Common Fund Data Ecosystem (CFDE) Council of Councils Working Group

Final Report

May 2022

National Institutes of Health
Office of Strategic Coordination–The Common Fund

Table of Contents

Common Fund Data Ecosystem (CFDE) Council of Councils Working Group Final Report

Executive Summary

The challenges that the CFDE faces are shared by data managers and researchers everywhere and can be grouped into three categories: 1) Enhance the value of data by fostering interoperability and reuse of Common Fund data, including integration with external data sets; 2) Ensure sustainability and accessibility of valuable data resources; and 3) Train users to work with Common Fund data in the cloud where the data have the potential for greater interoperability and reuse, and where increased cloud computing is expected to lower costs and increase equitable access to resources.

Establishing the CFDE as a trans-Common Fund infrastructure to facilitate consistent data management across many programs was an important step. The pilot phase of the CFDE has been impressive. It has provided a portal through which users can discover Common Fund data, established a core metadata model to enable users to find related data across data sets, piloted a cloud-based workspace, trained users, supported small grants to explore integration and analysis across data sets, and established a community of investigators, who are effectively collaborating to form the data ecosystem. The Working Group recognizes and commends the importance of CFDE to the future of biomedical research and the role of the Common Fund in defining the next generation of data resources at NIH.

The driving objective behind the CFDE is to foster scientific discovery through the reuse of data. The chief metric of success for the CFDE is discovery: if the CFDE is successful, in five years, many investigators will be using Common Fund data for new discoveries and new purposes. With this in mind, recommendations for each of the goals of the CFDE can be summarized as follows:

1. Supporting interoperability and the reuse of data, including integration with external data sets - Support queries across data sets through use of existing community-based standards, development of robust use cases and workflows that anticipate interests of diverse user types, scale-up of the metadata model, development of innovative methods for data search, expanded use of knowledge graphs that provide relationships across data sets, citation of data and tool contributors by CFDE users, and enhanced support for users. User experiences should shape further development. Recognizing the nature of ecosystems, the Common Fund should be prepared to support the active evolution of the CFDE through resource development that supports new use cases and new users, including making Common Fund data sets easily accessible in a cloud environment. Continued attention to the NIH-wide and other data communities is essential; the CFDE should participate in NIH-led collaborations with external entities when possible. With its diverse data types located in several different platforms, the CFDE is positioned to be an important contributor to developing best practices for cross-platform interoperability.
2. Sustainability and accessibility - This is a critical issue for Common Fund data since the programs that generate the data will be supported for a maximum of ten years. Since the Common Fund supports the development of new repositories as well as new data, the sustainability of these repositories is a particular concern. Public repositories offer an increasingly attractive option for sustaining data. The development of best practices for repositories and for repository management is an NIH-wide concern that exceeds the scope of the CFDE. The CFDE is an

important client of these repositories and should participate in efforts to define how repositories will be managed and supported. However, it should not become a long-term repository for Common Fund data sets. Repositories developed and tested through Common Fund programs will continue to transition to support from Institutes and Centers or other entities if they prove useful, as they have in the past.

3. Training – Transitioning to a cloud computing environment is a substantial cultural shift for the research community. The benefits of computational speed, data security, and long-term costs make the challenge of this transition worth the effort. Training users should be a greater emphasis for the CFDE moving forward, focusing on enabling the use of CF data sets specifically. NIH should consider diverse training mechanisms that have a high capacity to reach many people when and where they are ready to access them. Smaller capacity training efforts such as summer fellowships, courses, codeathons, and small grants are also valuable to democratize access to CFDE resources and could be targeted to early-stage investigators and lower-resourced institutions.

Introduction and Charge of the Working Group

The NIH Common Fund supports bold scientific programs that catalyze discovery across all biomedical and behavioral research. Common Fund programs create a space where investigators and multiple NIH Institutes and Centers (ICs) collaborate on innovative research addressing high-priority challenges for the NIH as a whole and make a broader impact in the scientific community. Approximately two-thirds of Common Fund programs generate large-scale data resources and digital tools that are intended to be used by researchers across the entire spectrum of biomedical research.

As described in the NIH Strategic Plan for Data Science, these data resources provide unprecedented opportunities to understand biological mechanisms, interrogate complex biological systems, deliver new types of discoveries, and rapidly advance novel treatments and cures for many diseases. However, multiple challenges in data collection, curation, storage, management, and sharing must be addressed to realize the full potential of the "big data" revolution in biomedical research. As described in this strategic plan, NIH is taking steps to modernize the NIH-funded biomedical data-resource ecosystem. In alignment with the NIH plan, the Common Fund is addressing data science opportunities and challenges related to Common Fund programs.

The Common Fund Data Ecosystem (CFDE) is an infrastructure investment made by the Common Fund to address the growing challenges facing scientific programs that leverage data-intensive strategies. To support these programs and downstream data users, the CFDE is helping to ensure that all Common Fund data sets are Findable, Accessible, Interoperable, and Reusable (FAIR), providing training for users to operate on the data in a cloud environment, and ensuring that Common Fund data continue to be available after individual programs are completed. The CFDE will amplify the impact of many Common Fund programs by enabling researchers to interrogate multiple disparate data sets, and thereby make new kinds of scientific discoveries that were not possible before. The CFDE is also being designed in parallel with NIH IC data platforms to enable crosstalk between Common Fund and IC data sets and address NIH-wide data management objectives described in the NIH Strategic Plan for Data Science.

The CFDE was approved as a three-year pilot program by the NIH Council of Councils in September 2019. In May 2021, a Working Group of the Council of Councils was approved to provide expert assessment of the CFDE's progress to date and provide recommendations for the CFDE's future. NIH expects to use these recommendations, presented to the Council of Councils for approval in May 2022, to guide a concept for the next phase of the CFDE. This concept will be presented to the Council of Councils in the fall of 2022.

In July 2021, Dr. James Anderson, Director of the Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI), charged the Working Group as follows:

- Review the current scope and goals of the CFDE as well as the progress to date
- Make recommendations for future scope and goals in the following areas:
    i. Findability and accessibility of data
    ii. Data harmonization and interoperability
    iii. Cloud workspaces
    iv. Sustaining access to data and tools after Common Fund programs end
    v. Training and outreach to enhance access to, and use of, the data
    vi. CFDE scope and strategy in the context of related NIH activities

Working Group Methods

Members of the Working Group brought expertise across a wide range of issues related to data science, including, but not limited to, data management, bioinformatics, data science training, proteomics, genomics, imaging, and data science infrastructure (see Appendix A). The Working Group met monthly from July 2021 to April 2022, with each meeting lasting 120 minutes. All meetings were virtual. Working Group meetings from July 2021 – January 2022 consisted of presentations from NIH data science experts and CFDE awardees, followed by discussions between presenters and Working Group members. Presentations were followed by closed discussions between Working Group members and NIH staff managing the CFDE. Meetings in February 2022 – April 2022 focused on outlining and drafting the report, then preparing for the report presentation to the Council of Councils. Agendas for the meetings and summaries of the main points from the presentations can be found in Appendix B.

CFDE Goals and Current Status

The CFDE has three overarching goals:

- Enhance the value of Common Fund investments by enabling users to query across and use Common Fund data sets.
- Ensure that Common Fund data and tools are sustained after individual programs end.
- Train users to work with Common Fund data in a cloud environment.

