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
- 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.
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:

<table>
<thead>
<tr>
<th>Applications</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open source (Including CRAN mirroring)</td>
<td>Original</td>
</tr>
<tr>
<td>Commercial (Group discounts)</td>
<td>(protected read only)</td>
</tr>
<tr>
<td>Database options</td>
<td>Shared space</td>
</tr>
<tr>
<td></td>
<td>Private work area</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>PIs</th>
<th>All users</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current</td>
<td>22</td>
<td>108</td>
<td>46</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>317</td>
<td>88</td>
</tr>
</tbody>
</table>

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
Current content includes Data Core datasets and ARCH Research Data Repositories

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.