



Memorial Sloan Kettering  
Cancer Center

# Data Catalogs, Metadata, and Data Discovery:

---

How data catalogs extend and enhance the IR ecosystem.

Anthony Dellureficio,  
Associate Librarian for Data Management Services

Steering Committee Member, Data Discovery Collaboration

MIRL Symposium  
November 17, 2021

# Data Discovery Collaboration (DDC)

## Community of Practice

- Multi-institutional consortium,
- Institutional and individual members,
- Platform-agnostic,
- Support FAIR data principles in the broader context of library and institutional systems,
- Focus on the data catalogs and other metadata expertise,
- Support research data discoverability through:
  - metadata,
  - outreach,
  - software development.



Memorial Sloan Kettering  
Cancer Center™



DONALD AND BARBARA  
ZUCKER SCHOOL of MEDICINE  
AT HOFSTRA/NORTHWELL.

We welcome members without data catalogs!

# What is a Data Catalog?

MSK Data Catalog - <https://datacatalog.mskcc.org/>

Includes datasets, code, analytical tools, or other research outputs not traditionally included in bibliographic catalogs.

## Features:

- Metadata only - not repository,
- Curated, enhanced metadata,
- Wayfinder describing access and restrictions,
- Connect researchers,
- Standardized/normalized schema,
- Institution-specific - not aggregator,
- Identify relevant analytical tools,
- Track reuse, publications, and funding,
- Highlight otherwise undiscoverable datasets,
- Accommodate concerns over PHI exposure.

Flexibility to integrate with local infrastructure!

**Molecular portraits of tumor mutational and micro-environmental sculpting by immune checkpoint blockade therapy**  
UID: 10509  
Author(s): Riaz, Nadeem\*, Havel, Jonathan J\*, Makarov, Vladimir\*, Desrichard, Alexis\*, Chan, Timothy A\*  
\*MSK affiliated

**Description**  
Summary from the GEO: Immune checkpoint blockade (ICB) has demonstrated significant promise for the treatment of advanced malignancies. Anti-CTLA4 and anti-PD1 therapy can activate the immune system and result in durable control in diseases such as melanoma and non-small cell lung cancer. 109 RNASeq samples (58 On-treatment and 51 Pre-treatment) from 65 patients.

**Subject of Study**  
Human

**Subject(s)**  
Carcinoma, Non-Small-Cell Lung  
Immunotherapy  
Melanoma

**OncoTree Cancer Type(s)**  
Melanoma  
Non-Small Cell Lung Cancer

Non-Small Cell Lung Cancer  
Code: NSCLC  
Main Type: Non-Small Cell Lung Cancer  
Tissue: Lung  
Parent: LUNG  
NCI #: C2926  
UMLS #: C0007131  
More at OncoTree

**Access via GEO**  
Raw, FPKM, and RLD sequencing (CSV) and plain text cytolytic score data  
Accession #: GSE91061

**Access via SRA (NCBI)**  
RNA Sequence reads for 109 samples.  
Accession #: SRP094781

**Access via BioProject**  
Additional information about overall initiative.  
Accession #: PRJNA356761

**Access Restrictions**  
Free to All

**Access Instructions**  
The NCBI Gene Expression Omnibus, SRA, and BioProject databases provide open access to these files.

**Associated Publications**  
Campeato LF, Budhu S, Tchalicha J, et al. Blockade of the AHR restricts a Treg-macrophage suppressive axis induced by L-Kynurenine. Nat Commun. 2020;11(1):4011. Published 2020 Aug 11. doi:10.1038/s41467-020-17750-z  
Riaz N, Havel JJ, Makarov V, Desrichard A, Urba WJ, Sims JS, Hodi FS, Martin-Algarra S, Mandal R, Sharfman WH, Bhatia S, Hwu WJ, Gajewski TF, Slingluff CL Jr, Chowell D, Kendall SM, Chang H, Shah R, Kuo F, Morris LGT, Sidhom JW, Schneck JP, Horak CE, Weinhold N, Chan TA. Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. Cell. 2017 Nov 2;171(4):934-949.e16.

