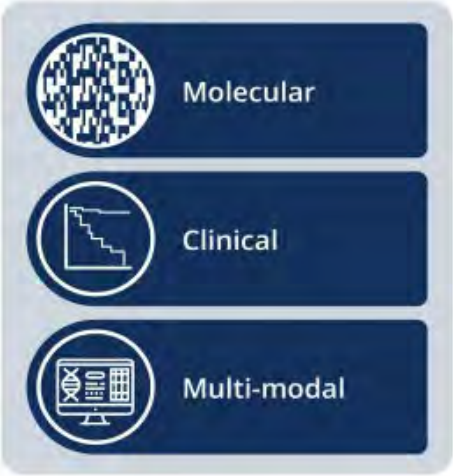
The CRDC Submission Portal:

- Submission Request
- Data Submission

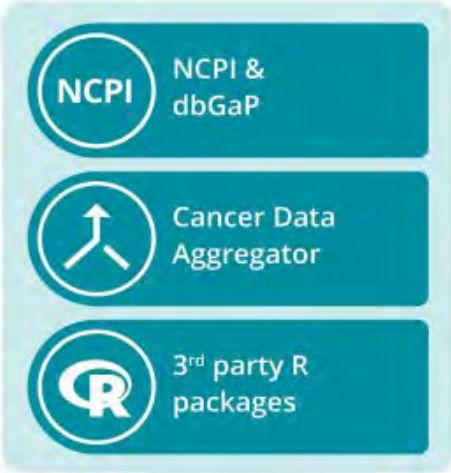
FAIR Principles

FINDABLE
Faceted Search and Key Word Search

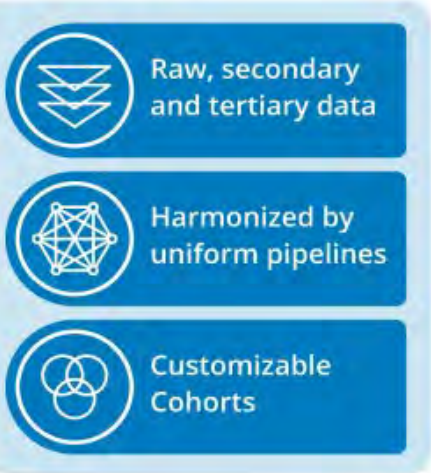
ACCESSIBLE
Online Analysis and Visualization



INTEROPERABLE
APIs and Standardized Metadata

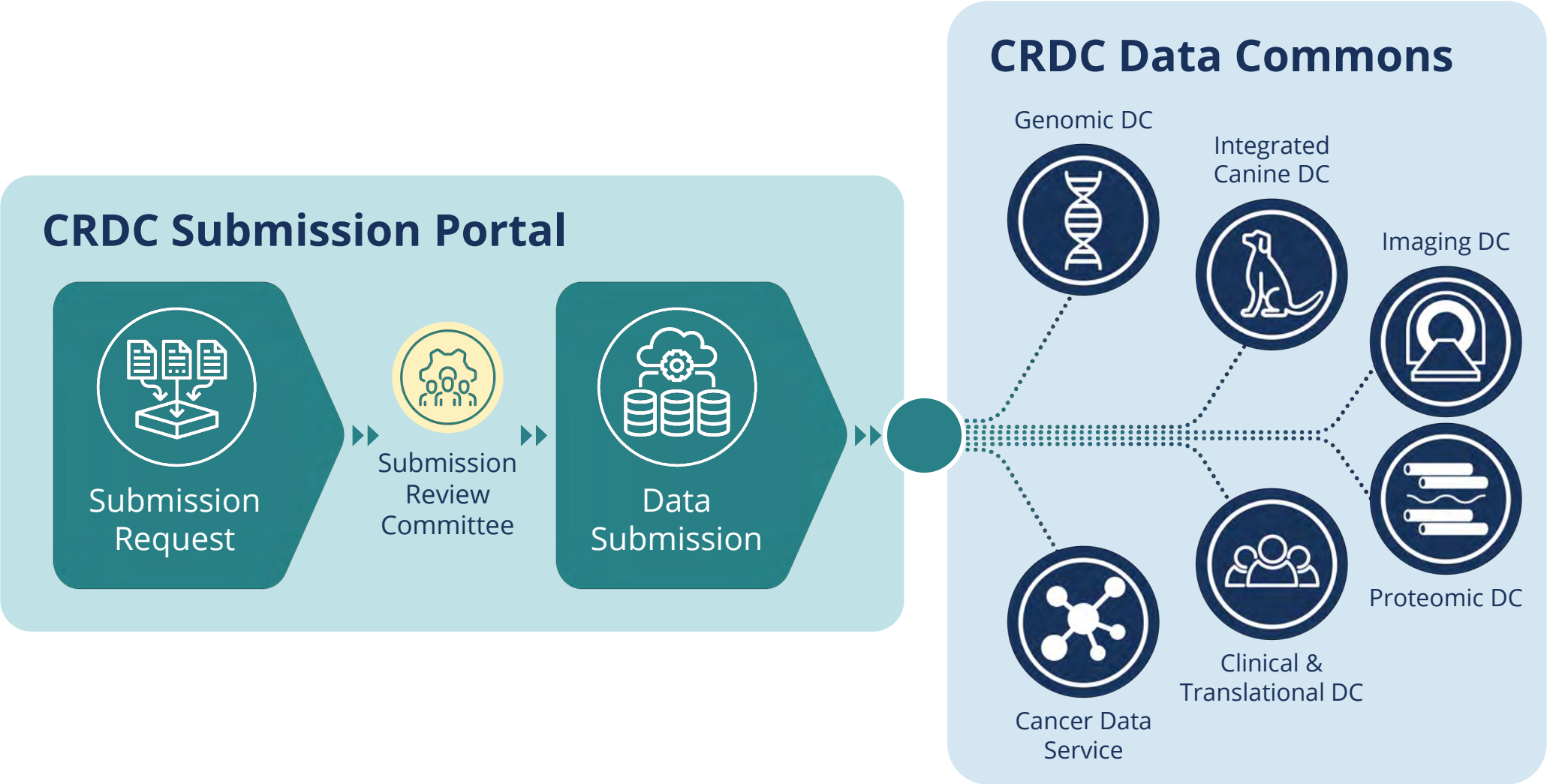


REUSABLE
Rich Metadata and Harmonized Scientific Data

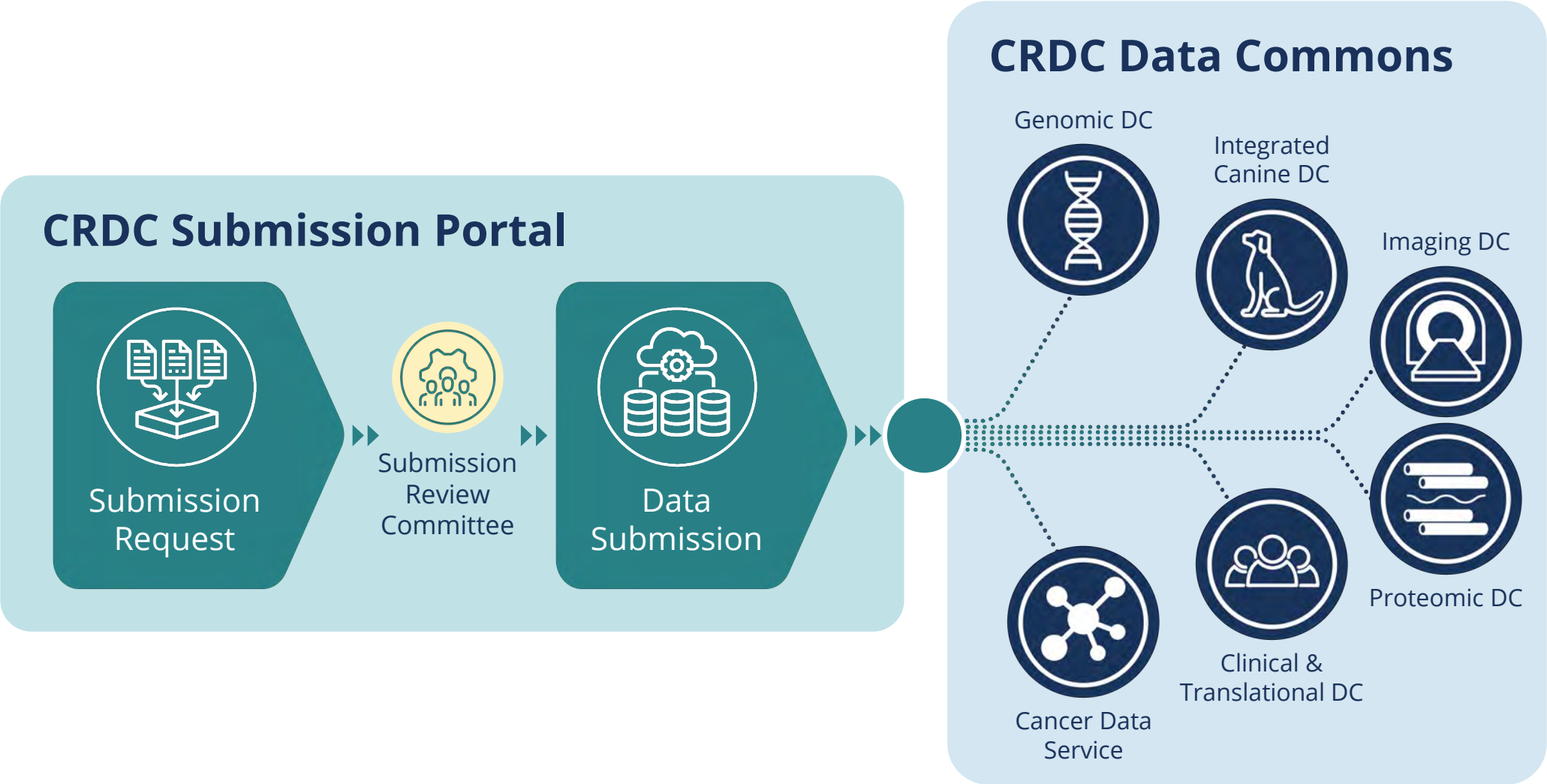


At the bottom are six specialized data commons (GDC, PDC, ICDC, CDS, IDC and CTDC). Selected CRDC features are used to demonstrate the implementation of FAIR principles.