The CFDE is comprised of the CFDE Coordinating Center (CFDE CC) and participating Data Coordinating Centers (DCCs) from various Common Fund programs. The CFDE CC manages and coordinates activities across the CFDE; develops and maintains the CFDE portal, a search platform that enables users to query

and identify relevant data sets across multiple Common Fund programs; helps ensure data and tools are sustained; and manages the CFDE user engagement and training center. The CFDE CC also works with participating DCCs to enhance FAIRness of data sets, capture best practices for Common Fund programs to leverage, and harmonize metadata to optimize cross-data set search. The participating DCCs work with the CFDE CC to make data available through the CFDE portal and work with other DCCs to identify and implement partnership activities addressing scientific opportunities across Common Fund data sets. The CFDE also supports investigator-initiated R03 projects to enhance the utility of Common Fund data sets.

The CFDE currently includes DCCs from the following Common Fund programs:

- 4D Nucleome (4DN)
- Extracellular RNA Communication (ExRNA)
- Gabriella Miller Kids First Pediatric Research (Kids First)
- Genotype-Tissue Expression (GTEx)
- Glycoscience
- Human BioMolecular Atlas Project (HuBMAP)
- Human Microbiome Project (HMP)
- Illuminating the Druggable Genome (IDG)
- Library of Integrated Network-based Signatures (LINCS)
- Metabolomics
- Stimulating Peripheral Activity to Relieve Conditions (SPARC)

PARTIES WITH AN INTEREST IN CFDE

There are many different parties who have an interest in the CFDE. Major interested parties include:

- Data generators – researchers, primarily supported by Common Fund programs, who generate data that is or will be included in the CFDE.
- Tool developers – researchers who build, test, and deploy novel tools to manage and analyze CFDE data. Developers may also be data/tool users.
- Data/tool users – researchers who use the CFDE data and/or tools; these researchers could be supported by the Common Fund or by NIH Institutes and Centers, other federal agencies, or other sources.
- Biomedical research community – researchers everywhere who will benefit from the novel discoveries and knowledge generated by CFDE users, even if they do not directly use CFDE data, tools, or resources. For example, a new biological insight generated through analysis of CFDE data may lead to new avenues of research for many additional scientists, even if these scientists do not use CFDE data in subsequent research efforts.

The CFDE portal allows users to find relevant Common Fund data sets by searching across metadata. DCCs submit metadata in alignment with the CFDE's Cross-Cut Metadata Model (C2M2). Currently, C2M2 covers several domains: assay and file type, anatomy, biosample, subject, and taxonomy, with plans to include disease phenotype information, gene IDs, and clinical metadata in the next year. C2M2 is built on the consensus of the DCCs, which allows all groups to have input but also requires time to reach consensus.

Many challenges faced by the CFDE are common across the biomedical research enterprise. To define a feasible set of activities the CFDE should undertake for highest impact and greatest chance of success, the Working Group provided the following recommendations about the overall scope and strategy of the CFDE moving forward.

Recommendation 1.1: The initial scope of the CFDE is broad and will need to be focused.

The Working Group emphasized that progress to date for the CFDE has been impressive. The CFDE is attempting to address a wide variety of high-priority challenges in data science and are tackling challenging issues that many institutions and research teams are grappling with. Although the CFDE will not solve these challenges alone, the CFDE may pilot solutions or novel approaches that could be adopted as part of a broader NIH strategy. In particular, the Working Group noted that the DCCs are working together in useful and impressive ways, and demonstration of interoperability, integration across initiatives, and attention to use cases and usability were appreciated. However, the Working Group also noted that the ambitious nature of the CFDE may need to be scaled back to focus on successful implementation of a smaller number of activities. Priority areas for this work focus on several general themes: supporting queries across data sets through use of community-based standards, development of robust use cases and workflows, innovative search methods, and enhanced user support; novel approaches to training; and sustainability of data and tools. The goal of these recommendations is to ensure that current and future CFDE data are actively used and available for continued study. Specific recommendations within these priority areas for the CFDE are articulated throughout this report.

Recommendation 1.2: Continued attention must be devoted to incentivizing data generators, infrastructure engineers, and CFDE data/tool users to ensure a healthy, sustainable ecosystem.

In addition to overcoming technical challenges and barriers, modern data management strategies need to address evolving cultural practices and traditions related to data science. Data generators and managers need to work together in new and highly collaborative ways to facilitate the broad use of diverse data sets. The CFDE has made significant efforts to bring together data generators from disparate programs to undertake joint projects and analyses, and it provides a framework for central coordination of activities that can rapidly adapt to dynamic program goals with minimal disruption. New incentive structures to enhance contributions to as well as the use of data repositories will help facilitate the growth of CFDE user bases. For example, the CFDE should encourage and train users to cite CFDE data and tools in their publications. Considering challenges holistically and enabling users to move easily across different parts of the data ecosystem will enhance usability of the data.

Recommendation 1.3: The CFDE should both enable and encourage data users, demonstrating utility through publications and use cases.

An important focus for the CFDE should be to both enable and encourage data users, going beyond technical interoperability. While the enabling component is going well, encouraging data use will require additional effort. Demonstration of data and resource utility through publications and concrete use cases will help increase awareness of the CFDE and encourage organic growth of a CFDE user community. For example, tracking and highlighting publications resulting from DCC collaborations or CFDE data users would help to establish utilization and provide a means for recruitment and training.

Findability and Accessibility

As findability and accessibility of data assets are pillar stones of FAIR-ness, portal work of the CFDE CC has focused primarily on findability and accessibility of Common Fund data sets during the CFDE pilot phase. These efforts to date were summarized as background material and discussed with the Council of Council Working Group (see Appendix B). The CFDE CC is compiling a catalog of available Common Fund data sets and making it available through its portal. Users can access the data sets of interest to them at

the DCC sites. Additionally, the CFDE CC portal will provide a shopping cart feature that will enable users to query for Common Fund data sets and then port them to their workspace of choice directly from the portal. The following recommendations provide prioritized activities and principles to guide future CFDE efforts for findability and accessibility, emphasizing what is feasible for the CFDE to undertake as well as the importance of focusing on user needs and experience.

Recommendation 2.1: Findability and accessibility is primarily the responsibility of the Common Fund project generating the data and should be addressed starting at the beginning of each program.

FAIR-readiness needs to happen close to the projects and data generation; hence it is mainly the responsibility of data generators. By the time there is a request for the CFDE to make data sets and tools FAIR, it is often too late. For the CFDE to be successful, the role of the individual initiatives that contribute data or tools needs to be clear, and data generators need to enhance FAIR-ness of their data sets and tools to a level that makes it sustainable by the CFDE. The CFDE can define an acceptable level of FAIR-ness for a given initiative to participate in the CFDE and then help with planning to achieve that goal. The CFDE can assist by identifying, contributing to, and raising awareness for existing standards. The goal is to ensure that the CFDE is not required to continually create new standards or to become a standards body. Instead, the CFDE should focus on enabling access using current standards to the broadest section of useful data possible.

Recommendation 2.2: It is important for the CFDE to emphasize data integration; engaged DCCs should not only think about their own data alone but plan for integration with other engaged Common Fund programs.

