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

October 2019

Introduction

The Common Fund Data Ecosystem (CFDE) project's charge is to collaborate with Common Fund programs to improve the value of Common Fund data sets by enabling re-use of datasets by typical biomedical researchers. In 2019, we are profiling the content and status of nine Common Fund Programs that were identified as priorities for engagement by Common Fund leadership. This is the second of a three-report series. Much of the initial report emerged from site visits to the Data Coordinating Centers (DCC)/Data Resource Centers (DRC) of four Common Fund Programs- HMP, LINCS, Kids First, and GTEx - and in that report we provided an initial set of recommendations for 2020 activities.

In this update, we report on site visits to two additional programs (HuBMAP HIVE and the Blackfynn DRC for SPARC). In the main text, we have provided summaries of both programs and highlight new opportunities and challenges that have emerged from our visits. We have worked to synthesize the insights gained across site visits into actionable challenges and describe our plans to begin addressing them. These seven approaches are not full-fledged solutions, but rather first steps in operationalizing the CFDE and moving toward a data

ecosystem. In Approach 1: Data Federation, we describe how simple tools such as asset inventories and manifests can improve data FAIRness as well as help the Common Fund to participate in any future trans-NIH interoperability initiatives. These future considerations are more fully addressed in Approach 5: Federating with Data Resources External to the Common Fund.

Approach Two: the CFDE portal, will allow anyone, for the first time, to see a full inventory of Common Fund datasets, and the full spectrum of possibilities for data re-use. Training, and Addressing Data Incompatibility, our third and fourth approaches, will help to ensure that end users are able to make use of the current Common Fund assets, and that they are ready to dive into whatever possibilities the future holds once the portal makes it feasible for users to actually find and reuse data.

Approaches Six and Seven focus on interactions between CF Programs, the CFDE, and the Common Fund leadership with the goal of improving the flow of information and building community. Approach Six seeks to answer specific questions of how Programs decide on cloud or local infrastructure, and how these choices might impact sustainability. Approach Seven describes how the tenor of site visits have shifted over time, and how we expect they might continue to evolve.

In our third and final report for 2019, we expect to include site visit reports for the remaining three Common Fund programs in the CFDE pilot scope - MoTrPAC, 4D Nucleome, and Metabolomics, as well as updated perspectives on our 2020 plans.

Assessment: Recent Deep Dives

With the addition of HuBMAP and SPARC, the CFDE Engagement team has now met with six Common Fund Programs. These six programs span the entire breadth of the Common Fund funding lifecycle, ranging from HMP (funding ended) to HuBMAP (just now releasing their first data sets).

We have continued to use the same “deep dive” format for our site visits, and our site visits continue to be both incredibly informative and very productive. Although we continue to use the same general agenda to organize each meeting, the content and tenor of each engagement is unique and the style of our meetings has evolved over time; we discuss the implications of this below, in Operationalizing the CFDE.

HuBMAP. Of all the Programs we have visited, HuBMAP is both the earliest in their lifecycle, and the most organizationally complex. The HuBMAP Integration, Visualization, and Engagement (HIVE) Collaboratory, the main organizational unit for integration and dissemination of HuBMAP data, is itself a coalition of five organizations. Together, they oversee the work of a Tissue Mapping Centers and Innovative Technologies Groups, all working to create a variety of cell maps and analysis approaches. The HIVE has been working diligently for

about a year to set up working groups, determine governance, and set up the infrastructure that is vital to the organization. While HuBMAP is still in early stages, they expect to transition to hosting data in the coming year, and their infrastructure design is unique among the sites we have visited so far. As members of the HIVE have a great deal of expertise in running High Performance Computing Centers, they have chosen to use a hybrid infrastructure where much of the data and compute is local to the Pittsburgh Supercomputing Center, with the ability to 'burst' into the cloud for larger jobs. HuBMAP expects that their hybrid system with mostly on-premises compute will save hundreds of thousands of dollars over the life of their program. The complete HuBMAP report can be found in Appendix A.

SPARC. The SPARC data portal, designed and operated by Blackfynn, is conceptually very different from the portals of other Common Fund Programs, and tries to incentivize data creators to deposit data as it is generated. While the portal can be used to discover datasets, it is primarily designed as a data management system. Users can store not only the raw data files, but analysis, notes, presentations, and almost anything else. By positioning themselves as a place where data generators can organize data, metadata and supporting documentation for their own day-to-day use, SPARC hopes to make all of their data more FAIR and encourage good data stewardship. All of the portal infrastructure and approximate 10TB of data is already hosted on the Amazon cloud (AWS), and they expect to have over 100TB of data by the end of 2020. SPARC expects that the reliability and flexibility of using cloud services will let them scale their program in a sustainable way, and allow them to focus their time and energy on creating innovative user experiences rather than maintaining servers. The complete SPARC report can be found in Appendix B.

Opportunities and Challenges, Summarized

The opportunities and challenges reflected in this section are a synthesis of what we have learned from the deep dive sessions described in the July report as well as the two recent visits with SPARC and HubMAP.

FAIRness of data is not inherent in hosting data on the cloud. The main outcome of our July assessment was the clear realization that the datasets hosted by the DCCs are not inherently interoperable, and **placing their assets in the cloud does not intrinsically solve the problems of findability, accessibility, interoperability, and reusability**. What is clear now, is that a progressive series of challenges must be addressed in order to achieve the goal of making Common Fund data more FAIR. The first challenge is that there are no clear guidelines for how data can be made FAIR. "We believe in FAIR", said one member of a DCC, when asked what they thought about the term -- but it was obvious from the response that while their daily lives revolve around increasing all aspects of FAIRness of their data, they sincerely did not have any other than their own set of subjective measures for FAIRness of their data. This challenge is addressed in CFDE Operationalization Approach 1, where we will provide clear, objective metrics for all Programs to follow in order to increase FAIR levels of Common Fund data.

Another challenge is to make the data derived from the portfolio of Common Fund programs more findable and accessible. Each of the Common Fund programs we visited has (or will have) large quantities of high-value data assets that can be found via their website. Those assets can be viewed, analyzed, and downloaded at each of their individual portals. For example, GTEx has many tissue-specific RNAseq data sets that can be used to compare gene and isoform expression between normal tissues. However, no end user, or even NIH program manager, is able to locate all Common Fund assets in a single system, nor is there an easy way to determine whether any given dataset exists. For example, for a Kids First user to examine a gene's expression in tumor tissue relative to normal, the Kids First user would first need to know that GTEx has relevant normal tissue samples in order to look for them. Knowing what data exists, and where to find it, would be a huge breakthrough for many Common Fund researchers. This challenge is addressed in Approach 2: CFDE Portal Implementation

A third challenge is that once we overcome issues relevant to findability and accessibility of data, interoperability is contingent on the data sets being combined between at least two Programs, and the ability of users to transport those datasets to analysis tools. This challenge is addressed in Approach 2: CFDE Portal Implementation, which will allow users to combine datasets from multiple Programs, and in Approach 3: Training, which will instruct users on how to use several analysis systems.

A final challenge is that interoperability is not always desirable. Many data curators are wary of efforts that might make incompatible datasets interoperate, and have raised major concerns (see especially the GTEx and HuBMAP deep dives). The two major concerns are that:

1. Not all data sets may be usefully interoperable, and the cost of making them interoperable may be prohibitive, especially in the absence of well-defined use cases.
2. Successful data integration requires a talented and motivated user with a deep understanding of the data, which would necessitate working with the original data sources. Or, to rephrase, the further analysts are from the origin of the data, the more likely they are to misuse it.

These concerns will be addressed in *Approach 4: Addressing Data Incompatibility Concerns.*

<u>More researchers must be enabled to reuse Common Fund data.</u> One person can make anything work, if they try hard enough and/or have enough tech support. The value of operationalizing FAIR is that it will enable many people to do analyses that were previously only available to expert bioinformaticians, and will advance the economy of scale that comes from investing in solving problems that affect many people. Our FAIRness metrics should reflect this. For example, a small number of individual high-impact papers is less valuable to the community than many papers that make opportunistic (and perhaps small) use of Common Fund data sets. The challenge of enabling many researchers will be addressed in *Approach 3: Training*.

Common Fund must prepare for a future of federation and interoperation. There are a number of initiatives and opportunities for interoperation, including within the Common Fund (GTEx/Kids First) and outside the Common Fund (GTEx/Kids First/ANViL, HuBMAP/HCA). These are areas where the Common Fund can prepare for the future by ensuring that efforts and standards emerging from the Common Fund are not incompatible with likely NIH (ODSS), national, and global (GA4GH) standards. More, the Common Fund should work towards implementations of these same standards in current efforts where possible, so that over time Common Fund programs gain interoperability with each other as well as more globally. This challenge will be addressed in *Approach 1: Data Federation and Approach 5: Interoperability Beyond the Common Fund.*

The Common Fund must plan for "catastrophic success". A continued message from the Programs is that increasing data reuse will lead to an increased support burden as well as increased costs for compute. For this reason, we expect to need to elaborate our recommendations in this area in the future, including recommendations for broad training, and tier 1 help desk support, as well as on flexible compute options that do not burden the data centers with increased costs as their user base increases. We will need in-depth usage information from data coordinating centers and a full release of the CFDE portal to determine how best to build out these recommendations, therefore these activities will occur beyond FY2020. We will work to understand the scope of this challenge through *Approach 3: Training*, and *Approach 6: Assessing the Optimal Balance of Cloud Versus On-Premises Computing.*

The balance of cloud-based capability versus local (on-premises) computing is unclear. New Common Fund Programs are increasingly faced with the decision of whether to build an on-premises solution or develop in the cloud. For example, HuBMAP and SPARC provided quite different perspectives on how to host infrastructure. HuBMAP is using on-premise infrastructure to provide a cost-effective hosting solution, while the SPARC Data Core is completely committed to cloud hosting. This is a complex decision that relates not only to the needs of each individual program, but also to the long-term sustainability of the data resources, and NIH plans. A particular challenge is that although the trend in the Common Fund is towards cloud based solutions, it is unclear whether it is mandated, and the costs of switching a project like HuBMAP to the cloud mid-project would be immense.

The SPARC platform produced by Blackfynn is entirely hosted on the Amazon cloud (AWS). The SPARC Portal is hosted through Heroku (https://heroku.com), which in turn is leveraging AWS. They chose to use a 100% cloud-based model for a number of reasons, mostly having to do with the flexibility of the cloud. Disk drives inevitably fail, and the amount of extra processing power, download or upload speed, and disk space that the project needs fluctuates depending on what users are doing. Determining up front how to account for those needs is difficult and error prone, however, when using cloud services, malfunctioning disk drives automatically fail over to working ones, and processes can scale almost infinitely. They also don't have to worry that their internet connection is stable or has enough bandwidth — a large proportion of

Amazon's servers around the world would have to be impacted before it would affect the SPARC Portal.

HuBMAP's data center, based at the Pittsburgh Supercomputing Center which has long been the home of XSEDE, mostly uses computers that are hosted at their own local facility (as opposed computers hosted at Google or Amazon). At their center, they have many years of expertise in both designing and maintaining High Performance Computing Clusters, and so are able to dramatically lower their computing costs by leveraging those abilities. They will still have some cloud computing capability, in particular they plan to 'burst' into the cloud for very large compute jobs, and they are hoping their service will be seamless. That is, a user will find working on HuBMAP's servers functionally indistinguishable from working in something like AWS, and if their job expands or shifts into the cloud, their experience will remain completely the same.

Although HuBMAP's hybrid infrastructure is likely to work well for them, it is not likely to work for all Common Fund programs. While getting contracts for inexpensive storage is relatively easy, the services required to keep them secure and running smoothly require specialized knowledge of the underlying site infrastructure -- expertise that is not required if DCCs use Google or Amazon -- and the equipment (e.g., servers, backup systems and disk space) depreciate over time and must be replaced. In general, for smaller systems at organizations without an existing infrastructure, cloud solutions provide a cost-effective way to implement a robust system without a large, upfront hardware and IT investment.

There is also a question of long-term data maintenance. As we have seen with the Human Microbiome Project and LINCS, data stored on local servers are subject to the infrastructure demands of the facility they are stored at. If the data is to remain local, as infrastructure is retired, new servers will need to be purchased to replace the old, and it's not clear who pays for this once a Program has reached the end of it's funding, or if the hosting organization would even allow the NIH to effectively rent space there for a Program that is no longer active. Of course, the challenge of long-term support is also true for cloud hosted data, which requires constant payments to keep running. The difference lies in the logistics of data access. Data that is already in the cloud is generally more expensive to maintain over its lifetime, however it is relatively easy for the NIH to take over custodianship of at the end of a given program, even if the program is unexpectedly cut short. Local data will be less expensive, but difficult or impossible to maintain on-site at a de-funded program facility, and will take time, likely weeks or months, to migrate to the cloud for NIH custodianship.