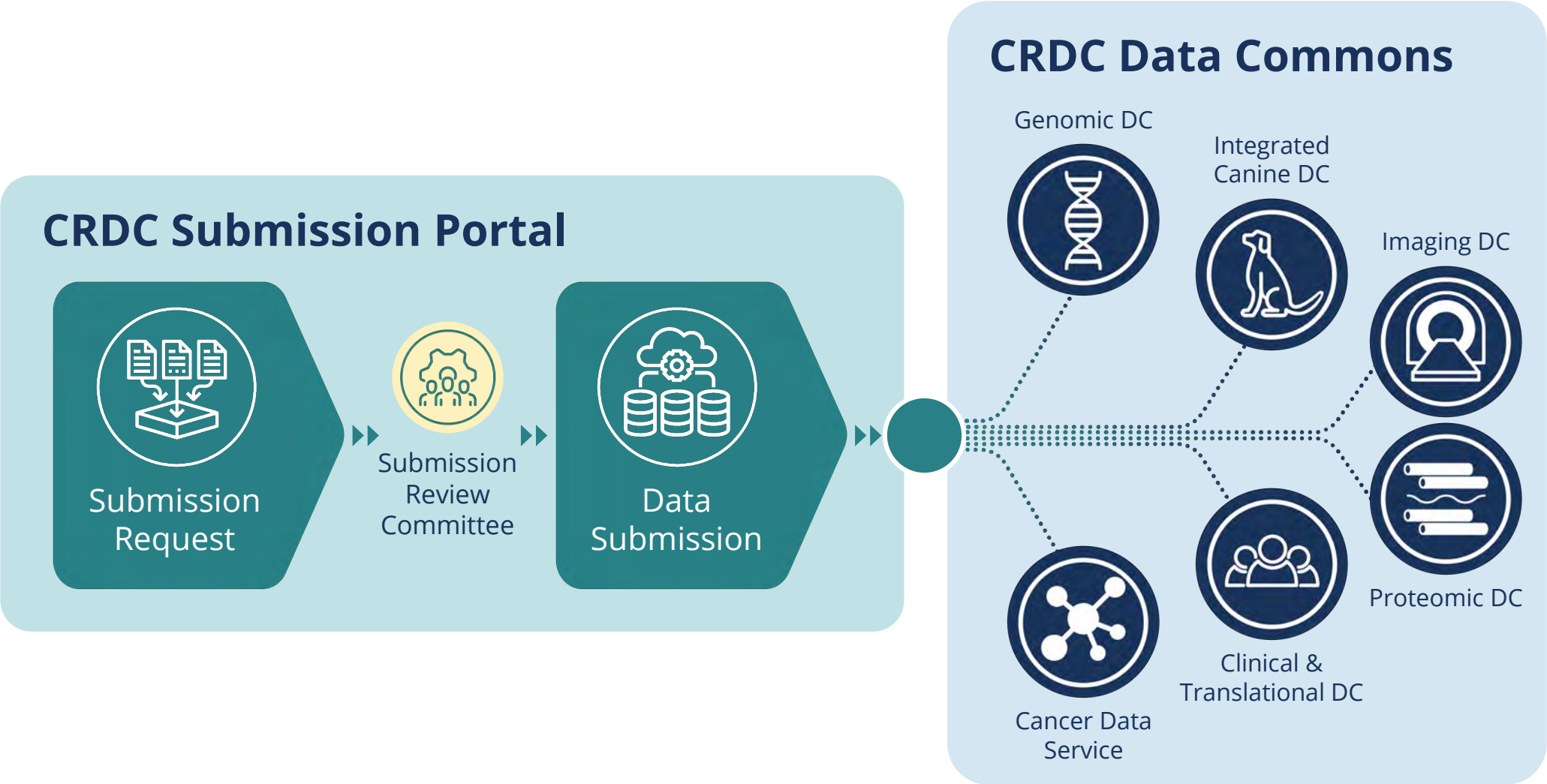
CRDC's Vision for Data Submission



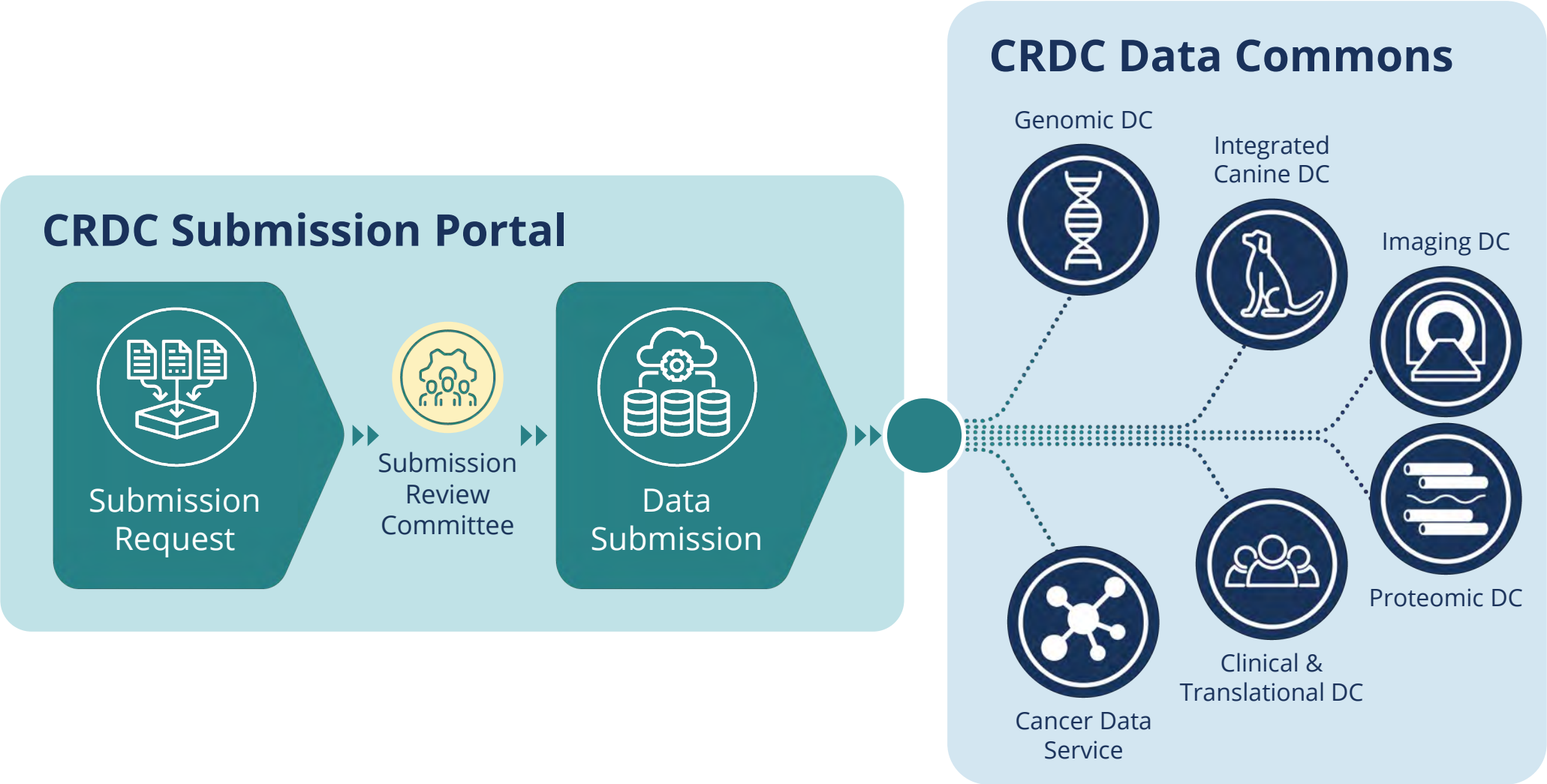
CRDC's Vision for Data Submission



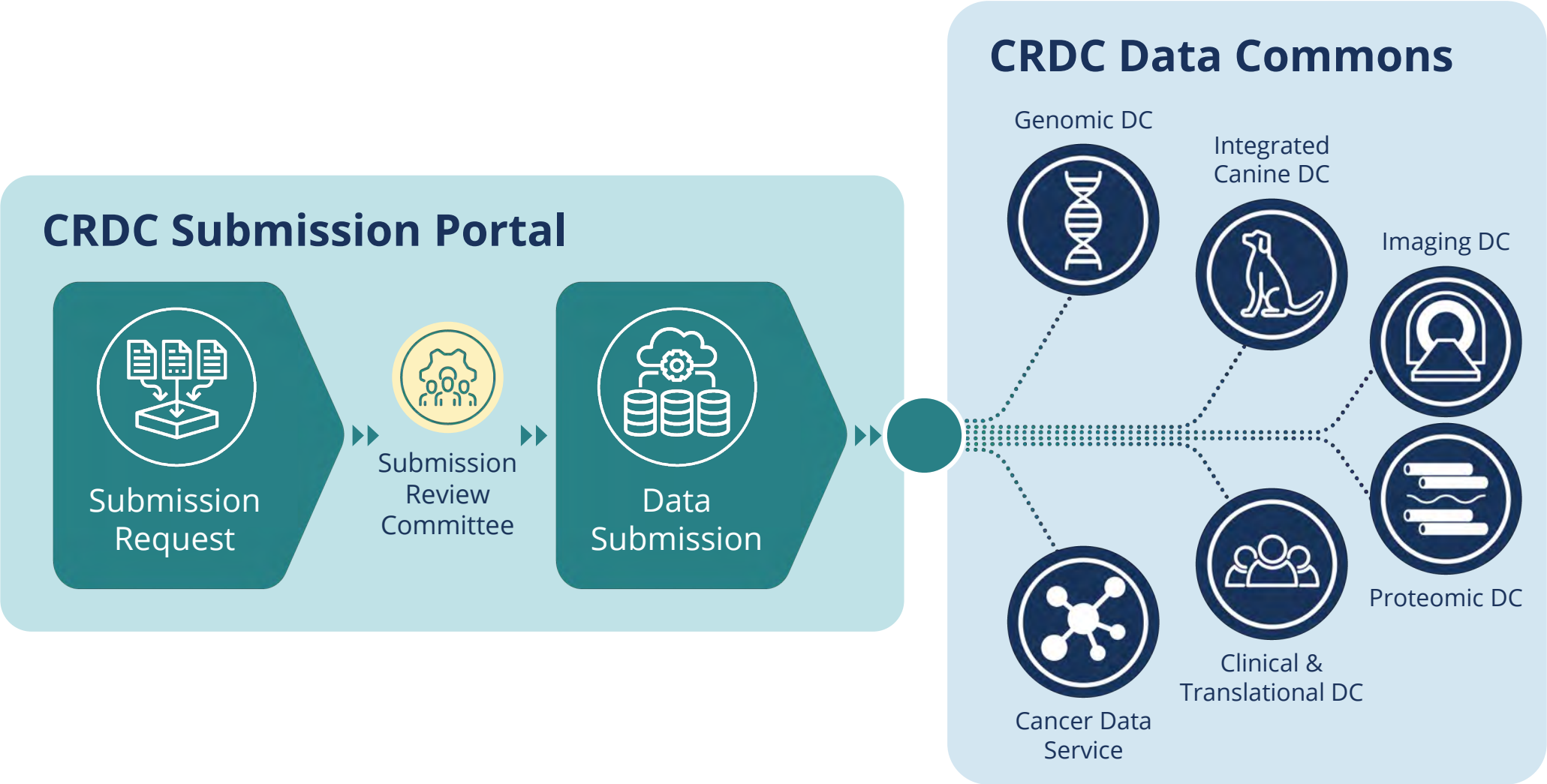
CRDC's Vision for Data Submission



CRDC's Vision for Data Submission



CRDC's Vision for Data Submission



CRDC Submission Portal: Impact

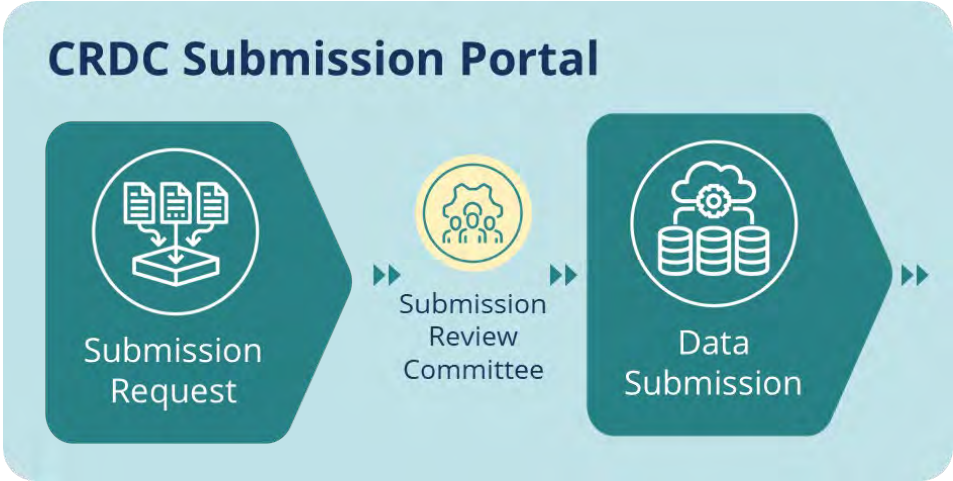
For Data Submitters

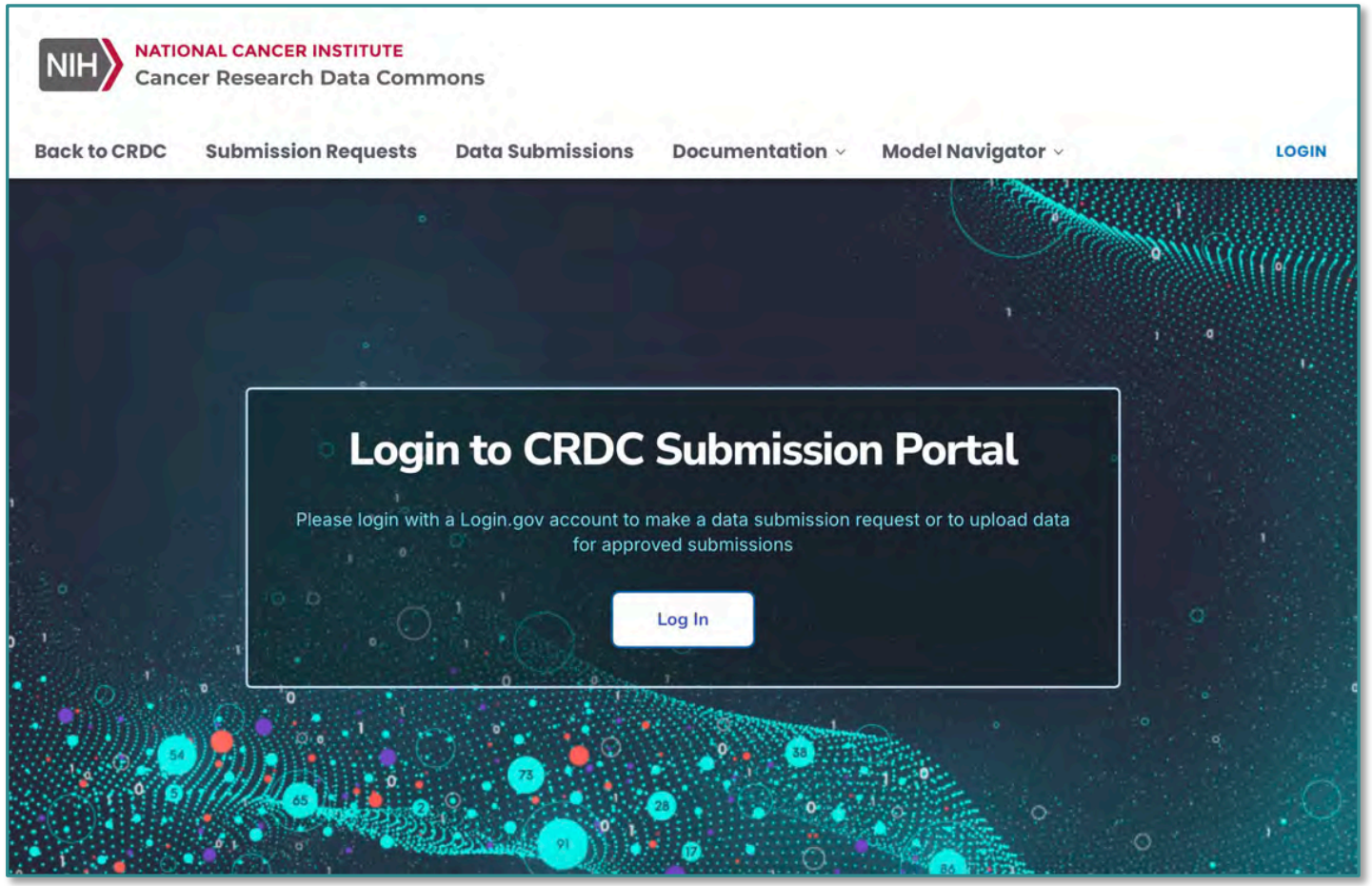
- Supports data sharing journey through clear, easy, transparent governance and data submission processes
- Supports researchers in complying with NIH's
  - Data Management and Sharing (DMS) Policy
  - Genomic Data Sharing (GDS) Policy

For Research Community

- Users can find and access high quality research data for reuse
- Data is accurate, complete, consistent, valid, and searchable
- Data is ready to be analyzed using CRDC resources including:
  - Cloud compute, analytical workflows, AI/ML models to draw new insights

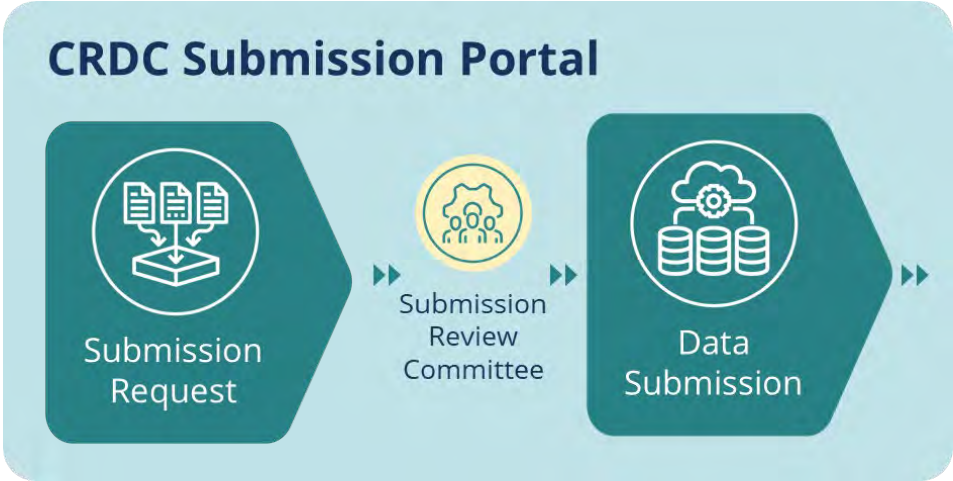
CRDC Submission Portal: How to get started

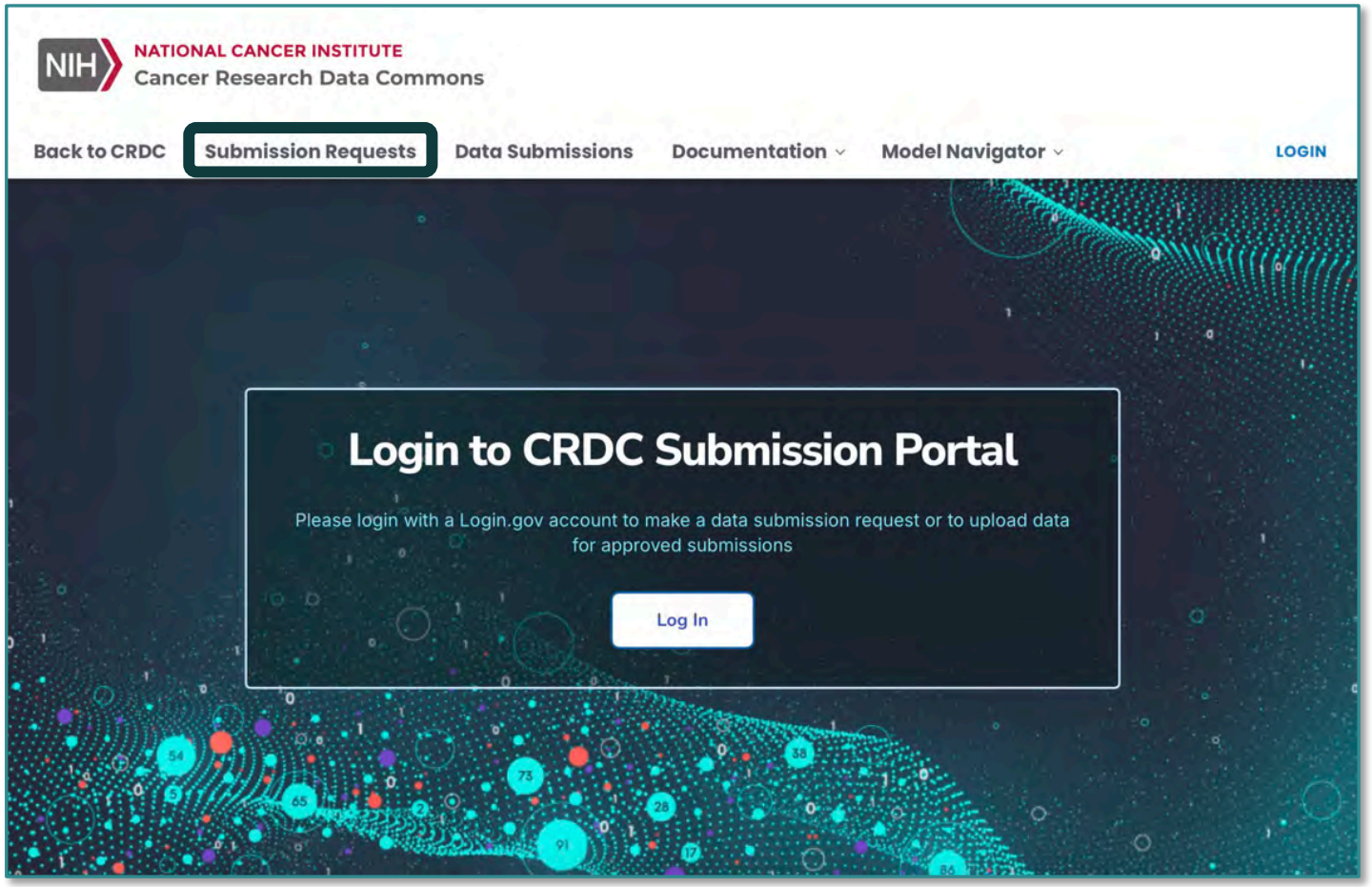




https://hub.datacommons.cancer.gov/

CRDC Submission Portal: How to get started



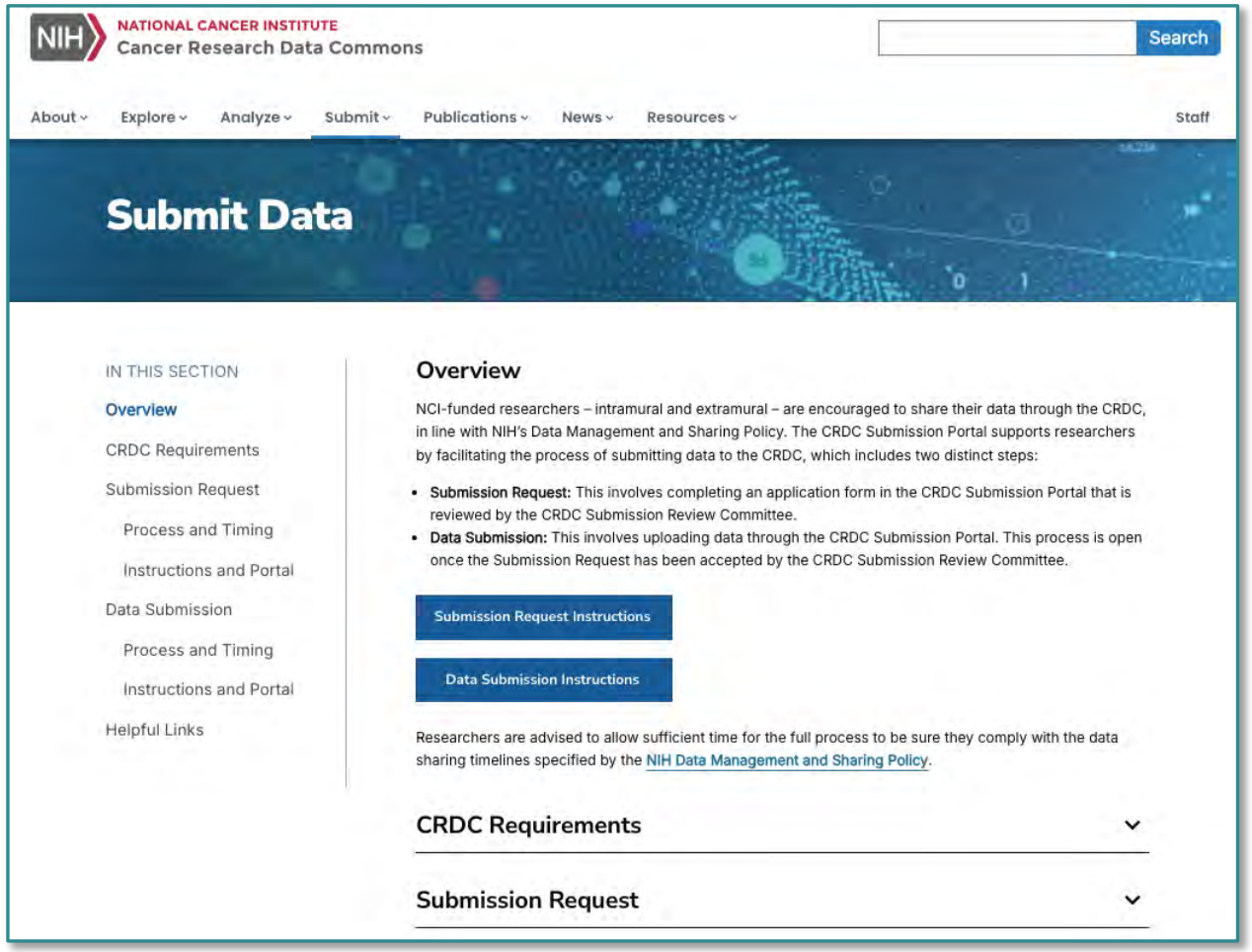


https://hub.datacommons.cancer.gov/

Submission Request: Resources

CRDC Submit Data page provides information about the Submission Request Form and associated process, including:

- CRDC Requirements
- Details about the Submission Request process and associated step-by-step instructions
- Links to important resources and Frequently Asked Questions (FAQs)



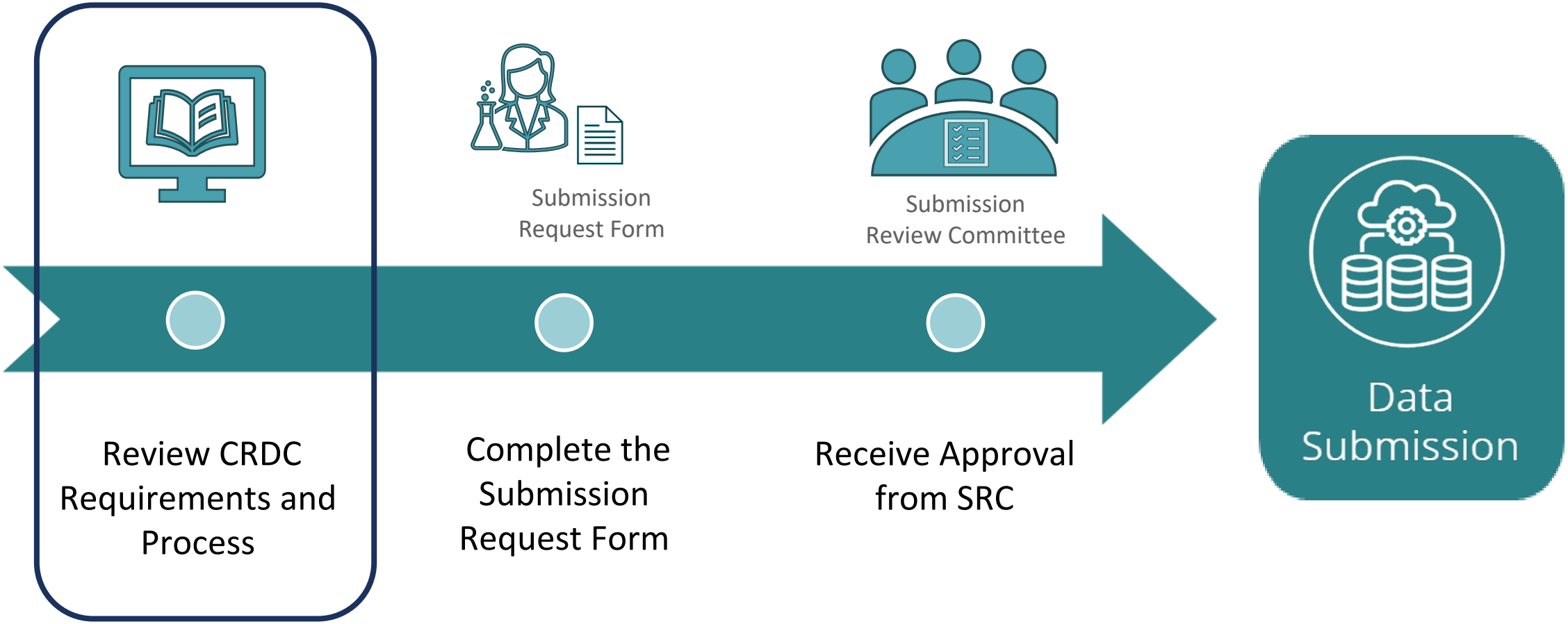
Submission Request: CRDC Requirements

Submitters will need to review and consider key CRDC requirements:

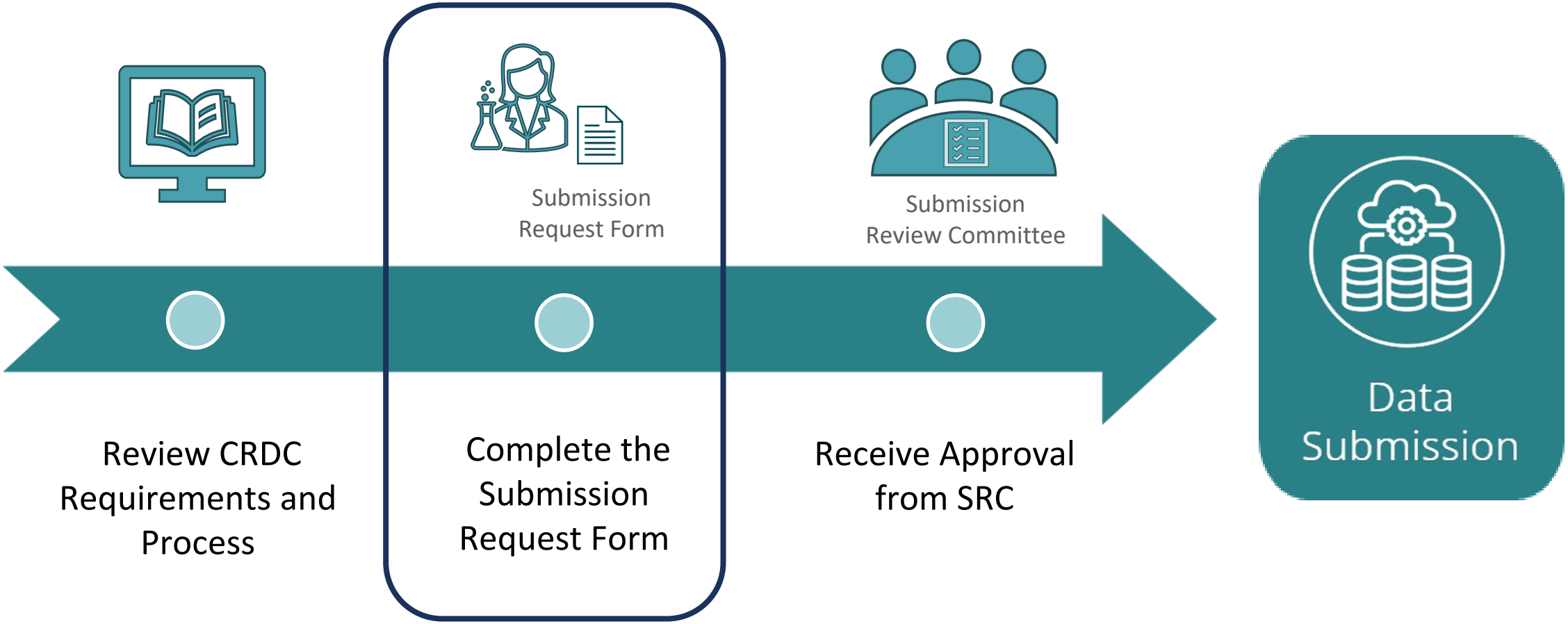
- Studies are NCI-funded or NIH-funded*
- The study data are fully collected
- The study data are ready to be shared
  - Data has been de-identified of PII
  - Legal permissions/data agreements have been completed
  - If the study contains controlled access data, it is registered in dbGaP
- Accepted species: human, canine, mouse and zebrafish*