Data integration needs to be done holistically within the CFDE, which requires agreement on standards. Engaged DCCs must plan for integration with other engaged Common Fund programs as well as thinking about their own data. One possible solution could be for data generators to plan for the ontologies, i.e., hierarchically organized controlled vocabularies, and schema that will be used to organize and map their data sets. It is important that data generators adopt standardized ontologies and schemas, ideally at the outset of a program. Requesting all Common Fund programs use particular agreed-upon ontologies and schemas is less important than adoption of commonly used standards by the programs, as there are ways to translate various ontologies or schemas. For existing and completed Common Fund programs, it may be too late to prospectively organize the data sets and resources around standardized ontologies or schemas; however, efforts to do so retrospectively may be worthwhile albeit resource intense. In some fields, such as metabolomics and proteomics, there may not be consistent schemas. For those fields, at a minimum, data should be mapped to an ontology so that for every mappable variable or data type, identifiers can be used to enable the integration of an appropriate ontology of choice. These mappings will simplify integration work. The CFDE needs to lay out general principles for every project to follow, but this mapping of ontologies and schemas needs to be done by the data generator and user communities. CFDE guidance for choices should represent the field and can use the CFDE working groups to determine some guidelines. Appropriate and consistent use of standards and ontologies will improve data integration, as well as enhance the user experience and data use.

Recommendation 2.3: Make the CFDE metadata model, C2M2, scalable.

The CFDE has developed a metadata model, C2M2, to populate the catalog displayed in its portal. C2M2 may have been developed in a way that did not include shared ownership by the DCCs to address their

differing data type needs. A future direction could be modularizing C2M2 such that the model could bring in other data types faster while still adhering to the CFDE's needs to make data findable. This may require relaxing constraints on the model to make it more flexible for multiple types of data generators, tool developers, and data/tool users.

Recommendation 2.4: Improve user experience with search for CFDE data assets.

Data sets and resources should be not only findable and accessible, the CFDE user environment should also be easy to use; in this light the user experience is a key factor. Emphasizing improved user experience and providing easier ways to move across different repositories to access Common Fund data could be highly impactful. The CFDE CC has made good strides in this direction but needs additional clarity and emphasis from the Common Fund. One of the CFDE CC's tasks should be driving toward a consistent user experience and the ability to look across multiple Common Fund programs. Simple but powerful faceted search should be possible. Fuzzy searchability would be useful. Simplifying search is nearly always beneficial for users.

Recommendation 2.5: Show the usefulness of the data and what users can do with them through powerful use cases.

The CFDE has been making Common Fund data sets and tools available on cloud platforms and in cloud workspaces. The cloud is a good way forward because it will increase the findability and accessibility of Common Fund data sets and resources and will enable future flexibility and innovation. However, the notion that researchers will be able to leverage data once they have found them may be too optimistic. Additionally, while many useful tools are available, several of these resources are not known to the sub-communities who could benefit from their use. Usefulness of the data and what users can do with them may need to be shown through powerful publicly available use cases. Additionally, finding the data sets is only the first step. The CFDE portal will also need to have flexible export functionality – potentially including export of raw or processed data, enabling users to pull the identified data sets easily into their analysis platform or to a workspace using the NIH Researcher Auth Service (RAS).

Harmonization and Interoperability

As the CFDE constitutes multiple independent Common Fund studies, incorporating multiple organisms, study designs, -omics measures and platforms, and research communities, the harmonization and interoperability of these factors will impact the approachability and functionality of this resource to the scientific community. Data harmonization refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies. Interoperability allows investigators, computers, and networks to discover, access, integrate, and analyze biological data. Both harmonization and interoperability will allow a more robust and diverse data ecosystem and facilitate data exploration, competition between analytical tools, and quicker scientific discovery. Harmonizing and enabling interoperability of the diverse and extensive data sets and resources available through Common Fund programs is an enormous undertaking. Therefore, to prevent expansive and extensive efforts with limited return on investment, the Working Group recommends the following guidelines for prioritization:

Recommendation 3.1: Similar to findability and accessibility, harmonization incorporating community-supported standards and other data sets is primarily the responsibility of the Common Fund project generating the data and should be addressed starting at the beginning of the program.

As feasible, adopting/recommending/requiring standardized ontologies and schema early in the initiation of a Common Fund program will make it easier to harmonize across Common Fund programs and enable new Common Fund programs to join the CFDE with far less effort. The CFDE should engage Common Fund programs early in their lifecycle to discuss standardized ontologies and schemas that will enable integration of their data into CFDE and provide recommendations and guidance as appropriate.

Recommendation 3.2: The CFDE can aid Common Fund programs by highlighting essential metadata and pointing to pipelines, workflows, and standards adopted by NIH Institutes and Centers, or when lacking, to other international collaborations or standards bodies.

Common Fund studies should strive to make data findable outside of the specific program. To do this, the CFDE should strongly encourage that community-supported metadata, standards, and workflows are in place, consistent, and enforced for data or tool submission.

Recommendation 3.3: When considering the future of the C2M2 model, it should be in a format that will promote the use of knowledge graph models to analyze and explore Common Fund metadata.

Data lead to information, and information leads to knowledge. Knowledge graphs (KG) are a powerful strategy for the integration of seemingly disparate information that may enable the emergence of new knowledge. For optimal utilization, KG rely on the use of ontologies to allow the mapping of diverse data sets that may use different identifiers into a semantically correct, machine-readable, and human interpretable architecture. We recommend that CFDE models enable extensibility allowing for the use of, or integration with, KGs.

Recommendation 3.4: *When developing harmonization and interoperability within the CFDE, the foundational principle should be: "If you build it and it is confusing, people will leave."*

Resources that are difficult to understand, poorly documented, or use uncommon standards will not be readily adopted by the user community the CFDE wishes to engage. Avoid barriers such as the need to go to multiple different sites, the need to run multiple often unintegrated analyses, and results that are difficult to interpret. Data generators and users of the CFDE should be consulted when developing these criteria and subsequent resources.

Cloud Workspaces

Utilization of cloud resources can increase and democratize the findability of Common Fund resources as discussed in the Findability and Accessibility section, especially Recommendation 2.5. The availability of multiple cloud platforms also allows flexibility. It may take time to realize the full potential and advantage of cloud computing, but long-term there is value in the CFDE making resources available in the cloud and providing access to cloud workspaces through which a user can access all CFDE resources. To ensure development and use of cloud resources within the CFDE will be successfully achieved and be maximally useful for a diverse user community, the Working Group provided the following recommendations.

Recommendation 4.1: The CFDE should work to make Common Fund data sets and resources easily accessible in cloud workspaces.

Making CFDE data sets and tools available in a cloud workspace will enhance access to CFDE resources and simplify management and maintenance of data and tool resources. The move to the cloud is a slow

push, and therefore continuing to allow users to work locally or in the cloud will be important for the next few years. The CFDE can either set up its own workspace or utilize existing ones. Drivers for these choices are different and not necessarily in competition with each other. One possibility is a hybrid approach of using cloud-based resources, such as CAVATICA, in collaboration with the NIH Cloud Platform Interoperability (NCPI) initiative while coordinating with other platforms. To the extent possible, making CFDE resources available as Docker or Singular containers so they can be run in both the cloud and locally will enhance usage and usability. Also, the data ecosystem ideally should ensure that data exposed through cloud environments and repositories will have a consistent look-and-feel (i.e., platforms such as CAVATICA; Analysis, Visualization, and Informatics Lab-space (AnVIL/Terra); and other NCPI partners should be consistent and invisible for a user).