Some approaches to address these concerns are in *Approach 6: Assessing the Optimal Balance of Cloud Versus On-Premises Computing.*

Outreach and Ecosystem Building require careful social interactions. Our July report outlined the critical role that our site visits play in understanding the opportunities and challenges for each Program, as well as establishing a relationship with the DCC/DRC. The site visits continue

to be important! However, as our July report is publicly available and many of the PIs had taken the time to read it, our hosts often had many more questions for us than in previous engagements, and were more likely to tailor their presentations to reflect on how their programs compare to what we had written.

Unexpectedly, this pre-knowledge of our recommendations served as both an expedient and a hindrance to each meeting. In cases where the challenges of our hosts were already reflected in our July report, their questions tended towards those issues and how our hosts could get involved. In cases where our report did not resonate with our hosts, the meetings were largely focused on exploring and clarifying our recommendations. Moreover, there is a tendency to regard our July recommendations as relatively fixed. This challenge will be addressed in *Approach 7: Changing Role of Site Visits.*

The cost of hosting and managing data must be addressed. The SPARC Data Core told us that the biggest threat to sustainability is the misconception that "Open Data is Free Data." There will always be costs associated with data download, storage, transfer and analysis, and while it is possible to make data available for free to users, that increases the cost of hosting the data, which is of particular concern after the Common Fund Program that generated the data has ended. Sustainability is defined by finding the best way to distribute these costs. Given the scale of the current and planned Common Fund data sets (10s of TB to PB of data), data storage, downloads and computing will be very expensive. Multiple Common Fund programs are struggling with issues of egress charges for the cloud, how much compute cost to provide for free (and for whom), and how to enable inexpensive compute close to the data for those they cannot support.

One specific example is GTEx, which provides 100 TB of raw RNA-seq data in their v8 release. There are three modes that users can interact with this data: 1) to use the visualization tools on the GTEx website; 2) to download the data to perform local analysis; and, 3) to transfer the data to another cloud-based system for analysis. The GTEx visualization tools are very sophisticated, however they answer specific questions, and unfortunately many users still must resort to option 2, to download GTEx data to their local computers. The egress charges for such a download are approximately $8500, which is prohibitive for many researchers. Option 3, performing analysis on the cloud, avoids egress charges for most users, and compute costs could be offset by providing free or hosted compute to internal program users, while external users would pay for their own compute. However, these solutions do come with their own challenges. One significant challenge is that relatively few biomedical data scientists have experience in doing their analysis in the cloud. Training is vital for these researchers, as trivial scripting errors can sometimes result in huge compute costs. A related challenge is that easier to use platforms for analyzing data in the cloud, such as Terra or Cavatica, may not be robust or flexible enough to meet users' needs. And, last, avoiding egress charges still means that researchers must pay for cloud computing costs to perform analysis, and so far, universities generally do not have good mechanisms for paying these costs.

Members of the SPARC Data Core noted “Accessible to us means the user should be able to get many Terabytes of data available to them in an easy and scalable way.” This concern must be addressed by arriving at a robust mechanism to either assist users with cost of data egress, or to provide users with some form of inexpensive computing analysis capability. This has been addressed by at least three Common Fund DCCs who are enabling users to access large-scale data assets and provide their own cloud-based analysis services to their users. For example, the Kids First DRC is making use of Cavatica, GTEx provides their pipelines via Terra, the SPARC Data Core has initiated an effort to link their data with Amazon. HubMAP is also planning flexible and robust compute via their data center as well as export to cloud services. These solutions will help, but they will always require an investment of funds from NIH.

A final important cost issue is demonstrated by GTEx, which hosts protected data that can only be served from a FISMA-compliant repository with appropriate authentication and authorization. While those costs have been shifted to NHGRI/ANViL, the burden of maintaining FISMA compliance for other Common Fund data sets will be an on-going concern.

The CFDE tech team can assist the Programs by assessing costs for different sources of computing, encouraging the data coordinating centers to pool resources to lower overall costs, training users to use cloud based systems so they avoid egress charges, and engaging closely with larger cost-saving initiatives such as STRIDES. We will also consult closely with Common Fund leadership to develop clear guidelines, and budgetary policies to enable the Common Fund Programs and end users take better advantage of cloud resources. However, there is little that the CFDE tech team can do to help with the challenge of computing costs, especially since costs are unpredictable due to variability in compute needs and downloads. In any scenario going forward, *NIH will always need to shoulder the cost of computing and storage, and to provide resources to train users to adopt cheaper cloud-based analysis systems.*

Approaches to Operationalize the CFDE

Approach 1: Data Federation

The Common Fund Data Ecosystem (CFDE) will be based on a collection of inventories derived from data that are being hosted on cloud based systems by a number of DCCs. The inventories will describe all the assets at each Program, with this information available via a central catalog to enable discovery of the assets. The advantage of this approach is that formation of the ecosystem does not require the data assets themselves be moved to a central repository, only inventories describing those assets are centralized. Cataloging all of the Common Fund assets is a simple and effective means of liberating data from what would be several siloed repositories, and therefore greatly increases Findability, Accessibility, Interoperability and Reusability of all Common Fund data. This form of data federation can also be extended to programs funded by other institutes, and easily linked to other NIH ecosystems: once an inventory system is available, it can be used by anyone.

There are several very important outcomes for the CFDE federated approach.

The CFDE is future-proofing the Common Fund for interoperability. Interoperability between data silos continues to be of significant interest at the NIH, which recently held an "NIH Workshop on Cloud-Based Platforms Interoperability". The meeting had representation from 4 NIH ICs (Common Fund, NCI, NHGRI, NHLBI). The report from this meeting proposed four major thrust areas to improve interoperability between several groups which included use of the federated asset catalog approach developed by the CFDE tech team.

Common Fund's data portfolio is diverse, and these resources have significant value as individual assets. However, integration of this data with assets at DCCs from other ICs is of critical importance. For example, sequence based technologies for variant detection, whole genome/exome analysis, single cells and human cell atlas, as well as epigenomic analysis will increasingly be used by future Common Fund programs, as well as programs at other ICs. HubMAP also reported a need to integrate their data with sites such as the Human Cell Atlas, LungMAP, and Allen Brain Atlas. The CFDE work on interoperability will ensure Common Fund is able to significantly add value to data assets at each of its programs, and increase their ability to make use of data generated by other Institutes.

The CFDE is defining and measuring FAIR, to guide systematic improvement of Common Fund asset FAIRness. One of the CFDE's missions is to guide improvement of Common Fund asset FAIRness by providing consistent definitions, metrics, and reports across the Common Fund. Under the CFDE, each data center's inventory will be evaluated consistently based on FAIRshake, and we will work with the individual Common Fund programs to improve their FAIR measures as well as adjust FAIRshake to meet the needs of the Common Fund (see Appendix C).

By applying the same objective measurements to each Program, we will establish an even playing field across all of the sites. This will incentivise sites to improve individually and learn from each other, and at the same time will lead to a more specific, consistent, and sophisticated set of FAIRness metrics for the CFDE. More importantly, the improvements to each site and across the ecosystem will enhance user abilities to find and make use of Common Fund data.

This approach overcomes a major obstacle for the Programs, because the Programs can not easily work with other groups to align around a common set of metrics. In this scenario the asset inventories generated by each DCC are created by adhering to a common standard coordinated by an external group (i.e., the CFDE tech team).

Approach 2: CFDE Portal Implementation

The CFDE tech team will provide a portal that will enable users to search all of the federated data assets at each Common Fund Program. The CFDE portal will increase a user's ability to

find these important resources, as well as mix and match sets of data from each site to use in subsequent analysis. We refer to lists of assets as *manifests*, which are similar in function to a shopping cart on a commercial web site. Generation of user-specified manifests will enable users move information off the portal for use in the analysis tool of their choice. For example: 1) users will be able to "send" search results to cloud-based workspace environments such as Terra, avoiding data egress charges; 2) manifests can be used in "notebook apps" such as Jupyter Notebooks. Notebook apps are documents that contain a combination of human-readable documents and computer code. These systems are very powerful, in that they let users describe how analyses are performed, and make it easy for users to perform their own analyses; 3) assets can be combined into other downloadable objects that are easily incorporated into popular analysis tools running on the users' own cloud instance or local compute.

The portal will provide several important and unprecedented functions. For example, for the first time end users to be able to answer the question: "where are all the RNAseq datasets associated with all Common Fund programs?". Similarly, Program Officers at Common Fund will be able to go to a single website and view the growth of data from their program over time, to review objective FAIR metrics for these assets, and view the degree of harmonization of these data in comparison to other sites. In the coming years, we intend to include additional usage information from each of the programs. For example, we plan to request and display portal metrics such as the number of users that register at each of their sites, and how often their data is downloaded or analyzed. Once this capability is established, an important outcome of the CFDE will be to give Common Fund leadership the new ability to objectively review the overall use of resources at each data center, and to easily perform that review in comparison to all other Common Fund data centers. We anticipate this type of information could assist with making better informed decisions with respect to maintaining and prioritizing which Common Fund datasets should be expanded over time.

Approach 3: Training

Our training program for the next year has three key efforts.

First, we will teach people how to use the portal. This will be important both for bringing users to the portal, and for observing what new functionality in the portal is needed. Over time, we also expect the training materials to serve as an entry point to the portal for external researchers seeking interesting and relevant data sets.

Second, we will run training to enable biomedical data scientists to find and analyze large amounts of data close to the data. Our initial training will focus on (1) using the Terra platform to access GTEx data and analyze RNAseq data using the GTEx pipelines, and (2) using the Cavatica platform to access Kids First data and analyze genomic and transcriptomic data. The curriculum and training programs will involve GTEx and Kids First staff. Both the Terra and

Cavatica platforms run in the cloud and address the training needs around bringing analyses to the data.

Third, we will run training for clinicians on using the Kids First DRC portal to discover data sets and analyses. This training will focus on exploring the Kids First portal functionality and browsing pre-analyzed data for variants and expression information. We expect this training to both enable more clinicians to make use of the Kids First portal and also to help expand and refine the Kids First portal functionality.

Collectively, these training efforts will allow the CFDE to develop pilot materials that can be expanded further, create assessment instruments to evaluate current efforts and guide future efforts, and expand training functionality at the individual DCCs.

Approach 4: Addressing Data Incompatibility

The first challenge the Common Fund faces is in making Common Fund data from individual programs findable and accessible in practice. True interoperability within the ecosystem would mean that compatible data sets would not only be findable and accessible and reusable, but that metadata and provenance would follow the data between platforms. In an idealized situation, compatible data sets would be presented contextually, and incompatible data sets would be flagged as incompatible.

Thus we would say that **findability** and **accessibility** of data are prerequisites for data **reuse**, while **interoperability** is contingent on the data sets being combined. A fully mature data ecosystem will include a large number of interoperable data sets to allow reuse ***across*** the ecosystem, while a nascent ecosystem like the CFDE may focus on findability and accessibility. This perspective provides us with progressive stepping stones to guide CFDE development.

We are building a portal to improve findability and access of Common Fund data, as preconditions for improving data reuse and interoperability across the Common Fund. This is because our deep dives suggest that while a specific, familiar data set is easy to search, it is effectively impossible to discover new data across the entire Common Fund. This approach enables talented and motivated users to find relevant data and directs them to the original data sources.

We will combat the inevitable increase in data mis-use and inappropriate combinations of data sets with user training. The majority of our proposed training efforts with GTEx and Kids First focus on enabling *sophisticated biomedical data scientists* to use flexible cloud-based platforms to analyze data. This approach allows expert biologists and clinicians with hypothesis driven questions to lead the scientific inquiry rather than having to delegate to a more technically savvy, but computationally experienced bioinformatician. We will lower technical barriers such as difficult data access and lack of defined workflows, while allowing experts to bring their own biological expertise and questions to their data analysis. Our other training efforts for 2020 (Kids

First clinician-centric training) focus on training highly motivated users in how to effectively search existing data analyses and results to answer specific questions: here the Kids First portal and data analyses have been provided by subject-matter experts at the Kids First DRC.

We will also pilot hackathons and data reuse postdocs. These activities provide talented and motivated users that are already close to the data, or can take the time to develop deep expertise in specific data resources. We expect these activities to both expand and refine our understanding of which data sets are interoperable

In sum, our 2020 activities acknowledge concerns around interoperability and expertise, and will expand responsible data reuse while increasing our set of available use cases.

