OSC (Common Fund)

The Common Fund

Concept Clearance: New opportunity within existing Common Fund Program

Targeted Needs at CF Data Coordinating Centers to Establish the Common Fund Data Ecosystem (CFDE)

Objective: **Enable DCCs to engage with the CFDE Coordinating Center to establish the Common Fund Data Ecosystem**

Funds Available and Anticipated Number of Awards: **~$2.5M/year for 3 years**

Award Project Period: 3 years

Council Action: Vote on support of CF DCCs to establish the CFDE

What is the CFDE?

Effort to advance scientific research by facilitating the use of data within and between Common Fund programs through:

- Infrastructure
- Collaboration
- Training
- Sustainability

NIH and Cloud

- **The way the data are stored and managed is unique to each NIH program**
  - (often) Not much attention is paid to data organization, structure, access, utility, findability, reusability
  - The focus and end goal are scientific results (which use the data) and journal articles
  - This results in reduced ability *(or inability)* to use or reuse the data within a program
    - *During or after a program's completion date*
    - *Often impossible to find or use data between programs*
- **NIH programs are (or planning) to use the cloud to store and compute on data**
  - *Large size (storage)*
  - *Analytics (compute)*
  - *Ability to share information between geographically distributed groups*

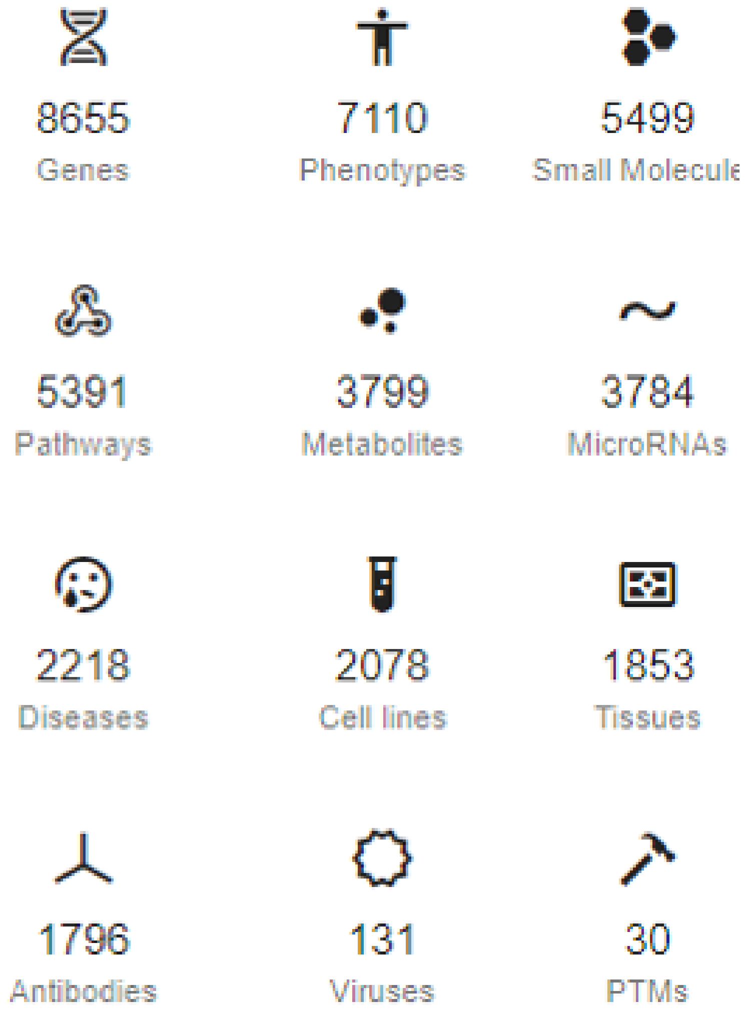
If everything is hosted on a cloud, why isn't it already interoperable?