Recommendation 4.2: Establish priorities for CFDE cloud workspaces based on clear use cases.

The CFDE will need to balance usability, accessibility, universality (platform independence) with cost, agility, and the maturity and needs of a given research community. This balance is dynamic, changing as the research evolves, the work force changes, and the data, tools, and technologies mature. Having a focus on meeting clear research community-driven use cases will help 'right size' the investment in any particular set of tools, features or workspace technology. There needs to be a balanced approach between having many workspace feature options to accommodate user preferences and building a simple interface that will not overwhelm novice users. A short-term need will be investing in workflows to both simplify and accommodate many data types, as it is not easy for users to bring workflows into the cloud. Similarly, developing new tools and ensuring a mechanism to use tools in the cloud without downloading is another critical issue that will help increase utilization and value of cloud platforms. The balanced approach needs to start with enabling access to data in different cloud environments. However, ensuring users can test tools and resources locally is important, too. Clear, transportable, and reusable use cases can help the CFDE balance and prioritize needs of a diverse end user community and promote utilization of cloud resources by sparking users' interest in cloud resources.

Recommendation 4.3: Establish partnership activities with other large-scale initiatives and programs for cloud workspaces for data types other than the genomics and transcriptomics data, for which there are extant workspaces.

While the focus, appropriately, has been on genomics and transcriptomics, workspaces for other types of data (e.g., image and video data) should also be addressed. The CFDE should work broadly with synergistic data initiatives, including NIH initiatives such as Cancer Research Data Commons and NCPI effort, to understand how they are handling imaging data interoperability. The Cancer Imaging Archive and the Imaging Data Commons are actively looking for data integration use cases and may be receptive to partnering with the CFDE. Cancer Dependency Map's holistic approach to metabolomics and proteomics data is showing the utility of housing and allowing access to such data is another example to consider. A key to this type of framework is that data can be easily visualized within the platform for new users and data sets can be downloaded and processed for more experienced users. As evidenced by the success of research proteogenomic projects (for example, the Clinical Proteomic Tumor Analysis Consortium [CPTAC]), the CFDE may want to increase its outreach efforts to bring together the genomics and proteomics communities to facilitate meaningful dialog rather than developing independent customized workspaces for different data types. Collaboration on cloud workspaces requires a shared vision and investment, difficult in the short-term but valuable and sustainable in the long-term.

Recommendation 4.4: Assess the user base in the data ecosystem to assess the impact of the training and identify groups that are not currently being served by the training.

The CFDE user base is anticipated to be complex, ranging from novice to expert users in one axis to data generators on another axis, and tool developers and other contributors on a third axis. Each user will have different needs, and the CFDE platform needs to enable the entire user matrix. The expert users will push the technical limits of the ecosystem, whereas new and computationally inexperienced users will need more resources and training to successfully develop their own tools or take advantage of pre-assembled workflows. With good use cases and examples, training can enable novice users to use available tools and answer sophisticated questions. In the ecosystem, expert users could build shareable lessons that would enable others to learn from their reproducible workflows. Investing in training, making processes clear, developing illustrative use cases, and enabling any user to contribute reproducible workflows will create a vibrant, sustainable ecosystem. The CFDE should adopt an iterative, user-centered design process for developing and disseminating use cases to inform training activities.

Recommendation 4.5: Cloud credits could be beneficial to recruit users to the cloud workspace.

Providing opportunities for users to work in the cloud, perhaps through the provision of cloud credits, may help create a new cloud-savvy workforce. It is important that potential new users can try cloud computing before fully investing. Providing good training materials, relevant analysis/visualization examples, and use cases are a start, along with advertising where these resources are available in environments where new users can be reached. While large operations might support their work on grants, early-stage projects or those in smaller labs will benefit from "credits" for different types of projects, as each of which would add value in a different way.

It is very important to prioritize cloud credits to small undergraduate institutions or institutions with limited research funds, and to all Minority Serving Institutions (MSIs). The CFDE should engage in targeted outreach to these communities. The goal is to engage groups that otherwise might not have access to the resources and discounts that researchers with long standing NIH-funding receive. Cloud computing can democratize access to CFDE resources, but that democratization will require attention, nurturing, partnership, and training. The CFDE should also consider the K and F awards to identify, advertise, and target incentives for trainees.

Sustainability

The CFDE's primary sustainability challenge is determining how best to allocate resources to align the availability of data assets with their utility to the biomedical research community. Common Fund programs sunset after ten years, so funding to sustain data assets is uncertain after that time. Potential sources of funding for sustainability include NIH Institutes and Centers, the Common Fund, public data repositories, and a fee-based system where end users pay for the maintenance of the data and tools. To address the challenge of sustainability, the Working Group made the following recommendations.

Recommendation 5.1: The CFDE should use uniform metrics across data sets to assess the appropriate level of availability.

These metrics would inform what level(s) (raw, derived) of data to preserve and for what time period. Some types of raw data occupy substantial amounts of disk space and consequently require significant funds to remain readily available. If usage metrics indicate little to no usage of such data, the data could

be moved to cheaper, less available modes of storage. Unused data sets and tools could be moved to cold storage or taken offline altogether, with clear and consistent rules on the sunset of data and notice given to the community.

Recommendation 5.2: Data generators should be responsible for getting data and tools into a shareable format but should not be responsible for sustainment.

As mentioned in previous recommendations (Findability 2.1 and Harmonization 3.1), aligning data and metadata with Common Fund and community standards (making these data, tools, and metadata FAIR) is best performed by the data generators. Once aligned, sustaining data sets ought not be the ongoing responsibility of the program which generated the data. Some data sets may transition to an IC-sponsored repository for long term maintenance. Others may best be preserved in a public repository for data of the corresponding type. Generally, the CFDE will not assume responsibility for maintaining the data and repository infrastructure. In each scenario, the role of the CFDE is to help the data resources implement a sustainability plan which preserves availability and minimizes cost.

Recommendation 5.3: Decisions about the source of funding for sustaining data and tools should be driven by identifying the solution that is least burdensome to the data users and data generators.

Burdens to data access include approval processes, difficult to navigate user interfaces, unharmonized metadata such that searches do not return all data sets that match a query, account creation requirements, institutional affiliation requirements, user fees, and more. Such barriers, while often required by data policy or institutional security requirements, decrease access and usage of data. The CFDE, while remaining compliant with legal and NIH policies, ought to seek sustainability solutions that minimize these barriers to maximize the accessibility of data to users.

Recommendation 5.4: A sustainability model relying solely on user fees for storage and infrastructure should generally be avoided.

User fees should generally be avoided as they can reinforce funding disparities, undermine overall usability and data access, and reduce ad hoc training potential. Fees would run counter to the goals of the CFDE resources and if used should be significantly subsidized. Access to data should follow the path recently taken by scientific journals, which are removing paywalls and providing more free access to their content. Alternative models could include:

- Cloud credits for data use (See Recommendation 7.3 in Context of other NIH Activities section for NIH's use of STRIDES to cover cloud costs);
- Centralization Common Fund data storage analogous to the solution used by the National Library of Medicine; or
- Transfer of data to appropriate public repositories.