Approach 5: Federating with Data Resources External to the Common Fund

Several other efforts are underway to establish cloud-based data platforms across NIH (e.g., NHGRI-Anvil, NHLBI-STAGE, and NCI-TCGA), and attendees at the recent NIH Workshop on Cloud-Based Platforms Interoperability agreed that adopting approaches similar to what we are using for the CDFE could greatly assist in expanding capability for all users. Final resolution of standards for federation across all of NIH will take some time, but several steps can be taken to ensure the internal standard adopted by the CFDE is either compatible with (or serves as a prototype for) a system that could be used by many other external resources. The following steps will be taken to ensure federation compatibility with ecosystems external to the CFDE.

First, we will keep in touch with national and international interoperability efforts. At the national level, our primary effort will be to connect with the NIH Interoperability working group, that represents four ICs (NHGRI, NHLBI, NCI, and Common Fund) and in particular includes the NIH ODSS. At the international level, we will connect with GA4GH which is the main standards body for genomics data.

Second, we will work with Common Fund programs to help them adopt and operationalize standards, and help channel feedback from individual programs about drawbacks and incompatibilities of emerging standards. This will help ensure that future standards are not incompatible with Common Fund program needs.

Third, we will work with users to identify challenges and opportunities with emerging standards. For example, if a standard provides users with opportunities to improve data discovery and reuse, we will provide training materials showcasing this. Conversely, if an emerging technical consensus blocks a specific use case, we will bring this use case back to the standards developers.

And fourth, we will work to expand the scope of existing interoperability efforts to automatically include Common Fund assets. For example, while our 2020 plans are focused on building a

portal, the Common Fund asset inventory underlying the portal will be directly usable by other efforts and other portals.

Approach 6: Assessing the Optimal Balance of Cloud Versus On-Premises Computing and Storage

One role for the CFDE will be to facilitate a discussion of on-premises, cloud, and hybrid models to weigh the relative strengths and weaknesses. For example, the Human Microbiome Project was all local, LINCS is currently local but considering a transition to the cloud, and Kids First has been entirely cloud-based from the beginning. Although the current trend is towards cloud based resources, there may be many trade-offs in moving everything to the cloud.

The cloud is now more of a business model than a different technical configuration. All the technical advantages once provided by cloud computing (e.g., VPNs, containers, workflow, serverless solutions) are now easy to implement in an on-premises solution. In general, smaller projects are likely to save money by hosting everything in the cloud. As projects grow, however, the cheaper option becomes an on-premises solution with cloud-computing available for bursts where additional compute capacity is needed with no permanent cloud storage. However, there are many other considerations, both for the day to day workings within a data center, and the implications for long term sustainability. The CFDE will work to get perspectives from within the Programs as well as Common Fund leadership to provide guidance for new Programs deciding on the appropriate infrastructure for their project.

Approach 7: Changing Role of Site Visits

Connecting the Common Fund Programs into a thriving Data Ecosystem will require much more than merely solving technical challenges. The larger challenge of this project is social: the technical solutions rely on consensus building among the stakeholders within the CF, and such a consensus can only be reached by fostering a community where cross-program discussion and collaboration are incentivized and recognized by NIH leadership as important goals in and of themselves. The site visits and the establishment of long-term DCC engagement are both critical to consensus building.

As more Common Fund Programs engage with our reports and gain familiarity with the CFDE's proposals, we expect that there will be both more excitement, and more skepticism, about our work to increase interoperability across the Common Fund. Our interactions with Common Fund Program PIs continue to reinforce the utility of our in person meetings both to learn about the incredible work of each program as well as to build trusted relationships across programs. These relationships, that allow for both enthusiastic and dissenting opinions, are fundamental for creating a community where everyone knows their input is valued and is incentivised to work towards creating a thriving ecosystem. As our focus shifts from initial introductions to sustained engagement, we will work to create more spaces within the CFDE for Common Fund Program PIs to provide input about the direction and work of the CFDE.

Concerns, Risks and Threats

Sustainability and cloud costs

There are monetary concerns around accessibility. Data hosting and egress charges have traditionally been covered by the institution that is funded to share the data. However, repositories increasingly don't have the funding to do this at scale. The Sequence Read Archive (SRA), for instance, stopped accepting large sequencing project data years ago, mostly due to budgetary constraints. For data in the cloud, the largest cost is typically data download: "To mitigate these costs, many repositories either limit the size of datasets or limit the throughput on downloads. However, this goes squarely against the FAIR principles and results in repositories that have the notion of data-sharing while in fact the data is not truly available." The SPARC Data Core suggested that the ultimate fix for this problem needs to be user education and training. Researchers need to be aware of the time and money required for downloading data, and that working in the cloud is a much faster and cheaper option in many cases. However, to work in the cloud, users need training that isn't readily available right now.

Even with a better educated user base, SPARC told us that there needs to be a long-term sustainability plan for data storage and use costs, and cautioned against adopting a system where the NIH pays data egress fees directly, without oversight:

> "If you do provide a platform that enables scalable, high-throughput data access to very large amounts of public data, one needs to take into consideration the cost that could be incurred by users. For example, what if a graduate student writes a script to download the entire public repository each night. What if this is a student in a different country, what if this is mal intended? Given the high availability of the resource, it would be very easy to have hundreds of thousands of dollars in unexpected costs within a couple of days."

Finally, there are additional concerns with thinking not only about FAIR but also about making data publicly available. In a discussion about data ownership, the SPARC Data Core told us that one difficulty with making data public, is that often several entities claim rights to the same dataset, and have different views on where and how it should be stored and accessed:

> "We strongly believe that data on our platform is owned by the users of our platform and not Blackfynn. However, our experience with academic institutions is that they also claim rights to the data, even though the NIH mandates data sharing. It would be great to have a discussion on mechanisms to break through this impasse."

Appendices

Appendix A: HuBMAP Site Visit

Location: 300 S Craig St, Pittsburg, PA 19107
Date: Tuesday September 3, 2019
Attendees: Representatives from CFDE included Amanda Charbonneau (UCD), Brian O'Connor (Bionimbus), Steve Edwards (RTI), Titus Brown (UCD), and Owen White (UMB). Representatives from HuBMAP included Jonathan Silverstein (PI), Nick Nystrom (PI), and Robin Flaus Scibek (Project Manager). We were also intermittently joined over Zoom by the HuBMAP Project Officers and Project Lead from NIH: Tyler Best, Ajay Pillai and Richard Conroy.

Meeting Logistics

We held a meeting with the Infrastructure and Engagement Component (IEC) of the HuBMAP Integration, Visualization, and Engagement (HIVE) Collaboratory at the Pittsburgh Supercomputing Center on Tuesday September 3, 2019 for a day and a half. During the meeting, we used the agenda at the end of this document as an informal guide for structuring the day.

The engagement team began by reviewing their goals for the meeting. These goals include learning more about:

- Structure and goals of HuBMAP, including specifics about the data they host
- Information about training and organization
- Overall set of priorities for their group

Prior to our discussion regarding the HuBMAP program, the engagement team presented an overview of the proposed projects for the CFDE in 2020 as requested by our HuBMAP hosts. The HuBMAP team then provided a brief overview of the current structure of the HuBMAP program and how it has evolved since the launch last November. This was followed by a discussion of their proposed infrastructure for housing HuBMAP data.

Day 1 concluded with a brainstorming session on what role CFDE could play in supporting the HuBMAP program. Day 2 focused these discussions with the goal of establishing a set of concrete actions whereby the CFDE could add value to the HuBMAP program.

HuBMAP Overview

The goal of the Human BioMolecular Atlas Program (HuBMAP) is to develop an open and global platform to house a molecular atlas of the human body at cellular resolution beginning with foundational maps from 8 tissues: bladder, kidney/ureter, colon, lung, lymph nodes, spleen, thymus, and vasculature. HuBMAP complements an earlier effort to study gene expression across many tissues (GTEx) by generating single cell level data on a smaller number of subjects. The program is organized into three components:

1. Tissue Mapping Centers (TMCs)
2. HuBMAP Integration, Visualization and Engagement (HIVE) collaborative components
3. Innovative Technologies Groups (Transformative Technology and Development (TTD) and RTI)

The five TMCs (Caltech-UW, Stanford-WashU, UCSD, University of Florida, and Vanderbilt) are responsible for generating high-quality data at scale with standardized metadata annotations. This will include single cell ‘omics assays that compare chromatin-level changes with changes in expression of biomolecules at the cellular level. These studies will be combined with molecular-level analysis of tissue blocks using imaging methods such as fluorescent microscopy, sequential fluorescent in situ hybridization, imaging mass spectrometry, and

imaging mass cytometry. In all, there are over thirty data modalities that will be employed by the TMCs. The tissues are derived from healthy human organ donors with the caveat that death of the donor will impact even healthy organs depending on the cause.

The data generated from the TMCs will be integrated and disseminated by the HuBMAP Integration, Visualization, and Engagement (HIVE) Collaboratory, which consists of five components of three types:

Component	**Organization(s) Included**	**Responsibilities**
Infrastructure and Engagement Component (IEC)	Pittsburgh Supercomputing Center and University of Pittsburgh	Housing all data from the HuBMAP program, providing compute capacity for analysis of the data, and coordinating the efforts of the consortium.
Two Tools Components (TC)	1) Carnegie Mellon 2) Harvard Medical School	Providing tools that allow users to visualize and analyze the data generated by the HuBMAP program.
Two Mapping Components (MC)	1) Indiana University Bloomington 2) New York Genome Center	Developing the common coordinate framework to allow exploration and visualization of results from different individuals, technologies, and labs.

The original arrangement of the HIVE had the Collaboration/Engagement activities separated from the Infrastructure Component. These activities were merged in summer 2019 with the Pittsburgh Supercomputing Center and University of Pittsburgh becoming the new Infrastructure and Engagement Component (IEC). Efforts for all HIVE components are coordinated by the IEC to ensure a unified user experience when interacting with HuBMAP data. In addition to collecting data from the members of the HuBMAP consortium, the HIVE has also been tasked with integrating data from related projects such as the Human Cell Atlas with the data generated by the HuBMAP program.

Whereas the emphasis of the TMCs is to generate large amounts of data using state of the art technologies, the Innovative Technologies Groups are focused on identifying new approaches that will complement and extend the existing technologies. The Transformative Technology and Development (TTD) investigators at the California Institute of Technology, Harvard, Purdue, and Stanford will establish proof of principle and validation of next generation tools to expand throughput, multiplexing, and discrimination of biomolecules in human tissues. The Rapid Technology Integration (RTI) organizations will be launching soon and are responsible for implementing new technologies in the HuBMAP consortium.

Program Lifestage

The HuBMAP program is completing its first year, and production phase for data generation is scheduled for 2022. However, the consortium is planning an initial data release in mid-2020. The first year of the HIVE has mostly been used to create teams and working groups to manage the many moving pieces that comprise HuBMAP. The HuBMAP HIVE also has a mandate to collaborate with other atlas projects and integrate data from those projects with the HuBMAP data. As a result, the HuBMAP HIVE will likely have a CFDE-like role for cellular level body atlas data.

As part of their planning phase, members of the HIVE have visited all the TMCs to determine the datatypes expected from each and the estimated timeline for data availability. They have recently established data release teams for each datatype (e.g., RNA, Histology, proteomics) to coordinate the data ingestion, processing, and release and portal release teams to focus on developing the web site (UI), infrastructure, and APIs (Application Support Interfaces) that will allow automated data access. These teams each consist of 5-10 people and are currently establishing metadata standards and acceptable file formats. This effort is complementary with the activities within the MC regarding semantic and spatial descriptions of the anatomical locations including defined integration points between the efforts. The Portal Release Team has adopted a scrum-based software development approach with sprints beginning in September 2019.

The overall HuBMAP consortium is guided by a steering committee that meets once a month. Under that are five working groups focused on policy, communications, data science, technology, and tools with representation from each HuBMAP awardee on each group. Additional teams are set up that include ad hoc members from the TMC, TTD, and HIVE components to carry out the work. This results in 3-4 meetings of various types every week, across the formal consortium-wide calendar. High-level decisions are made at the working group level with oversight from the steering committee and implementation plans developed by the individual task teams. All workstreams flow into two structures: product owners who determine what should be developed and technical contacts who determine how it will be built. This structure has recently evolved toward the conclusion of the initial year of the project as the IEC took on coordination and engagement inside and outside the HuBMAP consortium.