**Data Type**  
Genomic

**Equipment Used**  
Illumina Genome Analyzer

**Dataset Format(s)**  
CSV, Plain Text, SRA, gzip

**Dataset Size**  
15.6 MB (FPKM), 3.4 MB (raw), 18.4 MB (RLD), 1.8 KB (TXT), SRA broken up into files of 2-3GB

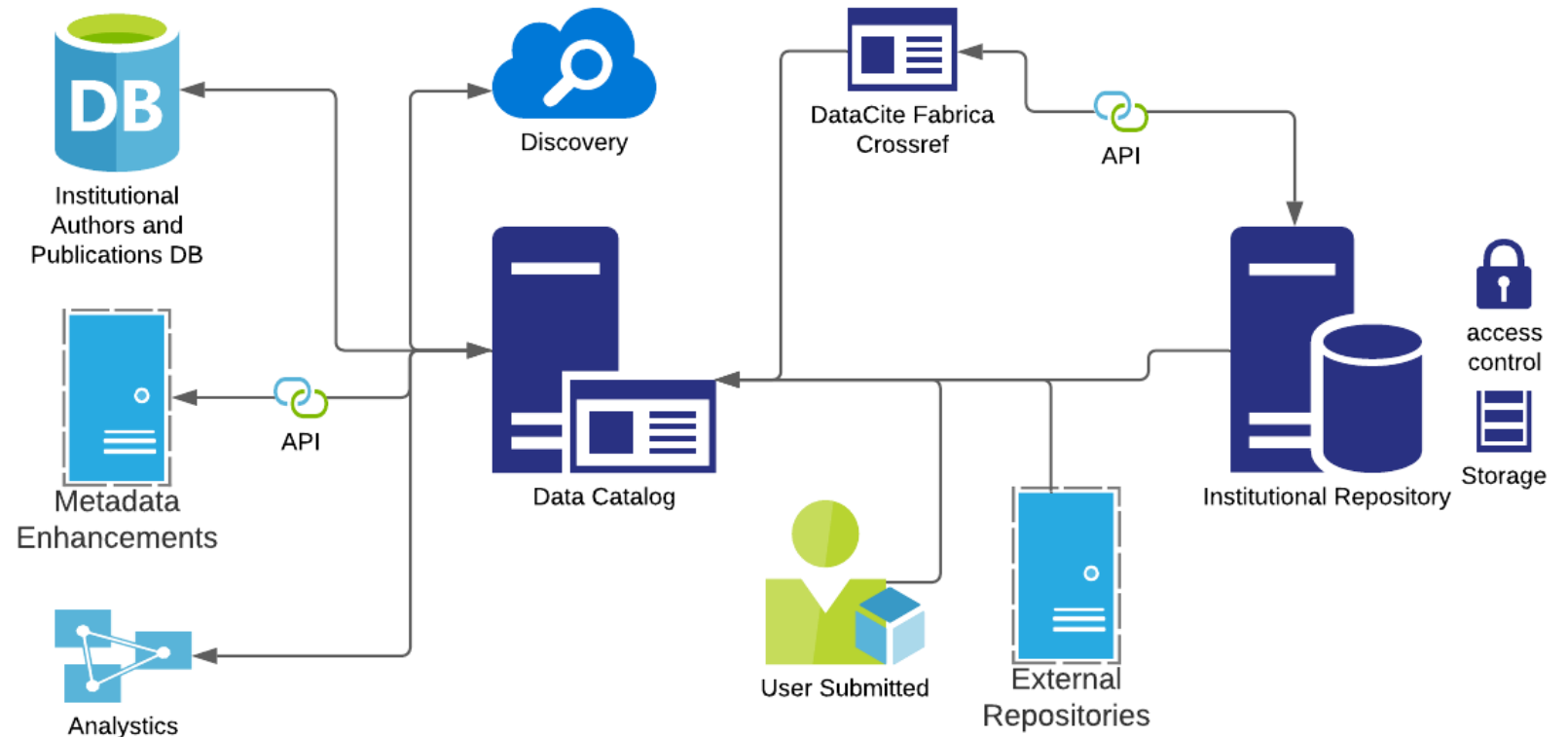
# Data Catalogs and Institutional Repositories

Reinforce FAIR data principles

Metadata ecosystem includes:

- Data catalog,
- Institutional Repository,
- External repos,
- Institutional author/pub db,
- PID minting applications,
- Discovery platform,
- External taxonomy/ authority dbs
- Other library resources.

Data Catalog - Institutional Repository Diagram



# Thank you!

## Resources and contact info

### Data Discovery Collaboration (DDC) Steering Committee:

- **Nicole Contaxis** (NYU HSL)..... [Nicole.Contaxis@nyulangone.org](mailto:Nicole.Contaxis@nyulangone.org)
- **Jason Clark** (Montana State University)..... [jaclark@montana.edu](mailto:jaclark@montana.edu)
- **Anthony Dellureficio** (MSK)..... [dellurea@mskcc.org](mailto:dellurea@mskcc.org)
- **Melissa Ratajeski** (University of Pittsburgh)..... [mar@pitt.edu](mailto:mar@pitt.edu)
- **Terrie Wheeler** (Weill Cornell)..... [tew2004@med.cornell.edu](mailto:tew2004@med.cornell.edu)

DDC Website: <https://datadiscoverycollaboration.org/>

MSK Data Catalog: <https://datacatalog.mskcc.org/>

Catalog code: <https://github.com/msk-library/data-catalog>

### Recent DDC publication on curation through data catalogs:

Sheridan H, Dellureficio AJ, Ratajeski MA, Mannheimer S, Wheeler TR. Data Curation through Catalogs: A Repository-Independent Model for Data Discovery. *Journal of eScience Librarianship* 2021;10(3): e1203. <https://doi.org/10.7191/jeslib.2021.1203>.