Cloud computing resources promise to democratize data by removing the need for significant local computing infrastructure. However, many cloud platforms charge usage-based fees to individual users whose usage exceeds a certain threshold. Users from well-resourced institutions with access to local computing resources are less hindered than users from under-resourced institutions. Some NIH programs offer cloud credits as a means of subsidizing or transferring the costs from the individual user to NIH. When the CFDE offers such credits, the program should prioritize users from under-resourced

institutions as well as new users. See Recommendation 4.5 under the Cloud Workspaces section for additional discussion.

Recommendation 5.6: Tools remain more accessible and usable when they are preserved in portable formats such as container images or open-source, version-controlled repositories.

Many analytical workflows for biomedical research are developed in a single lab for purposes primarily within that research group. The workflows work well in the hands of the tool's creator and close associates, but ease of use often fails outside of the local context in the absence of good documentation and direct interaction. Moreover, retrieving the tool from the generating lab's website is a barrier to findability and accessibility. An alternative approach is to format the tool as containerized images (Docker) and ideally deposit it in an open version-controlled repository (e.g., GitHub, CodeOcean). Containerization ensures that all necessary dependencies are maintained so that established tools can be robust to software updates that might otherwise jeopardize data reuse. Open-source repositories ensure that new and continuing developers can build upon the effort of previous CFDE projects. Tools made available in this manner are more likely to be maintained by the broader researcher community and remain available long after the original author has moved on to other projects. The CFDE should encourage authors to sustain their tools as container images in open tool repositories.

Training and Outreach

The CFDE undertakes training and outreach activities to bring new investigators to work with Common Fund data in the ecosystem and expand the capabilities of those engaged. Outreach efforts should focus on new opportunities provided to data and cloud computing. As part of the outreach efforts to demonstrate the utility of working across Common Fund programs and expose the resources to additional, often novel investigators, the CFDE provides support, via an R03 mechanism, for integrating data sets across three or more Common Fund programs. While integration between independent Common Fund training programs and the CFDE CC is lacking, the first integration steps are now being planned. A key goal of the CFDE training efforts should be enabling diverse users with varying needs and levels of expertise to find, access, and analyze Common Fund data sets, primarily in a cloud environment. The Working Group provided several recommendations for training and outreach to achieve this goal.

Recommendation 6.1: The CFDE should ensure training transitions to cross-program analyses and use cases.

Institutes, Centers, or trans-NIH data science-focused groups such as the Office of Data Science Strategy (ODSS) should provide foundational training. CFDE-delivered training should focus on basic cross-program analyses and use cases, with more specialized training coming later.

Recommendation 6.2: NIH should engage early career researchers, data scientists and engineers, as well as subject matter experts, to expose and train them to actively use biomedical data and tools.

Moving forward, training should be focused on ensuring CFDE tools and resources are easy to use, no matter the user's experience level. As training needs will evolve, it must be responsive and encompass a broad array of analyses that reflect the diversity of data types and workflows in the CFDE. This work will, in part, need to build on common ontologies and data harmonization mentioned above. Engagement with new users will need to include tutorials on data portal usage, directories of data types available,

simple walkthroughs, and basic information on how to search and use workspaces or repositories. Examples of how to adapt workflows and integrate data types with desired analyses along with open-source tools and consistent container images will allow for the flexible utilization of diverse data types within a common general framework.

Recommendation 6.3: CFDE should develop a learning framework with a strategic focus on supporting diverse, dynamic, and scalable training opportunities.

The training framework should outline a plan for ensuring that the trainings are responsive to the research community, with a special emphasis on outreach to early-stage investigators and MSIs (see also Recommendation 4.4 in the Cloud Workspaces section). Understanding usage of CFDE data and resources should inform training and outreach activities. This includes understanding the current state of usage and training, as well identifying new training activities and resources, such as novel use cases, that could encourage users to leverage CFDE data and tools in new ways. As training approaches are implemented, they should be assessed to determine which are successful and should therefore be continued.

Centralizing training will help ensure users can find and access the full spectrum of diverse training opportunities offered through CFDE. Many of the DCCs and CFDE have walkthroughs and tutorials for specific use cases but centralizing these diverse resources for access from a single point of entry would greatly help new platform users. This single point of entry method would also simplify training as shared data and tool access tutorials could be taught once instead of within the context of each sub-project or data set.

The CFDE should emphasize training approaches that are scalable; approaches such as videos, massive open online courses (MOOCs), Frequently Asked Questions, help desks, and other scalable tools should be more of a focus than activities like small workshops. This will help to ensure that trainings can reach the broadest possible audience. Posting classes, materials, videos, and outcomes of workshops to always available public platforms like YouTube, Google Groups, and StackExchange will be impactful.

Recommendation 6.5: The CFDE should consider supporting training needed immediately to ease the transition to cloud-based methods, which will also aid those whose institutions lack sufficient on premises IT infrastructure.

Cloud-based methods are accessible and can help ease the barrier to entry for large-scale analyses. Therefore, more effort needs to be spent on ensuring that sufficient training for new users is in place to improve interactions and the use of cloud resources. Examples of training and outreach that could be used to make cloud-based methods more accessible include codeathons and R25 Research Education Programs. Several DCCs and related NIH resources pointed out how these types of engagement events and options helped build new user bases and drive excitement for projects and analyses within the scope of CFDE resources.

Recommendation 6.6: The CFDE should consider creating a communication and outreach team to engage the research community on the resources and opportunities available in the CFDE.

Community-driven efforts must engage diverse communities to ensure buy-in and adaptive integration of ideas. A key to this and the new training modalities above would be a team that could identify new users, capture shortcomings in current training opportunities, and relate this to the CFDE and DCCs.

CFDE Scope and Strategy in the Context of Other NIH Activities

The CFDE is by no means the only NIH activity addressing data integration and storage. The National Center for Biotechnology Information (NCBI), ODSS, and many Institutes and Centers manage activities like those of the CFDE. While the CFDE is not unique in its mission to make data FAIR, there is no other entity focused on CFDE's mission or Common Fund data sets. In general, the CFDE should adopt established practices from other NIH activities and, where practical, collaborate with teams implementing common policies and procedures. For example, RAS is working to implement a secure, single-sign-on workflow for multiple NIH systems. The CFDE should implement RAS for user authentication and authorization services. To ensure that the CFDE remains aligned with ongoing NIH data science activities and strategically leverages existing or newly developed resources, the Working Group provided the following recommendations.

Recommendation 7.1: The CFDE should incorporate metrics used by other NIH activities to evaluate funding allocation to data resources.

NCBI has used metrics for evaluating their data resources for many years. Tracking publications for data repositories akin to RePORTER grant tracking could help pinpoint where data sets are used. Additionally, data set specific DOIs or other reference annotations could help determine when CFDE data sets are in use in newer applications or used in a publication.

Recommendation 7.2: The CFDE should leverage NCBI and other NIH data resources as long-term repositories for Common Fund data and tools.

Rather than building new resources where established resources of a given data type exist (e.g., NCBI), the CFDE should use the existing resources for long-term storage of its data assets. By doing so, the CFDE will avoid the need to attract users, maintain a unique data resource, and decide whether and when to stop funding the resource. In addition to the tool and data repositories mentioned in previous sections, this work will help to limit/streamline the tasks and effort of the CFDE. In cases where data resources do not yet exist, the CFDE should identify this lack of support and ensure that a community-driven approach is taken to maintain data FAIRness. One option is to explore potential partnerships with NCBI to adopt some of the data resources for long term sustainability.