*Other species or studies subject to review & approval

Submission Request Process



Submission Request Process



Complete the Submission Request Form: Step-by-Step Guide

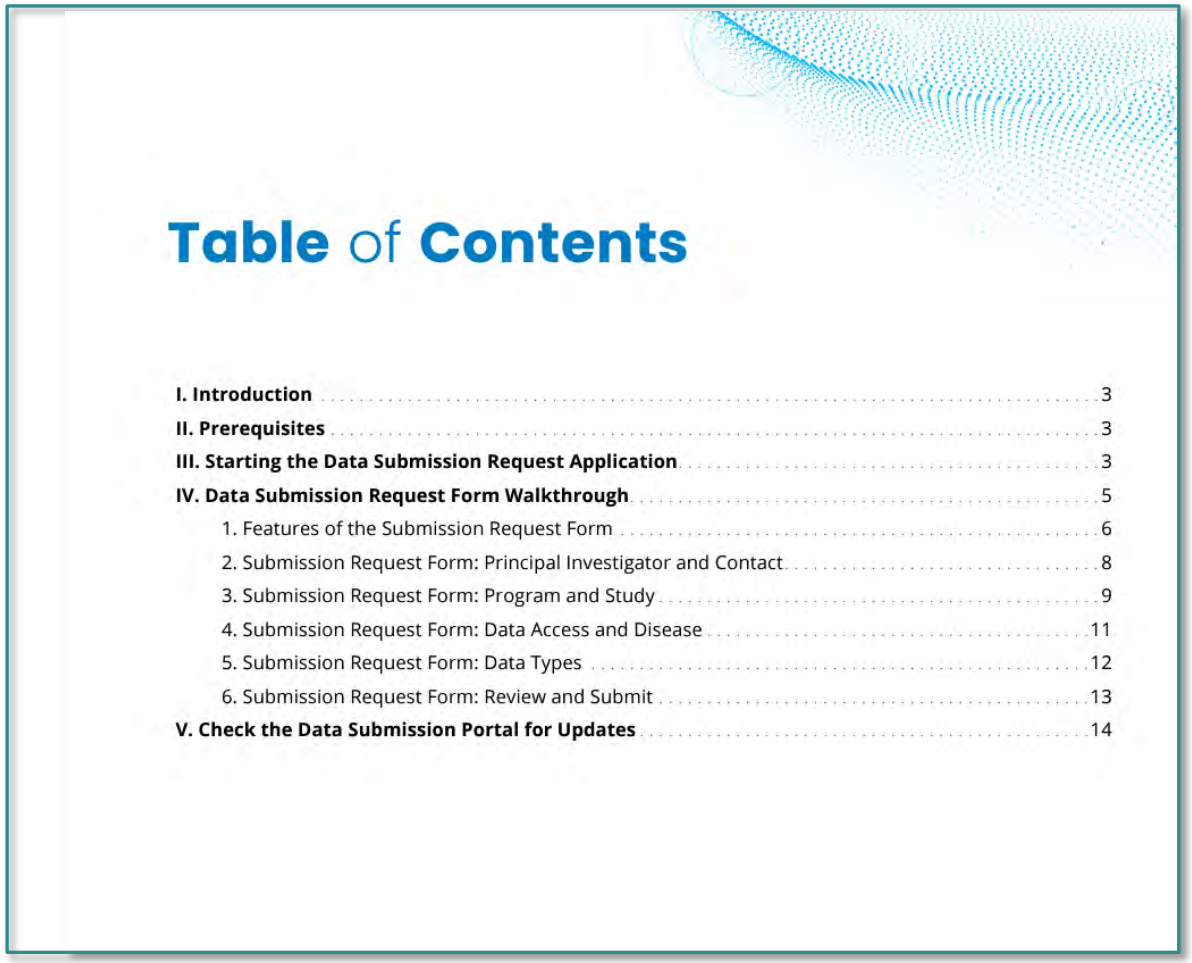
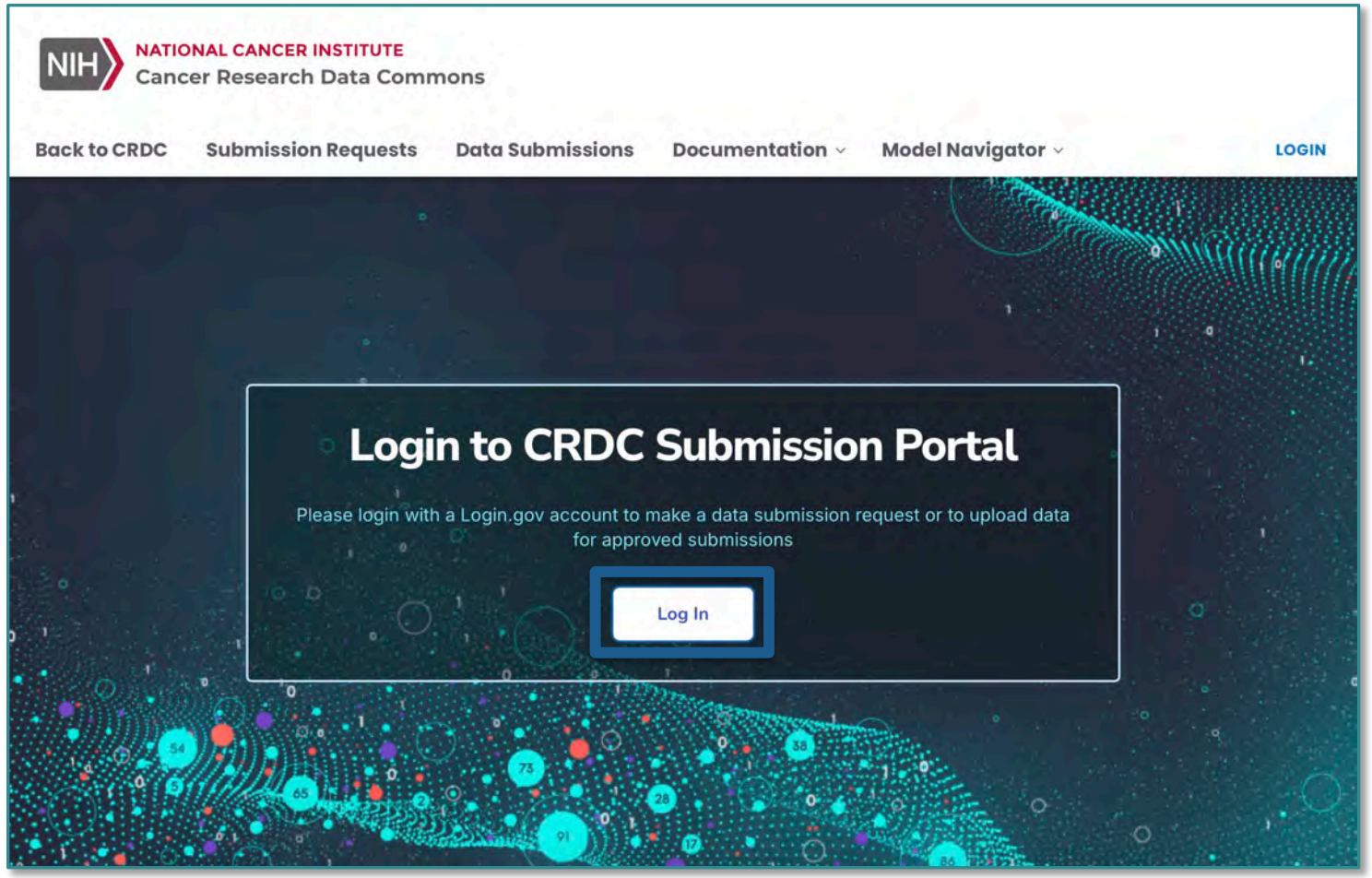


Table of Contents

I. Introduction 3
II. Prerequisites 3
III. Starting the Data Submission Request Application 3
IV. Data Submission Request Form Walkthrough 5
1. Features of the Submission Request Form 6
2. Submission Request Form: Principal Investigator and Contact 8
3. Submission Request Form: Program and Study 9
4. Submission Request Form: Data Access and Disease 11
5. Submission Request Form: Data Types 12
6. Submission Request Form: Review and Submit 13
V. Check the Data Submission Portal for Updates 14

https://datacommons.cancer.gov/submit

Complete the Submission Request Form: Prerequisite



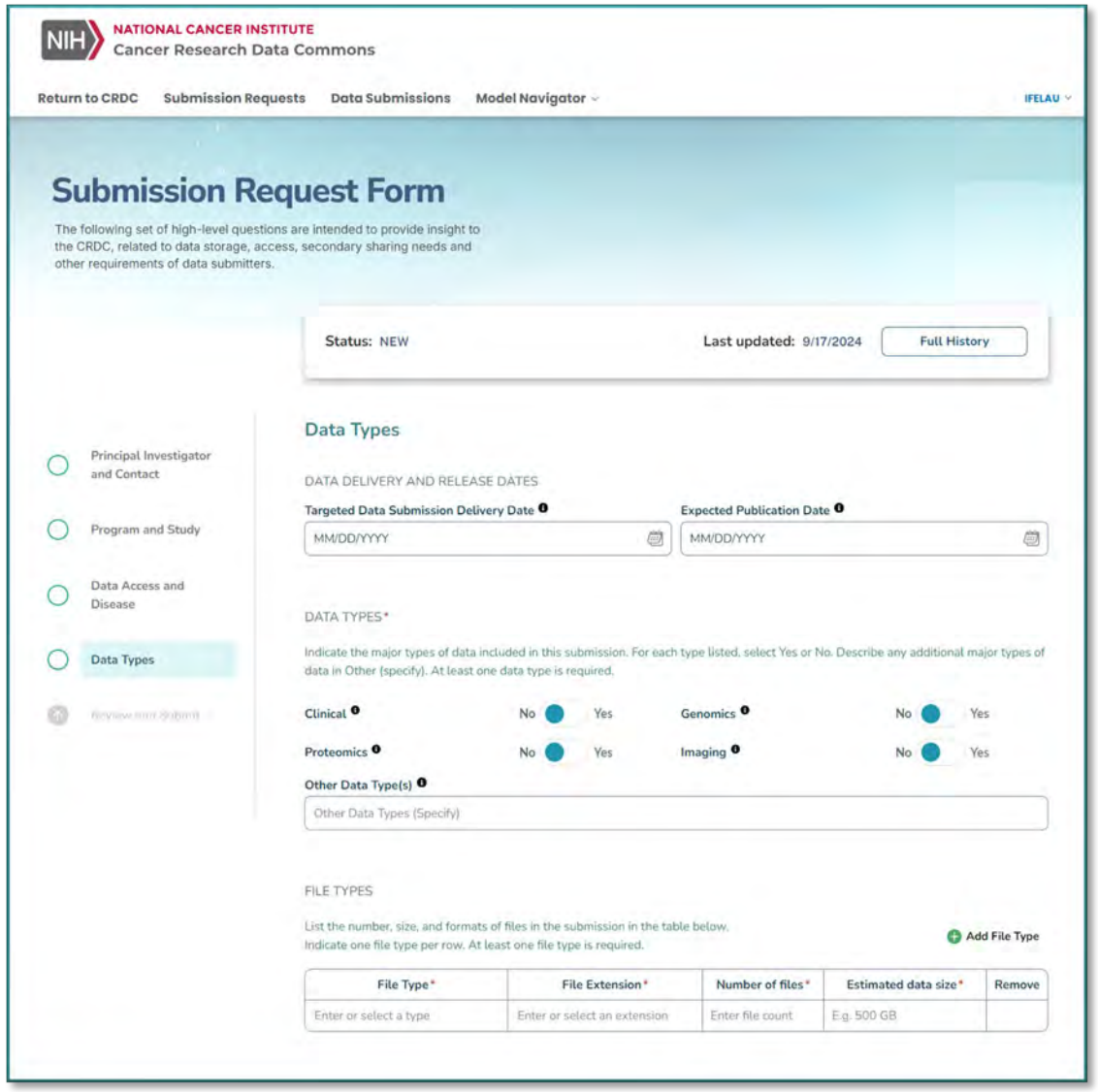
- Use login.gov
  - Grant permission to share the login information with NIH
- For NIH Staff, sign in using the NIH login ID or PIV card

https://hub.datacommons.cancer.gov/

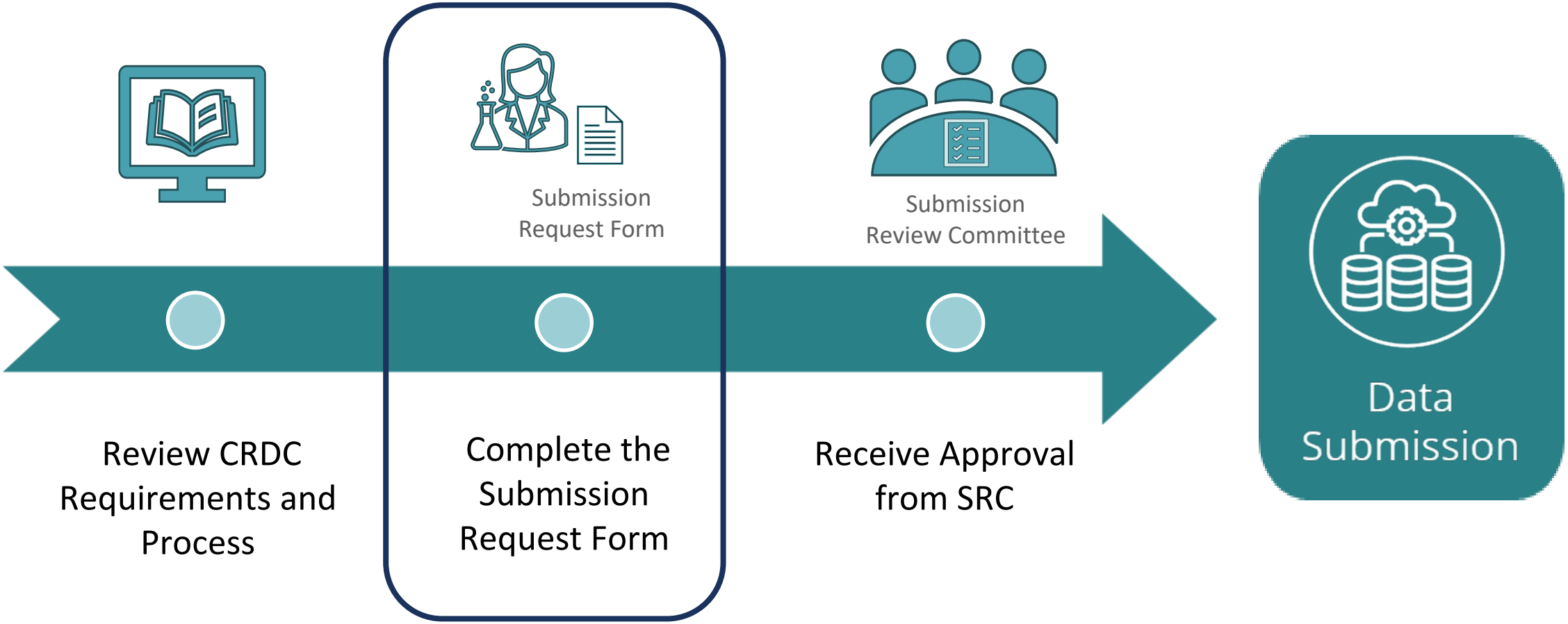
Complete the Submission Request Form

Researchers provide information about the study data they intend to submit:

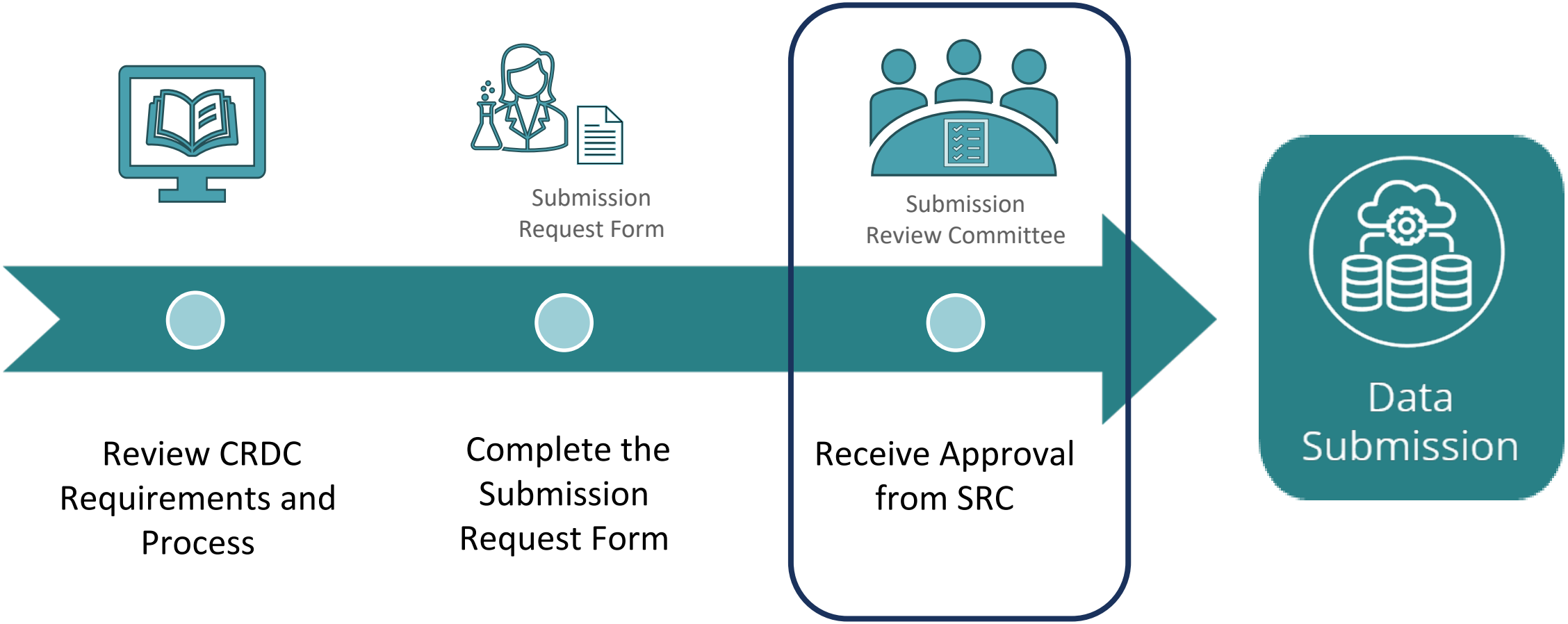
- Program and/or study information
- Open access and/or controlled access data
- Cancer types
- Data types
- Target data submission delivery date
- Expected publication date



Submission Request Process



Submission Request Process



Receive Approval: SRC Evaluation

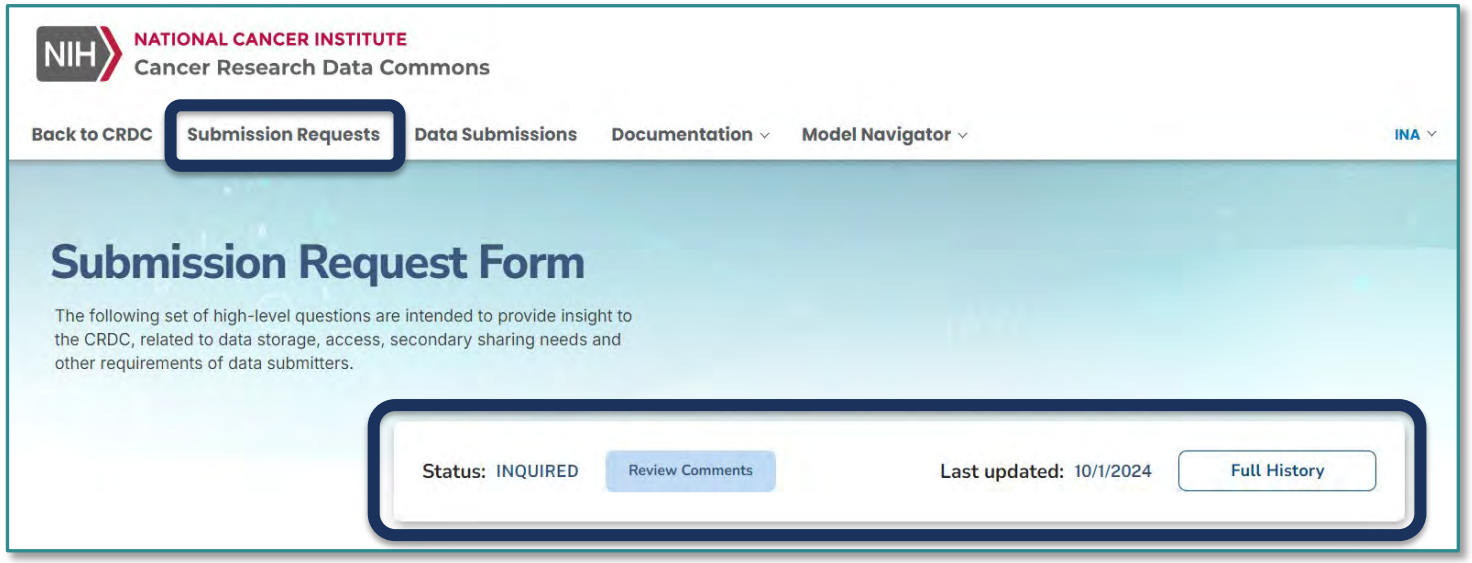
The CRDC Submission Review Committee (SRC) reviews Submission Request Forms and considers how the data and the study methodology contribute to the greater research community.