Infrastructure

The HuBMAP IEC is using a hybrid of on-premises and cloud computing for the HIVE. The Pittsburgh Computing Center is a leading partner in XSEDE (Extreme Science and Engineering Discovery Environment), the National Science Foundation (NSF) cyberinfrastructure program. With funding from the NSF, they have developed the Bridges resource that brings together high-performance computing (HPC), AI and Big Data to support applications including genomics, machine learning, graph analytics, image processing and materials science. They

are still working on FISMA certification, but they are certified on other projects and do not anticipate this being an issue.

This existing infrastructure allows them to provide some services at a small fraction of the cost of hosting on AWS. From the perspective of HuBMAP, a vast amount of compute capacity is freely available to academic and non-profit researchers through 2029 as part of the XSEDE program and the recent renewal of Bridges. Although this on-premises solution is local, it was built to be interchangeable with cloud computing systems, so HuBMAP can push jobs to AWS or other commercial cloud providers anytime additional compute capacity is needed. The IEC has had discussions with STRIDES regarding the fiscal management of the cloud component of their hybrid solution, but this is delayed due to factors outside of the control of HuBMAP.

The biggest savings for an on-premises solution is in storage. Given the scale of the existing resource, they can time purchases of new storage to match the anticipated receipt of data from HuBMAP researchers. This eliminates the typical advantage of on-demand storage from a cloud provider and makes hosting their own storage much more cost effective. The size of their operation also allows them to build in redundancy and mirroring backups that offset the risks associated with data loss from local storage vs. the cloud. They do not anticipate problems with the ever-growing storage demands, because the cost of storage tends to decrease in parallel with the increased storage requirements.

The data ingestion portal is being built now. The infrastructure behind this is a microservices architecture heavily using APIs and a combination of technologies: a Neo4j graph database for provenance, NOSQL for metadata indexing, and Globus for data management and security of APIs. The HIVE will collect raw data for the more common methodologies, but they may only collect processed data for the less common datatypes. For example, raw mass spectrometry files from the Vanderbilt group require specialized software to open and are not broadly useful until after processing. In this case, the HIVE will receive the results from Vanderbilt's initial pre-processing and analysis. For methods that are being implemented across multiple TMCs, however, there will be a common analysis pipeline within the HIVE to provide reproducible results across all the TMCs. The TMCs may still perform their own analyses for their publications, but the data provided by the HIVE will be consistent across TMCs. There is some tension between when/where to get their pipelines into production within the HIVE. The further away from data generation they get when running processing pipelines the more context gets lost. For this reason, the HIVE is paying close attention to tracking the provenance of all data both prior to and after deposition in the HIVE.

Analysis

The analysis tools for the HIVE are being developed by separate components, but the IEC is responsible for providing the compute platform and data in a computable form. The production phase of HuBMAP doesn't begin until 2022, so many of the issues surrounding this are still under discussion. Containers are likely to be a big part of the solution both for the initial

processing of the data by the IEC and the downstream analysis and visualization tools built by the TC. However, there is still a great deal of discussion within their consortium about data formats and specifications as well as what specific technologies to implement.
The Portal Release Team has been tasked with defining a vision for handling both initial data analysis by the IEC and analysis tools developed by the TC. They will evaluate resources such as Kubernetes and Airflow, but the general challenge is how to hook back to the compute allocations and not go over budget. There is a small amount of funding for this allocated in year 2 of the program.

One key specification for designing the workflows will be ensuring they work with the hybrid infrastructure. While some workflows can run entirely on the HIVE local infrastructure, that same workflow on a very large dataset may need to start locally and then move to the cloud for some steps of the analysis. HuBMAP workflows will need to account for the possibility of these ‘burst runs’ in the cloud and be able to push data into cloud temporarily and then vaporize. While the cost associated with moving data into the cloud is free, the time required for large datasets could be considerable. The workflows must optimize the stage in which data are transferred to the cloud to minimize the lag associated with large data transfers.

Access

All data received from TMCs will be processed and made immediately available back to the TMC, however the infrastructure will support embargos for other users based on the data sharing policies established by the HuBMAP Steering Committee, which are focused on public releases having proper quality assurance and moving data to public or appropriately restricted access based upon IRB/consent limitations as soon as possible. Data embargos will not be used to protect private research. Access to resources, including whether data can be downloaded or accessed, including open public information, is managed through Globus groups.

HuBMAP plans to offer their users some level of computing resources for free through the XSEDE program, but they do not have unlimited capacity. As free services will need to be limited, an open question is who should have access to these free services, at what level, and for how long. For example, HuBMAP plans to allow people to easily bring their existing data into the HIVE and do joint analysis, but users bringing in large datasets will require a lot of resources. The obvious solution would be to give each user some amount of credit and then charge for use, however this is complicated by HuBMAP’s relationship with XSEDE, as XSEDE is prohibited from charging users. One possible solution would be to allow users to use common workbenches such as CAVATICA and TERRA, which are linked to AWS billing, while another may be ensuring that users can export data with minimal friction to public cloud infrastructures for their own analysis.

Harmonization and Metadata

HuBMAP IEC is using the Unified Medical Language System and other open ontologies for canonical metadata. The IEC told us that Metadata standards within the HuBMAP program are a big enough challenge in the short term and trying to adhere to outside requirements (e.g. imposed by the CFDE) would be a distraction, so their main focus is how to accommodate the diversity of data and data formats coming from the TMCs. At the HuBMAP kickoff meeting, the NIH Director challenged the HuBMAP consortium with determining how to represent single cell data consistently across many different organs and tissues. There are ongoing discussions within the HuBMAP working groups to define the standards within the program, and they are trying to standardize the file formats to the greatest extent possible. In many cases, however, the file formats are determined by the equipment that generates the data.

The IEC indicated that they would be willing to participate in discussions with the other Common Fund Programs to promote interoperability, but they believe that the fundamental notion of defined metadata by consensus upfront is wrong. For example, they have experience translating multiple Electronic Medical Records stored in multiple different systems into a common metadata model. However, they find that the systems give different results over the same events, even though the underlying data model is identical, because translation to the model involves substantive data loss. This reinforces their position that “conference room” metadata harmonization ultimately fails for many purposes. Rather, flexible information sharing approaches (via APIs and other data interchanges) are the focus required to support users of those data to transmute those data for their interoperational purposes. In short, combined analysis will always require deep knowledge of the original dataset. Of course, integrated data will be presented but it must not be the exclusive goal.

Sustainability

The HuBMAP project is the newest of the Common Fund programs, so their perspective is quite different from many of the other Common Fund Programs that have been interviewed thus far. The HuBMAP program is currently funded through the OTA mechanism, and expects to be funded for eight years. It isn’t clear whether the 10-year limit on Common Fund grants will apply to this mechanism or whether there will be a future opportunity to continue support for the HIVE under a different mechanism. The IEC investigators are very happy with the OTA mechanism thus far. They noted that all changes in scope for the project have been coupled with supplemental funding to support the work. The largest of these was when the IC took over the engagement activities from the CC this summer. However, the OTA includes a much narrower description of what needs to be accomplished with the funds, so there is a little less flexibility in implementation compared with a grant. With the ongoing support for compute infrastructure from the NSF XSEDE program, they do not anticipate any issues with data access during the lifetime of the HuBMAP program. They plan to stand up a Data Access Committee to handle governance of the data dissemination once data starts coming in next year.

Training

Internal

During the opening presentation at the HuBMAP kickoff meeting, Richard Conroy (program officer) acknowledged the complexity of HIVE and the corresponding challenge of coordinating the multi-site effort: “It’s complex but that complexity is expected.” The first year has been heavily focused on getting committees and working groups in place to coordinate the disparate efforts while simultaneously trying to get the individual efforts started. They anticipate that merging the Infrastructure and Engagement components will streamline operations, and now have most of the committees and working groups established. Robin is now leading the engagement portion of the IEC, which includes all the training activities. The first year of the HuBMAP program does highlight the increasing complexity associated with the large Common Fund programs.

The HIVE is using the following technologies to coordinate activities and disseminate information:

- Zoom - Video conferencing
- GitHub - version control and issue tracking
- Slack - Team communications
- E-mail
- Protocols.io - Capturing experimental protocols and computational workflows
- Asana - track NIH milestones and project management

While the IEC has technologically solved the problem of communication across a distributed consortium, they are experiencing some common social communication issues, primarily in getting everyone to buy-in to the technology. For example, they noted that onboarding consortium members to services like Asana, and convincing them that the effort needed to learn that service is worthwhile, were ongoing pain points. The IEC also distributes a weekly newsletter to all members of the consortium, but struggles to get people to contribute content. This forces the IEC to devote extra time and resources to finding and compiling noteworthy news.

External

The HIVE goal is to provide sufficient compute options such that most users will use online tools rather than downloading the data. Since the instinct for users is to download data and work locally, this will likely be a significant challenge. The IEC told us that since the Tools Components are building the online tools, they expect that most of the training for those tools will be developed by the TCs.

Instead, the IEC plans to focus on documentation. A key question when designing tools and training to support data access is who the user of the data will be and what support will they need. The IEC pointed out that while providing tools that would allow anyone to meaningfully interact with the data is a laudable goal, they must face the reality that at a certain point they would invest more resources designing tools to support naive users than society would gain from what they are able to do with the data. They argue that anyone who is serious about successfully utilizing the data from HuBMAP will invest the time and energy to understand the data and how to access it regardless of how the data are shared, and that a deep understanding of the data is necessary when performing a secondary analysis to avoid making assumptions about the data collection that are incorrect. As such, they plan to build and document flexible APIs enabling motivated users to use tools that are as simple as fit the tasks.

Given the early stage of HuBMAP, they have not established a training plan yet, but they do have a lot of previous experience with training. They have previously developed training for onboarding to the system as part of the XSEDE program, but the login, set up, onboarding, etc. is sufficiently different that the HuBMAP training will be assembled from scratch. In terms of how to train, they have a great Wide Angle Classroom setup that allows them to have a single instructor to teach several virtual classrooms simultaneously. They plan to leverage their current XSEDE centers that already have TAs, rooms, etc. to pilot any future training initiatives. They are also considering Massive Open Online Courses (MOOCs) as another option.

FAIR

The HuBMAP HIVE is dedicated to the FAIR principles but have not yet documented clearly what that means in terms of implementation. For their part, all data in their portal will be accessible programmatically via well-documented APIs, and they are in the process of determining what minimum requirements for metadata would allow users to find and understand the data provided. They are also interested in ensuring that their workflows are reproducible and transparent. They noted that it would be useful to know how different Common Fund Programs are handling workflows and which ones are attempting to document and disseminate workflows for reproducibility purposes. They are considering Zenodo or Dockstore as potential repositories for sharing workflows.

The HIVE was given a mandate to interoperate with other consortia, and as the TMCs have a heavy focus on lung and kidney, the LungMAP and GUDMAP consortia are logical choices. However, the Human Cell Atlas (HCA) program is the highest priority source of non-HuBMAP data in the short term, and there is a joint meeting planned for HuBMAP and HCA from March 31 through April 1, 2020. They have evaluated the HCA annotated metadata model and what it would take to incorporate HCA data into the HIVE, however it is not the highest priority because the effort expended on this can’t pull resources away from the core effort required to launch the portal to support ingestion of HuBMAP data beginning next year. While the NIH wants to see cross-program analytics, the HuBMAP budget for this effort is only $75K at the moment.

In the short term, HuBMAP would rather focus on pairwise interactions with efforts such as HCA rather than trying to enable a cross-program multi-DCC portal. They emphasized the need to better explore how the data can be combined and used before trying to build tools that support the process. They expressed skepticism that users interested in superficially combining datasets across Common Fund Programs would gain much from a multi-DCC portal. In their view, successful data integration will require a talented and motivated user with a deeper understanding of the data, which would necessitate working with the original data sources.

Common Fund Program Cross-pollination

The IEC told us that they have limited time for general collaboration and would likely not attend Common Fund cross-pollination events without specific foci and benefits to participants. They expressed concern that a large event would not provide the right conditions for fostering relationships, and the time they invest in attending would be wasted. However, they said they would be very interested in meeting with individual programs and building pairwise collaborations, especially if they were given some outside direction about what programs would be best to work with.

They have a year 2 mandate to bring public data into HuBMAP, but have invested little time to research potential collaborations with other DCCs. The HuBMAP program is focused on normal tissue, and they would be most interested in opportunities to build a federation with repositories that include diseased tissues. They noted that they would even be interested in collaborations with sunsetting programs and would be willing to integrate older data into HuBMAP.

