Common Fund Data Ecosystem

Vivien Bonazzi
Senior Advisor for Data Science Technologies and Innovation
Office of Strategic Coordination, DPCPSI

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NIH National Institutes of Health
Office of Strategic Coordination - The Common Fund

NIH and Cloud

The Common Fund

- **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*

- **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 programs completion date*
    - *Often impossible to find or use data between programs*

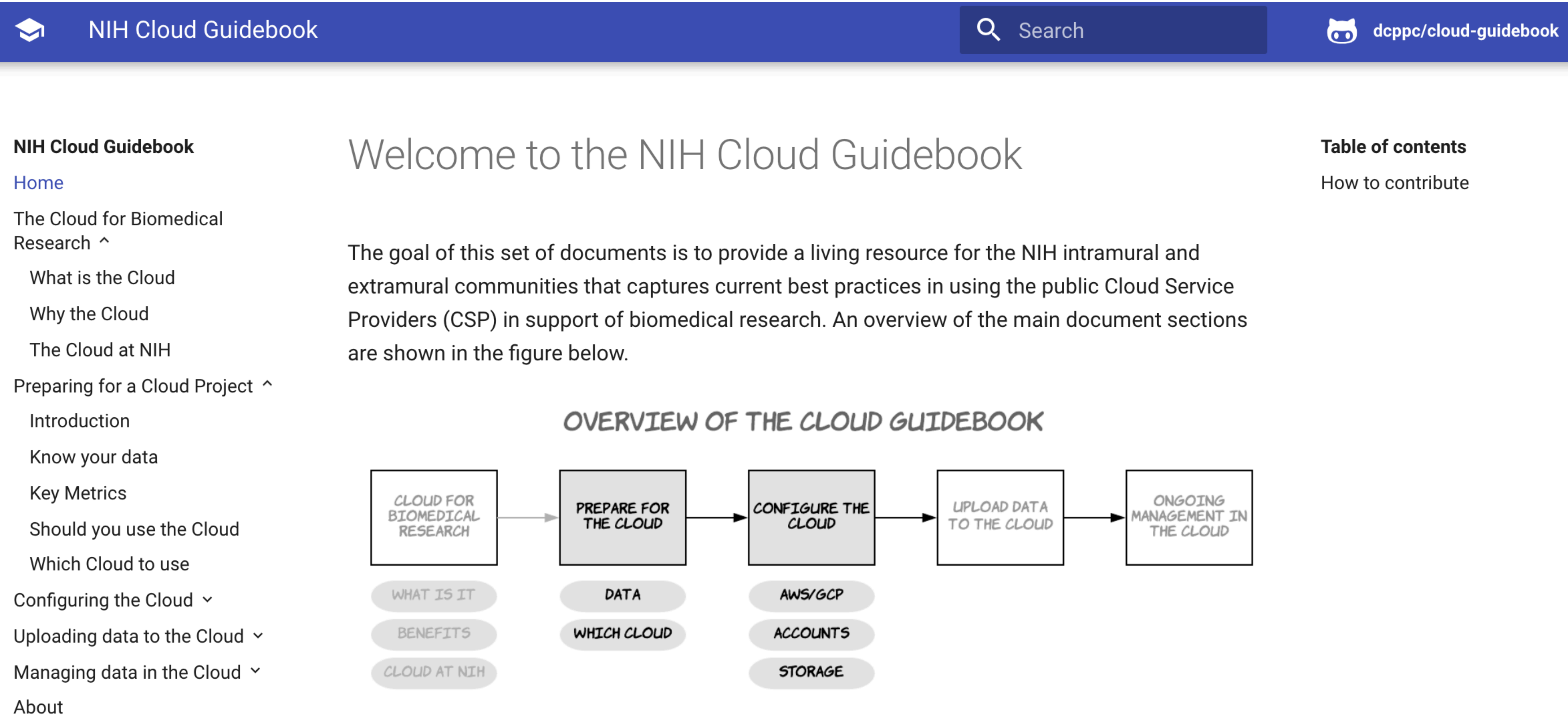
Common Fund Data Ecosystem: Goals

Extend from and leverage deliverables and lessons from the Data Commons Pilot Phase Consortium to enhance utility of Common Fund Data Sets

