Weill Cornell Medicine | NewYork-Presbyterian

High Fidelity:

Connecting information for Better Research Reproducibility

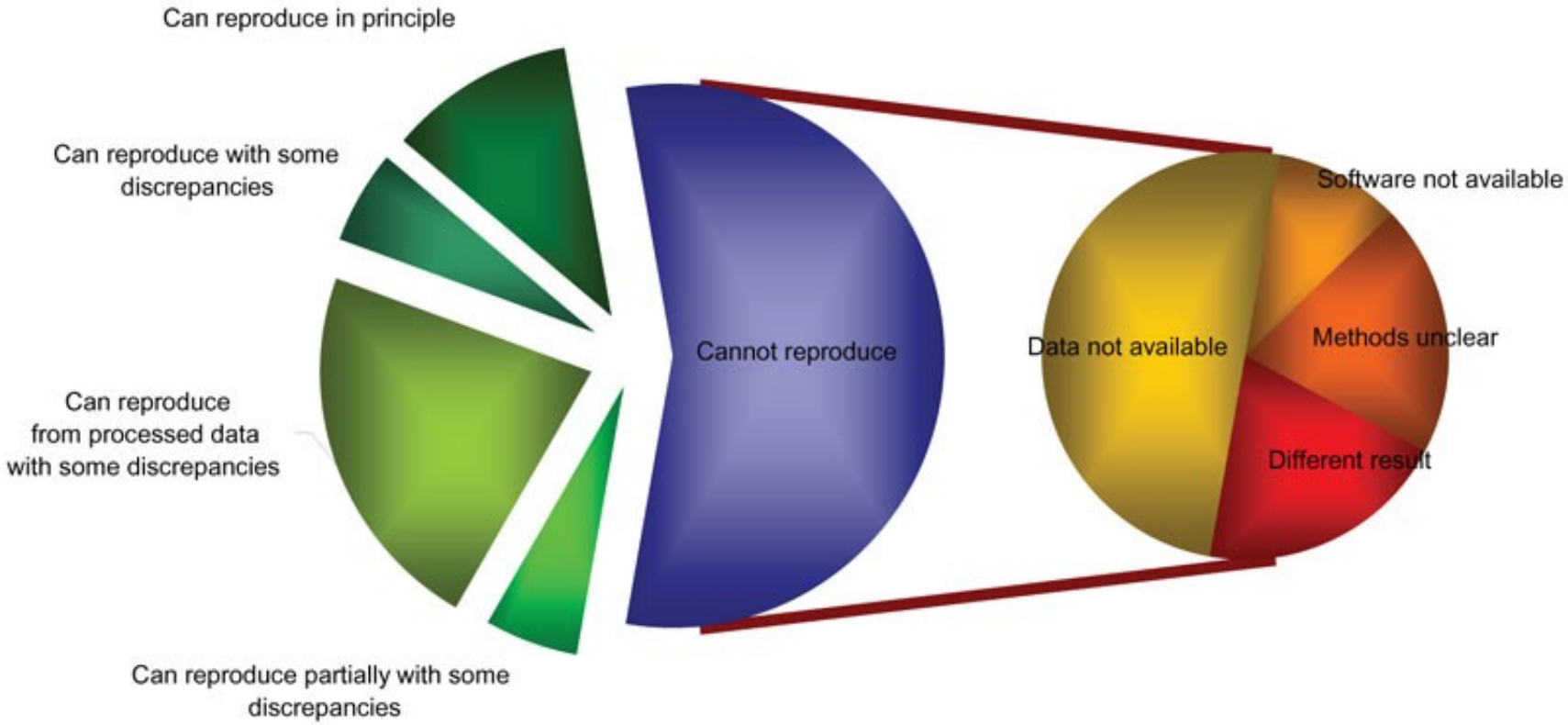
Terrie Wheeler, AMLS
Director
Samuel J. Wood Library

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

There are three fires burning
that motivate our actions towards better reproducibility

1. “Reproducibility Crisis”
2. Data Retention Mandates
3. Allegations of Misconduct

Attempting to reproduce research proves to be difficult, when you can't get the data…



Ioannidis *et al.* (2009) Nat. Gen. 41:149

“OMB Circular A-110 states that the retention period is three years from the date the final financial report is submitted.”