A number of reasons…

No matter how much a DCC may want to interoperate with another DCC…

They can not *unless other DCCs participate.*

In many cases, data are not
F: Findable
A: Accessible
I: Interoperable
R: Reusable

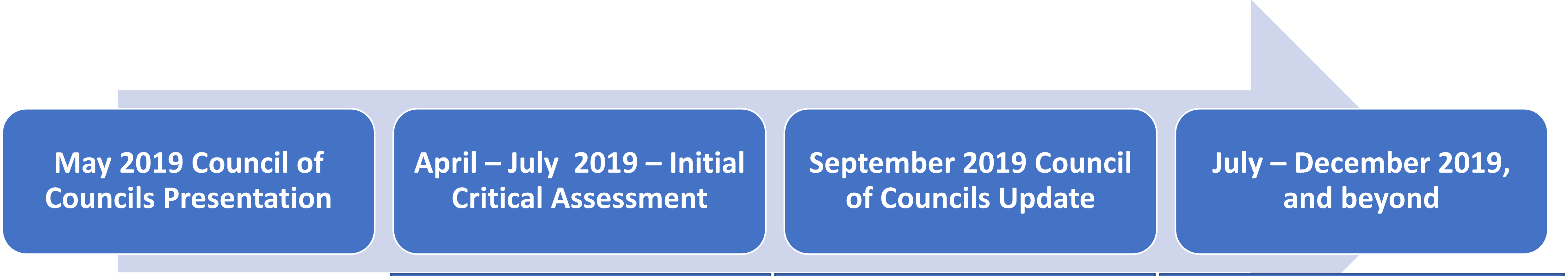


8655 Genes	7110 Phenotypes	5499 Small Molecule
5391 Pathways	3799 Metabolites	3784 MicroRNAs
2218 Diseases	2078 Cell lines	1853 Tissues
1796 Antibodies	131 Viruses	30 PTMs

Common Fund Data Ecosystem: Goals

- Making CF data sets more useful/usable ***within*** a program and ***between*** programs
  - *Improving **FAIR**ness: **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable*
- Capturing and developing best practices for new programs to leverage
- Enhancing the ability to ask scientific questions across data sets
- Increasing reuse of data *(and tools)* after a program ends
- Incorporating "old" data into new programs
- Data in the cloud is only part of the solution

Common Fund Data Ecosystem: Recent Timeline



May 2019 Council of Councils Presentation

April – July 2019 – Initial Critical Assessment

- Critical assessment of activities for selected CF programs, Kids First, GTEx, HMP and LINCS
- Obtain a deep understanding of the issues and needs by working *with* each CF Program
- Report to NIH on progress

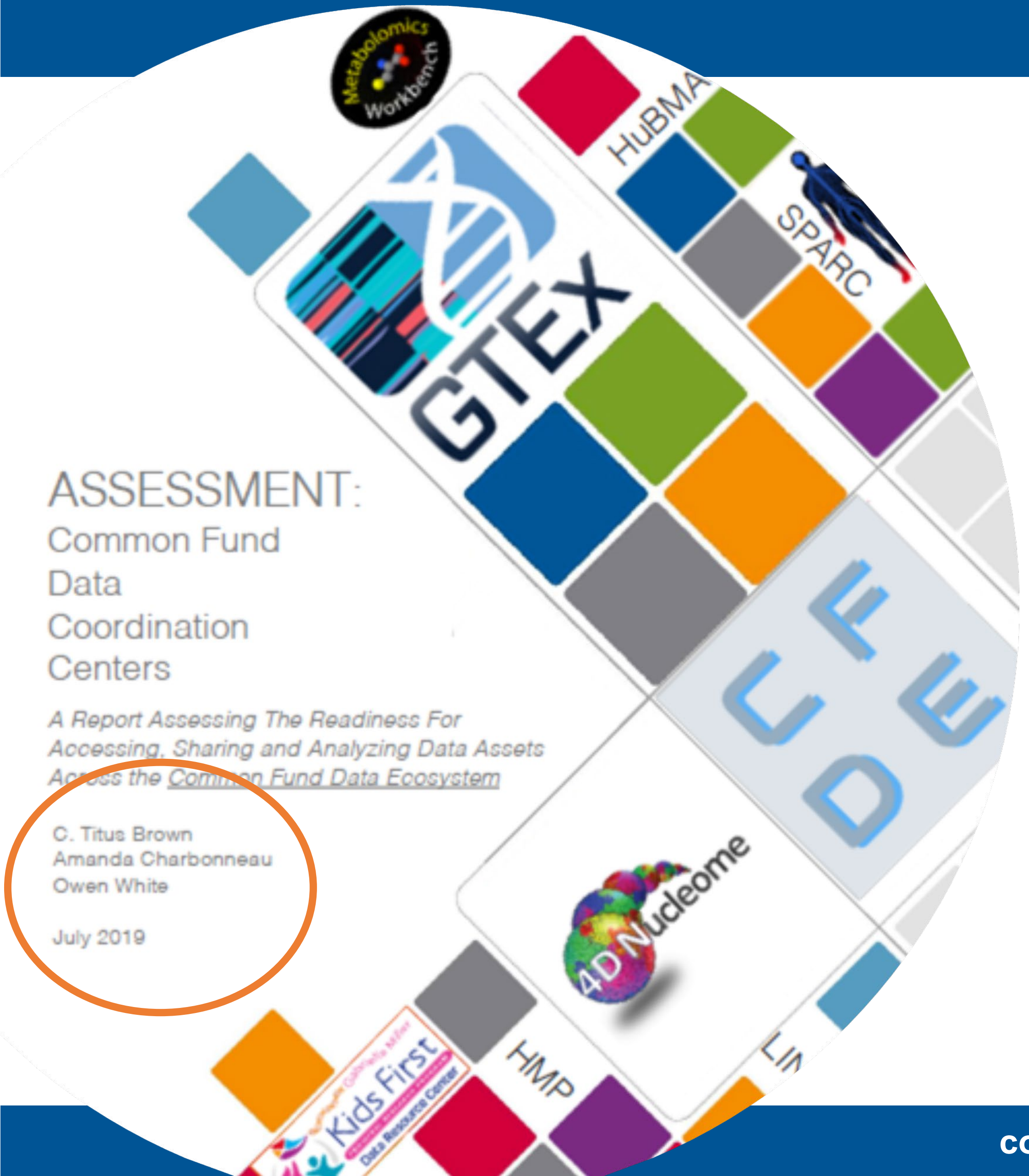
September 2019 Council of Councils Update

