The NIH *Commons*

Summary

The Commons is a shared virtual space where scientists can work with the digital objects of biomedical research, i.e. it is a system that will allow investigators to find, manage, share, use and reuse data, software, metadata and workflows. It will be a complex ecosystem and thus the realization of the *Commons* will require the use, further development and harmonization of several components.

Components of the Commons ecosystem include:

- A **computing environment**, such as the cloud or HPC (High Performance Computing) resources, which support access, utilization and storage of digital objects.
- Public **data** sets that adhere to *Commons* Digital Object Compliance principles.
- **Software** services and tools that enable;
    - Scalable provisioning of compute resources.
    - Interoperability between digital objects within the Commons.
    - Indexing and thus discoverability of digital objects.
    - Sharing of digital objects between individuals or groups.
    - Access to and deployment of scientific analysis tools and pipeline workflows.
    - Connectivity with other repositories, registries and resources that support scholarly research.
- A set of Digital Object Compliance principles that describes the properties of digital objects that enables them to be findable, accessible, interoperable and reproducible (FAIR).

A series of *Commons* pilots has been initiated to develop and test these components in order to understand and evaluate how well they will contribute to an ecosystem that will effectively support and facilitate sharing and reuse of digital objects.

The initial iteration of the *Commons* compute environment will be implemented using a federation of public and private computing clouds. As only a limited number of investigators today have access to such resources, it will be necessary to facilitate access to them in order to fully evaluate their use; accordingly, a *Commons cloud credits* business model is being tested, that is designed to provide unified access to a choice of *"Commons*-conformant" compute resources. This cloud credits model will offer individual investigators a choice of cloud provider so that the investigators themselves can select the best value for their individual research needs.

In addition to testing the cloud credit model that supports the use of cloud computing, several additional pilots have been initiated to implement and test other aspects of the *Commons* framework. These include:

- Model Organism Databases interoperability and cloud deployment: This pilot will test approaches to improve efficiency, cost and interoperability of essential data resources, which could help inform new approaches to long-term sustainability.
- BD2K Centers Commons Pilots: These pilots will test means of facilitating interoperability between and amongst different BD2K centers. The centers were chosen as a test group because of their already high level of computational skills.
- Human Microbiome Project (HMP) cloud deployment: approximately 20TB of HMP data will be made available on the AWS cloud, along with a suite of tools and APIs to facilitate their access and use. This pilot will expand upon earlier activities to make HMP data and tools more easily accessible to the broader research community, as any investigator interested in HMP data will have access to them. This pilot will also provide information on the utility of the *Commons* approach of making resources more readily findable, discoverable and usable for the typical investigator in a shared compute environment.
- NCI Genomic Data Commons (GDC) and Cloud Pilots. These parallel and complementary projects are designed to make cancer genomics data broadly accessible, computable, and usable by researchers worldwide. The GDC will store, analyze and distribute ~2.5 PB of cancer genomics data and associated clinical data generated by the TCGA (The Cancer Genome Atlas) and TARGET (Therapeutically Applicable Research to Generate Effective Treatments) initiatives. The NCI cloud pilots will make TCGA data available on the AWS and Google clouds, along with a suite of tools and APIs to facilitate their access and use.

These pilots are all designed to establish and test the workings of a *Commons* prototype and the ability of a range of investigators to use them to meet their computational needs better and more efficiently than can be done with current approaches. Assuming that these pilots are successful, additional pilots will be needed to develop successful ways of implementing the use of the *Commons* within the NIH grant system's processes.

Commons Framework

The Commons is defined as a shared virtual space where scientists can find, deposit, manage, share and reuse data, software, metadata and workflows - the digital objects of biomedical research. It is a digital ecosystem that supports open science and leverages currently available computing platforms in a flexible and scalable manner to allow researchers to transparently find and use computing services and tools they need, access large public data sets and connect with other resources associated with scholarly research (e.g. GitHub, zenodo, ORCID, Figshare, journal publishers etc.) Such a system must be adaptable to the different and evolving needs of research communities as well as the evolving technology innovations. Figure 1 shows the prototype *Commons* framework.