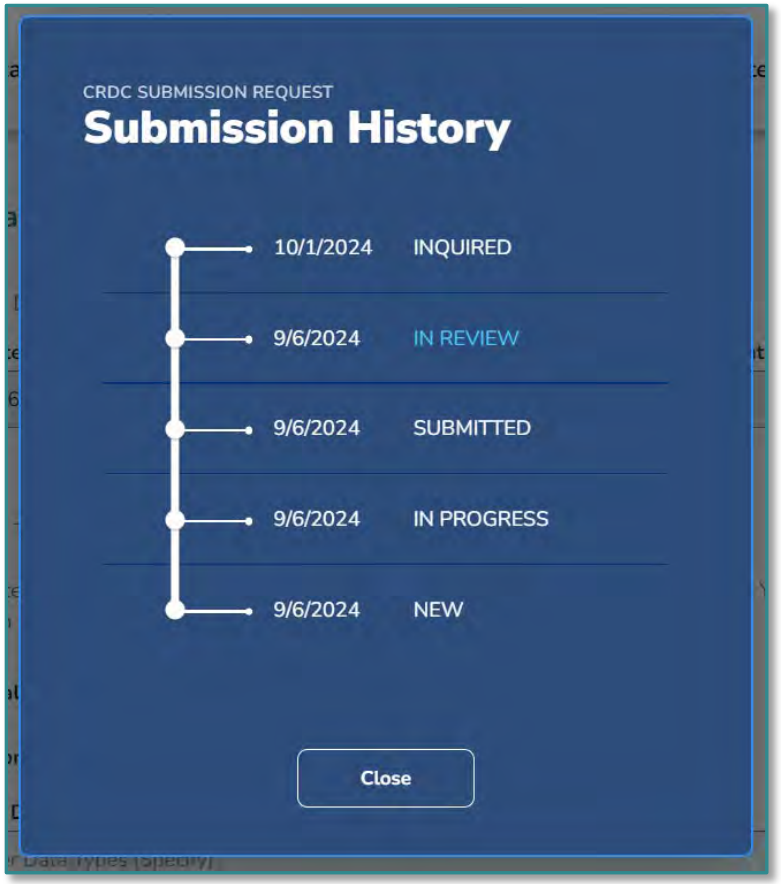
Submission Review Committee (SRC) Evaluation

- The committee is comprised of representatives from each CRDC Data Commons
- Together they serve as the decision body that evaluates any Submission Requests that come through the CRDC Submission Portal

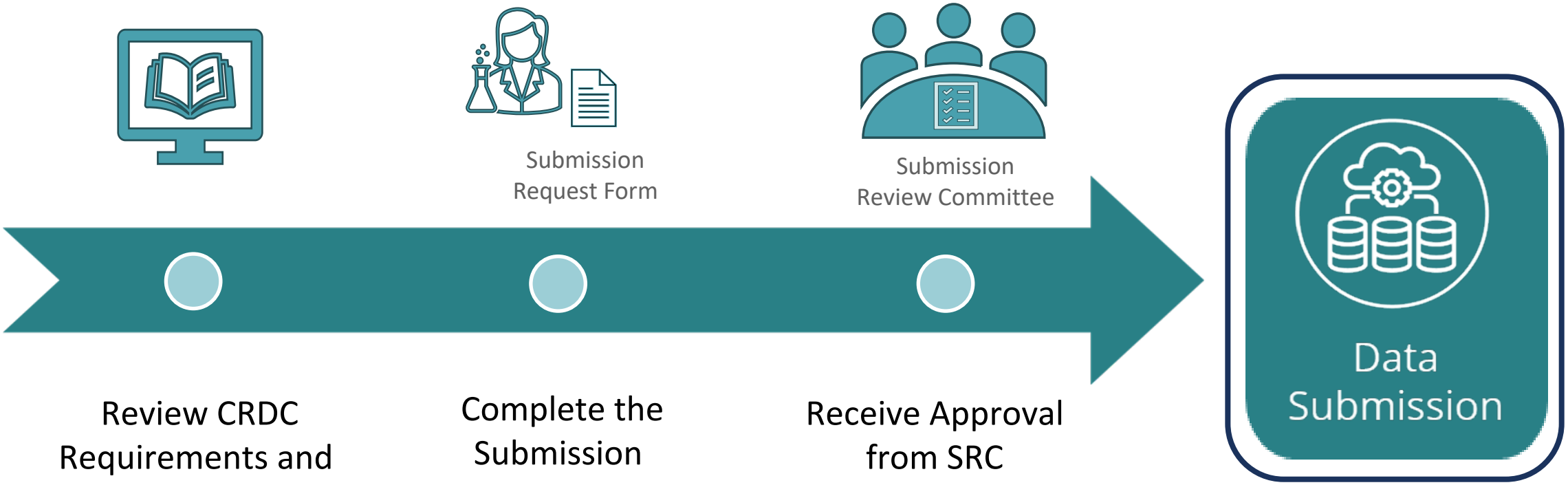
Receive Approval: Status

Researchers can monitor the status of their Submission Request by going to the CRDC Submission Portal:

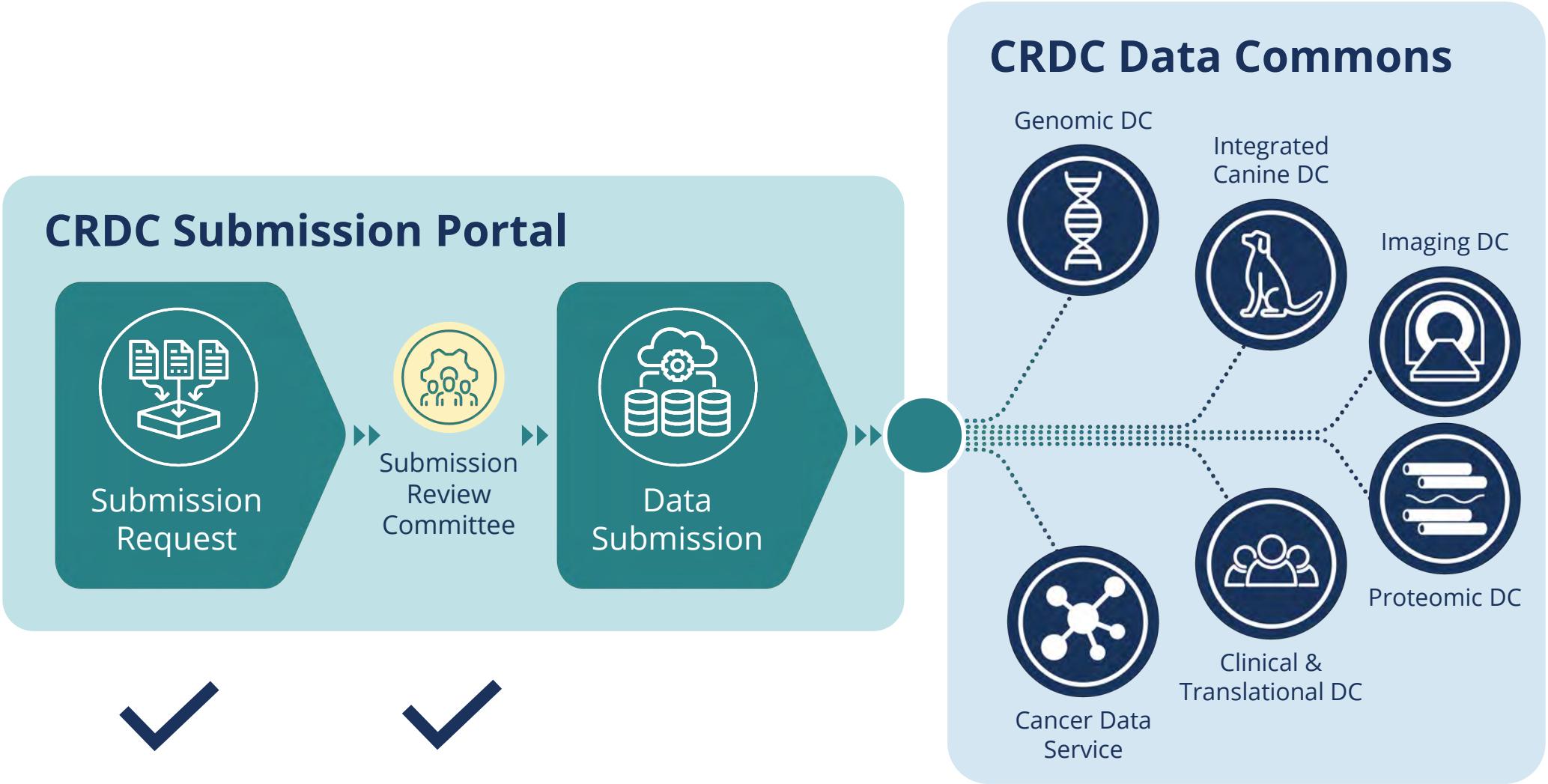




Submission Request Process



CRDC Data Submission



At this point during submission…

- CRDC Submission request approved
- All data de-identified of PII
- Legal permissions/data agreements complete
- Projects registered with dbGaP
- CRDC data and metadata standards have been considered
- Data is ready to upload

Part 2 - Agenda

- CRDC Data Submission Portal & Resources for Data Upload
- Data Submission Workflow
  - Upload & Validate
- Future Plans

What data can you submit to CRDC?

Cancer Research Data

- Genomic
- Proteomic
- Imaging
- Clinical Trial
- Immuno-Oncology
- Population Science

Data & Metadata

Supplementary Files

.RAR .PPTX .XLSX .DOCX .DMG
.APK .ZIP .PNG .PSD .TXT
.MP3 .MP4 .PPT .JPG .XLS
.AI .CSS .HTML .PDF .DOC

What data can you submit to CRDC?

open access

controlled access

New CRDC Data Submission Portal

- **Goal:** single point for *all* CRDC data submission activities
  - **Achieved:** Submission requests across all CRDC commons (Part 1)
  - **Working towards:** Data and metadata submission uploads (Part 2)
- Future integration: Genomic, Proteomic, and Imaging DC
- Currently accepting data/metadata upload submissions for:
  - Integrate Canine Data Commons
  - Clinical and Translational Data Commons
  - **Cancer Data Service (datatype agnostic)***

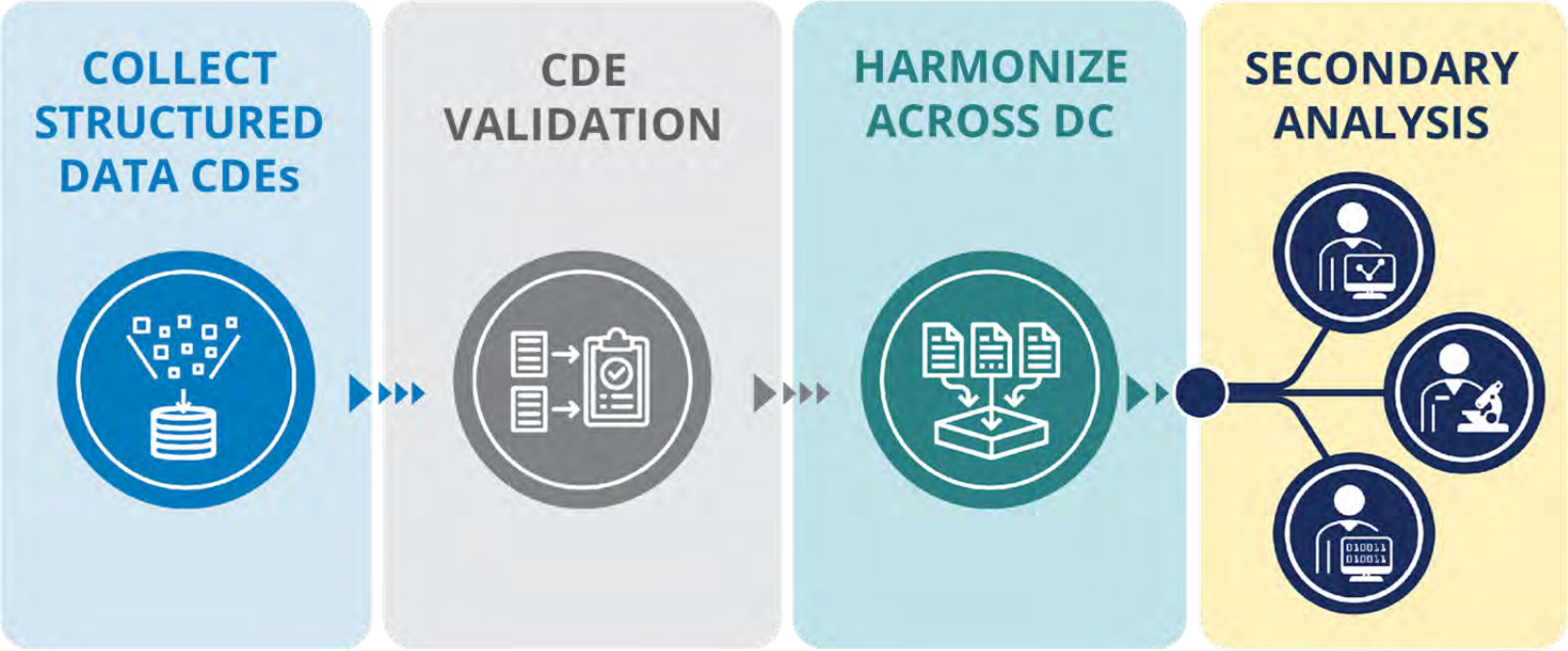
*Exemplar for today's talk

Resources: Get Data Ready for Upload

- Common Data Elements (CDEs)
  - Promote quality, consistency, and FAIRness
- Data Dictionaries
  - Provide definition and permissible values of CDEs
- Data Models
  - Define relationships of data
- Metadata submission templates
  - Guide submitters to structure metadata

Common Data Elements (CDEs)

- Standardize the way metadata is collected and shared
- New CDEs & updated permissible values are added regularly
- CRDC requires a set of CDEs such as age, gender, race, ethnicity



Data Dictionaries

- Describes metadata structure, content, permissible values
- Every Data Model comes with a Data Dictionary
- Required, Preferred and Optional data elements

Graph View | Table View

Dictionary Utils Viz dictionary has 9 nodes and 239 properties

Study

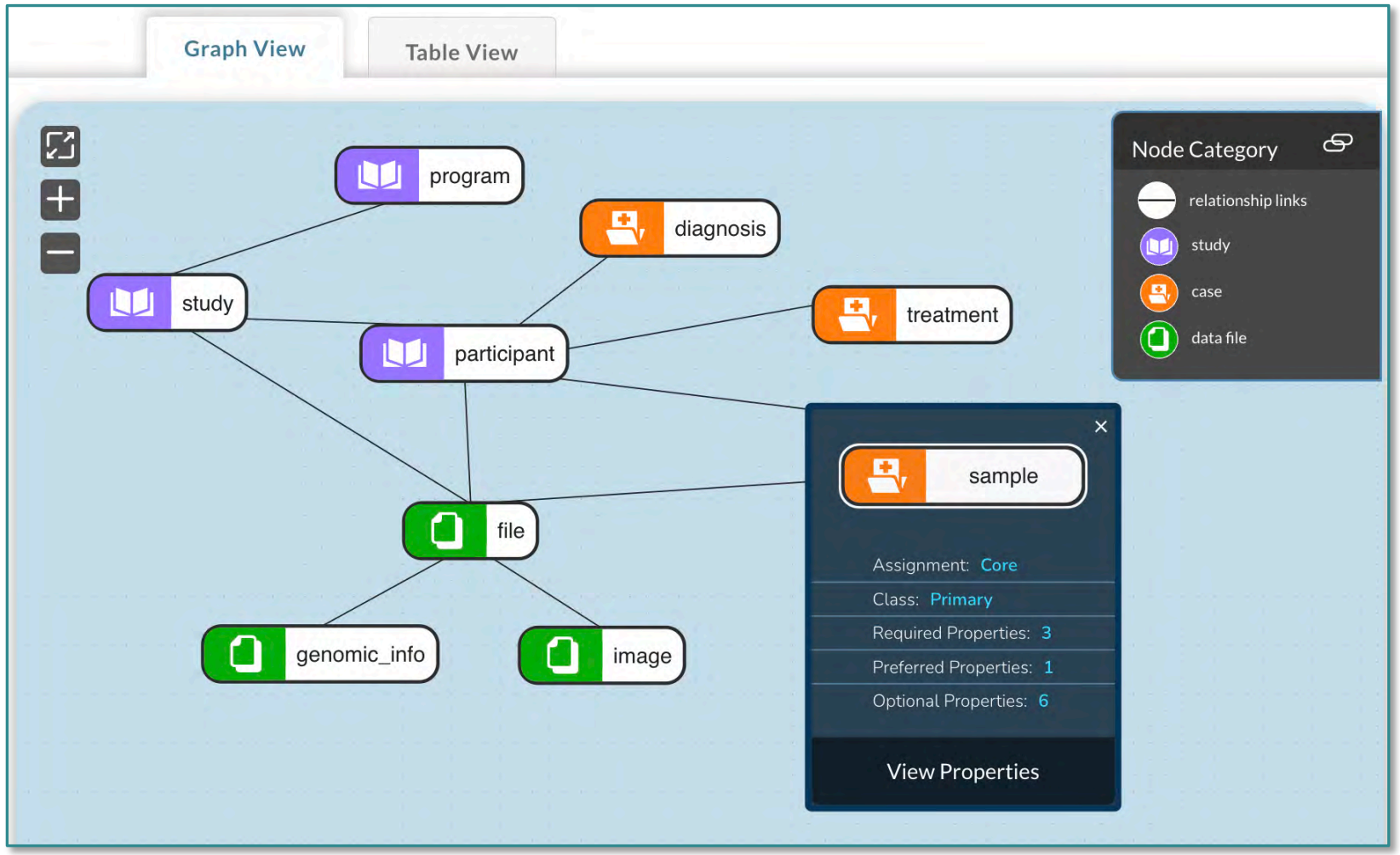
Program Program in the Cancer Data Service refer to a broad framework of goals under which related projects or other research activities are grouped. Example - Clinical Proteomic Tumor Analysis Consortium (CPTAC)

8 Properties | Assignment: Core | Class: Primary | TSV | PDF

Property	Type	Required	Description	Source
program_name	"string"	Required	The name of the program under which related studies will be grouped, in full text and unabbreviated form, exactly as it will be displayed within the UI.	"Internally-curated"
program_acronym	"string"	Required	The name of the program under which related studies will be grouped, expressed in the form of the acronym by which it will identified within the UI. This property is used as the key via which study records can be associated with the appropriate program during data loading, and to identify the correct records during data updates.	"Internally-curated"
program_short_description	"string"	Preferred	An abbreviated, single sentence description of the program.	"Internally-curated"
program_full_description	"string"	Preferred	A more detailed, multiple sentence description of the program.	"Internally-curated"
program_external_url	"string"	Preferred	The external url to which users should be directed in order to learn more about the program.	"Internally-curated"
program_sort_order	"integer"	Optional	An arbitrarily-assigned value used to dictate the order in which programs are displayed within the application's UI.	"Internally-curated"

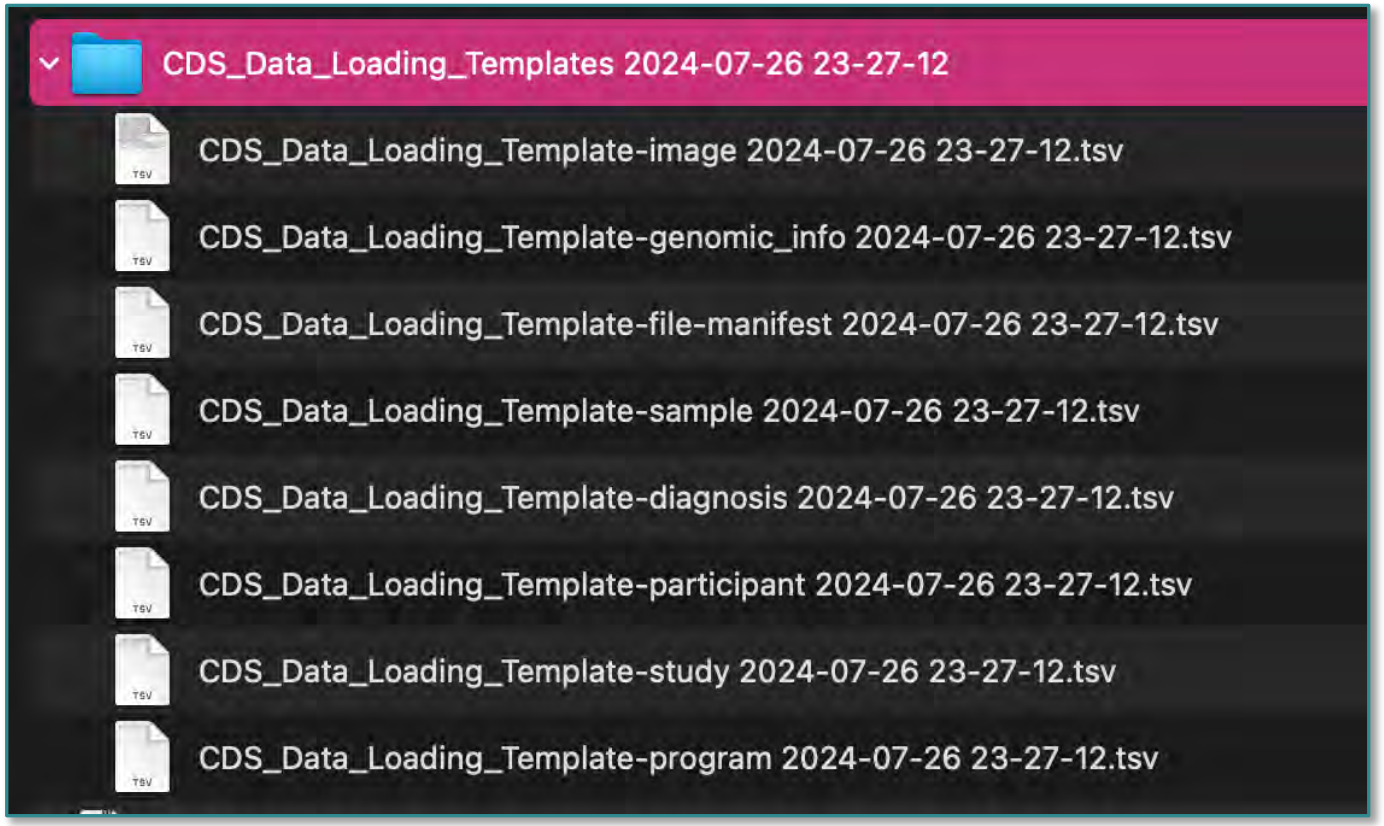
Data Models

- How data is organized and structured
- Ensures/facilitates accuracy and reusability
- Graph & Table views
- README



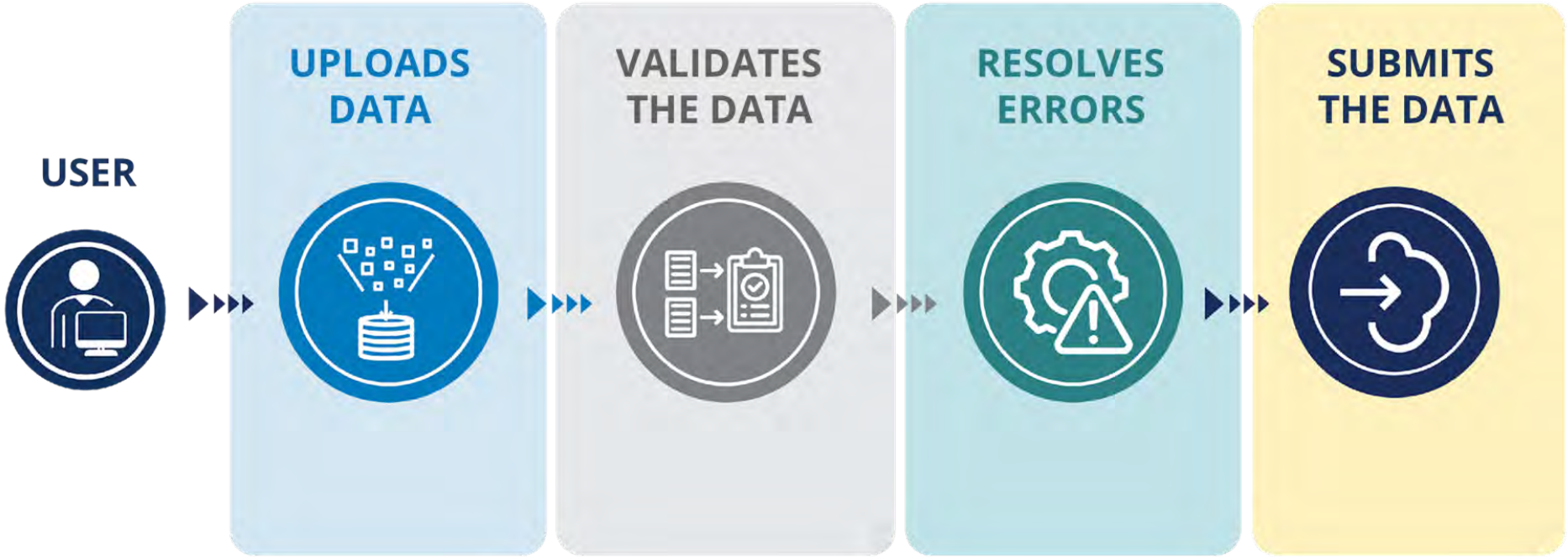
Metadata Submission Templates

- Collect information about metadata
- Consistent metadata = easy to search
- One template for each node of Data Model
- All Vocabularies
- File Examples
- Used to validate data



Data Submission Workflow

- Once you review the required standards it is time to upload
- After your data is validated and all errors resolved you are ready to submit



Get Started: http://datacommons.cancer.gov/submit

Data Submission

Process and Timing

Once the Submission Request has been approved by the CRDC Submission Review Committee, the submitter may proceed to data submission. The PI or primary contact will be assigned a data concierge who will work with them and provide assistance throughout the data submission process.