Recommendation 7.3: The CFDE should utilize resources and infrastructure from the Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) and NCPI initiatives.

As in Recommendation 7.2, the goal of the CFDE is not to reinvent working resources that exist elsewhere in the NIH data ecosystem. To this end, we further recommend that STRIDES and NCPI resources be assessed and, where possible, used within the CFDE. This could be as simple as modeling data structures or extend to using annotations and metadata types previously established. Use of STRIDES and NCPI would also promote reuse of cloud-funding models and established cloud platforms. The added benefit of this would be that resources could be easily transported between larger initiatives.

Conclusion

As biomedical research generates ever-increasing amounts of complex data, both the promises and the challenges of "big data" are expanding exponentially. The big data revolution will deliver novel discoveries across many biomedical research fields, from insights into the most fundamental workings of

the cell to innovations in patient care and treatment. However, to fully realize this potential, new data management approaches will need to be developed and implemented in a strategic, thoughtful, and equitable way.

The CFDE is a critical component of NIH's multi-pronged effort to develop data management strategies to support the broader biomedical data ecosystem. Through an emphasis on high impact, strategic activities to enhance usability and sustainability of Common Fund data sets, and provision of training and support for a diverse user community, the CFDE is poised to greatly enhance the impact of Common Fund programs. The recommendations provided in this report provide a guiding framework for the next phase of the CFDE. We look forward to seeing what the next phase of the CFDE will achieve.

Appendix A: Working Group Roster

WORKING GROUP CO-CHAIRS

Rick Horwitz, Ph.D.
Executive Director
Allen Institute for Cell Science

Elizabeth Wilder, Ph.D.
Director
Office of Strategic Coordination
Division of Program Coordination, Planning, and Strategic Initiatives, NIH

WORKING GROUP MEMBERS

Sergio Baranzini, Ph.D.
Professor of Neurology
Weill Institute for Neuroscience
University of California, San Francisco

Devin Schweppe, Ph.D.
Assistant Professor of Genome Sciences
University of Washington

Emiley Eloe-Fadrosh, Ph.D.
Metagenome Program Head
DOE Joint Genome Institute
Lawrence Berkeley National Laboratory

Bruce Weir, Ph.D.
Professor of Biostatistics, Epidemiology, and Genome Sciences
University of Washington

Warren Kibbe, Ph.D., FACMI
Professor in Biostatistics and Bioinformatics
Duke University School of Medicine

Cathy Wu, Ph.D.
Unidel Edward G. Jefferson Chair in Engineering and Computer Science
Director, Center for Bioinformatics and Computational Biology
University of Delaware

NIH STAFF

Stephanie Courchesne-Schlink, Ph.D.
Designated Federal Official
Senior Advisor
Office of Strategic Coordination
Division of Program Coordination, Planning, and Strategic Initiatives, NIH

George J. Papanicolaou, Ph.D.
CFDE Program Leader
Office of Strategic Coordination
Division of Program Coordination, Planning, and Strategic Initiatives, NIH

Christy Kano, Ph.D.
CFDE Project Manager
Office of Strategic Coordination
Division of Program Coordination, Planning, and Strategic Initiatives, NIH

Haluk Resat, Ph.D.
CFDE Program Leader
Office of Strategic Coordination
Division of Program Coordination, Planning, and Strategic Initiatives, NIH

Chris Kinsinger, Ph.D.
CFDE Program Leader
Office of Strategic Coordination
Division of Program Coordination, Planning, and
Strategic Initiatives, NIH

Appendix B: Meeting Agendas

KICKOFF MEETING

CoC Working Group for CFDE | July 8, 2021 | 11:00 am – 1:00 pm EDT

11:00 – 11:15	Introductions	CFDE CoC Working Group Members
11:15 – 11:40	Charge to the Group ● *Questions from the Group*	Jim Anderson, DPCPSI Director
11:40 – 11:50	Overview of the Common Fund	Betsy Wilder, OSC Director
11:50 – 12:15	Overview of the CFDE ● *Goals, Current Activities*	Chris Kinsinger, CFDE Program Leader
12:15 – 12:30	Initial thoughts from Co-Chair ● *Questions and Discussion from the Group*	Rick Horwitz, Working Group Co-chair
12:30 – 1:00	Moving Forward ● *Review/discussion of proposed schedule of work; discussion of how the group will operate*	Rick Horwitz and Betsy Wilder

Main points from presentations

- The CFDE Working Group of the Council of Councils (CoC) will produce a report and a set of recommendations to be delivered to the CoC in May 2022. The CoC will vote on whether to accept the report and recommendations; if accepted, the CoC will provide the recommendations to the NIH Director. The CoC does not change the report but can include additional comments.
- This Working Group will consider the scope and direction of the CFDE; the primary focus is on providing recommendations about the CFDE for the Common Fund, but there may be broader implications for data science and management across NIH.
- The Working Group charge is to review the scope, goals, and progress to date of CFDE and issue recommendations around findability and accessibility of data; data harmonization and interoperability; cloud workspaces; sustaining access to data and tools after CF programs end; training and outreach to enhance access to and use of the data; and CFDE scope and strategy in the context of related NIH activities.
- The Common Fund is intended to catalyze research across the NIH, and programs often develop tools and resources. The creation of large and powerful data sets is a common theme of many Common Fund programs. About 2/3 of CF programs involve generation of significant data resources.

- The CFDE aims to enable users to query across and use multiple CF data sets, sustain CF data and tools, and train users to work with CF data. The CFDE began with an initial award to the CFDE Coordinating Center (CFDE CC) in 2019 to the University of Maryland, Baltimore.
- In the fall of 2019, the CoC approved a concept for the CFDE to engage DCCs to build the ecosystem in a three-year pilot phase. In 2020, awards were made to eight DCCs, as well as 12 R03 grants to researchers to enhance the utility of CF data sets. The annual budget for the CFDE is approximately $14 million, with additional funds for R03 data utility projects added on from CF end-of-year funds.
- Internal governance of the CFDE includes a Steering Committee, composed of the principal investigator from the CFDE CC and each of the DCC awards, which coordinates activities across the CFDE. There are also Technical Working Groups that focus on different cross-cutting issues for the CFDE as a whole and provide recommendations to the Steering Committee.
- External governance of the CFDE includes the NIH CFDE Program Team, which coordinates the vision of the CFDE, represents the interests of stakeholders, and keeps the project on track. Stakeholders who also provide input into the CFDE (through the NIH CFDE Program Team) include the CoC and this Working Group, DPCPSI and OSC leadership, a trans-NIH Working Group of data science experts, and other NIH and CF programs.
- The main output of the CFDE is the CFDE portal, which allows users to find relevant Common Fund data sets by searching across metadata. CFDE is also making efforts to harmonize the actual data, beginning with the Kids First and Genotype-Tissue Expression (GTEx) programs (gene expression data) and Human BioMolecular Atlas Program (HuBMAP) and Stimulating Peripheral Activity to Relieve Conditions (SPARC) programs (imaging data).
- The CFDE is developing a systematic DCC program lifecycle. This lifecycle will include integration into the CFDE from program inception throughout the entire funding cycle and will also develop plans for sustaining data at the appropriate level of availability based on utility.
- Within the CFDE CC, there is a Training and Engagement Center tasked with leading efforts to train users to work with data in the cloud environment. CFDE is actively considering how to find the right balance of showcasing CF data sets within a general cloud platform webinar. Many DCCs are doing similar activities but historically have been operating independently, and the Training and Engagement Center is helping to create an environment to share wisdom and lessons learned across DCCs.