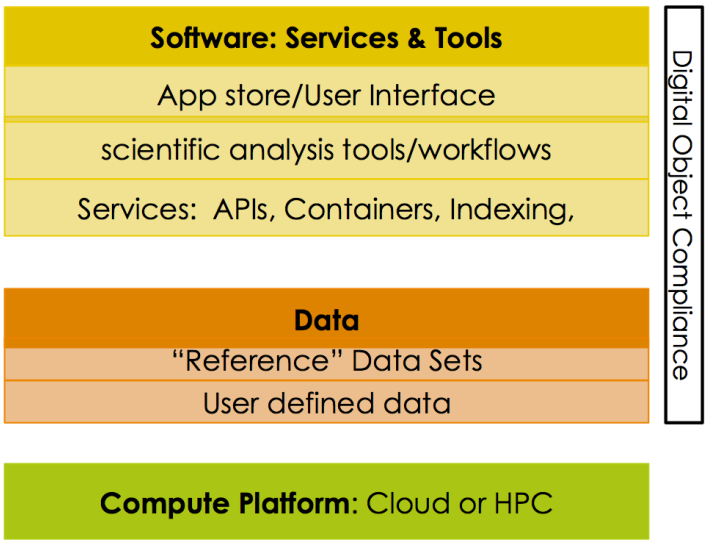

Figure 1: *Commons* framework

At its foundation the *Commons* framework requires a computing platform that, in its initial iteration, intends to take advantage of emergent public and private cloud computing capabilities and other capable compute platforms, e.g. university and national laboratory high performance computing (HPC) resources. Clouds are increasingly being used as a compute platform by biomedical researchers because they afford a high degree of scalability and flexibility in both cost and configuration of compute services.
The ability to co-locate public data sets, analytical tools and pipeline workflows, and to make the results accessible and shareable with others are key reasons to leverage cloud resources for *the Commons*.

Making public data, especially large commonly used data sets, easily accessible in the cloud will reduce the burden and cost of individual investigators independently moving these data sets to cloud, enable the ability to compute against data sets and permit new and novel uses across data sets. Adherence to a digital object compliance model will be essential in order to make these data sets indexable and easily discoverable.

Easily finding, deploying, linking and using computing services and analytical tools/workflows will promote rapid and flexible scientific discovery in the *Commons* and will make it easier for those with more limited computational skills to utilize the environment.
Services and tools can cover a wide range of applications. These include provisioning and deployment of compute resources, services that support access to and interoperability between digital objects (e.g. open APIs and containers), access to indexing services to find digital objects, simplified deployment of analysis tools and pipeline workflows and the ability to connect to other repositories, registries and resources that support scholarly research.

Digital Object Compliance for the *Commons*

Data sharing is a key objective of the *Commons,* and cloud computing directly supports this through the ability to co-locate data and tools, as well as to make data (including the processed data that is the result of analysis) and tools accessible and shareable by others. Access control can easily be implemented in the cloud so that data and tools can be appropriately and securely shared amongst groups authorized to use them, including the appropriate protections and access for human subjects data. Thus, while the cloud provides a computing environment to share data and tools in order to be able to effectively use these digital objects, they must have attributes that make them **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR). The *Commons* is intended to be a system that will do so.

A set of Digital Object Compliance principles that supports FAIR is currently under development. The Digital Object Compliance principles are expected to evolve over time as the ability to make digital objects meet the FAIR criteria increases, thereby improving the ability of digital objects to be shared and used more easily and effectively in the *Commons*.

To meet the most basic level of compliance, it is expected that digital objects would have the following elements:

- Unique digital object identifiers
- A minimal set of searchable metadata
- Physical availability through a cloud-based *Commons* provider

- Clear access rules and controls (especially important for human subjects data)
- An entry (with metadata) in one or more indices

At higher levels of compliance, digital objects would include additional elements that might include:

- Standard, community-based, unique digital object identifiers such as DOIs, ORCID IDs, Github IDs etc.
- Adherence to community-approved standard metadata for enhanced searching
- Data accessibility and exchange via open and/or standard APIs
- Software tool and pipeline workflow encapsulation using containers or other technology for easier deployment and use in cloud or other virtualized environments
- Availability in a *Commons*-compliant computing environment (i.e., digital objects are either directly copied into this environment (physically present) or exist outside of this environment but are indexed, findable and usable by being compliant with FAIR principles).

Commons Pilots

The primary goal of the *Commons* is to support the sharing and reuse of digital objects in a virtual space. Achieving this goal requires that all the elements described above in the *Commons* framework (a computing environment, data, software, and digital object compliance) work together effectively and efficiently. Each of the elements has its own complexities and it will take effort to develop the detailed principles and foundations needed for each element and then harmonize them in order to achieve a truly effective virtual research space.
A series of *Commons* pilots has thus been initiated to develop and test each of these elements in order to understand and evaluate how well they, together, will effectively support and facilitate sharing and reuse of digital objects.