The submission process involves uploading and validating a metadata manifest and the data files through the CRDC Submission Portal. When a dataset has passed all validations, the final submission is pushed to the Data Submission team for review before it is released to the appropriate CRDC Data Commons, which make the data available through their portals.

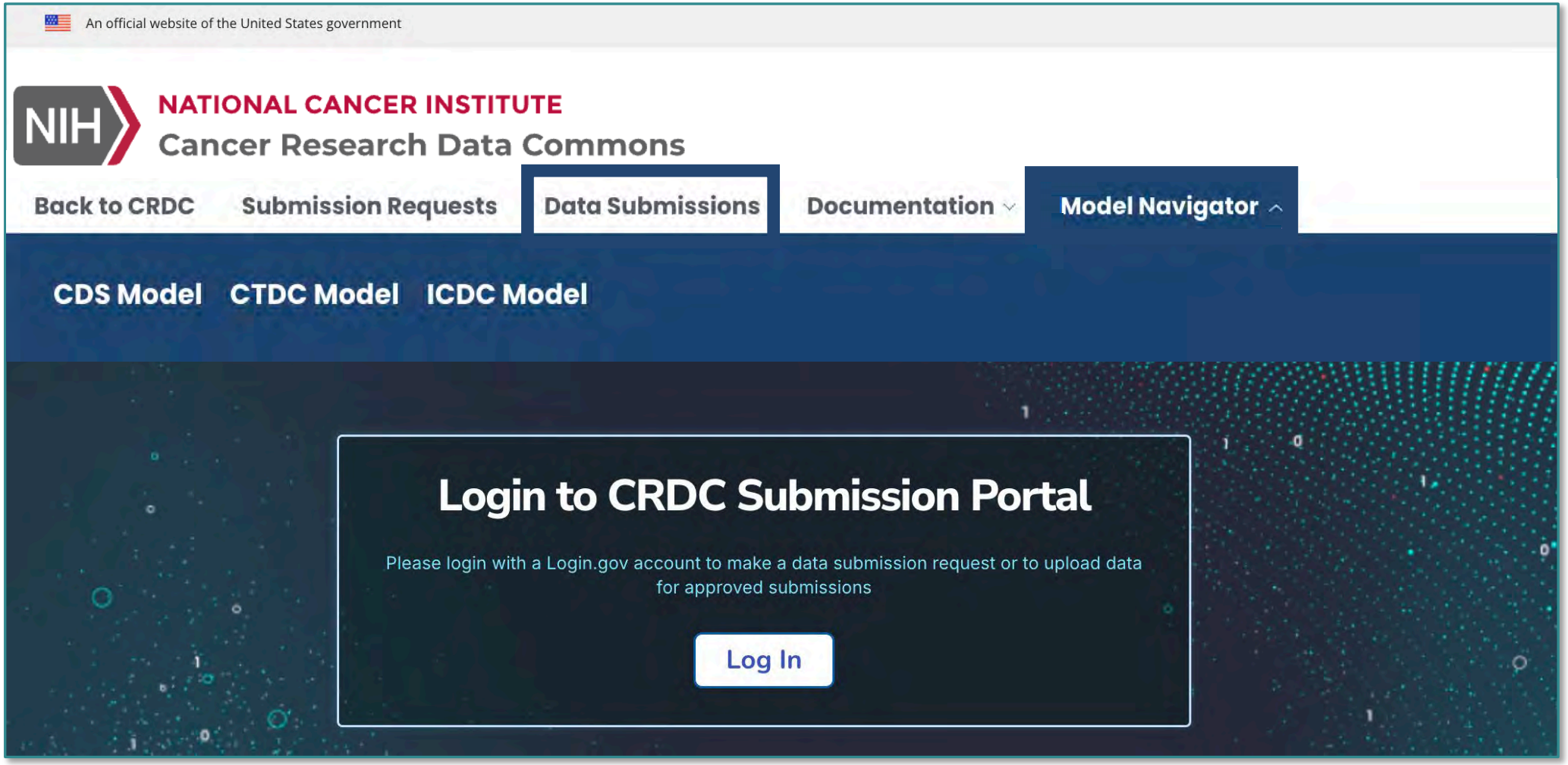
For users looking to align their data with CRDC standards before starting the submission process, the data model viewer is available to outline the types of data required. Users can also download the data dictionary and sample metadata templates to guide their submission process. These resources are available through the CRDC Data Submission Portal through the Model Navigator in the menu.

In addition, a comprehensive list of CRDC standard CDEs can be found at caDSR. Click the "CRDC Standard Data Elements" link in the *Links to Favorites* section or download them from the getCRDCList endpoint of the caDSR API.

Instructions and Portal

Data Submission Instructions	Detailed instructions are provided as a PDF.
CRDC Submission Portal	Link to the portal to start the Data Submission.

CRDC Submission Portal



Create a Data Submission

NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Back to CRDC | Submission Requests | Data Submissions | Documentation | Model Navigo

Data Submission List

Below is a list of data submissions that are associated with your account.
Please click on any of the data submissions to review or continue work.

Organization: NCI

Status: All

Submission Name	Submitter	Data Commons	Type	DM Version	Organization	Study	dbGaP ID
Demo STudy	Durga Addepalli	CDS	New/Update	4.0.1	NCI	STUDYABBREVIATION	

Create a Data Submission

Please fill out the form below to start your data submission

Submission Type*
New/Update | Delete

Data Type*
Metadata and Data Files | Metadata Only

Organization
NCI

Data Commons*
CDS

Study*
Select

dbGaP ID
Input dbGaP ID

Submission Name*
25 characters allowed

Create

Submission: Upload Metadata

Prior to beginning uploading process, read detailed instructions available in the Data Submission Instructions.

1 UPLOAD METADATA

Metadata Files | Choose Files | *No files selected* | Upload

2 UPLOAD DATA FILES

The CLI Tool is used to upload data files to CRDC Submission Portal and requires a configuration file to work. The CLI Tools is a one-time download however the configuration file needs to be customized for each submission. You can either edit the example configuration files found in the CLI Tool download, or you can click the button on the right to download a configuration file customized for this submission.

Download Configuration File

3 VALIDATE DATA

Validation Type: Validate Metadata | Validate Data Files | Both

Validation Target: New Uploaded Data | All Uploaded Data

Validate

Upload Activities | Validation Results | Data View

Batch ID	Batch Type	File Count	Status	Uploaded Date ↓	Upload Errors
1	Metadata	8	Uploaded	10-09-2024 at 10:26 PM	

Rows per page: 20 1-1 of 1 1

Submission: Upload Data

NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Back to CRDC | Submission Requests | Data Submissions | Documentation | Model Navigator | DURGA

User Profile | Uploader CLI Tool | API Token | Logout

Uploader CLI Tool

The Uploader CLI is a command-line interface tool provided for directly uploading data submission files from your workstation to the CRDC Submission Portal cloud storage. To download the tool and accompanying instructions, click on the Download button below.

Close | Download

API Token

An API Token is required to utilize the Uploader CLI tool for file uploads.

Each time you click the 'Create Token' button, a new token will be generated, and the previous token will be invalidated. A token expires 60 days after its creation.

Create Token

Close

Submission: Upload Data

Prior to beginning uploading process, read detailed instructions available in the Data Submission Instructions.

1 UPLOAD METADATA

Metadata Files Choose Files *No files selected* Upload

2 UPLOAD DATA FILES

The CLI Tool is used to upload data files to CRDC Submission Portal and requires a configuration file to work. The CLI Tools is a one-time download however the configuration file needs to be customized for each submission. You can either edit the example configuration files found in the CLI Tool download, or you can click the button on the right to download a configuration file customized for this submission.

Download Configuration File

3 VALIDATE DATA

Validation Type: Validate Metadata Validate Data Files Both

Validation Target: New Uploaded Data All Uploaded Data

Validate

Submission: Validate Data

Prior to beginning uploading process, read detailed instructions available in the Data Submission Instructions.

1 UPLOAD METADATA

Metadata Files Choose Files *No files selected* Upload

2 UPLOAD DATA FILES

The CLI Tool is used to upload data files to CRDC Submission Portal and requires a configuration file to work. The CLI Tools is a one-time download however the configuration file needs to be customized for each submission. You can either edit the example configuration files found in the CLI Tool download, or you can click the button on the right to download a configuration file customized for this submission.

Download Configuration File

3 VALIDATE DATA

Validation Type: Validate Metadata Validate Data Files Both

Validation Target: New Uploaded Data All Uploaded Data

Validate

Submission: Validate Data

Upload Activities | **Validation Results** | Data View

Batch ID: All

Node Type: All

- All
- genomic_info
- diagnosis

Severity: All

400 | 1 | 2 | 3 | 4 | 5 | ... | 20

Validation Issues

For Sample Node ID CDS1019_DNA

3 ISSUES

1. (Error) [cds_mock_data-sample.tsv: line 20] "DNA" is not a permissible value for property "sample_type".

2. (Error) [cds_mock_data-sample.tsv: line 20] "Abnormal" is not a permissible value for property "sample_tumor_status".

3. (Error) Related node "participant" ["study_participant_id": "phs000000_CDS1019"] not found.

Close

			Severity	Date	Issues
1		3	Error	09-13-2024 at 12:00 PM	Related node not found. See details.
1	Diagnosis	CDS-DX-136782	Error	09-13-2024 at 12:00 PM	Value not permitted. See details.

Submission: Data View

Upload Activities | Validation Results | **Data View**

Node Type: sample

Status: All

Submitted ID

- sample
- participant
- image
- genomic_info
- file
- diagnosis
- study
- program

Rows per page: 5 | 1–5 of 200 | 1 2 3 4 5 ... 40

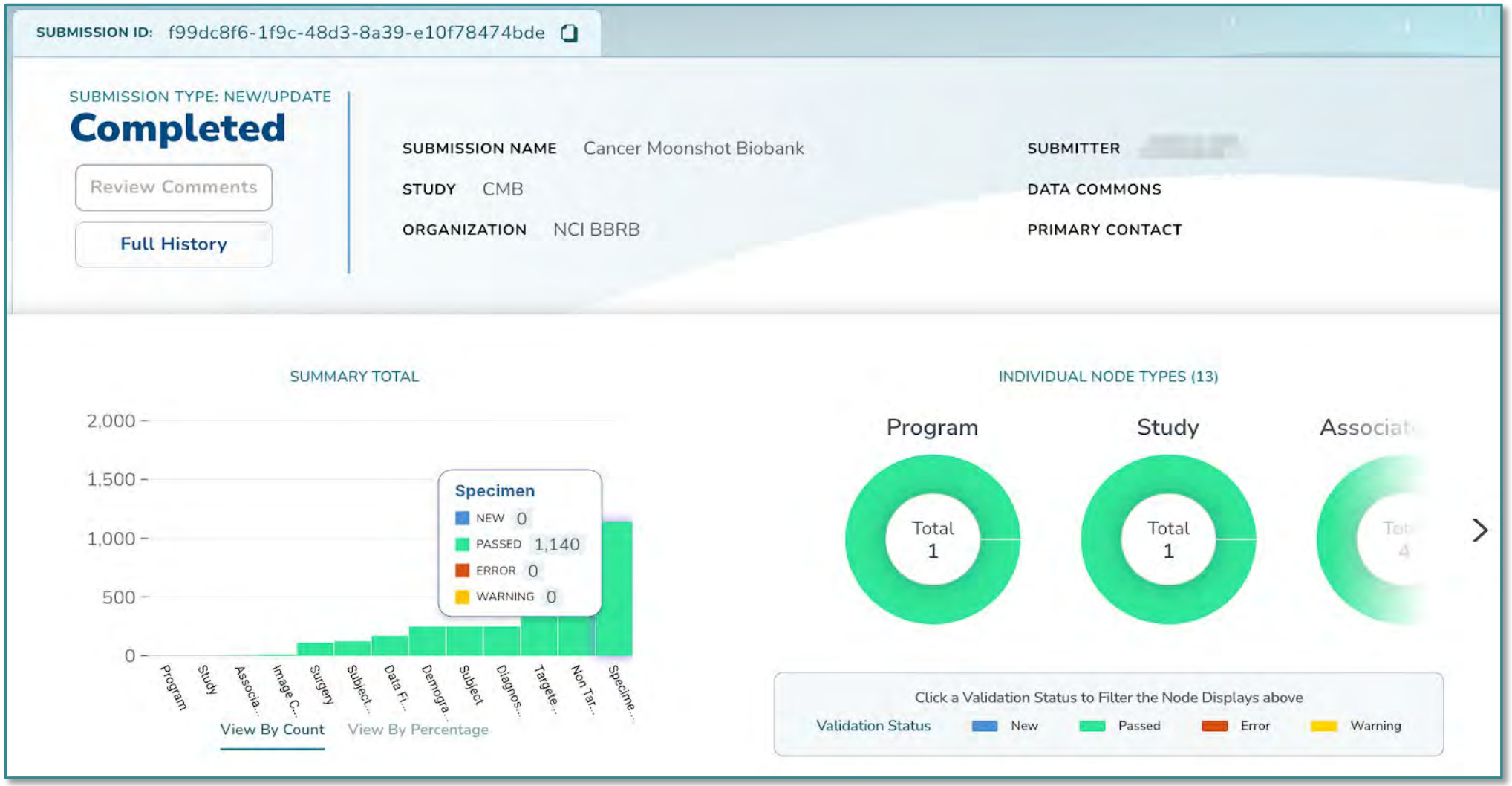
			sample_type	sample_anatomic_site	sample_tumor_status	participant
			RNA	tumor	Tumor	phs00000
			RNA	tumor	Tumor	phs00000
CDS1198_RNA	New		RNA	tumor	Unknown	phs00000
CDS1197_RNA	New		RNA	tumor	Unknown	phs00000
CDS1196_RNA	New		RNA	tumor	Unknown	phs00000

Rows per page: 5 | 1–5 of 200 | 1 2 3 4 5 ... 40

Submission: Validation on the Portal

- Validations run by submitter on Portal:
  - Metadata
    - Required CDEs and permissible values
    - Files validated against the selected data model
  - Data Files
    - Duplicate files
    - Duplicate Samples and participant IDs

Submission: Complete

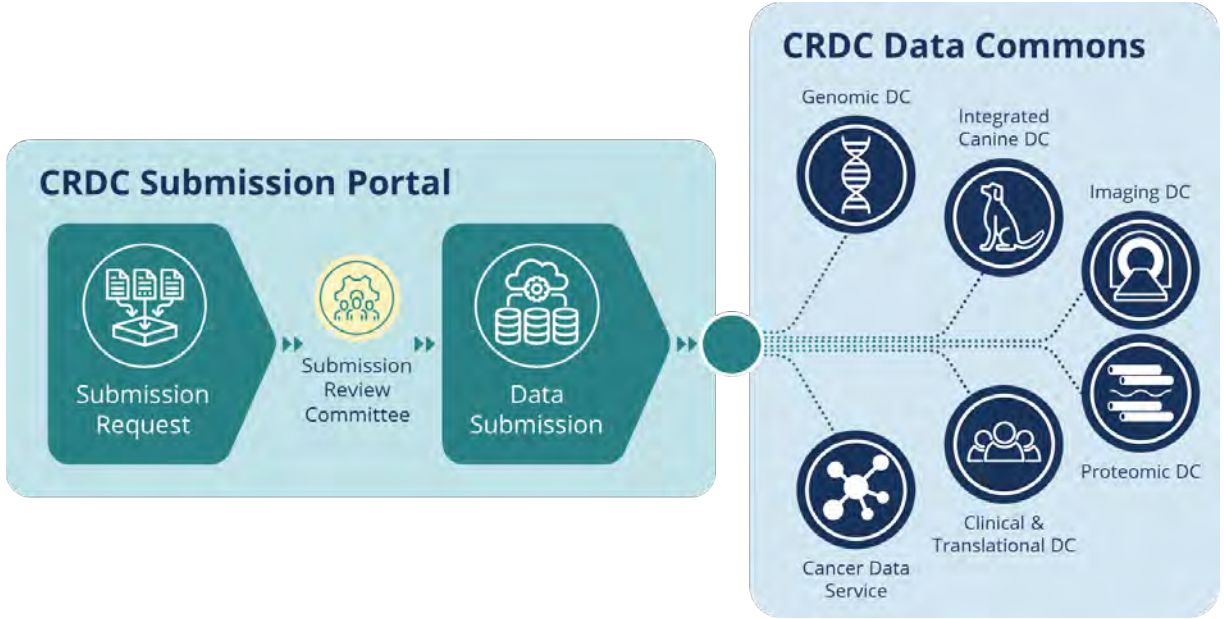


Post Submission Validation

- Validations run by CRDC Curators:
  - Data not covered by the Data Model
  - Cross validation within a study
    - Multiple active data submissions
    - Files listed on metadata templates and actual files
  - dbGaP validations
    - Study, Sample and Participant Ids are cross validated between dbGaP & CRDC

Future Plans

- Centralized submission of data across all CRDC
- Additional Data Commons integrated into the Submission Portal for data & metadata uploads



Where do I go if I have questions?

- The CRDC Submit Landing page has several helpful resources for answering questions and troubleshooting submission challenges, including:
  - Frequently asked questions
  - Information about relevant NIH policies and guidelines
  - Additional resources to learn more about the CRDC Data Ecosystem

Submitters can contact the CRDC Help Desk at NCICRDC@mail.nih.gov inquiries, including questions about the Submission Request Form or the Data Submission Process

Acknowledgments

NCI

- Durga Addepalli
- Emi Casas-Silva
- Heather Creasy
- Ina Felau
- Erika Kim
- David Sturgill
- Granger Sutton
- Xu Zhang

FFRDC/FNL

- Amanda Bell
- Kailing Chen
- Naila Gulzar
- Mark Jensen
- Todd Pihl

All CRDC team members

All partners throughout NCI/NIH and data contributors

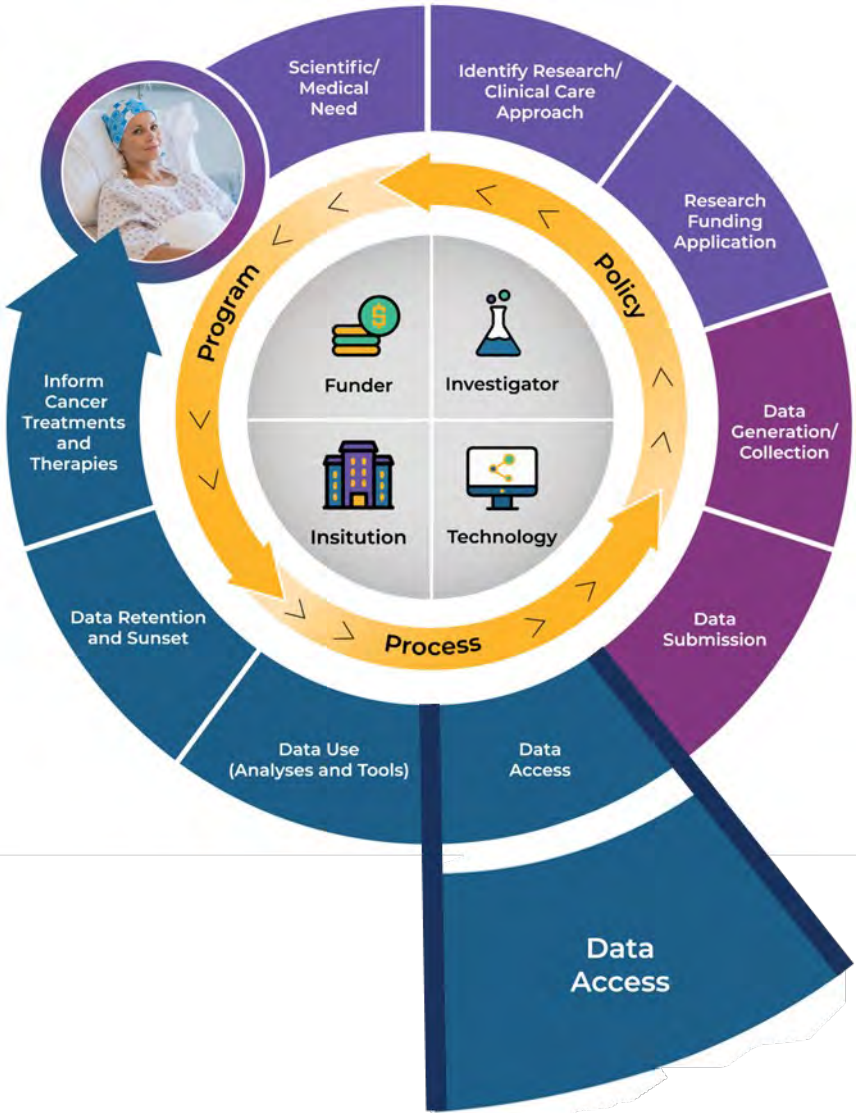
From Data to Insight: A Journey Through CRDC Resources

Heather Creasy & Erin Beck

October 16, 2024

NIH NATIONAL CANCER INSTITUTE

NCI Data Lifecycle: Data Access



Science-first

Hear a talk

- Where was data shared?
- How to access data?
- Is data in usable format?

Read a paper

- Is there a link to data?
- How to access data?
- Is data in usable format?

Data-first

Search Repositories

- Multiple interfaces
- How to access data?
- Is data in usable format?