CURRENT STATUS OF CFDE

CoC Working Group for CFDE | August 12, 2021 | 11:00 am – 1:00 pm EDT

11:00 – 11:15	Initial Discussion of Background Materials	CFDE CoC Working Group
11:15 – 11:30	CFDE Coordinating Center Activities and Status	Owen White, CFDE Coordinating Center Principal Investigator
11:30 – 12:15	DCC Engagement and Partnership Activities	
	● *Library of Integrated Network-Based Cellular Signatures (LINCS)*	Avi Ma'ayan, LINCS DCC Principal Investigator
	● *The Human BioMolecular Atlas Program (HuBMAP)*	Phil Blood, HuBMAP DCC Principal Investigator
	● *Gabriella Miller Kids First (Kids First)*	Adam Resnick, Kids First DCC Principal Investigator
12:15 – 12:30	CFDE Training and Outreach	Titus Brown, CFDE Training and Outreach Coordinator
12:30 – 1:00	Closed Session Discussion and Next Steps	CFDE CoC Working Group

Main points from presentations

- The CFDE Coordinating Center (CC) has developed a "socio-technological infrastructure" to rapidly adapt to program goals without disruption; they are providing a central mechanism to coordinate across independent Data Coordinating Centers (DCCs)
- The CFDE portal enables users to discover data across multiple Common Fund data sets; in the future, it is looking to build in resources to help users use the data. Several activities are being piloted – a DRS service with persistent IDs to fetch data, and a cloud workspace pilot.
- The CFDE portal allows users to find the relevant CF data sets by searching across metadata, which is encoded using its Cross-Cut Metadata Model (C2M2). Search tool uses faceted query.
- CFDE CC is planning to install shopping cart feature for the users to create logs of data sets identified in query. There are also plans for allowing the users to pull data from the shopping cart.
- CFDE is implementing the Researcher Auth Service (RAS) to access resources. RAS will feature a single sign-on, use authorizations from NIH dbGaP Data Access Committee (DAC) decisions, link and manage accounts from multiple identity providers, and use multi-factor authentication for data repositories that require a higher level of access security.
- The CFDE CC noted that they are planning to create a resource registry, which would allow external linking of tools. This registry is envisioned as starting fairly small, and would expand if feasible, but would need careful management to avoid becoming overwhelming to the user.

- CFDE is leveraging the NCPI/RAS/DRS framework for controlled access data.
- DCCs are partnering with each other and with the CFDE CC around specific projects. Examples include LINCS moving data and tools to the cloud, along with work to develop Appyters to develop and sustain bioinformatics tools and workflows; a collaboration between HuBMAP and Kids First to integrate data for gene burden testing; a collaboration between HuBMAP and SPARC to integrate spatial information to enable interoperability across anatomy and cell-type Common Fund resources; a collaboration between Kids First, HuBMAP, and GTEx to identify new, tumor-specific targets; a collaboration between Kids First, LINCS, and IDG to identify cellular signatures similar to disease signatures, then look at potential therapeutics based on LINCS application of drugs; and a Kids First effort to resolve difference in generation of RNAseq data to enable users to bring data from different sources together.
- The CFDE also conducts training and outreach as part of its scope (explored more fully in a later meeting). A key goal is to identify what motivates users, especially with respect to cloud use.
- Training and outreach incorporate social feedback cycles, where feedback is incorporated into future trainings, use cases, and portal function.
- All CFDE CC training materials are open access, and materials are mostly focused on the introductory level.

DATA HARMONIZATION AND INTEROPERABILITY

CoC Working Group for CFDE | September 9, 2021 | 11:00 am – 1:00 pm EST

11:00 – 11:15	Initial Discussion of Background Materials	CFDE CoC Working Group
11:15 – 11:20	Overview of Harmonization Approaches	Haluk Resat, CFDE Program Leader
11:20 – 12:05	Existing Harmonization Efforts within CFDE at the Data and Metadata Level	
	● *Efforts to Harmonize Genotype Tissue Expression (GTEx) and Kids First (KF) Data sets*	Francois Aguet, GTEx investigator
	● *Ontology Working Group*	Michelle Giglio, CFDE Ontology Working Group member
	● *Evolution of C2M2 Model*	Owen White, CFDE Coordinating Center Principal Investigator
12:05 – 12:30	Additional Approaches - Knowledge Graphs	Sergio Baranzini, CFDE CoC Working Group Member
12:30 – 1:00	Discussion and Next Steps	CFDE CoC Working Group

Main points from presentations

- For CFDE, harmonization has been focused on metadata/standards and data. Interoperability has focused on workflow and compute platforms but has not yet involved study design and raw data processing, or AI, but CFDE anticipates being involved in these in the future.
- Efforts to harmonize raw and called data (e.g., KF/GTEx RNA-seq) are complicated by differences in cell-type composition, cell state, as well as differences in molecular biology techniques; these differences must be computationally addressed.
- Support for harmonized data sets may be required in the future as common builds and annotations change, with users expecting/needing some level of support to aid analyses between 'omics and Common Fund programs.
- Ontologies should be: 1) stable, but not static, 2) actively developed, 3) have a mechanism for requesting new terms and ontology changes (e.g., GitHub), 4) be responsive to requests and questions, 5) have community buy-in, 6) conform to community conventions on ontology/vocabulary development, and 7) provide mappings to other related ontologies/vocabularies, as relevant
- The evolution of C2M2 has primarily been driven by use cases and asking DCCs what they are interested in
- C2M2 has been updated to include genes, phenotype, clinical data, event modeling, provenance and will next incorporate modeling gene-disease-phenotype-anatomy relationships, clinical metadata, analysis methods, genes, and data use restrictions

- Data sets like the ones generated by Common Fund are amenable to integration into knowledge graphs through the levels of biological complexity, sample by sample across data sets, and finally, abstractions at the information level

CLOUD WORKSPACES

CoC Working Group for CFDE | October 14, 2021 | 11:00 am – 1:00 pm EST

11:00 – 11:15	Closed Initial Discussion of Background Materials	CFDE CoC Working Group
11:15 – 11:20	CFDE Cloud Workspace Pilots Introduction	George Papanicolaou, CFDE Program Leader
11:20 – 11:30	CFDE Cloud Workspace Pilot Activities ● *Cloud Credits Lessons Learned* ● *Cavatica in the Context of CFDE and End User Experience*	Adam Resnick and Jack DiGiovanna, Kids First DCC Principal Investigators
11:30 – 11:40	Research Auth Service Initiative Implementation	Rebecca Rosen, National Institute of Child Health and Human Development, NIH
11:40 – 11:50	Overview of the NIH Cloud Platform Interoperability Effort ● *Overview* ● *Mission and Objectives* ● *Future Directions*	Valentina di Francesco, National Human Genome Research Institute, NIH
11:50 – 12:30	Discussion	CFDE CoC Working Group
12:30 – 1:00	Closed Session Discussion and Next Steps	CFDE CoC Working Group

Main points from presentations

- Benefits of a workspace include minimizing data egress, providing access to Common Fund tools in an optimal computational setting, making standardized analytical workflows available, ability to highlight harmonized data sets for further inquiry, providing an environment for easy provenance tracking, and incorporating a pre-structured cost structure.
- To explore whether providing a workspace platform for analysis would increase the utility of Common Fund data sets and tools, CFDE is supporting a pilot workspace through Kids First's on the Cavatica platform.
- Based on Kids First experience, for researchers that do use cloud, they do it because it saves time, it scales, and it is a managed environment that obviates the need to download large amounts of data from multiple sources. Of these, the biggest driver is saving time.