- 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

Common Fund Data Ecosystem: Activities needed

- **Onboarding data to the cloud in a consistent manner**
  - Using NIH STRIDES billing agreements
  - Ensuring the data is stored and organized optimally for each CSP (Cloud Service Provider)
  - Versioning and upkeep of data
  - Cost management and accounting
  - Documentation for use of cloud with NIH data



Welcome to the NIH Cloud Guidebook

The goal of this set of documents is to provide a living resource for the NIH intramural and extramural communities that captures current best practices in using the public Cloud Service Providers (CSP) in support of biomedical research. An overview of the main document sections are shown in the figure below.

Common Fund Data Ecosystem: Activities needed

- **Developing data management plans, best practices and use cases for each CF program**
    - Self governed metadata standards: *findability and reuse*
    - Harmonized data *combined analysis*
    - Cross cutting metadata models: *data querying within and across CF programs*
    - FAIR assessment: *tools to assess and improve FAIRness of data (common metrics)*
    - Authentication/Authorization: *permissions to use controlled access data*
    - Data Dashboards: *programs can monitor data management activities (*internal*)*
    - CF Data Portal: *directory to Common Fund data sets (*external*)*
    - Data Platform: *cloud platforms supporting end user interactions (SaaS)*
    - Training: *materials for end users to help use and understand CF data*

Common Fund Data Ecosystem: Approach

- **Community Engagement**: ***Collaborating*** **not mandating**

  CF Programs (PIs and POs);

  - Know their project and data well
  - May have existing:
    - *Data management plans or components*
    - *Use cases for cross cutting analysis*
  - Provide additional input on data management topics
  - Established collaborations with CSPs through awardees

Common Fund Data Ecosystem: Next Steps

The Common Fund

- **Critical assessment of described activities for a few (initial) Common Fund programs**
  - Kids First, MoTrPAC, HMP and iHMP
  - Obtain a deep under standing of the issues by working *with* each CF program
  - Collaborate in building a Common Fund dash board and portal
  - Identify additional needs in collaboration with each CF Program
  - Extend the assessment over time with additional Common Fund programs
    - HuBMap, SPARC, GTEx, Metabolomics, LINCs
  - Refine the roadmap for Common Fund Data Ecosystem with specific actions to undertake

* ***Activities*** *(described in slide 6)*

- *Onboarding data to the cloud | Self governed metadata standards | Harmonized data*
- *Cross cutting metadata models | FAIR assessment | Authentication/Authorization*
- *Data Dashboards |CF Data Portal | Data Platforms | Training*

Common Fund Data Ecosystem: Timeline

NIH The Common Fund

- **April – July 2019**
  - Critical assessment of described activities for a few (initial) Common Fund programs
  - Kids First, MoTrPAC, HMP and iHMP
  - Obtain a deep under standing of the issues by working *with* each CF program
  - Collaborate in building a Common Fund dash board and portal
  - Identify additional needs in collaboration with each CF Program

- **July – December 2019**
  - Extend the assessment over time with additional Common Fund programs
    - HuBMap, SPARC, GTEx, Metabolomics, LINCs
  - Refine the roadmap for Common Fund Data Ecosystem with specific actions to undertake

Common Fund Data Ecosystem: Team

- DCPPC Awardees whose work is being re-scoped
  - Owen White *(U Maryland)*
  - Avi Ma'Ayan *(Mount Sinai*)
  - Carl Kesselman *(USC)*
  - *Others (funding pending)*

- NIH Common Fund Team
  - Vivien Bonazzi
  - Lora Kutkat
  - Jen Yttri
  - Michael Ojiere
  - Simon Twigger *(OSC and CIT/STRIDES)*

Common Fund Data Ecosystem

The Common Fund

Questions?