“NSF states in its General Grant Conditions that records must be retained for three years after the submission of all required reports”

“in the case of research misconduct involving NIH funding, records must be retained for six years after the final resolution date of the case.”

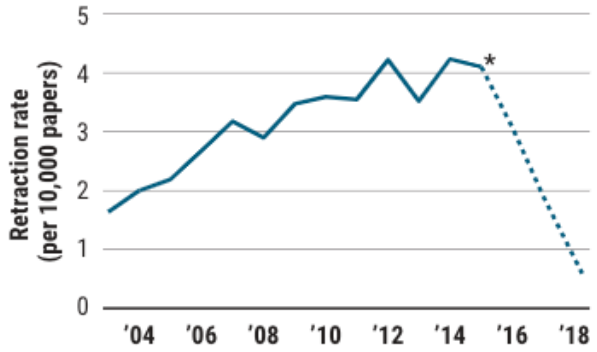
“retain research data pertinent to patented inventions for the life of the patent”

“about 2% of scientists admitted to have fabricated, falsified or modified data or results at least once”
“Up to one third admitted a variety of other questionable practices”

Fanelli (2009) PLOS One 0005738

“3.8% of published papers contained problematic figures, with at least half exhibiting features suggestive of deliberate manipulation”

Bik *et al.* (2016) mBio 00809-16



Brainard and You (2018) Science 00809-16

There are three fires burning that motivate our actions towards better reproducibility

1. “Reproducibility Crisis”
2. Data Retention Mandates
3. Allegations of Misconduct

Maximizing data value and ethical research conduct

We are not automatically making research reproducible, but providing the infrastructure that helps make research reproducible

Reporting Standards

PRISMA

MIAME

MIQE

Principles of reproducibility

FAIR

Literate Programming (Knuth 1992)

Compendiums (Gentleman and Lang, 2003)

Data Publication standards

Open Science

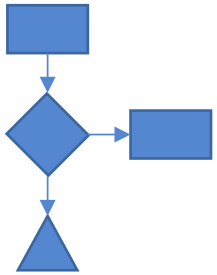
Communities of practice

The Turing Way

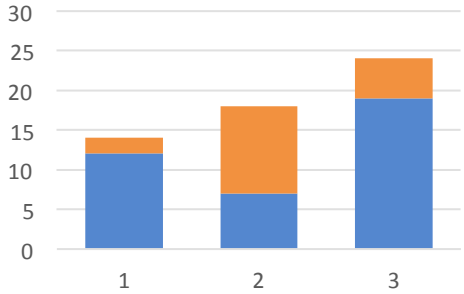
How can we help researchers capture the data and workflows that lead to publishable results?

010101000101010
111100001001101
001001000111111
001010001111010
001010100100100
100101111

DATA



WORKFLOWS



RESULTS

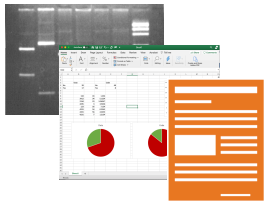
Hurdle #1: Dealing with confidential data

Hurdle #2: Researchers are very busy

Hurdle #3: Data and workflows are very diverse

Hurdle #4: Maintaining a high quality solution

Piece #1: Electronic Lab Notebooks capture (small) data and workflows, while remaining flexible to researcher needs



Direct storage of images, data files, analysis files, workflows

File versioning and immutable timestamps

Integration with Jupyter Notebooks (on roadmap)

Shareable and transferable

Piece #2: A file management system for tagging, tracking, and then archiving files on the institutional storage systems

