



Supporting Data Management within a secure enclave:

the story of the Weill Cornell Medicine Data Core



Peter Oxley, PhD
Associate Director for Research Services
Samuel J. Wood Library



Maximizing data value and protecting patients' data

FAIR

Findable

Accessible

Interoperable

Reproducible

Secure

Confidential

Consented

Authorized

Traceable

Controlled-
Disclosure

Data Core is your data and applications, in a Windows environment, *secured* in the WCM cloud



The Data Core interface (left) is a familiar Windows Desktop environment

One of few secure enclave environments described in the literature

Design and Implementation of a Secure Computing Environment for Analysis of Sensitive Data at an Academic Medical Center.

Oxley PR, Ruffing J, Campion TR Jr, Wheeler TR, Cole CL.

AMIA Annu Symp Proc. 2018 Dec 5; 2018:857-866.

Data Core's approach is to provide curated, online analytics



Secured

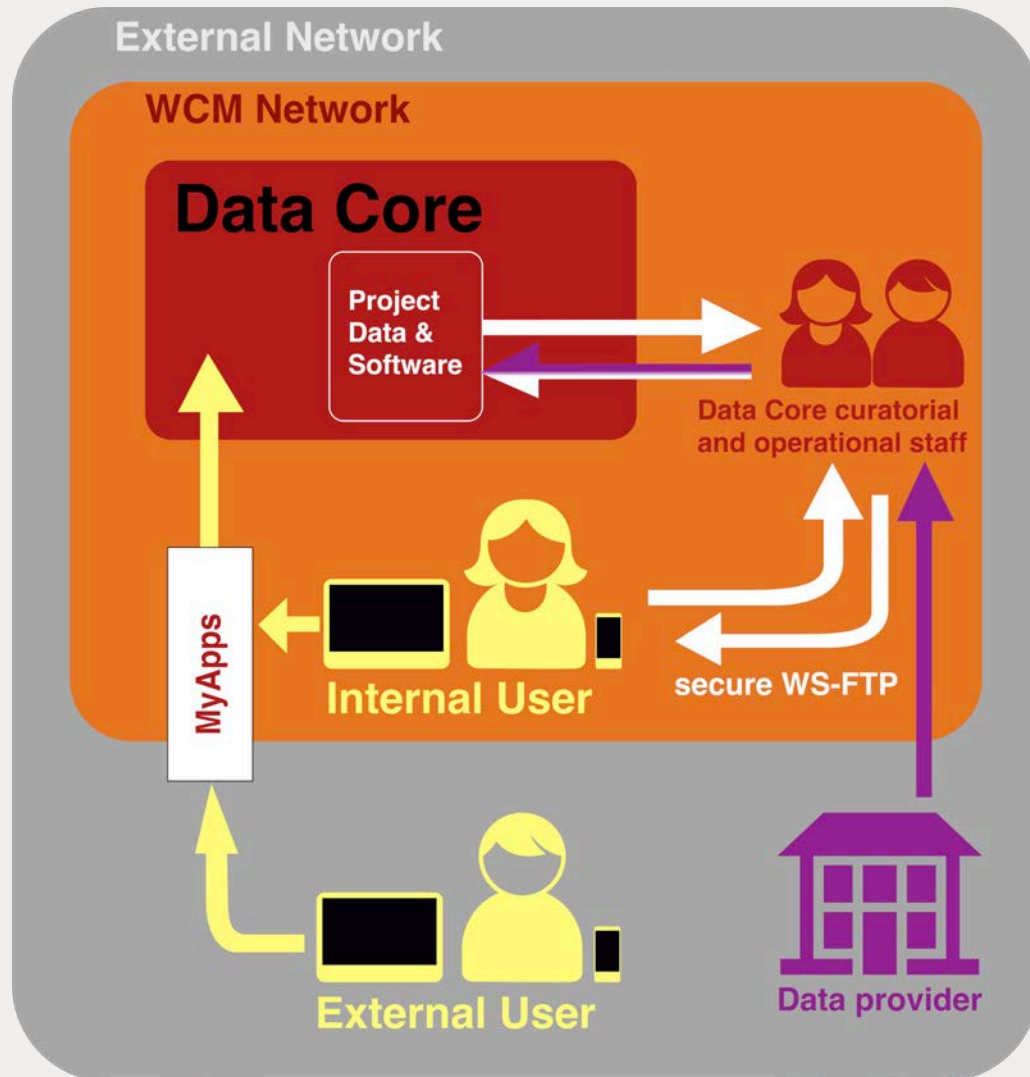


Collaborative



Flexible

Data Core security meets the requirements of major 3rd-party providers



Consistent interface and access allows collaborative workflow for researchers

Within each project, all users see the same:

Applications



Open source
(Including CRAN mirroring)



Commercial
(Group discounts)



Database options

Data



Original
(protected read only)