Cloud Credits Pilot: A Business Model to Support the Use of Cloud Computing for the Commons

For this pilot, the initial iteration of the *Commons* will be tested using a federation of public and private computing clouds, with the choice of cloud provider being made by each individual investigator who can select the best value for her/his individual research needs. This business model is shown in Figure 2.

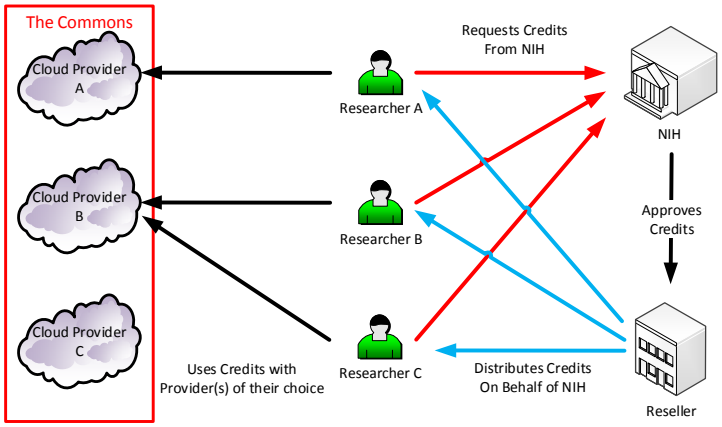

Figure 2: *Commons* Cloud Credits Business Model

In this pilot, the participating researchers obtain '*Commons* credits', dollar-denominated vouchers that can be used with the cloud provider of the investigator's choice. The involvement of multiple cloud providers will empower investigators by creating a competitive marketplace where researchers are incentivized to use their credits efficiently and cloud providers are incentivized to provide better services at the lowest possible price.

In order to participate in the *Commons*, a cloud provider must make its computing environment 'conformant', ensuring that it meets a set of NIH standards for capacity (storage, compute, network) and capabilities that enable scientists to work in such an environment. NIH would not directly distribute credits in this pilot; rather, it will contract with a third party, an HHS Federally Funded Research and Development Center (FFRDC) to manage the requests for and distribution of credits (shown as the 'Reseller' in Figure 2).

The business model being piloted offers several advantages. It is scalable and creates a competitive marketplace that, in conjunction with the elasticity of public clouds, should provide investigators with a cost-effective way of accessing cloud computing resources. Using this approach will provide the system with flexibility to respond to the computing needs of various groups at multiple scales. It should also reduce the barrier to uploading and sharing *Commons*-compliant digital objects to the cloud.

The use of a relatively small number of providers, coupled with a single reseller distributing credits provides NIH with an opportunity to assess the usage of digital objects that are being supported and maintained in the *Commons.*

The most significant disadvantage to this model is that it is pay as you go; that is, digital objects may no longer remain in the *Commons* if the NIH does not continue to pay for their maintenance. The proposed business model is primarily designed to increase the

ability of investigators to use the *Commons* cost-efficiently. The system will also allow the NIH to address another aspect of cost efficiency, namely the cost of maintaining the necessary digital resources long-term. The system will allow the gathering of data about both the use of digital objects in the *Commons* (not possible with current funding approaches) and the research output of that use, which then provides an opportunity to make data-driven decisions about which digital objects should be maintained and which are no longer cost-effective to maintain. An example of where this approach could be useful is with the Model organism Databases (MODs)

Model Organism Databases (MODs) Interoperability & Availability in the Cloud

Long-term use, re-use and sustainability of essential data resources are primary concerns to NIH, and therefore approaches that improve efficiency, cost and interoperability of such resources are a high priority. Model Organism Databases (MODs) provide a focused and discrete set of highly valued resources for development and testing of possible solutions. Many of the MODs already use or are considering development of ways to make themselves faster and more scalable, stable, interoperable, shareable, and cost-efficient. Such improvements would ultimately impact sustainability and interoperability of these resources, and many of these are in alignment with elements of *the Commons* framework described above. To further these nascent efforts and support their potential contribution to the *Commons*, a pilot has been initiated to support foundational work on how best to utilize cloud computing and the improvements needed to interoperate, and hence share, MOD resources.