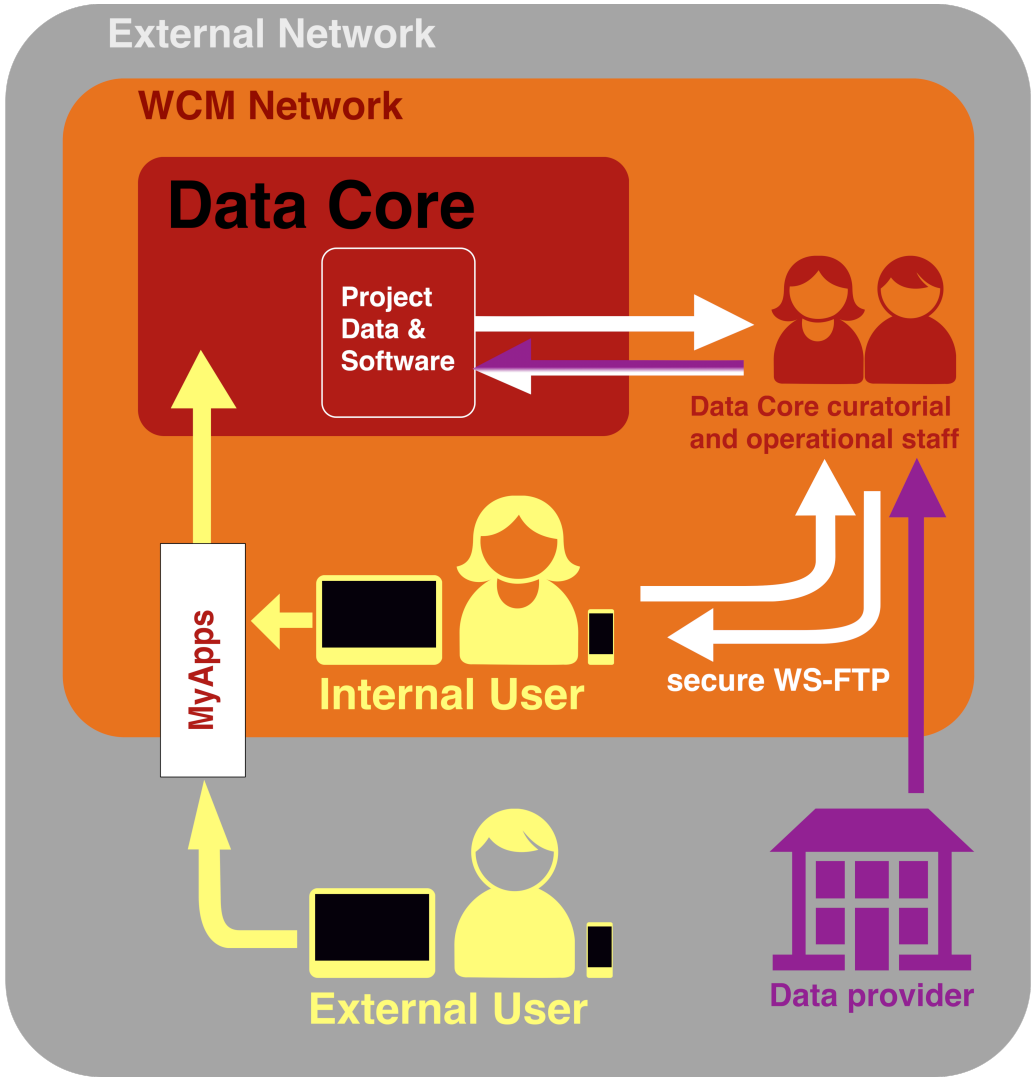
Data associated with a project/publication assigned a unique tag

File identity managed by hashing

File location tracked within institutional storage

Actionable scripts for marking project complete, and archiving

Piece #3: Secure file access management and computation through the institutional Data Core