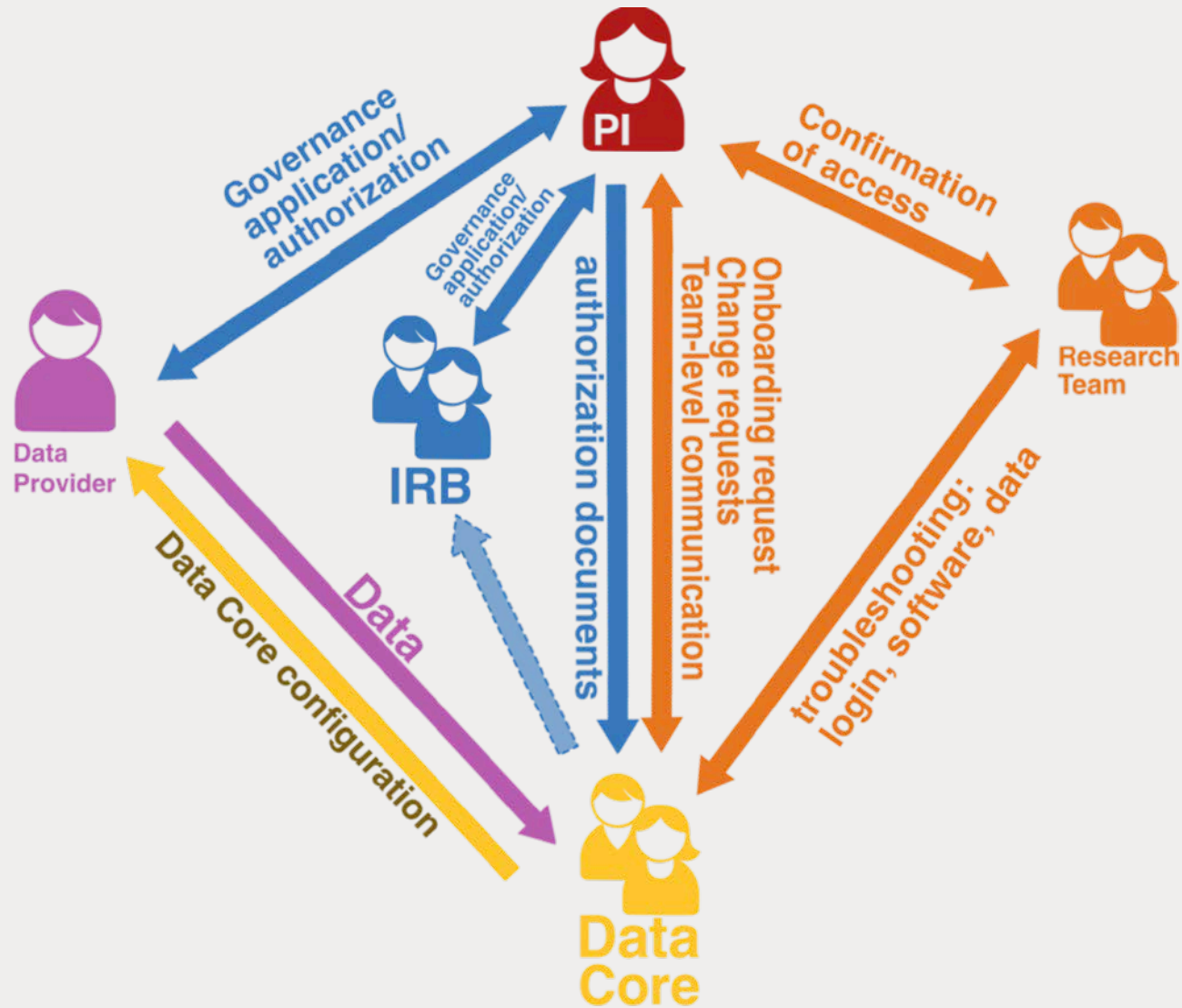


Shared space



Private work area

Data Core team assists in collaboration with external parties, and curating governance



Data Core is designed to be flexible

Interface

- Can connect from anywhere
- Can connect from Mac, Windows, or even Unix-flavor endpoints

Users

- WCM or external
- Central credentials (WCM ID and password)

Operation

- **Availability**
single 6h monthly maintenance window
- **Scalability**
Baseline of 4 CPU / 16 GB RAM
storage and computation can grow as needed

Uses

- Faculty research
- Student projects/theses
- Classes

Data Core has served many people and projects over the last 4 years

Consistent usage

	PIs	All users	Projects
Current	22	108	46
Total	34	317	88

Notable data sets/studies

CDRN (2 study-specific datamarts)

Pediatric Epilepsy LHS (multiple data providers)

SPARCS

Medicare

WCM IRB and NYP prefer Data Core for ePHI



We are actively developing a Data Catalog

<https://github.com/oxpeter/datacatalog>

Captures three primary components of PHI:

Dataset metadata

Governance

Access Requirements

The key feature of our data catalog is its ability to capture governance requirements

Descriptive metadata
Connect to datasets
Scope of authorization
Users authorization
Data Controls
Reuse scope

The screenshot displays a data catalog entry for the Healthcare Cost and Utilization Project (HCUP). The interface includes a navigation bar with 'Data Catalog (Beta)', 'View Details', and a search bar. The main content area shows the dataset title, OMB Control number, agency name, and a detailed description. A table lists key attributes such as start and end dates, the Principal Investigator (PI), and data management policies. Two highlighted boxes provide specific governance requirements for data storage and access.