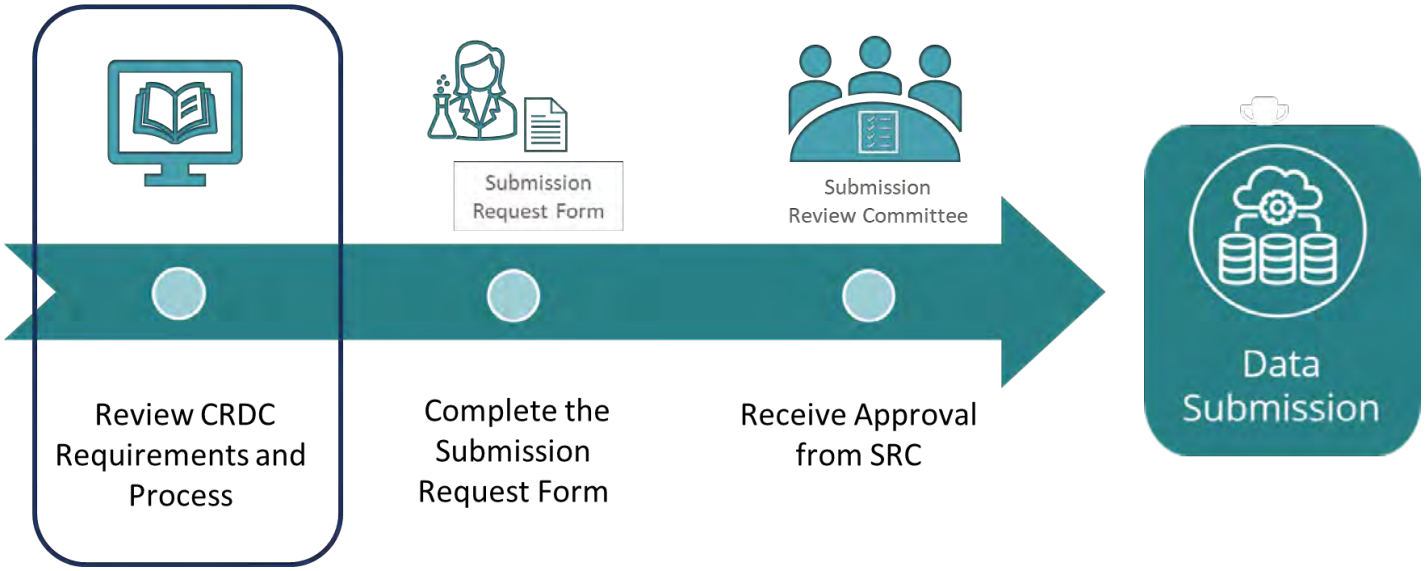
Search Harmonized Data Commons

- Common search terms
- Consistent access
- Standard file formats

Data is not useful unless it’s usable

This is why it’s important to:

- Share data in public repositories
- Be familiar with submission requirements & processes early on
- Follow data submission standards processes, including use of Common Data Elements (CDEs), ontologies & controlled vocabularies
- Ensure that data is accurate, complete, consistent, and valid



Data *within* CRDC Data Commons

Sex at Birth	Subject Reported Ethnicity	Anatomic Site	Disease Type
female	Hispanic or Latino	C17.9 : Small intestine, NOS	Neuroepithelial neoplasm, Glioma
Female	M	Retroperitoneum/Upper abdominal - Small Intestines	Low-grade glioneuronal tumor
000	Mexican, Mexican	Small Bowel	Low grade neuroglial tumor
F	Hispanic/Spanish	mucosa	Glioma, NOS
FEMALE	H	distal	Glial/glioneuronal neoplasm
1	Hispanic Latino	Small intestine	Glial tumor
U	Hispanic or Lati	Small intestine, NOS	Glial Neoplasm
	6	Small Intestine	Morphologically Consistent with Low Grade Glial Neoplasm
	Spanish Origin	Small intestine	
	HISPANIC OR LATINO	Other ill-defined sites	
	ETHNICGRP873		

Data *within* CRDC Data Commons

Sex at Birth

female
Female
000
F
FEMALE
1
U

caDSR Public ID: 7572817
NCIt concept code: C16576

Preferred Name:
Female

Subject Reported Ethnicity

Hispanic or Latino
M
Mexican, Mexican Hispanic/Spanish
H
Hispanic Latino
Hispanic or Lati
6
Spanish Origin
HISPANIC OR LATINO
ETHNICGRP873

caDSR Public ID: 2192217
NCIt concept code: C17459

Preferred Name:
Hispanic or Latino

Anatomic Site

C17.9 : Small intestine, NOS
Retroperitoneum/Upper abdominal - Small Intestines
Small Bowel
mucosa
distal
Small intestine
Small intestine, NOS
Small Intestine
Small intestine
Other ill-defined sites

caDSR Public ID: 14883047
NCIt concept code: C12386

Preferred Name:
Small Intestine

Disease Type

Neuroepithelial neoplasm, Glioma
Low-grade glioneuronal tumor
Low grade neuroglial tumor
Glioma, NOS
Glial/glioneuronal neoplasm
Glial tumor
Glial Neoplasm
Morphologically Consistent with Low Grade Glial Neoplasm

caDSR Public ID: 14905532
NCIt concept code: C3059

Preferred Name:
Glioma

Cancer Data Aggregator (CDA)

WHAT

Search using **harmonized, common language terms**

- Search for public information (**metadata**) on subjects, files, specimen across modalities
- Use CRDC common data elements (**CDEs,** cadsr.cancer.gov)
- Retrieve results in a standard format (tsv)

HOW

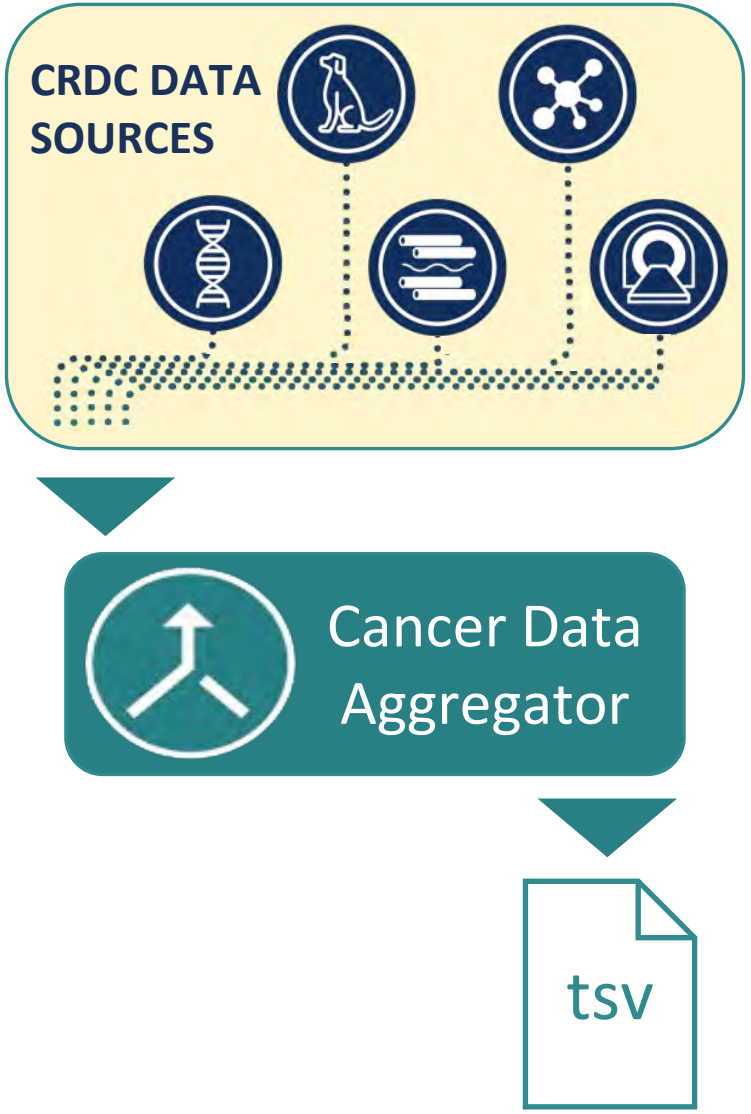
CDA is available as

- Application Programming Interface (API)
- Simple query language packaged in python - cdapython
- Interactive ipython notebooks powered by google colab

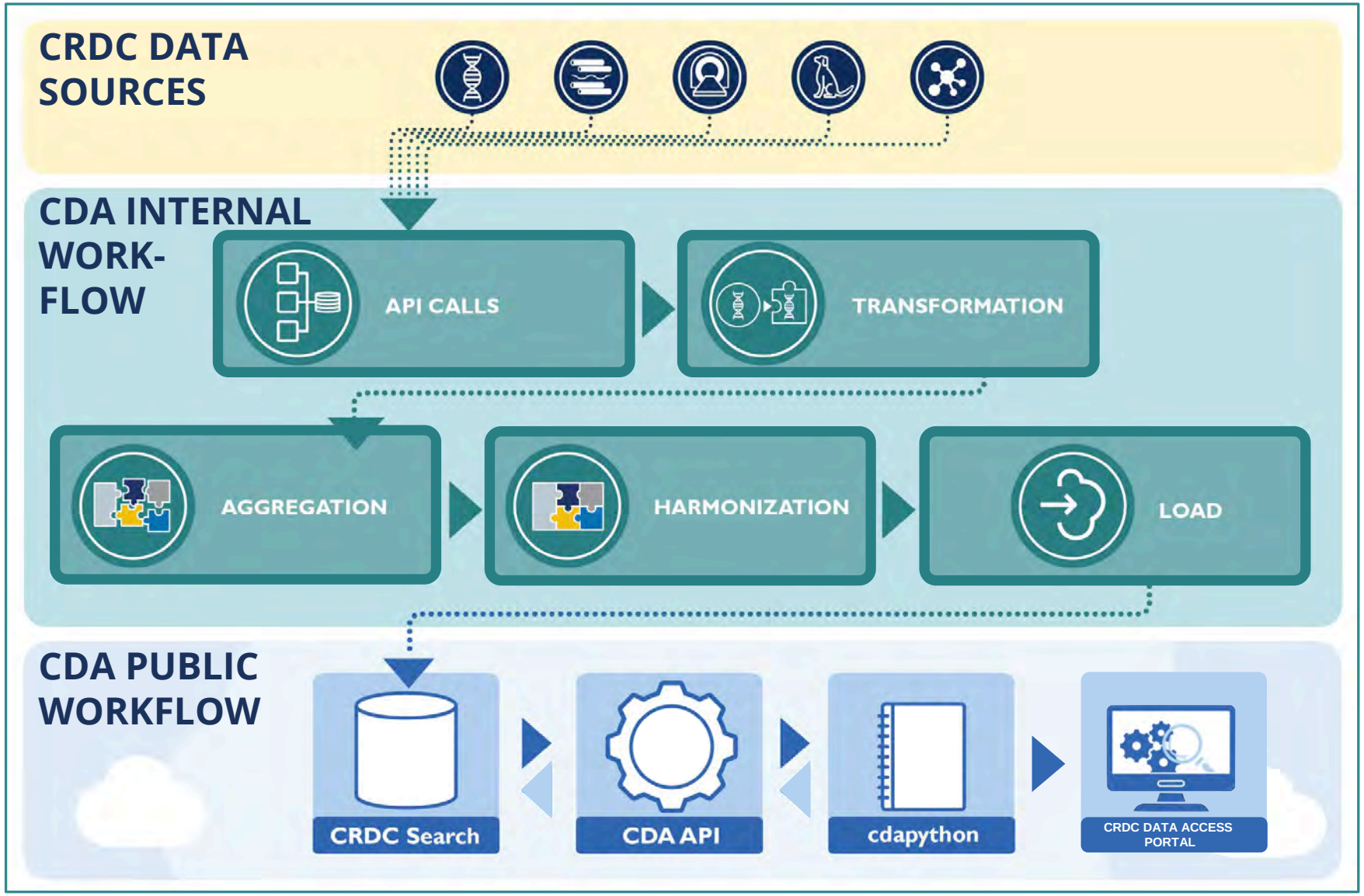
WHERE

Access CDA at **cda.readthedocs.io**

- No code flavor: interactive filtering tool
- Low code flavor: no install, CDA in the cloud
- Power user flavor: install cdapython and run within your own environment



CDA Enables the CRDC Data Access Portal

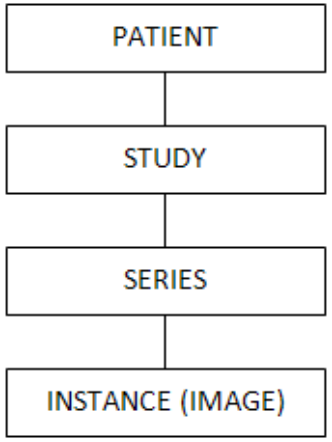


Aggregation & Harmonization of Data *across* CRDC DCs

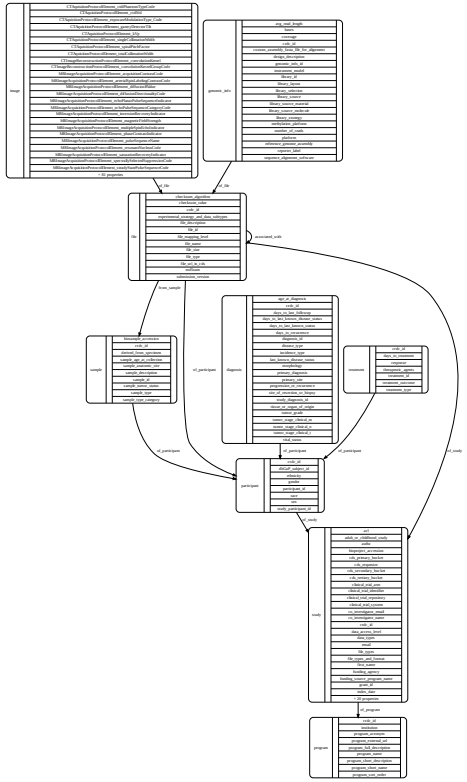
Genomic DC

Proteomic DC

Imaging DC



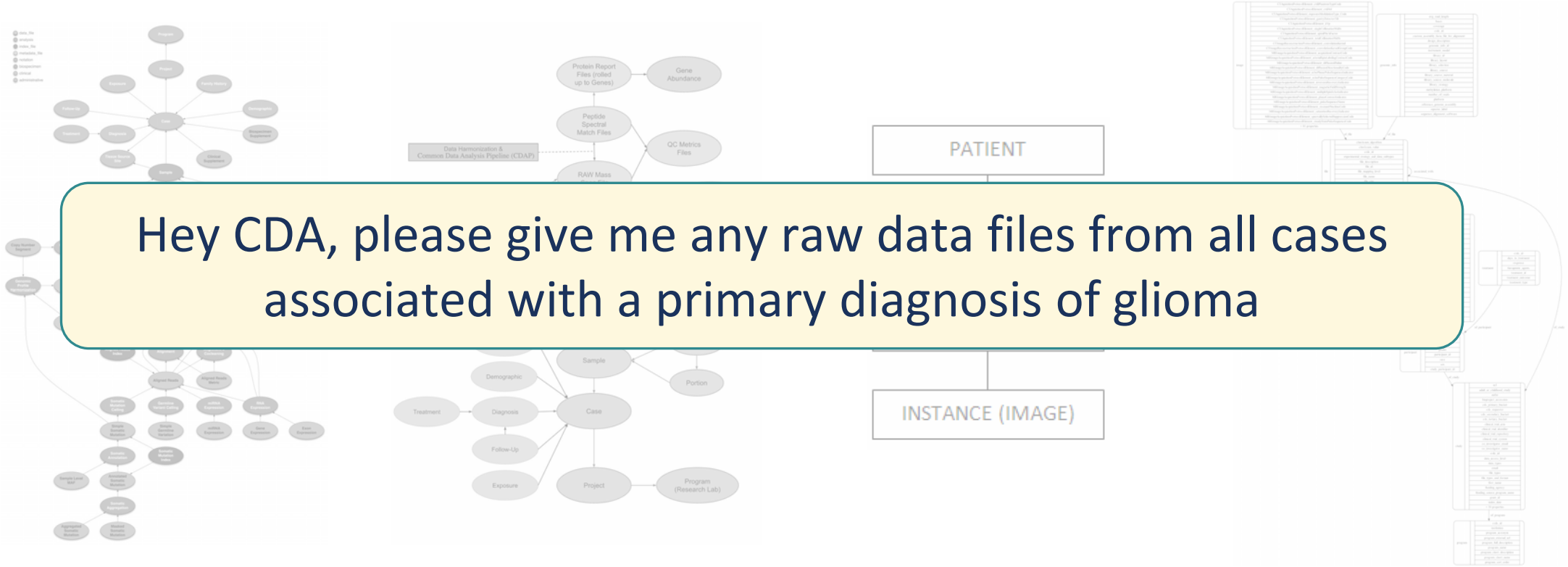
Cancer Data Service



* Fabricated example

Aggregation & Harmonization of Data *across* CRDC DCs

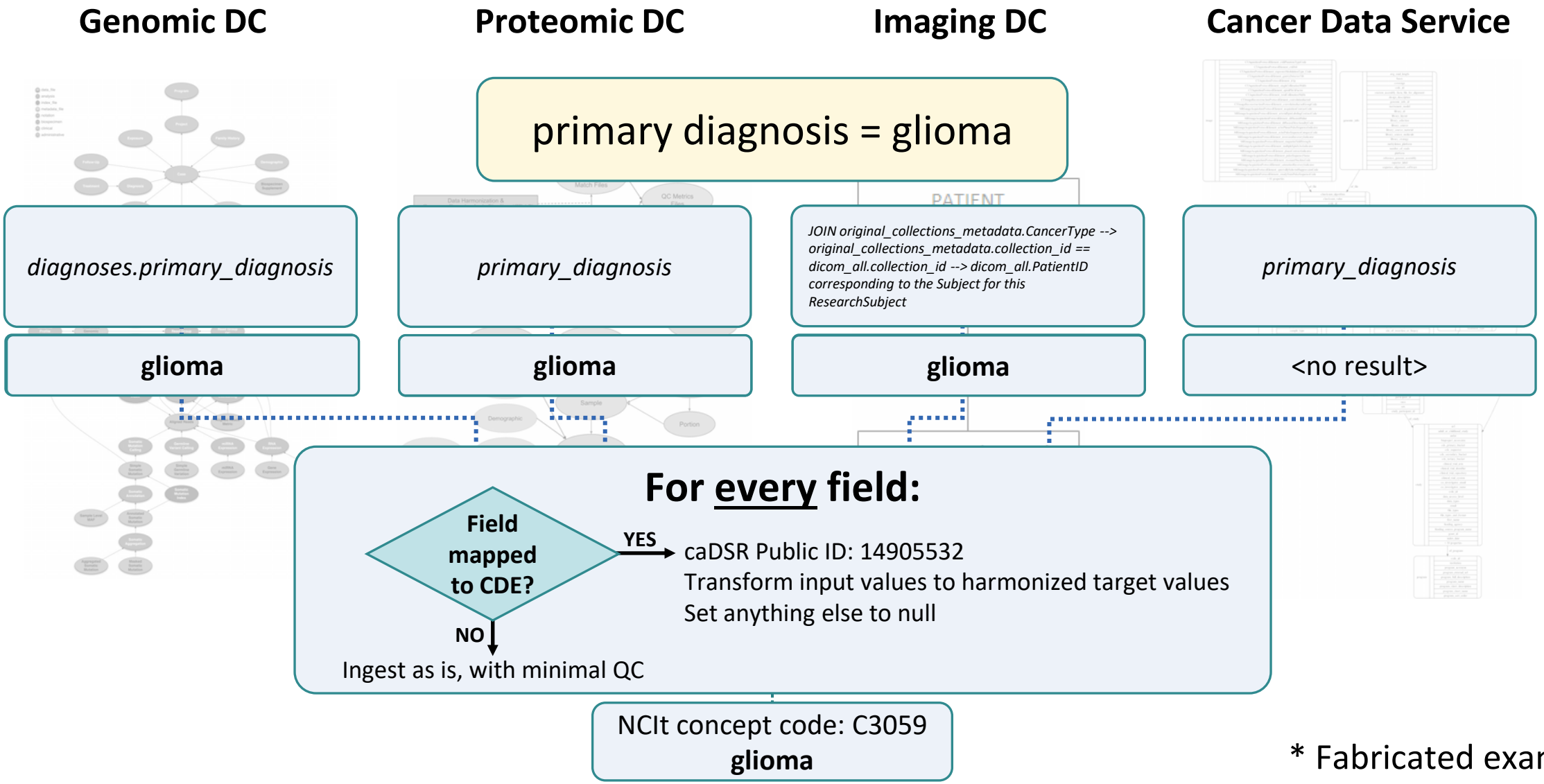
Genomic DC | **Proteomic DC** | **Imaging DC** | **Cancer Data Service**



Hey CDA, please give me any raw data files from all cases associated with a primary diagnosis of glioma

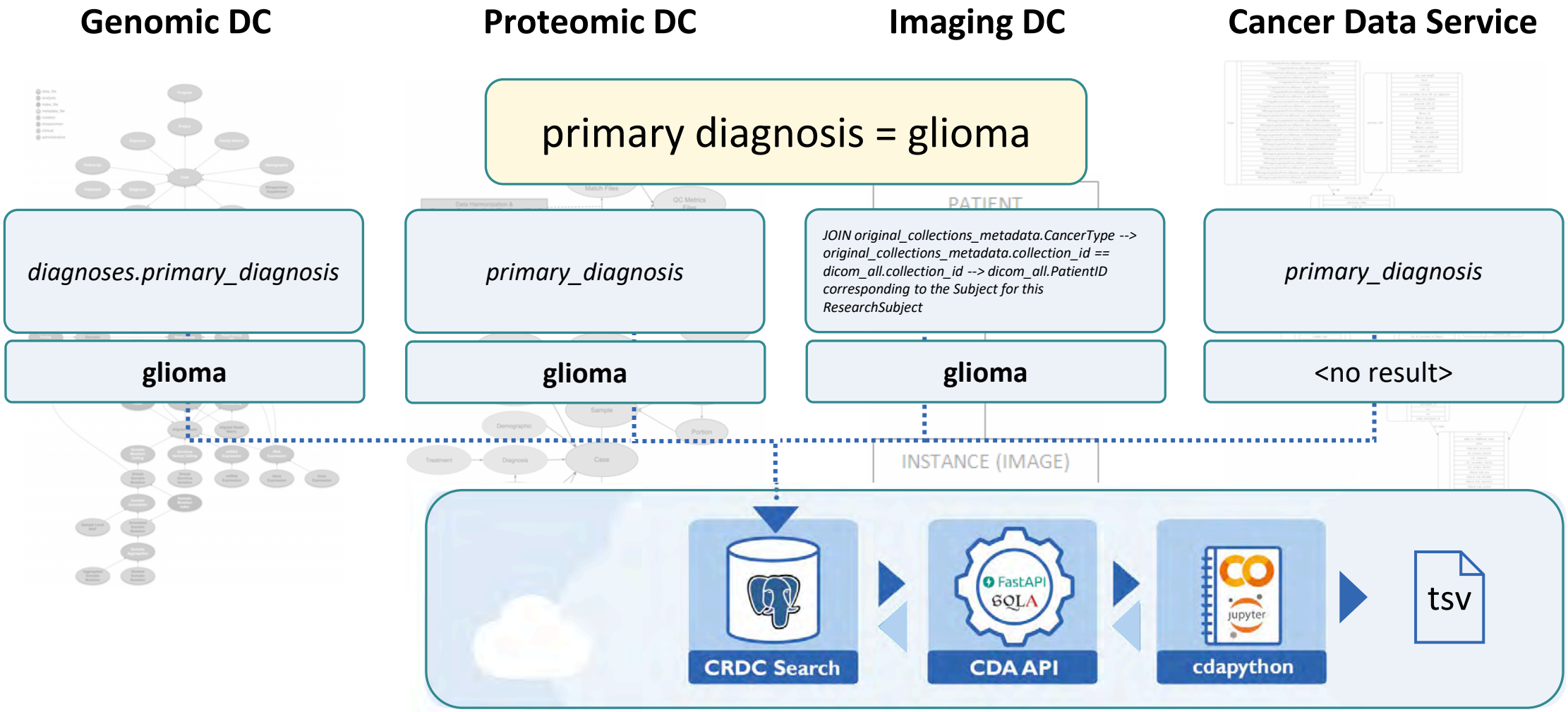
* Fabricated example

Aggregation & Harmonization of Data *across* CRDC DCs



* Fabricated example

Aggregation & Harmonization of Data *across* CRDC DCs



* Fabricated example

Accomplishments & Future Plans

- cdapython tool
- Aggregation
- No harmonization

CDA Public Release 4/24

- CDA in the Cloud
- ipython notebooks
- Integration with CRs
- Aggregation
- Limited harmonization