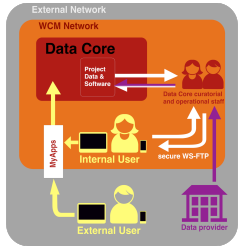


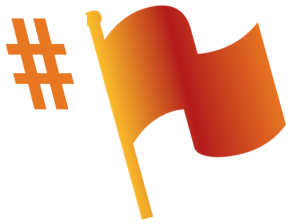
Secured, collaborative, flexible

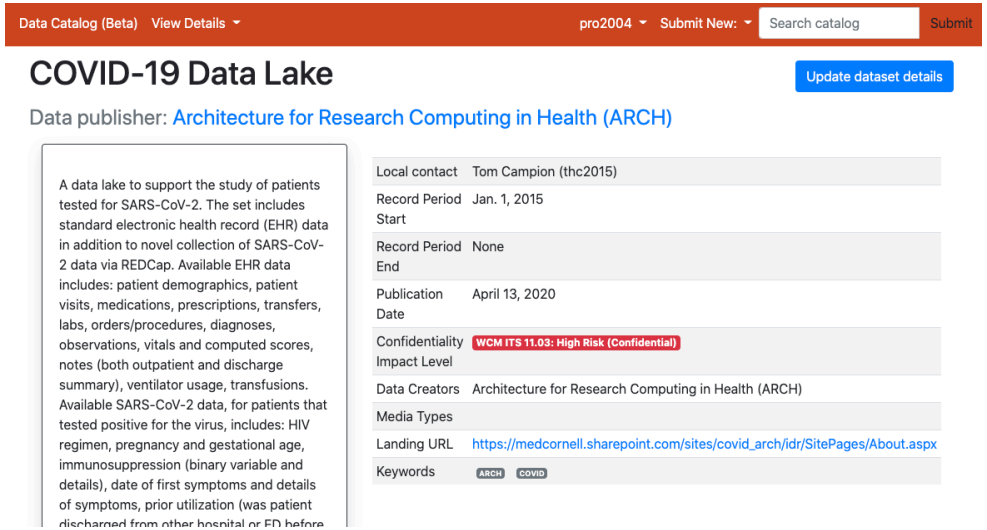
Project governance and monitoring

Data curation for import and export

Holding the pieces together: an institutional Data Catalog that connects data, workflows, governance and access conditions.







Discovery layer with the capacity to connect grants, data, publications

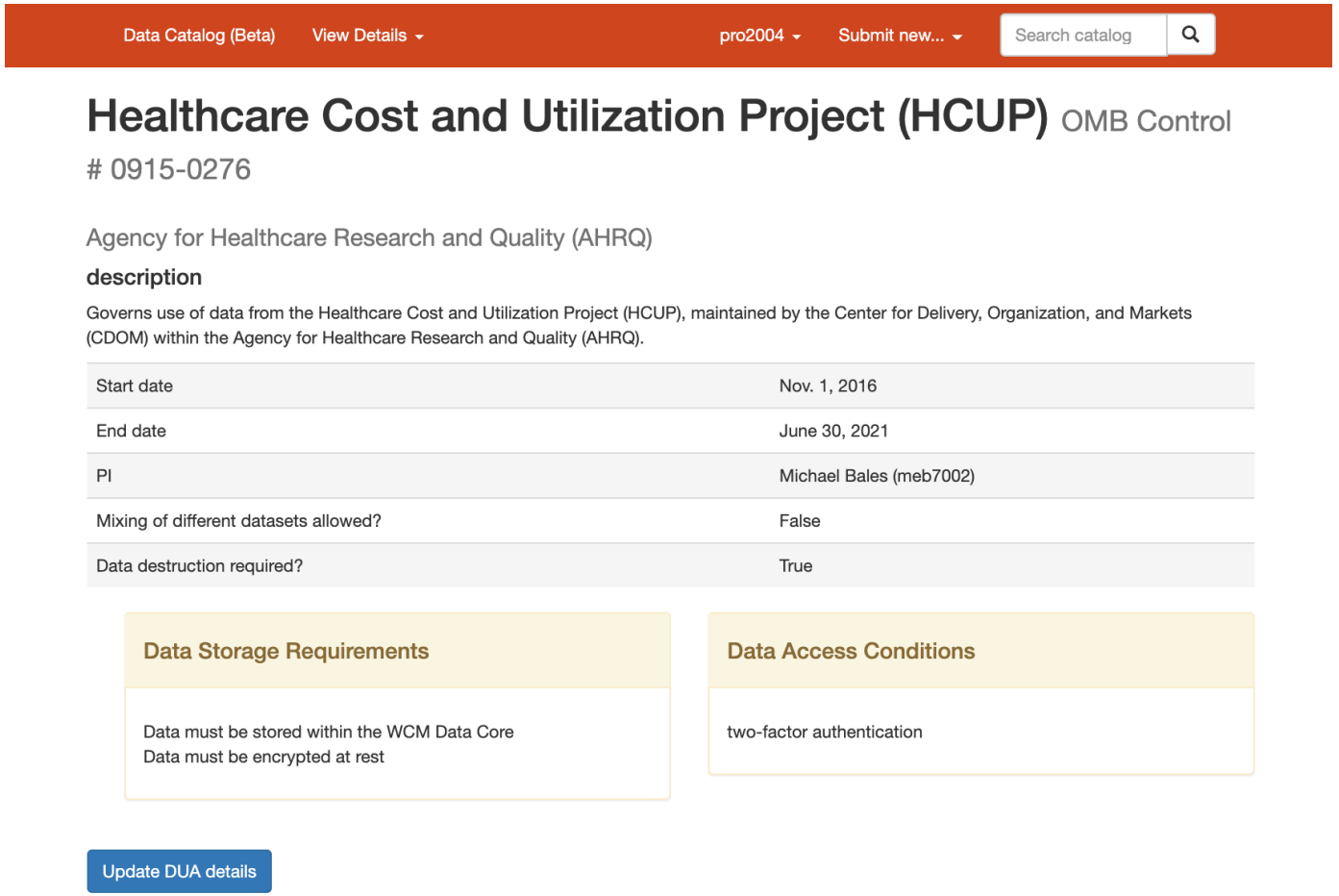
The WCM Data Catalog was built to manage data governance and access conditions