1. Update on the CFDE
2. Concept clearance on DCC participation in the CFDE

July – December 2019, and beyond

- Continue critical assessment activities with HuBMap, SPARC, MoTrPac, Metabolomics, 4D Nucleome
- Engage additional datasets/DCCs over time

- Support DCCs as active partners in the establishment of the CFDE
- Estimated at $250K/yr/award



ASSESSMENT:
Common Fund
Data
Coordination
Centers

A Report Assessing The Readiness For Accessing, Sharing and Analyzing Data Assets Across the Common Fund Data Ecosystem

C. Titus Brown
Amanda Charbonneau
Owen White

July 2019

Council Update:
CFDE Initial Report

Acknowledge:
Titus Brown, Amanda Charbonneau, Owen White, and the participating DCCs

Purpose of the Report

Assessment to understand the datasets, their use, and challenges to participation

1. Initial review of 9 DCCs

- Data types and size
- Metadata
- Storage
- Use and users
- Available resources
- Data access
- Security
- Initial FAIRness

2. “Deep dives” of 4 DCCs

- Use cases
- End-user needs (search, analysis, tools, training)
- Technical (metadata, software, tools)
- Implications of cloud use (human subjects, FISMA, A/A, migration)
- Costs and ideas on efficiencies

3. Game changers

- How to significantly advance science
- Improved DCC capabilities, barrier reduction
- Shared needs between DCCs
- Reduced costs
- User experiences