- Integration into CRDC Data Access Portal, point & click GUI
- Complete harmonization of legacy data
- Incorporate all current and future CRDC DCs

Where we started

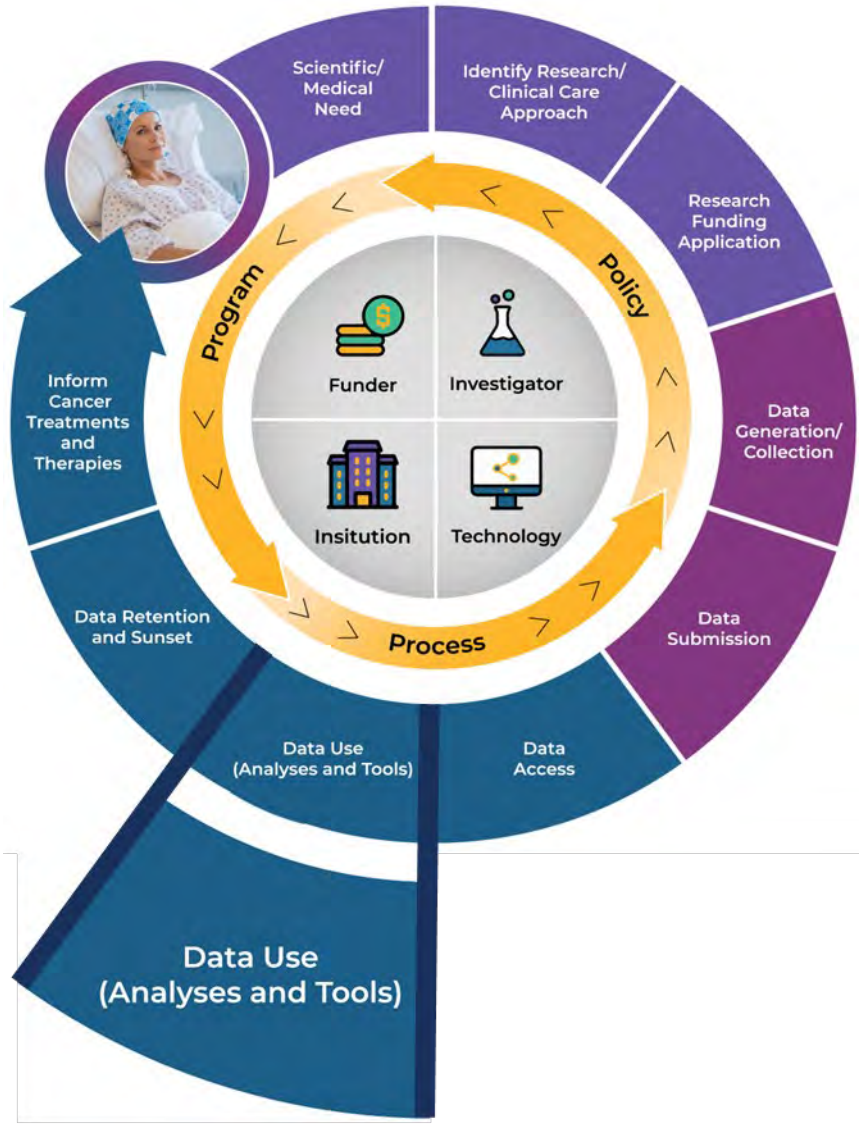
Where we are today

Where we arere going

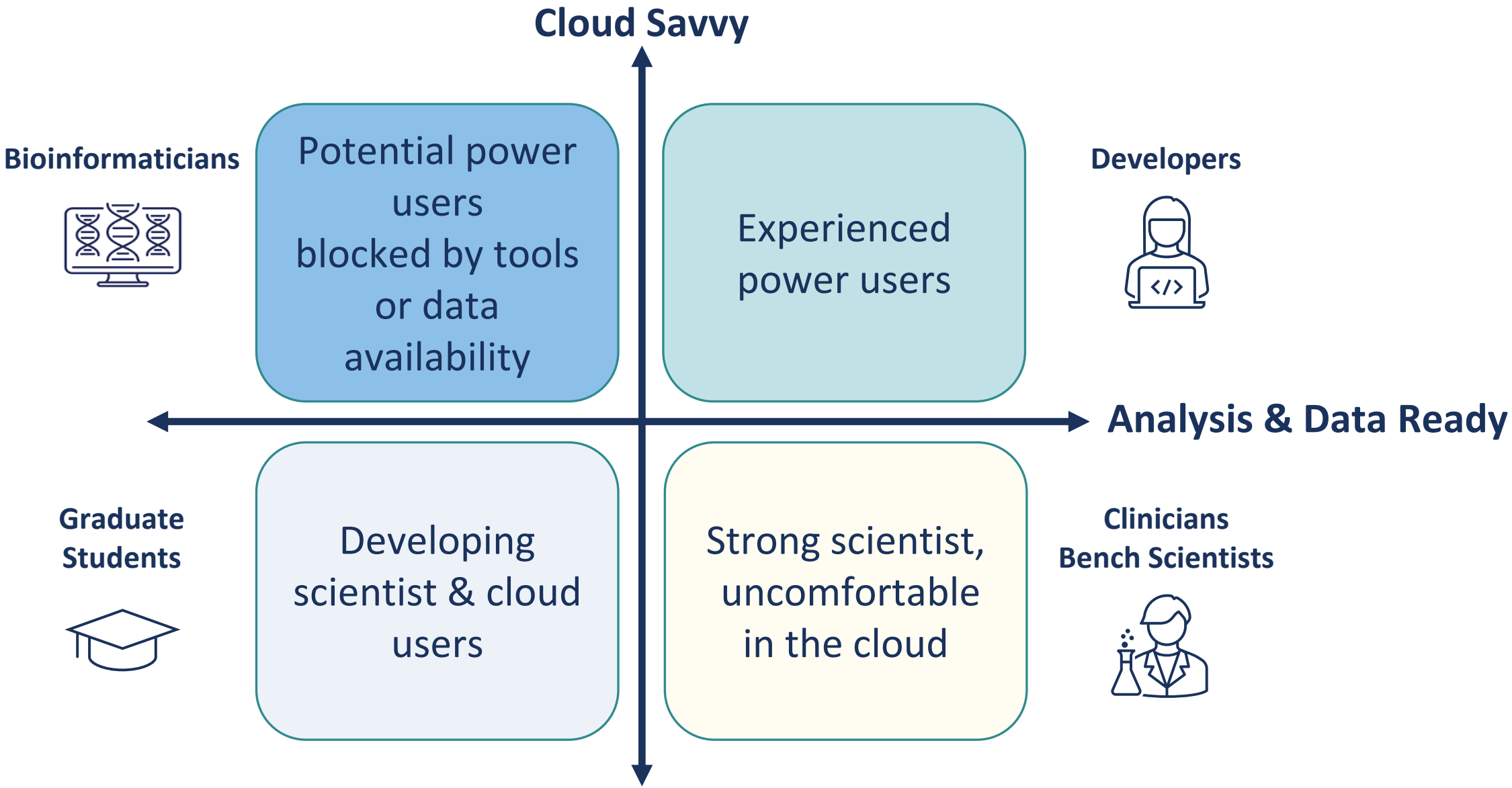
NCI Data Lifecycle: Data Use Analysis and Tools

The CRDC Data Ecosystem aims to provide:

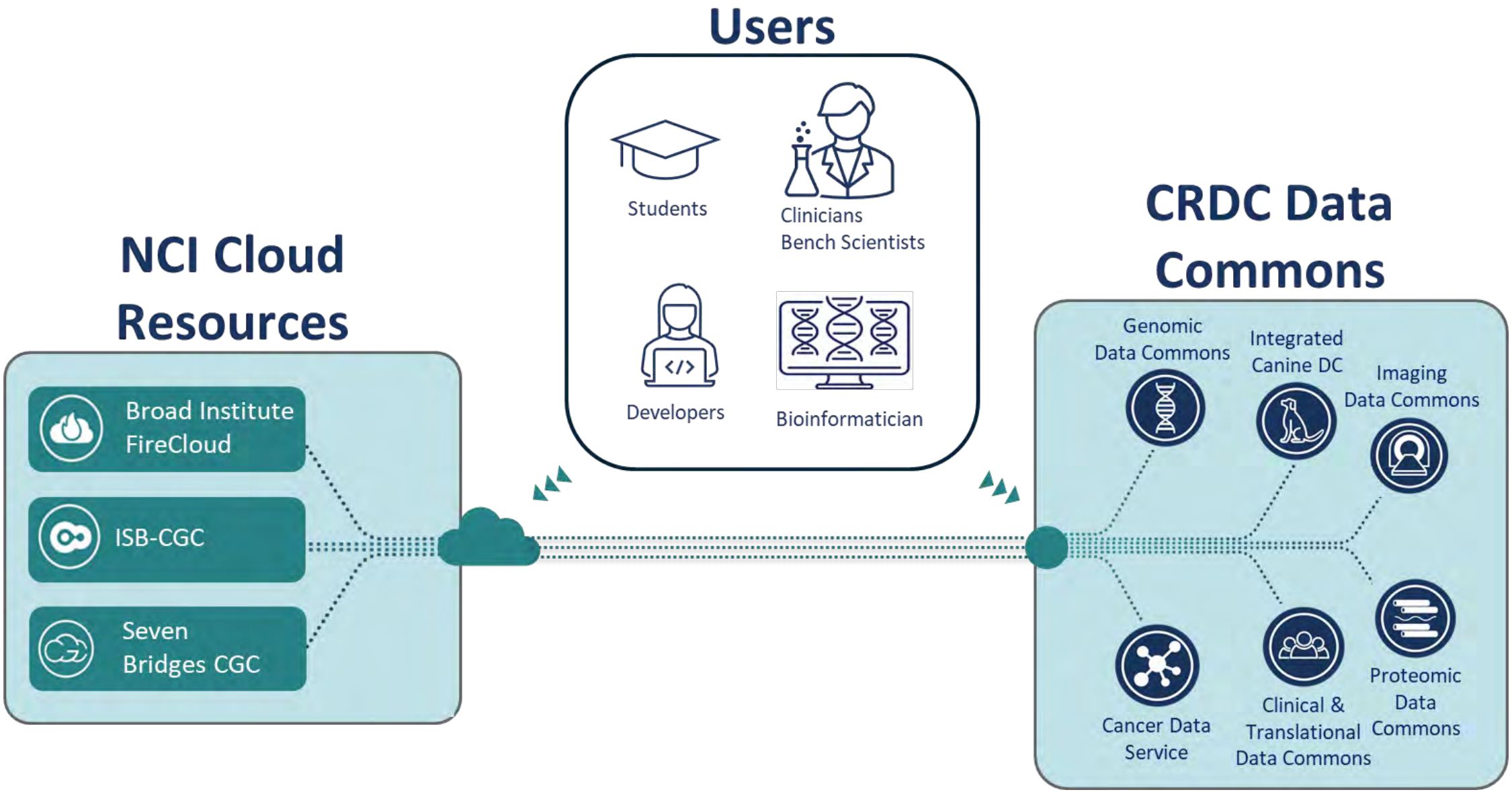
- Multiple workflows and access points for all types of users
- A flexible and customized analysis ecosystem



Diverse User Groups



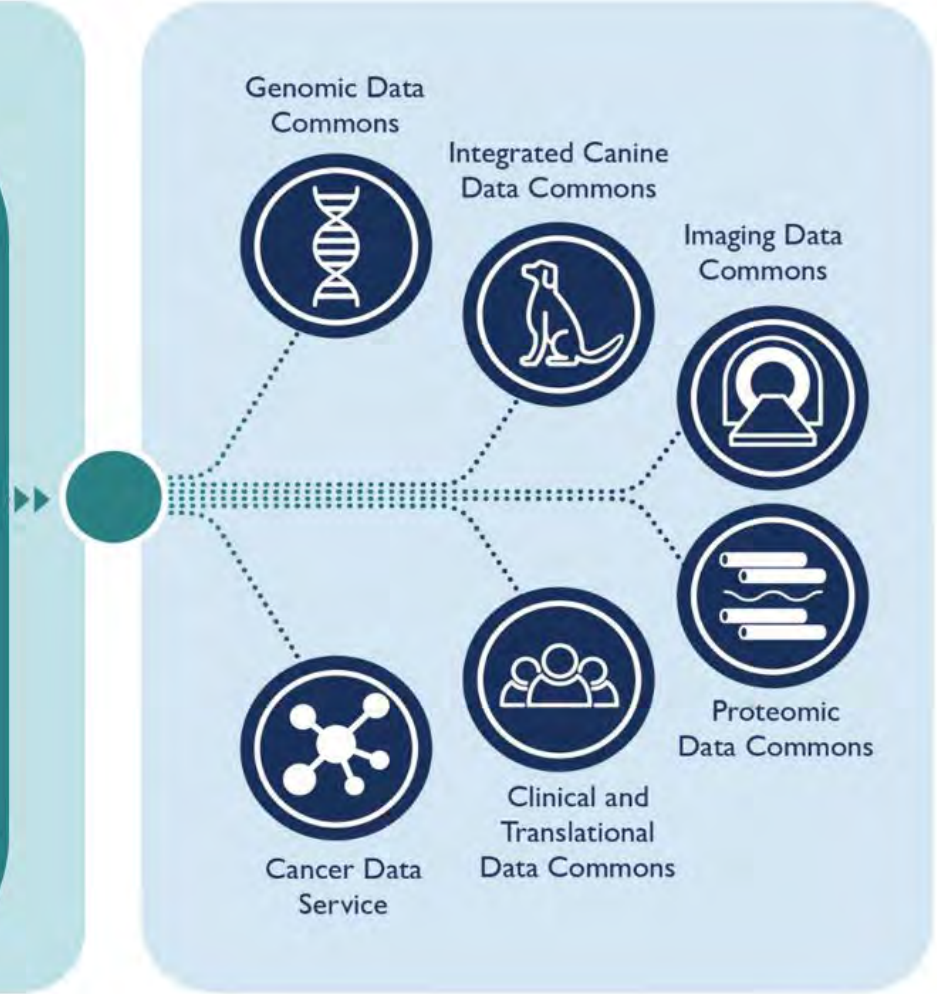
CRDC Data Access



CRDC Data Access: Data Commons Portals

Features

- Serves a specific research community
- Analysis tools specific to the data types stored
- Allows for more granular cohort building than CDA
- Instructions on how to transition from the portal to the CRDC Cloud Resources



CRDC Data Access: NCI Cloud Resources

Broad FireCloud, powered by Terra

- Based on Google Cloud Platform (GCP)
- Offers extensive repositories of pre-built tools
- Workflows in Workflow Definition Language (WDL)

ISB Cancer Gateway in the Cloud (ISB-CGC)

- Based on Google Cloud Platform (GCP)
- GCP native tools & BigQuery for big data analytics
- GCP Compute Engine for complex workflow execution
- Designed for users looking to use derived data

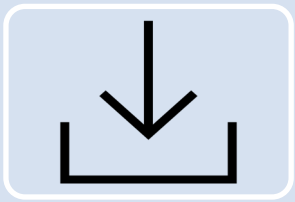
Seven Bridges Cancer Genomics Cloud (SB-CGC), powered by Velsera

- Based on Amazon Web Services (AWS)
- Offers a curated library of over 850 tools and workflows optimized for the cloud
- Workflows in Common Workflow Language (CWL)

Benefits of Cloud

- **Democratize access to data**
- Eliminate the need to download data
- Access to workspaces, analysis tools, workflows & pipelines
- Bring your own data and tools
- Collaborative pre-publication workspaces
- Integrate your data with other CRDC data and tools

Cloud Analysis Workflow



Search

Create cohort through a Data Commons

OR

The Cancer Data Aggregator

Download

Manifest of interested files from Data Commons or CDA

OR

Data Files Directly

(not available on all DC)

Import

Manifest into NCI Cloud Resource

OR

Data files into a cloud workspace

Difficulties of Being FAIR

Data Fragmentation

- Data scattered across platforms, databases, file formats
- Non-standardized metadata, inconsistent data organization

Interoperability Issues

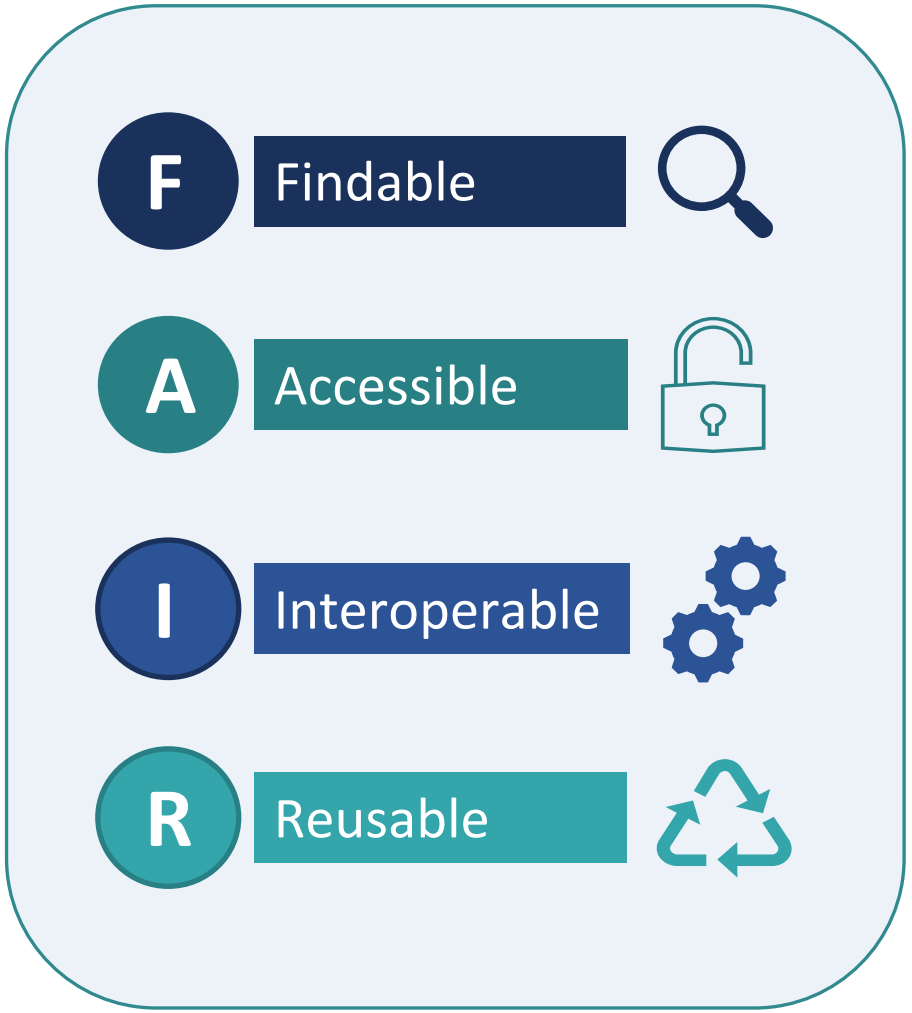
- Lack of standardized data models, ontologies, controlled vocabularies

Data Quality

- Incomplete metadata, inconsistent data formats

Infrastructure

- Lack of infrastructure, resources, technical knowledge



Standards and Policy

GA4GH: The Global Alliance for Genomics & Health

Build technical standards & policy frameworks/tools to expand responsible, voluntary, and secure use of genomic and other health data

Community Need	Solution	GA4GH Standard
Consistent access to data regardless of physical location	Persistent Digital IDs	Data Repository Service (DRS)
Know who a researcher is & what they are granted access to	AuthN and AuthZ	Passports & Visas

Data Commons Framework
University of Chicago's Center for Translational Data Science

GEN3 DATA COMMONS

CRDC's Data Commons Framework Services (DCFS) are an instance of Gen3 and provides a re-usable, expandable framework for the CRDC infrastructure through the implementation of modular components