Scope of authorization

User authorization

Data Controls

Reuse scope



Three triggers to prompt capture and storage

1. Project/Grant completion
2. Publication
3. Faculty member leaves the institution

Move files to archive
Register project, file tag, ELNs in Data Catalog

Acknowledgements:

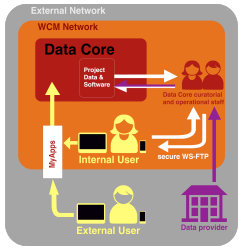
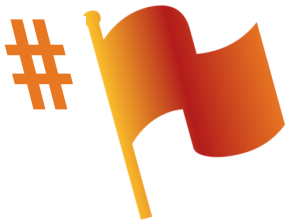
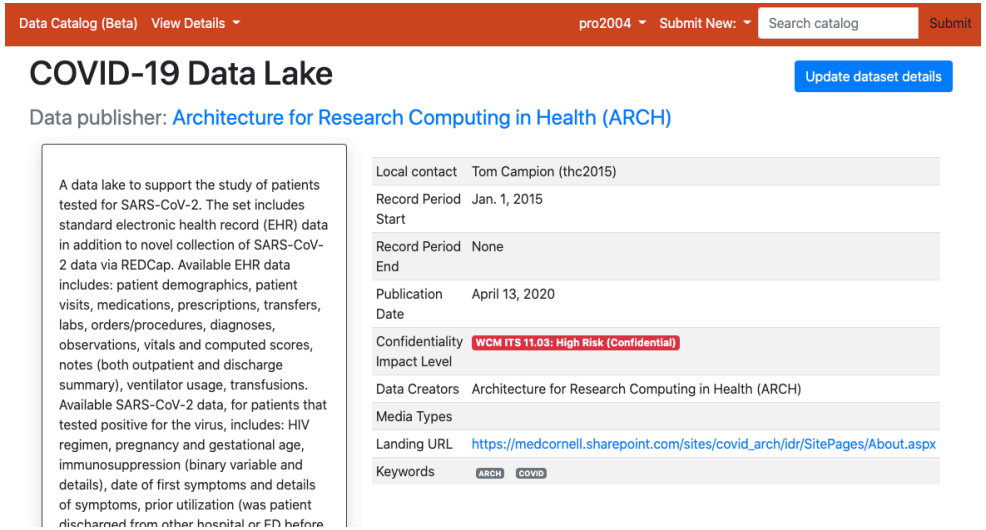
Curt Cole (WCM CIO)

Data Core
Alice Chin
Ana Proper
Frank Ashmun
John Ruffing
Jo Hargitai
Lucy Walle
Michael Bales
Heather Kleinschmidt

Electronic Lab Notebooks
Marie Linvill (LabArchives)
Cindy Chen
Danny Tan
Lucy Walle
Tony DiFazio

Data Catalog
The Data Catalog Collaboration (DDC)
Nicole Contaxis (NYU Langone)
Tom Campion

File System Management
Cindy Chen
Darrin Stivala (Starfish)
Doug Hughes (Starfish)
Jo Hargitai
Tony DiFazio

Peter Oxley pro2004@med.cornell.edu
Terrie Wheeler tew2004@med.cornell.edu

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.