BD2K Centers: Testing the Commons Framework

The 12 BD2K Centers of Excellence for Big Data Computing offer a unique opportunity to develop collaborative approaches to sharing and reuse of data objects within cloud environments amongst a relatively large, complex and diverse group of projects. BD2K Centers *Commons* pilots will develop and test elements of the *Commons* framework that will facilitate interoperability between and amongst different BD2K centers. Digital objects from the BD2K Centers should conform to the digital object compliance principles and indexed using methods developed by the BD2K Resource Indexing group (e.g. bioCADDIE).

Cloud-Based access to HMP (Human Microbiome Project) data and tools

The Human Microbiome Project (HMP) was funded by the NIH Common Fund to generate data and resources to characterize the commensal microbiota present in the human body. Mining these data offers great promise for understanding human health and for identifying new diagnostic and therapeutic targets. The initial phase of this project generated over 20 terabytes of data that was stored at NCBI and the HMP Data Analysis and Coordination Center (DACC). However, in order to analyze the HMP data, researchers currently must download data to, and install tools on, their local infrastructure, which requires significant computing resources that may not be available to all users. In response to these challenges NHGRI and NIAID in collaboration with Amazon Web Services (AWS) have made these data available on the AWS cloud.

The purpose of the HMP *Commons* pilot is to:

- Improve our understanding about accessing and sharing digital objects (HMP data and tools) in a cloud environment.
- Develop newer approaches for data access and for deploying tools in the cloud.
- Develop API access between the data, the tools and the web interfaces.
- Use unique identifiers for these shared digital research objects, which will facilitate finding and searching the data in the cloud.
- Update the AWS Cloud with the remaining HMP data (~20TB).

NCI Computation Genomics Initiatives

Two parallel efforts are underway in NCI to make cancer genomics data broadly accessible, computable and usable by researchers worldwide. The NCI Genomics Data Commons (GDC) is a data system at the University of Chicago that will store, analyze and distribute ~2.5 PB of cancer genomics data and associated clinical data generated by the TCGA (The Cancer Genome Atlas) and TARGET (Therapeutically Applicable Research to Generate Effective Treatments) initiatives. In addition, the GDC will also be a transactional system that will accept new cancer genomics data from NCI projects and from other researchers who wish to share their data broadly. In the NCI Cloud Pilots, TCGA data will be made available on the AWS and Google clouds, along with a suite of tools and APIs to facilitate their access and use. The NCI Cloud Pilots have been awarded to the Broad Institute, SevenBridges Genomics, and the Institute for Systems Biology. They are working closely with the NCI GDC to build a next-generation genomics and clinical data resource, including the possibility of extending the GDC into the commercial cloud. Both the Cloud Pilots and the GDC aim to implement accepted genomics standards, such as those being developed by the Global Alliance for Genomics and Health (GA4GH). Each of these NCI computational systems have NIH Trusted Partner status and will serve as demonstration projects that can inform the NIH *Commons* in

various ways, including how to interoperate between a dedicated resource and multiple cloud resources while maintaining the FAIR principles of Findable, Accessible, Interoperable, and Reusable for data, algorithms, and computation.

Commons Evaluation

The MODs, BD2K centers and HMP *Commons* pilots outlined above will run for 1-2 years, during and after which NIH will evaluate the effectiveness to meet the goals of the ADDS strategic plan. The NCI GDS and Cloud pilots will run for up to 5 years. The ADDS office is working closely with the NCI to collaborate on metrics and discuss outcomes so as to better inform the ADDS strategic plan.

We anticipate issuing contracts to have a 3rd party experts in evaluation approaches to address questions such as:

- Has the *Commons* promoted research object sharing beyond what would have happened with traditional computing environments?
- Does indeed the *Commons* business model lead to a better use of NIH funding?
- Are there indications that the *Commons* supports a more sustainable research environment than what we have today?
- Is their evidence that the *Commons* will accelerate scientific discovery beyond what we have today?

Summary Comments

This document is intended as starting point to foster further discussion about the concept and framework of the *Commons* and how to enable a digital ecosystem.

We fully expect the concept and framework to evolve and change over time and that the Commons pilots, which are the initial experimental steps, will provide one path towards a better understanding and evaluation of the Commons framework. Community engagement is key so we welcome comments and suggestions from the community.

For further information about the Commons please contact

Vivien Bonazzi: bonazziv@mail.nih.gov