TRAINING

CoC Working Group for CFDE | November 8, 2021 | 1:00 pm – 3:00 pm ET

11:00 – 11:15	Closed Initial Discussion of Background Materials	CFDE CoC Working Group
11:15 – 11:20	Training Overview and Challenges	George Papanicolaou, CFDE Program Leader
11:20 – 11:30	Training Activities at CFDE-CC	Titus Brown, CFDE Training and Outreach Coordinator
11:30 – 11:40	Training and Outreach Activities at DCCs ● *LINCS Massively Open Online Course (MOOCs)* ● *SPARC Code-a-thons*	 Sherry Xie, LINCS DCC Susan Tappan, SPARC DCC
11:40 – 11:50	R03 Awardee: Improving Deposition Quality and FAIRness of Metabolomics Workbench	Hunter Moseley, CFDE R03 Principal Investigator
11:50 – 12:30	Discussion	CFDE CoC Working Group
12:30 – 1:00	Closed Session Discussion and Next Steps	CFDE CoC Working Group

Main points from presentations

- Training in CFDE has strengths (i.e., partners have expertise on the content of the trainings using existing, familiar materials) and weaknesses (i.e., skill level, topic, and mastery progression are not coordinated across partners, and the trainings are not tailored to integrate cross-Common Fund data sets).
- Training at the CFDE CC is currently focused on biomedical data scientists (as they are readily able to explore new data integration opportunities and new use cases). It will next move to biomedical and clinical researchers, who will benefit from resource registries and directed types of data.
- LINCS conducts multiple training activities such as a ten-week summary research program for undergrads and master students and Coursera MOOCS (22K enrolled), one of which focuses specifically on LINCS data and tools.
- SPARC has hosted two codeathons, with prizes from $3K to 20K. Lessons learned include: 1) virtual formats allow for international collaborations, 2) short and intense activity is preferred, and 3) bootcamps for the computational environment can identify areas of difficulty and confusion, and 4) projects directed towards making the data more FAIR were well received by SPARC and participants.
- R03s provide opportunities for non-Common Fund scientists to use the data and provide opportunities in resource development, support for early and mid-career scientists and their grant-funded students, and feedback to Common Fund programs and NIH ranging from documentation, data cleaning, and how the data meets with FAIR principles.

SUSTAINING DATA AND TOOLS

CoC Working Group for CFDE | December 9, 2021 | 11:00 am – 1:00 pm EST

11:00 – 11:15	Closed Initial Discussion of Background Materials	CFDE CoC Working Group
11:15 – 11:25	Overview of Current Sustainability Plans for CFDE-Participating Common Fund Programs	Chris Kinsinger, CFDE Program Leader
11:20 – 11:30	Usage Metrics ● *LINCS* ● *ExRNA Communication*	 Avi Ma'ayan, LINCS DCC Principal Investigator Aleks Milosavljevic, ExRNA Communication DCC Principal Investigator
11:30 – 11:40	Generalist Repositories	Ishwar Chandramouliswaran, Office of Data Science Strategy, NIH
11:40 – 11:50	Pilot User Fees to Support CFDE Data & Resource Sustainability	George Papanicolaou, CFDE Program Leader
11:50 – 12:30	Discussion	CFDE CoC Working Group
12:30 – 1:00	Closed Session Discussion and Next Steps	CFDE CoC Working Group

Main points from presentations

- 4 models have been used to sustain Common Fund data to date:
  - An NIH Institute or Center (IC) steps up to maintain a data set
  - Data generators receive a competitive award to continue the resource
  - Data are transferred to public repositories
  - CFDE stores the data in the cloud
- Portions of the LINCS data are available through the DCC while others are available at data generator sites.
- The ExRNA sustainability plan relies on successfully competing for NIH funding for sustaining data resources.
- Reusability of tools requires access to both data and tools
- NIH encourages researchers to share data using domain-specific repositories when available. When these are not available, NIH is developing options to support data sharing, including through PubMed Central (stored supplementary materials and data sets), use of generalist repositories, and STRIDES cloud partners.
- The HEAL (Helping End Addition Long-term) program uses data stewards to work with data coordinating centers and data generators from across the HEAL ecosystem to find optimal repositories for individual data sets

CFDE IN THE BROADER NIH CONTEXT

CoC Working Group for CFDE | January 13, 2022 | 11:00 am – 1:00 pm EST

11:00 – 11:15	Closed Initial Discussion of Background Materials	CFDE CoC Working Group
11:15 – 11:20	Overview: CFDE in the Broader NIH Context	Haluk Resat, CFDE Program Leader
11:20 – 11:30	Office of Data Science Strategy (ODSS)	Susan Gregurick, Office of Data Science Strategy, NIH
11:30 – 11:40	National Center for Biotechnology Information (NCBI)	Kim Pruitt, National Library of Medicine, NIH
11:40 – 12:20	Discussion	CFDE CoC Working Group
12:20 – 1:00	Closed Session Discussion and Next Steps	CFDE CoC Working Group

Main points from presentations

- The Office of Data Science Strategy (ODSS) and the National Center for Biotechnology Information (NCBI, part of the National Library of Medicine) are not envisioned as the all-encompassing trans-NIH data repository or trans-NIH data platform
- ODSS is implementing the Strategic Plan for Data Science, which includes multiple initiatives around FAIR data sharing
- NCBI manages domain-specific repositories and generalist repositories as well as several metadata databases that support primary data.
- NCBI evaluates funding allocation to data resources based on usage metrics of the resources.

DISCUSSION OF REPORT OUTLINE

CoC Working Group for CFDE | February 10, 2022 | 11:00 am – 1:00 pm EST

11:00 – 11:30	General discussion of draft report outline	CFDE CoC Working Group
11:30 – 12:30	Discussion of specific issues needing clarification	CFDE CoC Working Group
12:30 – 1:00	Finalizing input and next steps	CFDE CoC Working Group

REVIEW DRAFT REPORT

CoC Working Group for CFDE | March 10, 2022 | 11:00 am – 1:00 pm EST

11:00 – 11:30	Draft Synthesis/Conclusions section	CFDE Working Group
11:30 – 12:30	Discussion of draft report	CFDE Working Group
12:30 – 1:00	Finalizing input and next steps	CFDE Working Group

PLAN FINAL REPORT PRESENTATION

CoC Working Group for CFDE | April 14, 2022 | 11:00 am – 1:00 pm EST

11:00 – 11:30	Finalize draft report	CFDE Working Group
11:30 – 12:30	Discuss presentation to CoC	CFDE Working Group
12:30 – 1:00	Final issues/Next steps	CFDE Working Group