SSO

HuBMAP uses Globus for all security, including Federated identity management, data movement, sign on for their portal, application, and even workflow, pipeline, and query API permissions (via passing authorization tokens). They have successfully used this for the XSEDE program in the past. Globus includes ORCID and ERA Commons as authentication sources, so they view this as meeting the NIH requirements for SSO.

Outcomes

Infrastructure Reuse

Although many Common Fund Programs are opting for cloud-based infrastructure, HuBMAP's hybrid cloud model employed by the HIVE could be more broadly useful within the Common Fund. The hybrid model works well for HuBMAP because they already have a great deal of expertise, infrastructure, and connections. From their work with XSEDE and the Pittsburgh

Supercomputing Center, the IEC staff have tens of years of experience in designing, maintaining and administering large complex computing systems at scale.

They also have the space to store these machines, as well as dependable high-throughput fiber internet connections, electrical systems that can reliably support thousands of supercomputing processors, and staff that can maintain it all. Finally, they have vendor agreements and connections in the industry that they can leverage to purchase software and hardware at volume for a substantial discount. By leveraging their existing expertise and resources from XSEDE, HuBMAP was able to dramatically reduce their computing costs.

Although HuBMAP's hybrid infrastructure is working well for them, it may not work for all Common Fund programs. While getting contracts for inexpensive storage is relatively easy, the services required to keep them secure and running smoothly need to be maintained and refreshed periodically, and require specialized knowledge, as well as a lot of underlying site infrastructure. In general, for smaller systems at organizations without an existing infrastructure, cloud solutions provide a cost-effective way to implement a robust system without a large, upfront hardware and IT investment.

As the size of the system increases, the cost considerations shift such that for larger programs, the initial investment in hardware and IT support is cheaper. In the case of HuBMAP, the size of the program makes it more cost-effective to host the system locally. In addition, both the infrastructure and expertise already existed at the Pittsburgh Supercomputing Center because of their ongoing work as part of the XSEDE program. It also is unclear how the hybrid infrastructure will impact data accessibility when HuBMAP funding ends, however their current system is built to seamlessly move data to the cloud for extra compute capacity, so it shouldn't be difficult to transition the data to the cloud if Common Fund decides to go that direction.

The wide-angle classroom technology that HuBMAP is planning to use may also be useful to other Common Fund Programs. This technology is similar to a webinar in that a single person can present a lesson, and it will be broadcast to many locations. However, the system is much more interactive than a typical webinar. Each classroom receiving the broadcast is outfitted with cameras which display back to the presenter, such that the instructor can see all of the learners and react to questions individually, much like they would in a standard classroom. Each receiving classroom also is staffed by a small number of teaching assistants who can help troubleshoot and fix problems locally.

By investing in creating compatible classrooms at several locations around the world, they have substantially reduced the number of teachers required to have a worldwide presence. The XSEDE program currently offers training on a regular basis in several countries, with just a single instructor.

Challenges

The HuBMAP team highlighted a number of challenges faced by the increasingly complex centers that coordinate these large Common Fund Programs, and we had a productive discussion regarding how CFDE could potentially assist in overcoming those challenges.

- HuBMAP is a multi-component, multi-organization, collaboration similar to CFDE in scale, but huge complex projects take much longer to ramp up, and there are thousands of decisions that go into building their infrastructure. This highlights the need for assistance during the first year, as well as a Common Fund Playbook and list of best practices to guide those decisions.
- While the FAIR principles are clear, there is concern that if the CFDE attempts to normalize the implementation across different programs, it will create an unneeded hardening of the requirements, which will reduce the flexibility for the individual programs to implement FAIRness.
- The HIVE IEC expressed concerns about the zeal to create tools for merging data across different programs. They worry that the further the analysts are from the origin of the data, the more likely they are to misuse the data. GTEx expressed similar concerns during their site visit.

Potential Solutions

- Training and Support
    - Create a playbook on challenges that result from the increasing complexity of the Common Fund Programs. Provide CFDE assistance to help new programs during their first year. Promote sharing of best practices among Common Fund Programs.
    - Contact the TCs to discuss training for their tools
    - Create a decision tree to assist new Common Fund Programs in selection of on-premises, cloud, or hybrid solutions based on the experiences from the existing Common Fund Programs. Provide CFDE support and connect new Common Fund Programs with existing Common Fund Programs as needed to provide additional details.
    - Establish practical guidance on the implementation of FAIR principles. Documentation for new Common Fund Programs and CFDE helpdesk support to field questions.
- Enable data aggregators
    - The HIVE has a mandate for incorporating single cell atlas data from outside as well as inside the HuBMAP consortium. CFDE could provide support for these applications and establish a community of Common Fund Programs to share best practices and identify partnership opportunities.
    - CFDE could establish a pilot project with the HIVE to establish a proof of concept for Common Fund data aggregators. The HIVE has a mandate to incorporate

HCA data, but the funding for this is minimal. CFDE could support the development of this resource and document the process for the benefit of future Common Fund Programs should they be given a similar mandate.

- Promote pairwise interactions
    - The CFDE can identify people from different Common Fund Programs who could usefully interact about specific topics, and provide introductions as a way of beginning to build community connections without requiring a lot of time or effort on the part of the Common Fund Program PIs.
- Pilot Projects
    - CFDE could promote pilot projects whereby two or more Common Fund Programs collaborate to demonstrate the value of integrating data from disparate sources. These should be science-driven and not too theoretical. The goal is to have pilot projects that are indicative of the types of questions researchers would ask rather than projects driven by a general desire to show utility. These pilot projects would result in the following benefits:
        - Demonstrate utility
        - Highlight the costs associated with specific analyses and help determine the best mechanisms for covering those costs.
        - Create detailed use cases to drive future CFDE tool development
        - Help to clarify policy questions about who gets access to a resource, what level of access to grant, who pays for it, etc.

Potential Projects

- CFDE could develop a set of case studies to illustrate science-driven data reuse/analysis. This could include small grants to the larger research community to support postdocs who want to analyze data from two or more separate sources. It would also include support for members of the chosen data sources to assist with the interpretation and use of the data.
    - HuBMAP is interested in connecting their data from normal humans with disease data. HCA has some disease data in addition to the normal core work.
    - Could also consider projects that bring in a small amount of data from an R01 or equivalent and then try to leverage data from these larger projects to provide additional context for those data.
    - Possibly model this after the NCI ICTR (https://itcr.cancer.gov/) program.
    - The Kids First/GTEx collaboration is a model for these case studies. Identify a dataset or two available in this time frame.

Game Changers

The OTA funding mechanism has allowed HuBMAP to adapt much more readily than other Common Fund Programs. This, coupled with the early stage of their program, meant that they are not in a good position to recommend game-changing opportunities for the CFDE at this

time. Most ideas stemming from our brainstorming are likely to benefit future Common Fund Programs more than the HuBMAP program. However, there was one new area where CFDE activities could enable HuBMAP to more easily incorporate non-HuBMAP data.

Enabling data aggregators. As noted previously, the HIVE was given a mandate to incorporate other human atlas data that would complement those being generated by the HuBMAP program. While they've received a small amount of funding to support these activities, their primary responsibility must be on preparing the HIVE to accept HuBMAP data as is becomes available. If CFDE could facilitate the incorporation of the HCA metadata model into the HIVE provenance model and thereby facilitate the ingestion of HCA data, that would greatly increase the probability of success for that effort while providing valuable information about how this should be done in general.

Creating a lifecycle support program for Common Fund Programs. While they have worked through the process of coordinating a large, complex effort such as HuBMAP during their first year, they acknowledged the potential benefits of having access to lessons learned from prior Common Fund Programs at the beginning. They are also looking ahead to the training required as the HuBMAP data become available to users. They would welcome assistance in this area and were enthusiastic about the idea of a tiered help system that would shield the HuBMAP training team from general questions about bioinformatics or data access. While end of life for the HuBMAP program is not an imminent need, they acknowledged the need for the CFDE to provide solutions that avoid any data from a Common Fund Program being lost to the research community. They also offered their platform as an option for programs that are reaching end of life who have data that would be compatible with the HuBMAP data models.

Building a Common Fund Program community. The HIVE IEC recognizes that even within the single cell body atlas space, it is highly unlikely that a single repository will ever house all the data. They are interested in collaborating with other related programs to build a data federation whereby the data from separate repositories can be assembled without unreasonable effort. They would be willing to participate in CFDE organized Common Fund Program communities with this goal in mind as well as to benefit from the collective knowledge of the community.

Agenda

Day 1

9-9:30am Introductions

Short introductions from engagement team members and attending DCC members. The overarching goal for the engagement team is to collect value and process data about the DCC. Values data will include things like: mission, vision, goals, stakeholders, and challenges. Process data includes: datatypes and formats maintained, tools and resources owned by the DCC that they would like to have broader use, points of contact for follow up on technical resources, etc.

9:30-10am DCC overview

Short overview of DCC. Can be formal or informal, choose 1-5 topics to cover. Suggested topics: What is your vision for your organization? What big problems are you trying to solve? What are your big goals for the next year? Who do you see as your most important users/stakeholders? What project(s) is currently taking up the bulk of your effort/time? What areas of your organization are you putting the most resources into? What is the rough composition of your user base in terms of discipline? Do you have any challenges that are blocking implementation of your current goals? What skill set would you like to add to your project? How do you engage with your users? What kind of sustainability issues are you confronting? Can you currently do combined analyses with external datasets?

10am-Noon Goals Assessment

An exercise to get an idea of what types of things are important, what types of things are challenges, what do you dedicate your time/resources towards, and what types of things are not current priorities. Given a list of common goals provided by the engagement team, plus any additional goals the DCC would like to add, DCC members will prioritize goals into both timescale: “Solved/Finished”, “Current-Input wanted”, “Current-Handled”, “Future-planned”, “Future-unplanned”, “NA to our org” and for desirability: “Critical”, “Nice to have”, “Neutral”, “Unnecessary”, and “NA to our org”. The engagement team will work to understand the reasons for prioritization, but will not actively participate in making or guiding decisions.

Goal List

- Increase end user engagement X% over Y years
- Move data to cloud
- Metadata harmonized within DCC
- Metadata harmonized with _____
- Metadata harmonized across Common Fund
- Implement new service/pipeline _____
- Increase number of visitors to your site
- CF Data Portal

- Single Sign On
- Pre-filtered/harmonized data conglomerations
- A dashboard for monitoring data in cloud
- User-led training for end users (i.e., written tutorials)
- Webinars, MOOCs, or similar outreach/trainings for end users
- In-person, instructor led trainings for end users
- A NIH cloud playbook
- Full Stacks access
- Developing a data management plan
- Increased FAIRness
- Governance role in CFDE

Lunch

1 - 3:30pm Open discussion (with breaks)

Using the results of the morning exercise and a collaborative format, iteratively discuss goals, blockers, etc., such that the DCC agrees that the engagement team can accurately describe their answers, motivations and goals. Topics don't need to be covered it order, we'd just like to touch on these types of questions.

Topics:

Infrastructure:

- Do you intend to host data on a cloud service?
- Have you already started using cloud hosting? If yes:
    - Approximately how much of your data have you uploaded? How long did that take? How are you tracking progress?
    - What challenges have you faced?
    - How have you dealt with those challenges?
- What potential future problems with cloud hosting are you watching for?
- Does your org use eRA Commons IDs? Do the IDs meet your sign on needs?
    - If yes, did you have/are you having challenges implementing them?
    - If no, what do you use? What advantages does your system provide your org?

Use cases

- What is the rough composition of your user base in terms of discipline?
- What if any, use cases do you have documented? Undocumented?
- What things do people currently love to do with your data?
- What things would people love to do with your data, but currently can't (or can't easily)?
- What pipelines are best suited to your datatypes?
- What are the challenges associated with those desired uses?
- What other kinds of users would you want to attract to your data?

Review of metadata:

- What's metadata is important for your org? For your users?
- Do all of your datasets have approximately the same metadata? Or do you have many levels of completeness?
- Do you have any data already linked to outside resources?
    - Did you find the linking process easy? Challenging? Why?

- What kinds of datasets would you like to link into your collection?
- What implementation and schemas do you already have (or want)?
- What standards do you have (or want)?
- What automated systems do you currently have for obtaining metadata and raw data?

Training:

- What training resources do you already have?
- What training resources would you like to offer? On what timescale?
- What challenges keep you from offering the training you'd like?

Policies:

- How do users currently obtain access to your data?
- What are your concerns about human data protection?
- What potential challenges do you see in bringing in new datasets?

FAIR:

- Has your org done any self-assessments or outside assessments for FAIRness?
- Are there any aspects of FAIR that are particularly important for your org?
- Are there any aspects of FAIR that your org is not interested in?
- What potential challenges do you see in making your data more FAIR?