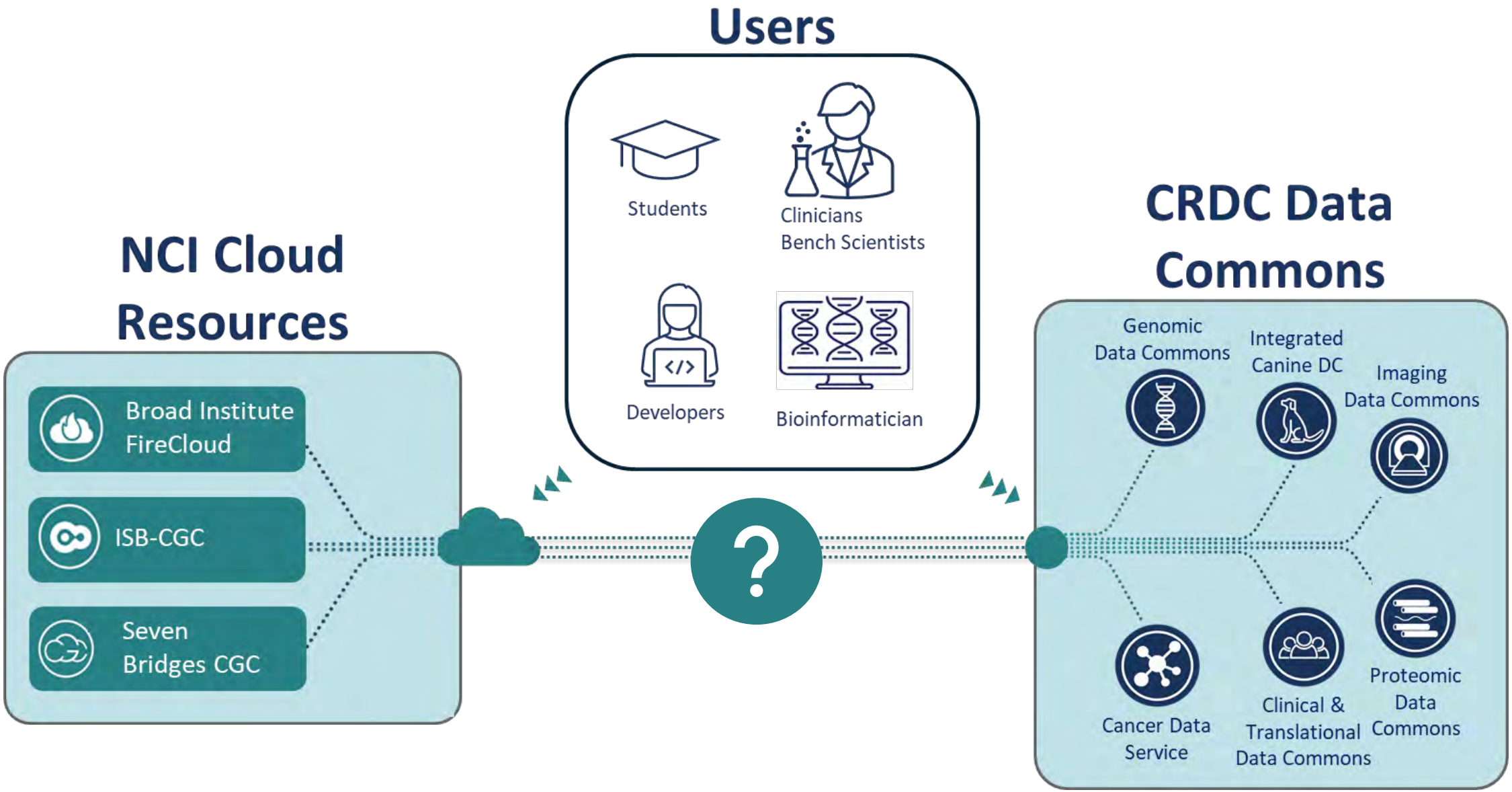
IndexD

Mints Data Repository Service (DRS) IDs for all CRDC datasets

Fence

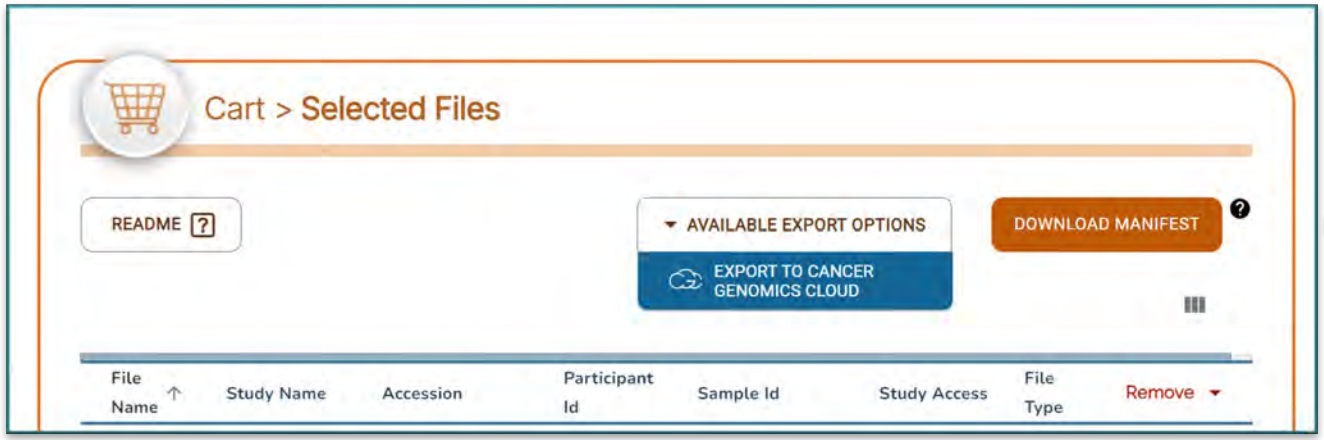
Authentication (AuthZ) and Authorization (AuthN) services

CRDC Data Access

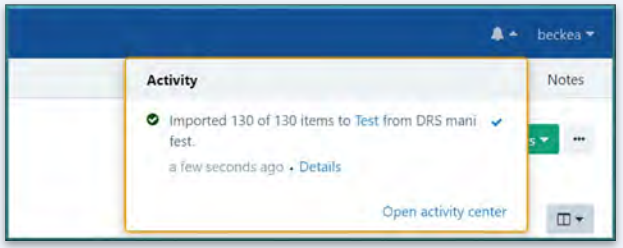


Interoperability: DRS Manifest

Data Commons Portal



Seven Bridges-CGC



Search Data Commons Portal → Add Files to Cart → Export to Cancer Genomics Cloud → Imports data files directly into cloud workspace

NIH Cloud Platforms for Interoperability (NCPI)

NIH's Researcher Auth Service (RAS)
Provides a single sign-on (SSO) experience for searching and accessing NIH's open and controlled data assets

AnVIL

NHGRI: Genomic Data Science Analysis, Visualization and Informatics Lap-space

BioData CATALYST

NHLBI: Tools, applications, and workflows in secure workspaces

NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

NCI: Accelerating data-driven scientific discovery through the Cancer Research Data Commons

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM

Common Fund: Alleviating suffering from childhood cancer and structural birth defects

dbGaP
GENOTYPES and PHENOTYPES

NCBI: Access to protected genomic, subject and sample data related to human studies

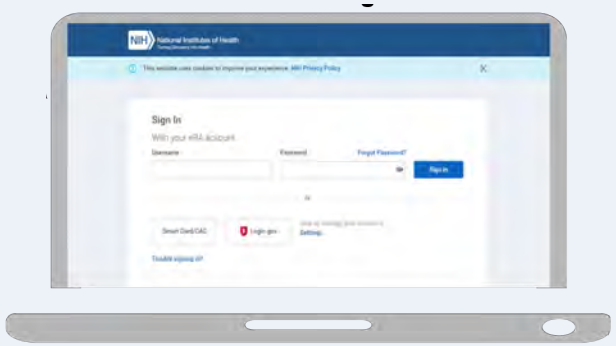
Mission: To create a partnership between multiple NIH-Supported systems by developing and implementing technical standards to enable interoperability and facilitate a federate data ecosystem

Working Together

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

- Office of Data Sharing (ODS)
- Cancer Research Data Commons (CRDC)

NCPI



Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

F Findable

A Accessible

I Interoperable

R Reusable

Acknowledgements

- The content of this presentation is based on the work of the following teams and organizations:
  - The CRDC Program
  - NCI Frederick National Lab Team
  - The Broad Institute
  - General Dynamics Information Technology, Inc
  - Institute of Systems Biology
  - University of Chicago's Center for Translational Data Science
  - Velsera
  - All Partners throughout NHI/NCI and data contributors

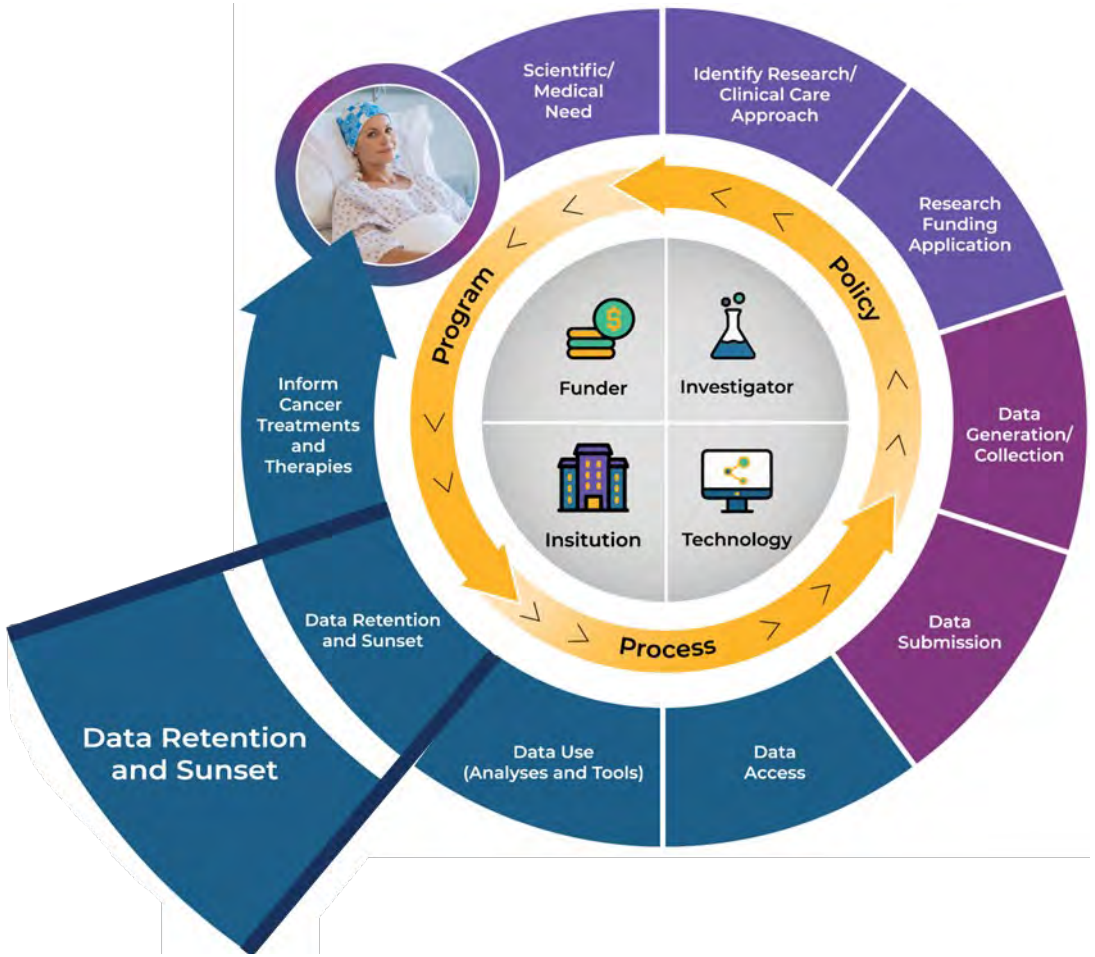
Data Retention and Sunset: Ensuring Data Accessibility and Integrity

Mike Warfe

October 16, 2024

NIH NATIONAL CANCER INSTITUTE

NCI Data Lifecycle: Data Retention and Sunset



- Cancer data is highly personal and crucial to preserve
- The CRDC contains study data that are impactful
- Data retention to ensure future access

CRDC Data Storage and Usage

- Our standards-based, scalable architecture empowers cancer researchers

354 STUDIES

57M OBJECTS

10PB+ DATA

82.3K UNIQUE USERS / YEAR

6 DATA COMMONS

CRDC is Based on Standards

FAIR

Infrastructure ensures FAIR principles are adopted across the CRDC.

GA4GH

CRDC participates and implements GA4GH standards such as DRS and Passports.

Open APIs

Open web standards for accessibility and a consistent user experience.

Cloud Native

Cloud services are utilized for scalability, flexibility, and cost-effectiveness.

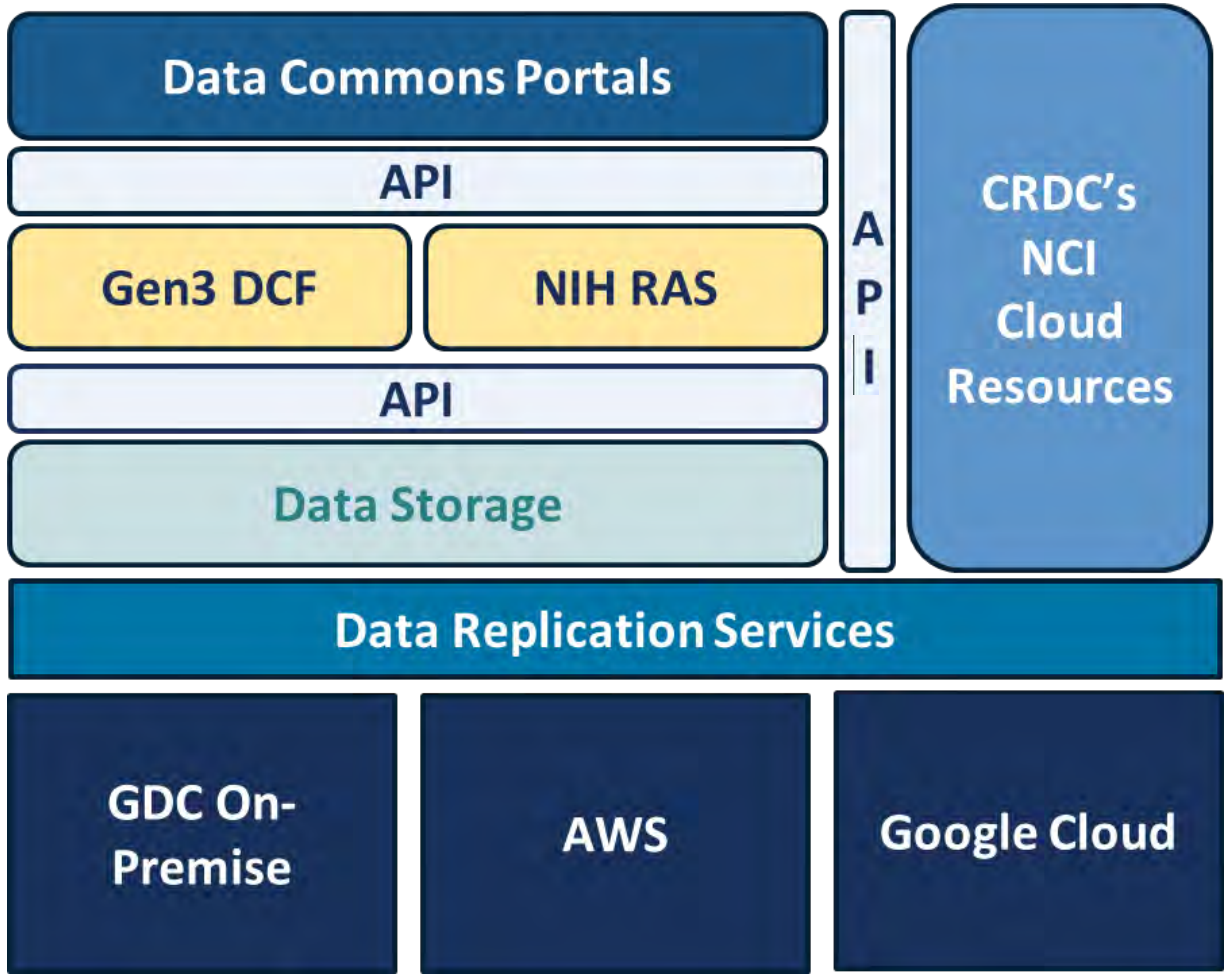
Data Models

Data models are published for interoperability across all Data Commons.

Governance

Data governance ensures data quality throughout its lifecycle.

CRDC Data Infrastructure



- Data Commons portals for cohort discovery
- Data accessed via API calls
- NIH approved access to Controlled Data
- Data is replicated to the cloud providers

Challenges of Operating CRDC at Scale

- Data Governance
- Data Management Infrastructure
- Data Infrastructure Economics

Data Governance and Sustainability

Key Activities

- Communicating change and publishing documentation
- Processing data for harmonization
- Data management tasks
- Reviewing and selecting appropriate technologies

Data Management Infrastructure

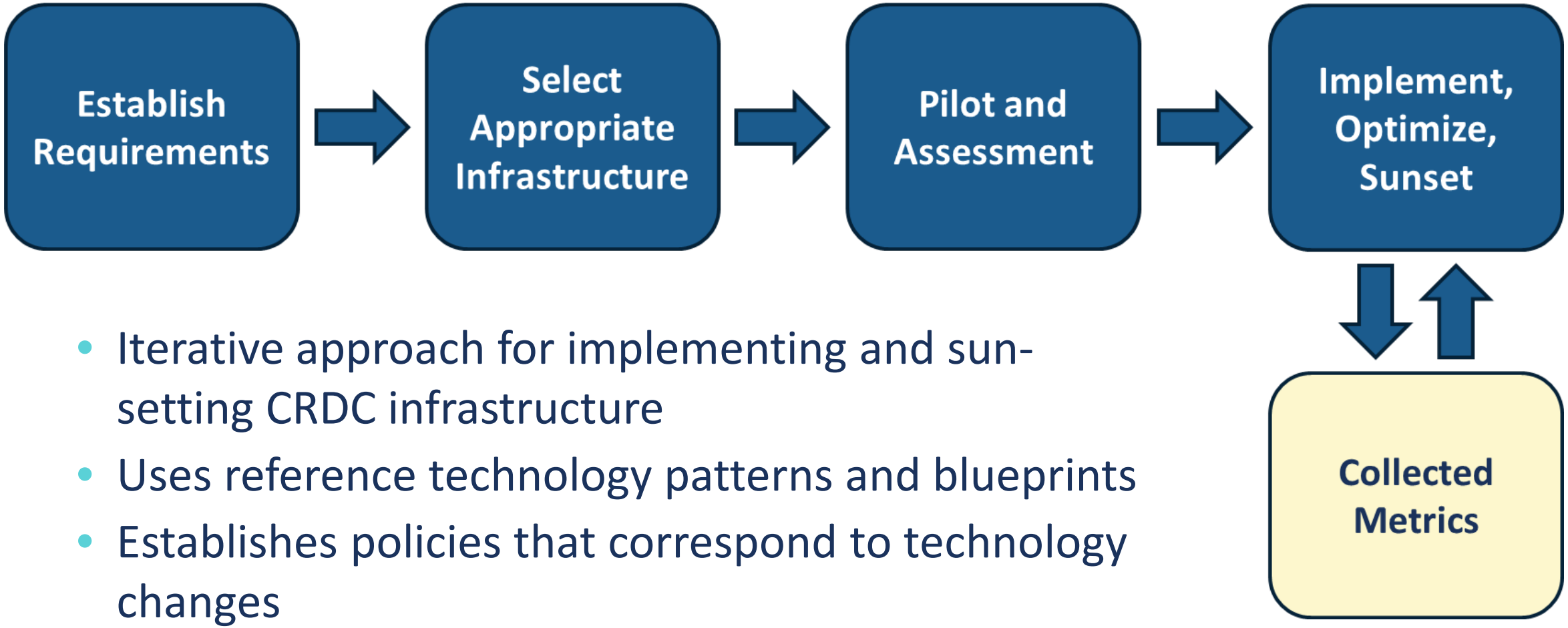
Multiple lifecycle phases: ingestion, harmonization, indexing, release, & eventual removal

Data is constantly being standardized with additional data types being added

Sustainability and Sunset: When is data removed and how - de-indexed, archived, or destroyed?

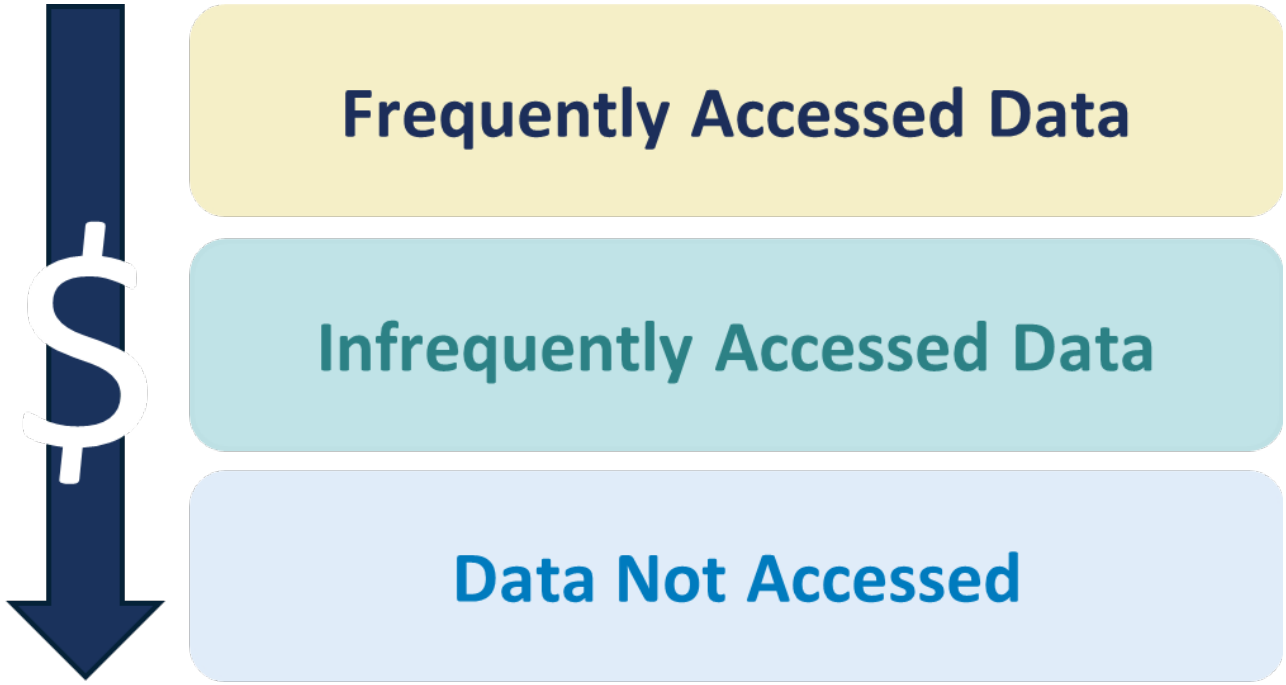
Continuous development of appropriate metrics on data access and operations costs

Infrastructure Lifecycle



- Iterative approach for implementing and sun-setting CRDC infrastructure
- Uses reference technology patterns and blueprints
- Establishes policies that correspond to technology changes

Cloud Storage Optimization Example



- CRDC Data needs to be accessible, performant, have integrity, and be cost effective
- Not all data is being accessed at the same time
- CRDC piloted AWS S3 Intelligent Tiering storage
- Implementation reduced storage costs by 60% without impacting access

Data Infrastructure Economics

- CRDC expenses for compute and storage are significant
- Data egress challenging due to networking speeds & cost
- **NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES)**
  - Initiative to reduce the cost of operating in the cloud
  - Provides training and “test” workspaces
  - All NIH awardees are eligible
  - All NCI Cloud Resources users benefit automatically
  - URL: https://cloud.nih.gov

Future Direction of CRDC Infrastructure

- Continued exponential growth anticipated
- Continued participation in standards communities
  - Data interoperability & technology development
- Implementing new services
  - Enable data discovery utilizing sustainability best practices
  - Compression transparent to the end user
- Formalizing policy, process, and technology for data sunset and other governance activities

Acknowledgements

- The content of this presentation is based on the work of the following teams:
  - The CRDC Program
  - Sustainability Implementation Plan Team
  - NCI GDC Team
  - NCI Frederick National Lab Team
  - NCI OCIO Cloud Team
  - NCI ODS Team