CRDC

NCI Cancer Research Data Commons

What is the **NCI CRDC**?

NCI's Cancer Research Data Commons (CRDC) is a cloud-based data science infrastructure that connects data sets with analytical tools, providing a foundation for the cancer research community to make new scientific discoveries and lower the burden of cancer. The CRDC is one major component of the broader National Cancer Data Ecosystem that supports the Cancer Moonshot℠ Blue Ribbon Panel recommendation to enhance data sharing. The CRDC includes the Genomic Data Commons (GDC), the Proteomic Data commons (PDC), the NCI Cloud Resources (CR), and the Data Commons Framework (DCF), and serves as a central location to support public data sharing for NCI-funded programs.

Currently, more than 23 data sets are available to the cancer research community (see page 3). Starting 2020, new data sets from Cancer Moonshot research projects, such as Human Tumor Atlas Network (HTAN) and Immuno-oncology Translational Network (IOTN), will be available through the CRDC for public data access.

What is a **Data Commons**?

Data Commons co-locates data, storage, and computing infrastructure in the cloud with tools for analyzing and sharing data to create an interoperable resource for the research community.

Goals of the **NCI CRDC**

- Enable the cancer research community to share diverse data types across programs and institutions
- Provide secure access to data
- Facilitate the generation of innovative tools
- Help NCI-funded Data Coordinating Centers sustain and share data publicly
- Build in an open and modular way to make components extendable and reusable
- Adhere to FAIR principles of data stewardship: Findable, Accessible, Interoperable, and Reusable

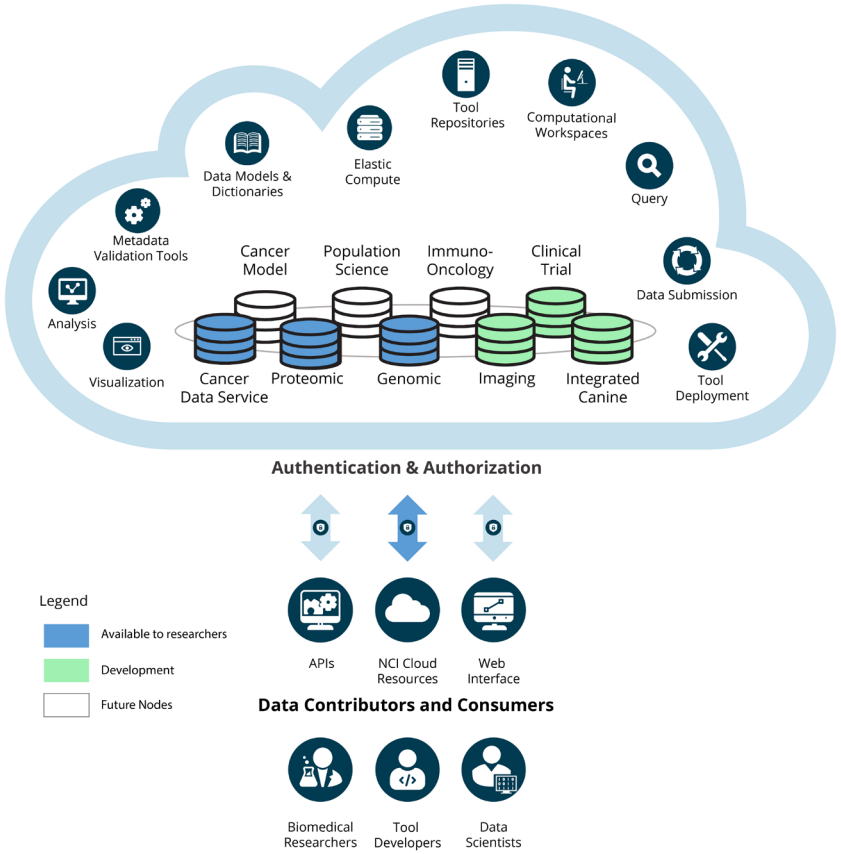
What Does the NCI CRDC Include?

Data Repositories

Genomic Data Commons (GDC)
Share, analyze, and visualize harmonized genomic data, including TCGA, TARGET, and CPTAC.

Proteomic Data Commons (PDC)
Share, analyze, and visualize proteomic data, such as CPTAC and International Cancer Proteogenome Consortium (ICPC).

Imaging Data Commons (IDC)
Share, analyze, and visualize multi-modal imaging data from both clinical and basic cancer research studies. The resource is expected to launch in mid 2020.



Cancer Data Service (CDS)
Share NCI-funded data that is currently not hosted by other repositories.

Integrated Canine Data Commons (ICDC)
Share data from canine clinical trials including PRE-medical Cancer Immunotherapy Network Canine Trials (PRECINCT) and Comparative Oncology Program. The resource is expected to launch in early 2020.

Clinical Trial Data Commons (CTDC)
Stores data from NCI Clinical Trials, such as Molecular Analysis for Therapy Choice (MATCH) Trial publicly available. The resource is expected to launch in early 2020.

