Supporting Open Data and Open Science With Data Commons: Some Suggested Guidelines for Funding Organizations

Robert L. Grossman
University of Chicago
and Open Commons Consortium

March 23, 2017
Draft 0.7

1 Introduction

With the growing importance of high throughput data and the rapidly falling costs, biomedical researchers today routinely produce large datasets. The challenge for them is how to analyze, integrate, and share the large datasets that they produce. The challenge for funders is how to encourage the researchers they support to share their data so that research discoveries in a field can be accelerated and so that the patients can benefit from research advances more quickly.

2 What is a Data Commons?

Advances in large scale cloud computing made by companies such as Google, Amazon and Facebook have disruptively changed many industry sectors, including how we communicate and how companies target us with advertising, but, with some exceptions, have not changed how we do biomedical research.

Over the last several years, large scale cloud computing has been applied to biomedical research [4, 3, 1]. One of the technology platforms that has been developed is called a data commons. *Data commons* co-locate data, storage and computing infrastructure with commonly used software services, tools and applications for analyzing and sharing data to create a resource for the research community [2].

3 The Benefits of a Data Commons

In this section, we review some of the benefits of using a data commons to manage, integrate, analyze and share data.

1. The data is available to other researchers for discovery, which moves the research field faster.

2. Some diseases are dependent upon having a critical mass of data to provide the required statistical power (e.g. to study combinations of rare mutations in cancer).

3. Data commons support repeatable, reproducible and open research.

4. With more data, smaller effects can be studied (e.g. to understand the effect of environmental factors on disease).

5. Data commons enable researchers to work with large datasets at much lower cost to the funder than if each researcher set up their own local environment.

6. Data commons provide a funder's donors with greater impact for the dollars they contribute.

7. Data commons allow the data generated by your researchers to impact the research in other fields beyond a funder's specific diseases of interest, providing a broader societal impact.

8. Data commons generally provide higher security and greater compliance than most local computing environments.

9. Data commons support large scale computation so that the latest bioinformatics pipelines can be run.

10. Researchers usually "move on" once their data is analyzed. When projects continue to maintain and update their data, there is a much greater research impact. A good example is provided by the large amount of research that has been generated by The Cancer Genome Atlas (TCGA) project, which has been managed and maintained as an atlas long past the original data has been collected and analyzed.

4 Three Guidelines to Those that Fund Researchers

Guideline 1. Require that researchers supported by your organization share the data generated by research that you fund. This is the first and most important guideline and provides two critical benefits: First, creating a critical mass of research data co-located with useful tools and services accelerates discovery for researchers. Second, bringing together data about health outcomes contributes towards a learning health system with enough statistical power to make the best available decisions about patient care.

Guideline 2. Funders should provide the computing infrastructure and bioinformatics resources that is required to support data sharing. Researchers don't have the IT and bioinformatics experience to build data sharing platforms. It can also be a lot of work to contribute data to data sharing platforms (it data is just dumped, its not very useful to anyone). Funders should build their data commons or contribute to existing data commons and either provide research projects with sufficient funding so that there are bioinformaticians available to curate and add data to data commons or fund, or contribute funding, to a centralized resource that can help researchers prepare their data so that it can be uploaded to data commons.

Guideline 3. The data commons supported by funders should themselves share data. With the appropriate software services, a data governance structure, and an operations center, different commons can themselves share data so that a researcher using one commons can access data from another commons transparently. This is sometimes called *data peering*.

Here is a guideline for the future:

Guideline 4. Plan for a future when patients are empowered to contribute their own data to data commons. Within several years, patients will be able to direct their health related data from a provider to a research organization of their choice. Blockchain and other emerging technologies will enable this to be done in a secure and privacy preserving fashion.

5 Setting Up and Operating a Data Commons

Figure 1 shows one approach for setting up data commons sponsored by one or more funders. This approach uses the open source Bionimbus software stack, the same software stack that was used to developed the NCI Genomic Data Commons [3] and the BloodPAC Data Commons [1], and a framework that has been developed over the last two years by the Open Commons Consortium.