Other:

- What search terms would make your data stand out in a shared DC search engine?
- Does your org have any dream initiatives that could be realized with extra resources? What resources would you need?
- If you had free access to a Google Engineer for a month, what project would you give them?
- Any other topics/questions the DCC would like to cover

9-10am Review of goals and CFC involvement

A quick review of what topics are priorities for the DCC with suggestions from engagement team on how we can help.

10-noon Open Discussion

DCC reflection on suggestions, open discussion to find shared solutions.

Lunch

1-2pm Thoroughness checking

Touch on any questions not covered previously, ensure we have:

- Action Items for us, and rough timelines for getting back to DCC on them
- Tools/resources the DCC thinks might be useful for the overall project
- Points of contact "Who is the best point of contact for your metadata schemas, your use cases, the survey of all your datatypes?"
- Who would like to be added to our governance mailing list?
    - Or contact info/instructions on how to get that information offline.

Appendix B: SPARC Site Visit

Location: 1218 Chestnut Street, 8th Floor, Philadelphia, PA 19107
Date: Thursday September 5, 2019
Attendees: Representatives in attendance from the CFDE were: Amanda Charbonneau (UCD), Steve Edwards (RTI), Titus Brown (UCD), and Owen White (UMB) and Anup Mahurkar (UMB). The representatives from SPARC were Chris Baglieri, Blackfynn's SVP of Product, Leonardo Guercio Blackfynn's Scientific Engagement Manager, and Joost Wagenaar, co-founder and VP, Scientific Applications of Blackfynn.

Meeting Logistics

We held a meeting with Blackfynn at their office in Philadelphia on Thursday, September 5, 2019 for a day and a half to discuss their work for the NIH Common Fund's SPARC program. During the meeting, we used the agenda at the end of this document as an informal guide for structuring the day.

The engagement team began by reviewing their goals for the meeting. These goals include learning about the structure and goals of SPARC, including specifics about the data they host, as well as information about training, organization, and the overall set of priorities for their group. In turn, Blackfynn provided us with a wide-ranging overview of their work on the SPARC program, demos of their data management system, and an insightful view into the intersection of government and industry in human research.

SPARC Overview

Stimulating Peripheral Activity to Relieve Conditions (SPARC) is a Common Fund program focused on accelerating the development of therapeutic devices that modulate electrical activity in nerves to improve organ function. SPARC is different from many Common Fund programs in that its Data and Resource Center is shared among three academic and corporate entities.

- Blackfynn, which serves as the Data Core, is a private company, and SPARC is its largest NIH grant.
- Simulation Core is headed by Niels Küster of the IT'IS Foundation in Switzerland.
- Map Core is run by Peter Hunter at the University of Auckland.

On this visit, we met with the Data Core. Currently, there are six full-time developers working on the SPARC project, plus a fluctuating number of people who contribute on an ad hoc basis. The platform is populated by data from sixty different labs. Blackfynn's mission, both for the SPARC initiative and their company, is to provide technology that enables basic research, optimizes clinical trials, and transforms the treatment of neurological diseases. Their largest engagement outside the NIH is with the Michael J. Fox Foundation, supporting their Parkinson's Progression Markers Initiative (PPMI: https://www.ppmi-info.org/).

They told us that the general challenge is that data is not accessible in practice, and they want to ensure that clinicians and scientists have access to data, and at the right level of detail, to answer both clinical and scientific questions. They were excited to discuss the role for companies in meeting this challenge and were interested in the opinions within NIH.

Program Lifestage

The SPARC Data and Resource Center (DRC) components were all funded beginning in September 2017 and had just begun their third year at the time of our visit. However, like several other Common Fund Programs, data collection by researchers had been funded for

several years before the DRC was formed. This means that the DRC already has a great deal of data, more than 8TB, and is primarily focused on creating technology that can be easily worked into the pre-existing workflows of their researchers rather than defining and imposing standards.

Data Platform

Blackfynn develops two interconnected platforms that are leveraged by the SPARC program:

- The Blackfynn Data Management platform (https://app.blackfynn.io/)
- Blackfynn Discover (https://discover.blackfynn.com/)

The former is a platform for uploading, curating, and managing datasets. From this platform, users can publish snapshots of a dataset to Blackfynn Discover, which is a public, open data repository focused on the Neurosciences.

The SPARC Portal is an independent open-source web portal (https://github.com/nih-sparc/data-portal) developed by the SPARC consortium to provide an integrated portal into the data, simulation, and mapping tools developed by the SPARC consortium. A large component of the SPARC Portal is powered by the Blackfynn Discover platform. That is, the SPARC Portal leverages the Blackfynn Discover APIs to show public datasets, support search, and allow browsing the contents of the dataset. The goal of the SPARC Portal is to share data, anatomical maps, and computational models generated by the SPARC-funded researchers.

In order for the data (and subsequently the maps and models) to be visible on the SPARC Portal, the investigators must first publish their datasets within the Blackfynn platform to the Blackfynn Discover repository. The Blackfynn platform allows researchers to “publish” their dataset, and includes a number of traditional publishing features and rules: datasets have authors and co-authors (who can associate their ORCID IDs); the dataset is published with a markdown abstract to describe it, which is editable by submitter.

Datasets can have multiple versions with multiple DOIs that are managed by DataCite.org. Once the data is in the SPARC Portal, there are several specialized analysis tools for running computational models (developed by the Simulation Core), as well as a collection of interactive anatomical maps for various species (developed by the Map Core).

Infrastructure

The Blackfynn platforms are entirely hosted on the Amazon cloud (AWS). The SPARC Portal is hosted through Heroku (https://heroku.com), which in turn is leveraging AWS. They chose to use a 100% cloud-based model for a number of reasons, mostly having to do with the flexibility of the cloud. Disk drives inevitably fail, and the amount of extra processing power, download or upload speed, and disk space that the project needs fluctuates depending on what users are doing. Determining up front how to account for those needs is difficult and error prone, however, when using cloud services, malfunctioning disk drives automatically fail over to working ones,

and processes can scale almost infinitely. They also don't have to worry that their internet connection is stable or has enough bandwidth — a large proportion of Amazon's servers around the world would have to be impacted before it would affect the SPARC Portal.

The SPARC Portal launched in July 2019, and, at the time of our visit, it contained approximately 10TB of data, about 3TB of which is out of embargo and publicly available. In the first eight months of 2019, 35 datasets have been published and each dataset has one or more versions.

SPARC expects their overall data collection to increase to at least 100TB over the next year. The Data Core hosts a wide variety of data types including imaging data (mostly microscopy), metadata (as CSV files), Word documents, PDFs, PowerPoints, -omics data, modeling and simulation data, and EEG traces. Currently, all data is submitted by SPARC-funded investigators; however, SPARC hopes to support community contributions in the future. The Blackfynn platform is HIPAA-compliant, and soon to be General Data Protection Regulation (GDPR) compliant, so the web user interface could support protected information. However, the SPARC Portal does not currently contain personally identifiable data.

Unpublished data on the Blackfynn Data Management Platform is stored differently than published data. Data and metadata that are still under embargo are located within the Blackfynn platform. Once the data and metadata have been reviewed and approved by the SPARC Data Curation Team (a subgroup of the Map Core), the dataset is published to Blackfynn Discover and the SPARC Portal. While submitted/embargoed metadata is initially uploaded as CSV files, once the dataset is approved for publication, it is stored in a Neo4j graph database on the Blackfynn platform, and the metadata is indexed using AWS ElasticSearch once published. Published metadata is exported as a mix of CSV and JSON files to allow users to utilize the platform of their choice when working with the data.

When a dataset is first published, it's "snapshotted" and given a DOI. Any user with manager access to the dataset can make changes, however only the dataset owners can publish new versions of the dataset with a new DOI. This system works well for most of the data types that SPARC supports, that is, anything that is uploaded as a physical file.

A challenge with linking to external repositories in the context of dataset publishing is that you lose control over the underlying assets and rely on the external resource to handle this. For example, if you create a snapshot of a dataset and assign a DOI, we can guarantee that the data will never change. However, if part of the data is a URL to an external resource, we cannot guarantee that the contents at that URL have not changed. For example, SPARC leverages Protocols.io to store protocols associated with a dataset. They rely on protocols.io and users to not update/change a protocol after the SPARC dataset is published.

Blackfynn told us that their two main goals are to ensure the scalability of the platform and to ensure a quality user experience. Most of their resources are being invested in scalable platform

development. Creating the infrastructure needed to support the work of a diverse range of researchers at the scale of petabytes of data, while adhering to security and global compliance requirements, is a gargantuan task. They are also putting significant effort into ensuring that their platform is friendly to both novices and technologically savvy users. This includes developing both a point-and-click user interface and enabling programmatic data access using a Python or MATLAB API or by using a Command Line Interface (CLI).

Analysis

SPARC does not have a dedicated workspace for users to do custom analysis, however it does host several specialized analysis tools for working with specific datasets, and a number of simulation models. Researchers publishing data to SPARC also can (and do) publish their pipelines or links to pipelines in their publication, and the Data Core is considering adding features such as embedded Jupyter notebooks to allow users to run custom, cloud-based analysis pipelines within the portal in the future.

One of the biggest challenges is how to fit a platform like this into researchers' existing workflows. Most of their users have established, custom workflows and have no idea how to even begin changing their workflow to take advantage of Blackfynn. Given the very large datasets that are made available through the SPARC effort, there is a need to educate investigators to run analyses in the cloud and avoid egress charges on the data.

Blackfynn has made instructions available to spin up AWS compute instances close to the data, but these are not intended for a general audience at this point. However, a significant effort is underway to leverage the public, standardized nature of the Blackfynn Discover platform to run analysis over public data using the users' own AWS account. By training users how to leverage their own AWS accounts to freely interact with published data, Blackfynn 1) reduces the likelihood of runaway analysis costs on the platform side, and 2) removes any security concerns that arise from running arbitrary code on behalf of a user in the platform.

Access

At the time of this writing, 23 datasets are openly available on the SPARC Portal, with about forty more embargoed until January of 2020. About 60 different data generating sites are currently active, and there are between 300 and 400 portal users. Most of those users are data uploaders, but the platform is slowly moving towards supporting workflows for combining data and modeling.

The SPARC platform supports several modes of data access. For datasets less than 5GB, users can directly download a zipped archive of the data and any provided metadata files for free. For larger datasets, users can transfer data from the SPARC AWS bucket to their own or access data using any of the AWS tooling. This is generally free within AWS regions. Users can also use an AWS account to download the data, or to move it to another location using AWS

tools or clients. The SPARC Data Core made the important point that “open data is not free data” at large scales. Their bucket billing system is currently set as ‘requester pays’, so users who download data, move it to another AWS region, or transfer it to another host will be required to pay data egress charges. SPARC public data is indexed through Google datasets.

Harmonization and Metadata

Blackfynn’s goal is not to build a platform to share data, but to build a platform for data management that makes it easy to share and publish scientific data. Their philosophy is that scientific data is more than just files, and that meaningful data sharing should include rich, descriptive metadata. For this reason, the Blackfynn platform doesn’t restrict users to a specific data model. Users are required to supply only five metadata values for the dataset to receive a DOI and be published: Title, Summary, Description, Contributors, and Tags. Users can supply as much additional metadata as they want to describe their datasets, using any standard or user-defined metadata model. Users are not required to fit their data to a schema, but they are given the option to map their data to one or more schemas, as they see fit. Users can upload data files in any format and can include any number of supplementary files (such as PowerPoint presentations) or link to outside resources to enhance the reusability of their data. The SPARC program, however, does require that users follow a pre-fixed data structure, ontology/data standard, and specific file formats based on the BIDS (Brain Imaging Data Structure) specification (https://docs.sparc.science/submit_data.html#1-creating-a-draft-dataset).

Blackfynn’s strategy for harmonization has typically not been to define a strict metadata schema up-front and require users to fit their data into a specific schema. Rather, they work with users to define a schema that best fits their needs. Harmonization can happen retrospectively by smart mapping between the schemas. Rather than try to make data interoperable by mapping it onto the same metadata model, the Data Core envisions a future where data models are organically and progressively linked to ontologies. They have done some work on building classification algorithms that can do this. Currently, their software can take two datasets that use different metadata, but that were built from the same samples, and identify which metadata terms should be joined, with minimal user input.

Sustainability

Blackfynn told us that they were chosen as the Data Core for the SPARC program because of their company’s focus on data sustainability, security, and scalability. The company’s philosophy on data management is also important. They view data sharing as more than just making files available and are thinking strategically about how to interpret and contextualize these datasets and how to enable people to interact with them.