Attribute	Value
Start date	Nov. 1, 2016
End date	June 30, 2021
PI	Michael Bales (meb7002)
Mixing of different datasets allowed?	False
Data destruction required?	True

Data Storage Requirements

Data must be stored within the WCM Data Core
Data must be encrypted at rest

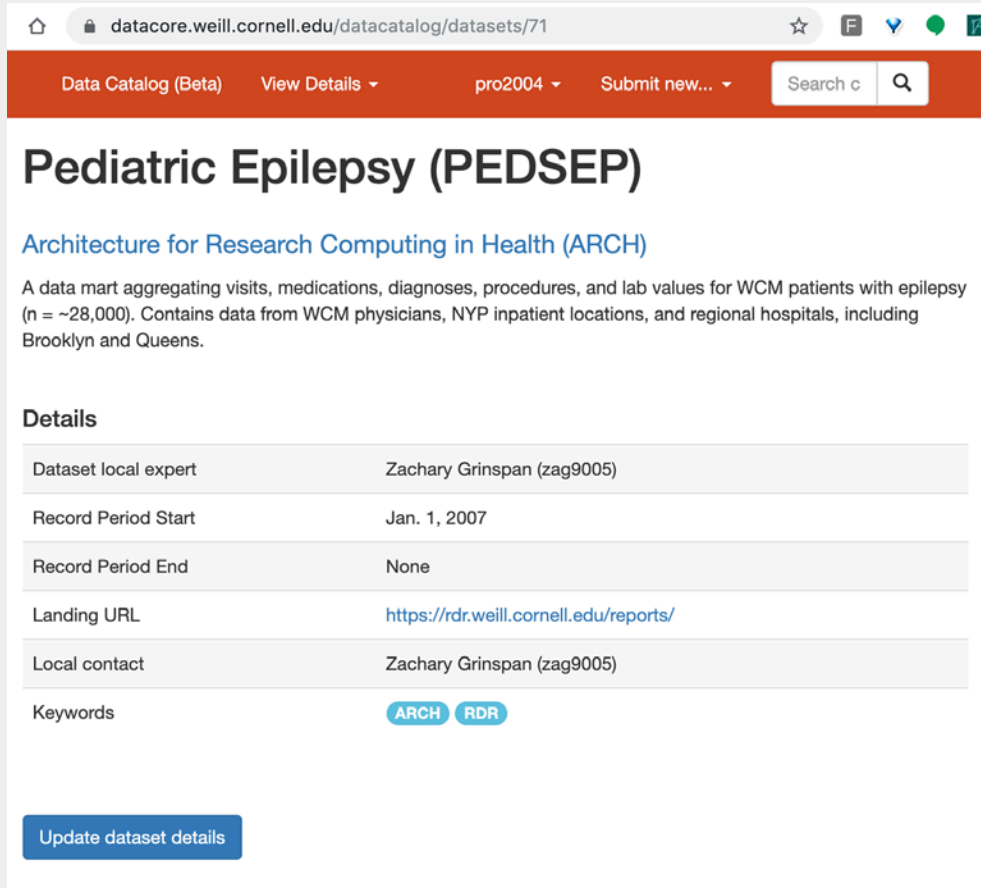
Data Access Conditions

two-factor authentication

[Update DUA details](#)

Datasets for OMB Control # 0915-0276: Healthcare Cost and

Current content includes Data Core datasets and ARCH Research Data Repositories



The screenshot shows a web browser window with the URL datacore.weill.cornell.edu/datacatalog/datasets/71. The page title is "Pediatric Epilepsy (PEDSEP)". Below the title, it is identified as an "Architecture for Research Computing in Health (ARCH)" dataset. The description states: "A data mart aggregating visits, medications, diagnoses, procedures, and lab values for WCM patients with epilepsy (n = ~28,000). Contains data from WCM physicians, NYP inpatient locations, and regional hospitals, including Brooklyn and Queens." The "Details" section includes the following information:

Dataset local expert	Zachary Grinspan (zag9005)
Record Period Start	Jan. 1, 2007
Record Period End	None
Landing URL	https://rdr.weill.cornell.edu/reports/
Local contact	Zachary Grinspan (zag9005)
Keywords	ARCH RDR

At the bottom of the details section, there is a blue button labeled "Update dataset details".

Primary records:

ARCH RDRs including
NYP shared data

Data Core-hosted data
from AHRQ, CMS, Aetna

Questions

Peter Oxley

pro2004@med.cornell.edu

Acknowledgements:

Curt Cole

Terrie Wheeler

Architecture, Design and Compliance

John Ruffing

Data Core curators:

Frank Ashmun

Michael Bales

Alice Chin

Heather Kleinschmidt

Technical Staff

Jo Hargitai

Danny Tan

Lucy Walle

Design and Implementation of a Secure Computing Environment for Analysis of Sensitive Data at an Academic Medical Center.

Oxley PR, Ruffing J, Campion TR Jr, Wheeler TR, Cole CL. AMIA Annu Symp Proc. 2018 Dec 5; 2018:857-866.