Infrastructure

Data Commons Framework (DCF)
Provides secure user authentication and authorization and permanent digital object identifiers for data objects.

Center for Cancer Data Harmonization (CCDH)
Provides semantic services and tools that facilitate interoperability of the data across the NCI CRDC.

Cancer Data Aggregator (CDA)
Enables users to query and connect data distributed across NCI CRDC for integrative analysis. The CDA is expected to launch in 2020.

NCI Cloud Resources (CR)
Provides access to cancer data sets to perform large scale analysis using the elastic compute of commercial cloud platforms. The three resources include:

- **Seven Bridges CGC (SB-CGC)** (http://www.cancergenomicscloud.org/)
- **Institute for Systems Biology CGC (ISB-CGC)** Genomics Cloud (http://isb-cgc.org/)
- **Broad Institute FireCloud** (http://firecloud.terra.bio/#)

Data and Tools

What data sets are available in the NCI CRDC?

A variety of data sets are available in the NCI CRDC. Users can bring their own data to combine with the existing data to perform novel analyses through the NCI Cloud Resources.

Data sets	Description	Cancer Types	Cases
The Cancer Genome Atlas (TCGA)	TCGA is a landmark cancer genomics program that molecularly characterized more than 20,000 primary cancer and their matched normal samples spanning 33 cancer types. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga	33	11315
Therapeutically Applicable Research to Generate Effective Treatments (TARGET)	The TARGET program applies a comprehensive genomic approach to determine molecular changes that drive childhood cancers. The goal is to use data to guide the development of effective, less toxic therapies. https://ocg.cancer.gov/programs/target	5	5062
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	CPTAC integrates proteogenomic and imaging data from 11+ cancer types to identify proteomic-centric subtypes and improve understanding of cancer-relevant pathways through posttranslational modifications. https://proteomics.cancer.gov/programs/cptac	11 (as of 2019)	2208 (as of 2019)

What tools and resources are available in the NCI CRDC?

The NCI CRDC makes a wide range of popular and innovative analytic and visualization tools available for all stages of data analysis. Users can also bring custom tools to the NCI Cloud Resources.

The three NCI Cloud Resources provide search, compute, and analytical tools, with more than 1000 tools and workflows including RStudio and Jupyter notebooks.

		Access	Analysis Tools and Workflows							
	Resources	Portal & Workspace	Variant Calling	Mapping	Transcriptome	Epigenomic	Imaging	Proteomic	Multi-omic	Data Visualization
Data Repository	GDC	x	x	x						x
	PDC	x		x				x		x
	IDC	P					P			P
Cloud Resources	SB-CGC	x	x	x	x	x	x	x	x	x
	FireCloud	x	x	x	x	x	x	x	x	x
	ISB-CGC	x	x	x	x	x	x	x	x	x

Currently available features are represented by 'X' and features that are planned are indicated by 'P'.

Submitting Data to the NCI **CRDC** And Other **NIH** Repositories

The NCI CRDC provides repositories to store diverse data types. Currently, data submission requests are made by completing an application at each target repository's website or by contacting repository's helpdesk.
Table summarizing target repositories for respective data types as of 11/2019.

Data Domains	Subcategory	Target Repository*	Submit
Genomic	Non-harmonized genomic data	CDS	CDSHelpDesk@nih.gov
	Harmonized genomic data (e.g., tumor-normal pair)	GDC	https://gdc.cancer.gov/data-submission-request-form
Proteomic	Mass Spectrometry	PDC	https://pdc.cancer.gov/pdc/submit-data
	Other (e.g., Reverse-phase protein arrays)	CDS	CDSHelpDesk@nih.gov
Imaging	Radiology	TCIA** (IDC in 2020)	https://www.cancerimagingarchive.net/primary-data/
	Digital Pathology	TCIA** (IDC in 2020)	https://www.cancerimagingarchive.net/primary-data/
	Other (e.g., Cellular imaging)	CDS (IDC in 2021)	CDSHelpDesk@nih.gov
Clinical	Demographics, phenotypes, etc.	NCBI dbGaP CRDC as appropriate	https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/ ncicrdc@mail.nih.gov
Other	Flow cytometry, videos, etc.	Figshare *(up to 20 GB)*	https://nih.figshare.com/f/about

*CRDC is under active development and target repositories may change as new components become available. Check the NCI CRDC website for the latest information.
** All clinical imaging data will need to be de-identified through The Cancer Imaging Archive (TCIA) before storing on the IDC.

Data Sharing Policy Resources

NCI has policies and guidelines regarding sharing NCI-funded data including NCI's Genomic Data Sharing (GDS) Policy. NCI's Office of Data Sharing oversees and coordinates the sharing of NCI-funded data.

Learn more about

NCI GDS Policy Page: https://www.cancer.gov/grants-training/grants-management/nci-policies/genomic-data
GDS FAQs: https://osp.od.nih.gov/scientific-sharing/genomic-data-sharing-faqs/
Moonshot supplemental guideline: https://datascience.cancer.gov/data-sharing/policies