Their platform is a great example of this philosophy in action. It is designed as a data management system that happens to make it easy for users to share their data, rather than as a data sharing platform. This distinction is important, because their platform is primarily about

making their users daily data workflow easier, by giving them simple tools to organize their datasets, from the moment the data is collected. The platform also allows users to store related files like PowerPoints, or any other associated data, so it's all in one place. This means that when it is time to share a dataset, the data, and all the other files the researcher has been using, are already uploaded and can be published together. This helps to ensure that data published on the SPARC Portal is complete, and easily reusable by other researchers.

By creating the platform as a data management tool, Blackfynn also hopes they can encourage people to adopt shared standards and other best practices. For instance, providing the appropriate text suggestions when implementing autocomplete capability for metadata fields, similar to predictive text options in a Google search, can help make metadata more consistent across datasets and improve interoperability. The predictive text shows the researcher similar terms that already exist on Blackfynn in other datasets and lets them easily use those same terms. The user can still choose to finish typing their own custom terminology, but by offering easy selections, the system encourages users to all adopt the same set of metadata terms. Blackfynn also hopes to create incentives for their users to foster the use of data repositories and to make data sharing a common practice within the Neuroscience community.

Training

Internal

The SPARC Data Core did not express any dissatisfaction or problems related to their internal communication methods. Given that the DRC is distributed around the world, it will be interesting to talk to the Map Core and Simulation Core to see how well it is working across the consortium.

External

About 80% of SPARC users are academics in the Neurosciences and the other 20% are clinicians. The SPARC award does not have a training mandate, however, Blackfynn provides user documentation for their web application (help.blackfynn.com) and developer documentation (developer.blackfynn.com) for programmatic access to the platform, and they are planning to add short tutorial videos in the next year. Furthermore, Blackfynn manages the SPARC Program documentation page (https://docs.sparc.science) detailing the overall workflow of the collective DRC. They also hold a ~monthly webinar, and occasional outreach, however they would be interested in improving and increasing the frequency of these events. The team was also very excited about collaborating with the CFDE to host hackathons.

The SPARC Data Core told us that their support burden, and that of the Map Core, for these users is primarily about big data. Most users want to download data, and don't understand that downloading terabytes of raw data is not efficient or sustainable. These users also usually don't know how to use the cloud or realize that it is a better choice. Blackfynn also told us about problems with data upload: "We have had users drag 50,000 files into a browser window

expecting that data would automatically upload instantly, and we have had users upload single files that were larger than 130GB."

The team also told us that they anticipate that providing the right tools to enable cloud-based analysis of data will be the next big challenge for their platform, and that they would like to offer the ability to run custom cloud-based analysis pipelines. The challenge is figuring out what the researchers actually want to do and creating both the tools and training. The goal would be to bring biomedical data scientists with varying levels of technical expertise closer to the ability to use Blackfynn effectively.

FAIR

The Data Core told us that they are 100% committed to making a FAIR resource. They are members of DataCite, ORCID, and the Research Data Alliance and actively work with those groups to further this mission. They currently mint DataCite URLs for data published in their platform and follow previous guidance from the Data Commons program regarding identifiers for their data. In our meeting, the Data Core took a very pragmatic stance on FAIR, and talked about it mostly in terms of data ownership and the practicalities of data movement.

When talking about Findability and Accessibility, the SPARC Data Core differentiated between how these words are used in theory, and what they should mean in practice. Blackfynn pointed out that just putting data in a repository doesn't necessarily make it findable or accessible in a practical sense. "Accessible to us means the user should be able to get many Terabytes of data available to them in an easy and scalable way."

Data may be very difficult or impossible to find inside a given repository for any number of reasons. For instance, sensitive data, like that at dbGAP, has many metadata terms that could be of interest to a user, but that are hidden until the user has been granted access, which makes discovery of datasets impractical. Or, some data may not be well suited to the faceted search at its home repository, which keeps it from being found by most users. Even when a user finds data, it is only accessible if the user can actually use the data. A very large dataset at a repository that only supports data download may be functionally inaccessible to users, as they are restricted by both their local infrastructure and internet bandwidth.

There are also monetary concerns around accessibility. Data hosting and egress charges have traditionally been covered by the institution that is funded to share the data. However, repositories increasingly don't have the funding to do this at scale. The Sequence Read Archive (SRA), for instance, stopped accepting large sequencing project data years ago, mostly due to budgetary constraints. For data in the cloud, the largest cost is typically data download: "To mitigate these costs, many repositories either limit the size of datasets or limit the throughput on downloads. However, this goes squarely against the FAIR principles and results in repositories that have the notion of data-sharing while in fact the data in not truly available." The Data Core suggested that the ultimate fix for this problem needs to be user education and training.

Researchers need to be aware of the time and money required for downloading data, and that working in the cloud is a much faster and cheaper option. However, to work in the cloud, users need training that isn't readily available right now.

Even with a better educated user base, SPARC told us that there needs to be a long-term sustainability plan for data storage and use costs, and cautioned against adopting a system where the NIH pays data egress fees directly, without oversight:

> "If you do provide a platform that enable scalable, high-throughput data access to very large amounts of public data, one needs to take in consideration the cost that could be incurred by users. For example, what if a graduate student writes a script to download the entire public repository each night. What if this is a student in a different country, what if this is mal intended? Given the high availability of the resource, it would be very easy to have hundreds of thousands of dollars in unexpected costs within a couple of days."

Finally, there are additional concerns with thinking not only about FAIR but also about making data publicly available. In a discussion about data ownership, the SPARC Data Core told us that one difficulty with making data public, is that often several entities claim rights to the same dataset, and have different views on where and how it should be stored and accessed:

> "We strongly believe that data on our platform is owned by the users of our platform and not Blackfynn. However, our experience with academic institutions is that they also claim rights to the data, even though the NIH mandates data sharing. It would be great to have a discussion on mechanisms to break through this impasse."

The SPARC Data Core also told us that although they believe they have made their stance on data ownership clear, they still frequently encounter resistance from academics who are wary of hosting their data on a portal ostensibly owned by a corporation.

DCC Cross-pollination

The SPARC Data Core told us that they would be very interested in participating in a Common Fund Program community at many levels. While they have not identified any potential partners, they would be interested in participating in paired initiatives for demonstrating data reuse or building integrated datasets. As the Data Core has a deep interest in data ownership and sustainability, they also expressed excitement about the possibility of participating in CFDE governance and helping to establish best practices and standards.

They also expressed interest in less focused community engagement such as participating in discussions and workshops at meetings for Common Fund Project PIs. In particular, they suggested that a "bring your analysis to the data" workshop with one or more Programs, would

be an interesting way to test scaling analyses out into the cloud. Additionally, they suggested that NIH attendance at these events would be a useful way for the Programs to engage the NIH on complex, cross-program issues.

SSO

Blackfynn provides its own user management and is planning to roll out single sign on (SSO) for their web applications in the near future. They expressed concern about the use of other single sign on systems for several practical reasons. As the SPARC Portal is hosted and managed by Blackfynn, and is part of their corporate constellation of systems, it falls under their service level agreements, that is, they guarantee a certain, small, maximum level of downtime, and have a number of processes dedicated to ensuring their system is always available. If they were to use an SSO service provided by another entity, such as the eRA Commons, they would lose that control over the system. Anytime the eRA Commons was down, either for maintenance, or due to technical issues, the SPARC Portal would be inaccessible. Worse, SPARC (and any other Common Fund Program using that SSO) would have no ability to fix or mitigate the problem.

Outcomes

Infrastructure Reuse

The SPARC data management model of building a portal (see the *Sustainability* section) is a good candidate for technology/philosophy that could be integrated into the larger Common Fund Program community. Many other Common Fund Programs have told us that their data creators are willing to submit data back to the coordinating center, but that getting those creators to submit the needed metadata in a useable format is much harder. The SPARC Portal encourages users to put all the data and files (including things like PowerPoint presentations) related to a project in a single place as they are created, which lowers the burden on the user at the time the data is shared. It also minimizes the metadata requirements while allowing users flexibility to include additional metadata as desired.

Although it is still in the planning stages, Blackfynn also told us about an open source ETL (extract, transfer, load) pipeline that they would like to build, which would be useful across the Neurosciences, and potentially the entire Common Fund. This pipeline would essentially be a cloud-based file format converter that could be used to make data more broadly reusable. Since many laboratories have their own data management systems and file formats, this would provide them with an easy way to get data into data sharing platforms or get data from these platforms into their workflow. It would also allow the community to make new file format converters that leverage the power of the cloud instead of relying locally run scripts.

Challenges

The Blackfynn team described a number of challenges with two main foci: user training and sustainability.

- Users on the SPARC Portal frequently do not understand how to work with data at a large scale. Blackfynn related many stories of users attempting to upload or download extremely large datasets that overwhelmed the users' web browser, as well as problems with users attempting to upload or download hundreds of smaller files simultaneously, with similar results. However, there is no clear articulation of what SPARC-specific training should be — "There is a lot of urgency as long as you're in a conference call."
- The SPARC Data Core told us that the biggest threat to sustainability is the misconception that "Open Data is Free Data." There will always be costs associated with data download, storage, transfer and analysis, and while it is possible to make data available for free to users, that increases the cost of hosting the data, which is of particular concern after the Common Fund Program that generated the data has ended. Sustainability is defined by finding the best way to distribute these costs.
- Data ownership ambiguity is also a sustainability threat. The NIH funds the projects, and so ultimately owns the data; the data creators, and their institutions also often claim ownership rights. This ambiguity impedes data sharing and FAIRness.
- Blackfynn also noted some practical concerns regarding the implementation and meaning of FAIR principles. For example, if FAIR means that each file needs a DOI, there are significant complications for the service that mints these DOIs (e.g., DataCite.org). If all Common Fund Programs started minting each file, the DOI services would all come to a standstill. For SPARC, each dataset will get a single DOI per version, and each file within a dataset is assigned a unique ID but does not need a global DOI.

Potential Solutions

- Training and Support
    - Run hackathons for SPARC tools and interface enhancements. This would both help with refining the SPARC tools and with discovering new use cases that are important to their users. There is value in getting technical folks and users in the same room to better understand how the tools are being used and what is needed on the infrastructure side.
- Sustainability
    - Increased interaction between Common Fund Programs, and Common Fund Programs and the NIH. Blackfynn was very interested in bringing the NIH into workgroups and brainstorming sessions to help address the many challenges to sustainability that they see both within their program and across the Common Fund.

- From our discussions with the programs, we suggested that SPARC and LINCS might mutually benefit from discussing their approaches to data and potentially doing some analytic collaborations.

Potential Projects

A theme for training could be, “bringing your analysis into the cloud/close to the data.” This would help to address both training (“open data is not free data”) and sustainability (transfer costs). Potentially, these could be a CFDE-wide pilot workshop with KF/Cavatica, GTEx/Terra.

Game Changer

Open Source ETL (extract, transfer, load) Pipeline

If Blackfynn were funded to create the ETL they envision, it would be a game-changing development towards interoperability. They want to design and build a cloud-based file format converter that could be used to make data more broadly reusable, as it would allow a user to translate a file into a wide range of other file types and formats. They estimate that creating this pipeline would require around 500K of funding per year for 2-3 years, as well as a consortium of stakeholders who would like to participate. They already have good connections with Neurodata Without Borders, the INCF, and other academic partners that would be willing to participate.

Agenda

Day 1

9-9:30am Introductions

Short introductions from engagement team members and attending DCC members. The overarching goal for the engagement team is to collect value and process data about the DCC. Values data will include things like: mission, vision, goals, stakeholders, and challenges. Process data includes: data types and formats maintained, tools and resources owned by the DCC that they would like to have broader use, points of contact for follow up on technical resources, etc.

9:30-10am DCC overview

Short overview of DCC. Can be formal or informal, choose 1-5 topics to cover. Suggested topics: What is your vision for your organization? What big problems are you trying to solve? What are your big goals for the next year? Who do you see as your most important users/stakeholders? What project(s) is currently taking up the bulk of your effort/time? What areas of your organization are you putting the most resources into? What is the rough composition of your user base in terms of discipline? Do you have any challenges that are blocking implementation of your current goals? What skill set would you like to add to your project? How do you engage with your users? What kind of sustainability issues are you confronting? Can you currently do combined analyses with external datasets?

