Research Data Lifecycle

Objective 1: Build a DM program
Objective 2: Engage & FAIR principles
Objective 3: Storage solutions

Share & publish; Discover, reuse and cite: Data Catalog
Manage, store and preserve: Data Retention tool
Collaborate and analyze: Data Core & Software Hub

WCM Data Management Services

- Librarians Curate Data
- Data Core
- Scientific Software Management
- Data Catalog
- Data Retention – started July 2022
- Data Management & Sharing Plan
- Data Repository
- Data Integrity

Objective 1: Build a DM program
Objective 2: Engage & FAIR principles
Objective 3: Storage solutions
Some history

- Institutional demand for a secure enclave
- Sensitive big datasets
- PHS
- Compute and storage
- One to multiple users
- Collaborations
- Department subsidy
Secure

- Secured
- Collaborative
- Flexible

<table>
<thead>
<tr>
<th>Volume</th>
<th>PIs</th>
<th>Users</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>62</td>
<td>312</td>
<td>83</td>
</tr>
<tr>
<td>Total</td>
<td>82</td>
<td>679</td>
<td>181</td>
</tr>
</tbody>
</table>

WCM IRB and NYP prefer Data Core for ePHI

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.
The New York INSIGHT Clinical Research Network (CRN) COVID study
NLP project (Llama 2) automated clinical note summarization
Pediatric Epilepsy LHS project
Data Core environment
Data Core actors

- Infrastructure/Storage team (2)
- Security team (2)
- Operations team (2)
- AWS Cloud team (2)
- Application hosting (1)
- ITS service architect (2)
- Scientific computing team (4)
- Librarians (4)
Data Core team

Data Core services

Terrie Wheeler
Director

Sarah Ben Maamar
Data Core Manager

John Ruffing
Architecture, design & compliance

Curation

Data custodial & curation services
governance oversight & certified deidentification (1 certified member)

Technical

Scientific computing
Infrastructure/Storage
Application hosting
Operations
Security
AWS Cloud

Patrick Chen
Erick Lojano

Alice Chin
Data Core operations

**Inputs**
- Authorization documents (IRBs, DUAs)
- Timeline, project requirements
- Change requests, data
- Records of project status
- Closure request

**Outputs**
- Operational Environment
- Communications/actions to maintain authorized user access and environment requirements
- Record of authorization checks
- Removal of user access, deletion of environment, data disposition

**Life cycle processes**
- Onboarding
- Operations
- Review
- Offboarding

**Project elements**
- Data
- Users
Features and options

• Integrated with Service Management System (ServiceNow)
• Data Core admin web application, “MARIGOLD”
• Cost recovery:
  – # users, Software, Compute
• Grant applications
• Basic licensed statistical software
• Option to add:
  – More memory
  – More CPU
  – GPU
  – Databases (postgresql)
  – Storage
  – Additional licensed software
• Connected to Data Catalog, Retention tool (deep archive)
Data Core charge back model

Monthly cost for a project is typically:

- $0 - 85
- $2.50 - 25/100GB
- $250 (32CPU, 128GB RAM)
- $250/user

**TOTAL:**
$502.5 - 610
Status

• Center for Medicaid Services (CMS) certification:
  – On-premise & AWS Cloud environments

• Cloud initiatives:
  – AWS (Windows, Linux, GPU)
  – Azure – testing phase

• Packages repositories:
  – CRAN (R), active
  – Python, in progress
Challenges

- Moving target
- Training
- Engagement
- Cost

Share & publish; Discover, reuse and cite: Data Catalog

And successes

- Collaboration
- Library/ITS
- Highly motivated staff
- Leadership
- Help from researchers

Manage, store and preserve: Data Retention tool

Objective 1: Build a DM program
Objective 2: Engage & FAIR principles
Objective 3: Storage solutions

Collaborate and analyze: Data Core & Software Hub

Thank you!

Samuel J. Wood Library

Sarah Ben Maamar,
Associate Director for Research Services
sbm4003@med.cornell.edu

Terrie R. Wheeler,
Library Director
tew2004@med.cornell.edu