The main components of the OCC Framework are:

1. A Project Sponsor who is responsible for establishing the charter, coordinating, funding, and related activities. For example, as a sponsor of d data commons, a funding organization can set up and operate their own commons, join other funders to set up and operate a commons, contribute to an existing data commons, or fund a third party to develop a data commons for them.

2. A Commons Services Operations Center, or CSOC that provides the services for setting up and configuring the commons, operating commons, including monitoring the security and compliance, and for customizing the commons as required. The idea of CSOC is relatively new and was developed by the University of Chicago that runs a CSOC that manages six data commons currently.

3. A not-for-profit neutral party to manage the Data Commons Framework, including managing the agreements required for getting the data, setting up the commons, managing the data contributor agreements and the data use agreements, the IP required to keep the data commons software stack open source, etc.

4. A Cloud Services Provider (CSP) that provides the underlying cloud computing infrastructure required by the commons. Clouds can be run using on-premise academic clouds, public clouds such as provided by Amazon's AWS, Google's GCP, or Microsoft's Azure, or a hybrid model of both.

5. Data contributors who provide data to the commons and sign a Data Contributors Agreement (DCA).

6. Users of the data commons that a sign a Data Use Agreement (DUA) and Commons Services Agreement (CSA).

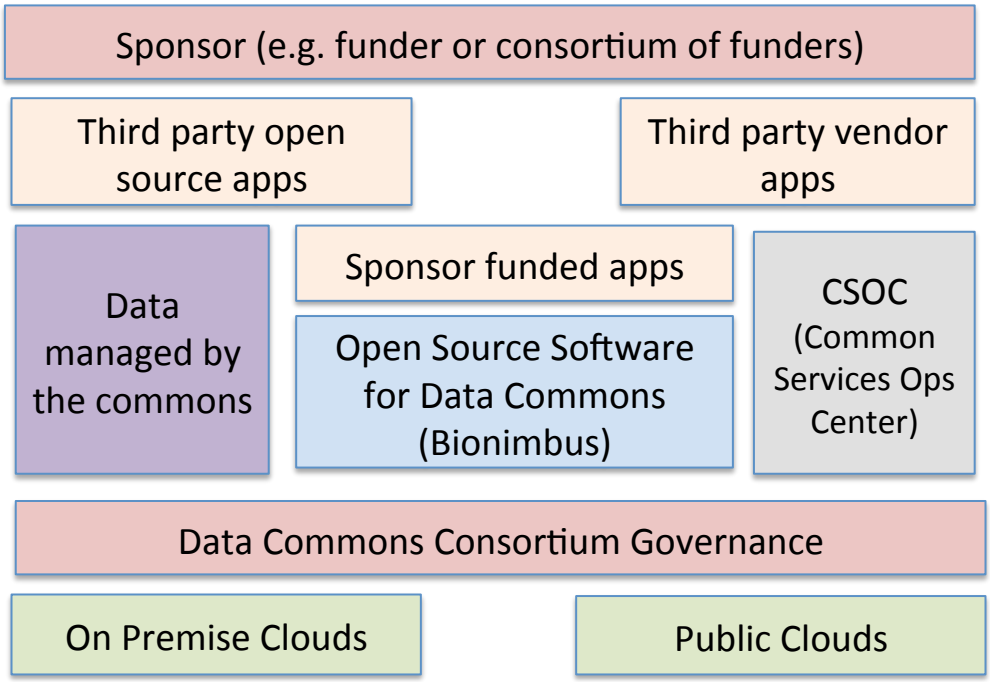

Figure 1: The figure shows a model for a data commons developed by the Open Commons Consortium.

6 Conclusion

We conclude with our four main points: First, data sharing and open science will accelerate research and improve patient outcomes. Second, data commons are a proven technology to support open data, open science, and data sharing. Third, foundations have a critical role to play and can disruptively change the open data and open science landscape. Fourth, the Open Commons Consortium (OCC) can set up and manage data commons for your researchers and the University of Chicago Commons Services Operations Center (CSOC) can operate them.