Report Highlights

The Common Fund

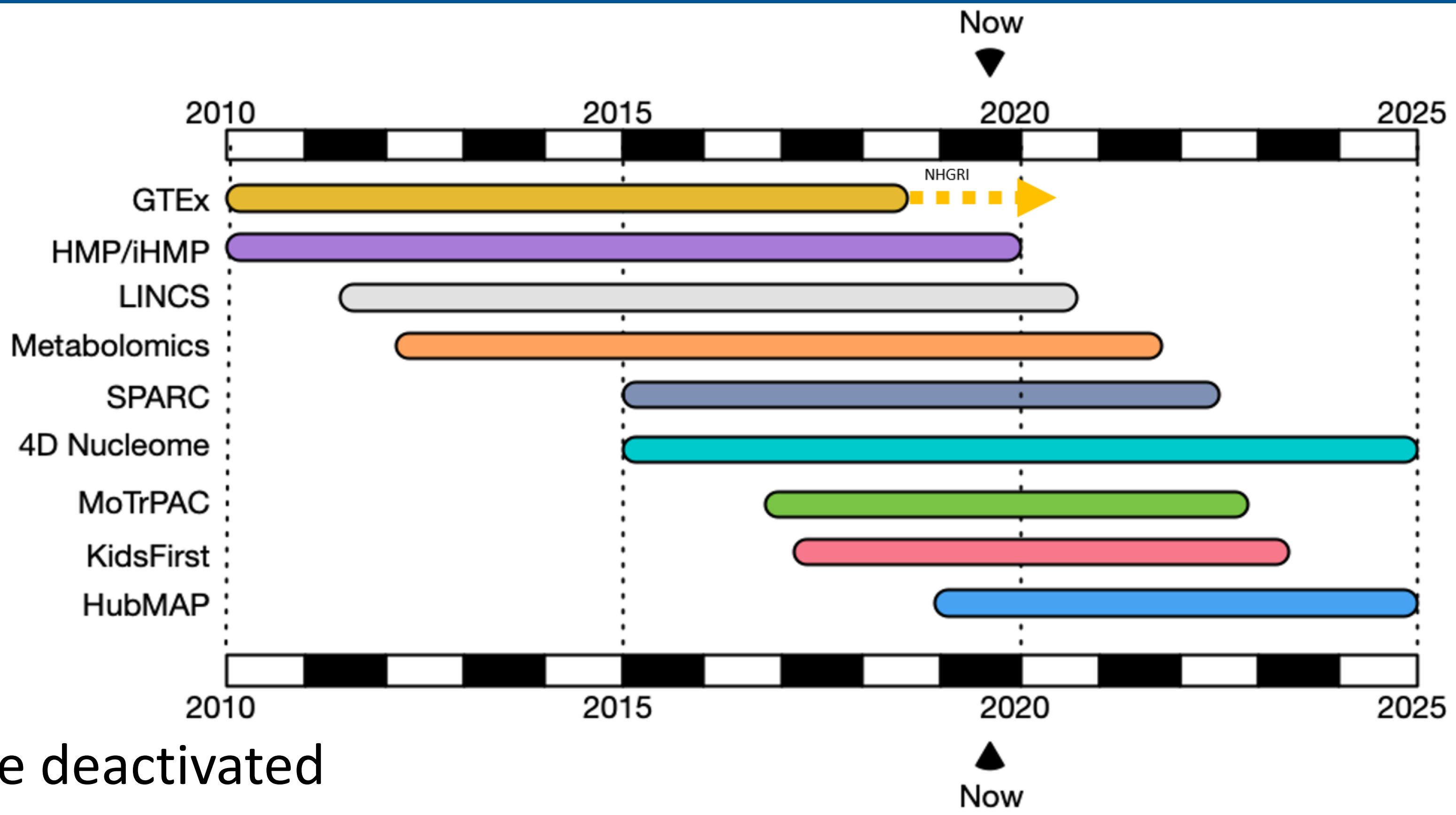
- Infrastructure: Cloud storage (e.g., .STRIDES) necessary but not sufficient. No matter what, will need other infrastructure to link between data.
- Collaboration: DCCs expressed a strong desire for collaboration with other DCCs, but time, $, and partners an issue. It takes (at least) Two to Tango.
- Collaboration: Want their datasets to “interact” with other data sets and are enthusiastic about doing this. Need harmonization for linkages and access.
- Training: Need support for end-user training (e.g., bioinformatics research, uncompressing files, clinical research). Demand is high but DCCs have resource constraints.
- Sustainability: Concerns over long-term sustainability, data life-cycle.

Sustainability

The Common Fund

DCCs at beginning of life cycle

- Startup costs associated with increasing expertise, building infrastructure
- Continuity of best practices
- Resource and time-saving



DCCs at end of life cycle

- Stewardship of assets as they are deactivated
- Permanence of assets on accessible cloud storage
- Painful loss of expertise/training
- Ageing datasets prematurely obsolete

Recommendations

- Targeted DCC investment – end of lifecycle support, $ and opportunities for CFDE participation, targeted training, STRIDES
- Cross-DCC investment – e.g., DCC-to-DCC joint exercise(s) for interoperability, collaboration opportunities, infrastructure re-use
- CFDE technical activities – portal (gateway to access datasets), DCC data dashboard, metadata, FAIRness, more DCC and NIH engagement
- Possible new transformative CFDE team activities – lifecycle support, best practices, authentication/authorization, etc.
- Longer term Common Fund Data Ecosystem RFAs -- future

future years

Concept

Targeted Needs at CF Data Coordinating Centers to Establish the Common Fund Data Ecosystem (CFDE)

- Support DCCs to understand their specific requirements to expand the science that can be conducted and speed the use of data in translational and clinical applications.
- Targeted, limited competition solicitation of applications from CF DCCs to engage with the CFDE and other DCCs to establish the CFDE.
- The initial duration of awards through this solicitation will be for 3 years.
- We expect to re-solicit applications as DCCs for new programs are established.
- Budgets are expected to vary depending on the specific requirements of each DCC but are estimated at $250K/year/award.

- **~$2.5M/year for 3 years**