10am-Noon Goals Assessment

An exercise to get an idea of what types of things are important, what types of things are challenges, what do you dedicate your time/resources towards, and what types of things are not current priorities. Given a list of common goals provided by the engagement team, plus any additional goals the DCC would like to add, DCC members will prioritize goals into both timescale: “Solved/Finished”, “Current-Input wanted”, “Current-Handled”, “Future-planned”, “Future-unplanned”, “NA to our org” and for desirability: “Critical”, “Nice to have”, “Neutral”, “Unnecessary”, and “NA to our org”. The engagement team will work to understand the reasons for prioritization, but will not actively participate in making or guiding decisions.

Goal List

- Increase end user engagement X% over Y years
- Move data to cloud
- Metadata harmonized within DCC
- Metadata harmonized with _____
- Metadata harmonized across Common Fund
- Implement new service/pipeline _____
- Increase number of visitors to your site
- Common Fund Data Portal
- Single Sign On
- Pre-filtered/harmonized data conglomerations
- A dashboard for monitoring data in cloud
- User-led training for end users (i.e. written tutorials)
- Webinars, MOOCs, or similar outreach/trainings for end users
- In-person, instructor led trainings for end users
- A NIH cloud playbook
- Full Stacks access
- Developing a data management plan
- Increased FAIRness
- Governance role in CFDE

Lunch

1 - 3:30pm Open discussion (with breaks)

Using the results of the morning exercise and a collaborative format, iteratively discuss goals, blockers, etc., such that the DCC agrees that the engagement team can accurately describe their answers, motivations and goals. Topics don’t need to be covered it order, we’d just like to touch on these types of questions.

Topics:

Infrastructure:

- Do you intend to host data on a cloud service?
- Have you already started using cloud hosting? If yes:
    - Approximately how much of your data have you uploaded? How long did that take? How are you tracking progress?
    - What challenges have you faced?
    - How have you dealt with those challenges?

- What potential future problems with cloud hosting are you watching for?
- Does your org use eRA Commons IDs? Do the IDs meet your sign on needs?
    - If yes, did you have/are you having challenges implementing them?
    - If no, what do you use? What advantages does your system provide your org?

Use cases

- What is the rough composition of your user base in terms of discipline?
- What if any, use cases do you have documented? Undocumented?
- What things do people currently love to do with your data?
- What things would people love to do with your data, but currently can't (or can't easily)?
- What pipelines are best suited to your data types?
- What are the challenges associated with those desired uses?
- What other kinds of users would you want to attract to your data?

Review of metadata:

- What's metadata is important for your org? For your users?
- Do all your datasets have approximately the same metadata? Or do you have many levels of completeness?
- Do you have any data already linked to outside resources?
    - Did you find the linking process easy? Challenging? Why?
- What kinds of datasets would you like to link into your collection?
- What implementation and schemas do you already have (or want)?
- What standards do you have (or want)?
- What automated systems do you currently have for obtaining metadata and raw data?

Training:

- What training resources do you already have?
- What training resources would you like to offer? On what timescale?
- What challenges keep you from offering the training you'd like?

Policies:

- How do users currently obtain access to your data?
- What are your concerns about human data protection?
- What potential challenges do you see in bringing in new datasets?

FAIR:

- Has your org done any self-assessments or outside assessments for FAIRness?
- Are there any aspects of FAIR that are particularly important for your org?
- Are there any aspects of FAIR that your org is not interested in?
- What potential challenges do you see in making your data more FAIR?

Other:

- What search terms would make your data stand out in a shared DC search engine?
- Does your org have any dream initiatives that could be realized with extra resources? What resources would you need?
- If you had free access to a Google Engineer for a month, what project would you give them?
- Any other topics/questions the DCC would like to cover

9-10am Review of goals and CFC involvement
A quick review of what topics are priorities for the DCC with suggestions from engagement team on how we can help.

10-noon Open Discussion
DCC reflection on suggestions, open discussion to find shared solutions.

Lunch

1-2pm Thoroughness checking
Touch on any questions not covered previously, ensure we have:

- Action Items for us, and rough timelines for getting back to DCC on them
- Tools / resources the DCC thinks might be useful for the overall project
- Points of contact “Who is the best point of contact for your metadata schemas, your use cases, the survey of all your data types?”
- Who would like to be added to our governance mailing list?
    - Or contact info/instructions on how to get that information offline.

Appendix C: FAIR Assessment Plan for 2020

Written by Avi Ma'ayan, Daniel J.B. Clarke, and Sherry Jenkins

What are FAIR assessments, and why are they needed?

It is accepted that we need to do a better job with making Common Fund (CF) datasets more findable, accessible, interoperable, and reusable (FAIR) [1]. However, exactly how to achieve this goal is challenging. The FAIR principles provide a framework that covers most of the general things that would need to be considered when going through the process of CF digital products FAIRification. Hence, the FAIR principles serve as a guide for making sure that we "don't forget anything". One way to achieve awareness of compliance with FAIR is to perform FAIR assessments. FAIR assessments can be performed by mapping compliance of a digital resource with a specific FAIR requirement, for example, whether a dataset can be accessed via a well-documented API, or whether the website that is hosting a dataset has a license that covers the terms in which the dataset can be used [2]. The process of FAIRification can then be coupled to an evaluation of FAIRness that measures whether the activities, services, and products generated by the CFDE project cover all the FAIR requirements. In addition, FAIR assessments can inform the CF programs' data coordination centers (DCCs), and the NIH, about existing gaps that need to be filled. These are gaps between the current state of the data, and other digital objects, on DCC portals, and the required upgrades to make these digital resources adhere to community standards that would render them FAIRer.

What was achieved so far by the CFDE in regards to FAIRification and FAIR assessment?

- We developed and published FAIRshake (https://fairshake.cloud) [2], a system to manually and automatically assess the FAIRness of digital objects including datastes, tools, and repositories.
- FAIRshake provides FAIR assessments of datasets listed on 7 CF DCCs. These FAIR assessments are visualized as an insignia. FAIR analytics are automatically calculated for each CF DCC as well as for the collective of all CF programs bundled together.
- The publication that describes FAIRshake was accepted for publication in Cell Systems and it is currently In Press. An older version of the article is available on bioRxiv at: https://www.biorxiv.org/content/10.1101/657676v1
- We developed scripts to convert metadata that describe CF datasets from 7 CF portals into two community accepted schemas: DATS and Frictionless.
- The conversion of the CF datasets into Frictionless enables the upload of these datasets into DERIVA [3]. The database engine that is behind the CFDE portal.
- We assessed the FAIRness of datasets from 7 CF programs with a customized rubric that was created from the list of case studies developed by CFDE members.

- We developed scripts to automatically assess the FAIRness of the 7 CF programs. This required mapping metadata elements from each DCC to FAIR metrics that belong to the customized rubric. During this process we manually linked some metadata elements to existing agreed upon ontologies and dictionaries.

What do we plan to achieve in year 2020?

- Improve the FAIRness of CF digital resources by manually adding links to ontologies and dictionaries.
- Streamline and harden data ingestion pipelines by documenting versions, creating a portal that enables non-experts to execute these scripts, and associate each step with a FAIR assessment.
- Assess the FAIRness of tools and workflows by establishing a tools and workflows registry.
- Display FAIR assessment insignias on the CFDE portal.
- Harmonize datasets at the data level by developing and cataloging data processing pipelines. Convert datasets into dataframes that can be loaded into Python and R for further analyses including application of machine learning.
- Develop data visualization components that are independent and can be used as plug-ins to enhance the user experience at the CFDE portal. Develop protocols to enable the community to develop and contribute such data visualization components to the CFDE.

Below we provide more details about each of these planned activities:

Plans to continue to improve the FAIRness of CF digital resources

The experience that we had converting DCCs datasets into the DATS and Frictionless formats, and assessing these datasets for their FAIRness with specific use cases in mind, pointed out that much manual work remains toward improving the FAIRness of these datasets. Hence, much effort is needed to map fields to ontologies and dictionaries, and harmonizing metadata elements across programs. While this activity can be done by the DCCs after some training, we have the expertise to do much of it ourselves. In fact, during the conversion process, we have already done some initial manual mappings and harmonization. In 2020, we plan to continue to guide as well as perform additional FAIRification of CF resources.

Plans to streamline and harden data ingestion pipelines

In 2019, we developed prototype scripts to convert DCC datasets into common catalog consumable schemas, i.e. DATS and Frictionless. This process is critical for the harmonization and presentation of the CF data and metadata on the CFDE portal as well as for performing FAIR assessments. In 2020, we plan to automate and harden these data processing scripts. Specifically, we plan to automate the pipelines that convert metadata from DCCs->DATS->Frictionless->DERIVA (including provenance of all steps). These scripts will be tightly linked to dynamic FAIR assessments throughout process. We expect that the FAIRness

will increase/decrease after each processing step. We will also document these data processing scripts, as well as make their execution available from a dashboard with button clicks. This will allow us to track FAIRness over time of the DCC resources in addition to having the capability to propagate changes to the CFDE portal.

Plans to add FAIR assessment of tools and workflows

The initial assessment of the FAIRness of the CF resources in 2019 fully focused on data and datasets. The FAIR assessment plan for 2020 includes the indexing of tools and workflows produced by the CF DCCs as well as by other related community efforts. This effort will produce a catalog of CF tools and workflows with FAIR assessments. Since the metadata of tools and workflow will be organized in a similar way as the metadata for the CF datasets, such catalog of tools and workflow will be made available for searching and browsing on the CFDE portal.

Plans to integrate FAIR assessment visualization into the CFDE query portal

Once the DERIVA portal is working and available, we will enable the display of the FAIR insignia near each digital object that will be hosted on the site. We have already created all the needed hooks to enable such visualizations and initially assessed the FAIRness of digital resources from 7 CF projects. Hence, this effort will require coordination and handshake between FAIRshake and the CFDE instance of DERIVA.

Plans to harmonize datasets at the data level and prepare these datasets for machine learning and other complex data analysis and integration tasks

In 2019, effort went into processing and harmonizing the metadata that describes the CF datasets but the data contained in those datasets was untouched. In 2020, we plan to begin the systematic processing and cataloging of the actual data by identifying data levels, processing scripts, and developing harmonization strategies for abstracting the data to a level where it can be integrated. For example, a GWAS study that called variants, can be integrated with RNA-seq data by converting each data type into gene sets. Metabolomics profiling can be compared to RNA-seq data by applying a model that converts metabolomics data to RNA-seq data and RNA-seq data to metabolomics. In most cases, tools and workflows to perform such abstractions already exist, but these need to be better organized and tested. High level processed data provided with rich metadata will be delivered in dataframe formats that can be consumed by data analytics platforms such as R Studio or Jupyter Notebook (Python). These are data science commonly used data analysis platforms that would benefit from having easy access to CF datasets. Hence, we will develop R and Python libraries specifically for easy access to CF data and metadata. These libraries will access the data and metadata via a well-documented API. Having the data ready for integrative analyses, we will prepare examples that show how machine learning algorithms can be applied to such data. For example, we plan

to test whether we can predict metadata elements directly from the data. This particular example will also inform and accelerate the manual FAIRification efforts of CF datasets.

Plans to develop data visualization components

Once the CFDE metadata and datasets are loaded into a catalog such as DERIVA, the data and metadata will be made available for query and download via a user interface. Since these data and metadata will be in a format that is well structured, it could be systematically visualized. Such visualization can be for the purpose of providing summary statistics of what is in the catalog, as well as for exploring the high dimensional structure within the data. Such data visualization capabilities are expected to significantly enrich the user experience. Our group has extensive experience in developing such UI components. In addition, existing UI components developed by the DCCs, and by others, could be integrated into the catalog user facing portal interface. Hence, we plan to develop new, and integrate existing, data visualization components for the CFDE portal. We will focus on data visualization components that are concerned with comparing gene sets and signatures, and querying such sets and signatures against public databases. For example, we developed Clustergrammer [4], application: https://amp.pharm.mssm.edu/clustergrammer/ source code: https://github.com/MaayanLab/clustergrammer to visualize heatmaps of any data matrix. We are currently developing ScatterBoard https://github.com/MaayanLab/react-scatter-board to visualize any dataset as scatter plots that place data objects in 2D or 3D based on their similarity. These components are developed with the React framework so they can be embedded in any website, Jupyter Notebooks, or other web-based system that host biomedical datasets.

References

[1] Wilkinson et. el. Sci Data. 2016 Mar 15;3:160018.
[2] Clarke et al. Cell Systems. 2019 Accepted
[3] Bugacov et al. Proc IEEE Int Conf Escience. 2017 Oct;2017:79-88.
[4] Fernandez et al. Sci Data. 2017 Oct 10;4:170151.