A Requirements for a Data Commons

In this section, we follow [2] and list seven core requirements for data commons. Note that the first three requirements are sufficient to support the FAIR principles for sharing data [6].

Requirement 1. Permanent Digital IDs. The first requirement is that the data commons must have a digital ID service and that datasets in the data commons must have permanent, persistent digital IDs. Associated with digital IDs are i) access controls specifying who can access the data and ii) metadata specifying additional in-formation about the data (see Requirement 2). Part of this requirement is that data can be accessed from the data commons through an API by specifying its digital ID.

Requirement 2. Permanent Metadata. The second requirement is that there is a metadata service that for each digital ID returns the associated metadata. Since the metadata can be indexed, this provides a basic mechanism for the data to be discoverable.

Requirement 3. API-based Access. The third requirement is that data can be accessed by an application programming interface (API), not just by browsing through a portal. Part of this requirement is that there is a metadata service that can be queried that returns a list of digital IDs that can then be retrieved via the API. For those data commons that contain controlled access data, another component of the requirement is that there is an authentication and authorization service so that users can first be authenticated and the data commons can check whether they are authorized to have access to the data.

Requirement 4. Data Portability. The fourth main requirement is that data be portable in the sense that a dataset that has been contributed to a data commons can be transported to another data commons so that in the future if it is hosted by the second data commons if required. In general, if data access is through digital IDs (versus referencing the physical location of data), then software that references data should not have to be changed when data is re-hosted by another data commons.

Requirement 5. Data Peering. By data peering, we mean an agreement between two data commons service providers to:

a) Support mutually agreed to services for Authentication, Authorization and Access Controls and Services.

b) Supporting mutually agreed to index services, metadata services, and data transport services so that data is subject to FAIR principles.

c) Agreeing to transfer data at no cost so that a researcher at data commons 1 can access data stored at data commons 2. In other words, the two data commons agree to transport research data between them with no access charges, no egress charges, and no ingress charges.

Requirement 6. Elastic Computing Infrastructure. With the explosive growth of data and the importance of machine learning, which can require large scale computing resources, a data commons requires on demand, elastic computing, storage and networking resources [5]. Cloud based computing is one way of providing this elastic infrastructure. The cloud computing infrastructure may be an on-premise cloud, a public cloud, or a hybrid cloud that integrates both [5].

Requirement 7. Pay for Compute. The final requirement is to co-locate with the data commons computing infrastructure that is available to researchers. Since, in practice, researchers' demand for computing resources is larger than the available computing resources, computing resources must be rationed, either through allocations or by charging for the use of computing resources. Notice that there is an asymmetry in how a data commons treats the storage and computing infrastructure. When data are accepted into a data commons, there is a commitment to store the data and make it available for a certain period of time, often indefinitely. In other words, the rationing decision for initial storage is made when data are accepted. In contrast, computing over data in a data commons is rationed in an on-going fashion, as is the working storage, and the storage required for derived data products, either by providing computing and storage allocations for this purpose or by charging for computing and storage. For simplicity, we refer to this requirement as "pay for computing," even though the model is more complicated than that.

References

[1] RL Grossman, B Abel, S Angiuoli, JC Barrett, D Bassett, K Bramlett, GM Blumenthal, A Carlsson, R Cortese, J DiGiovanna, et al. Collaborating to compete: Blood profiling atlas in cancer (BloodPAC) consortium. *Clinical Pharmacology & Therapeutics*, 2017.

[2] Robert L Grossman, Allison Heath, Mark Murphy, Maria Patterson, and Walt Wells. A case for data commons: Toward data science as a service. *Computing in Science & Engineering*, 18(5):10–20, 2016.

[3] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[4] Allison P Heath, Matthew Greenway, Raymond Powell, Jonathan Spring, Rafael Suarez, David Hanley, Chai Bandlamudi, Megan E McNerney, Kevin P White, and Robert L Grossman. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. *Journal of the American Medical Informatics Association*, 21(6):969–975, 2014.

[5] Peter Mell and Tim Grance. The NIST definition of cloud computing, special publication 800-145. http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, September, 2011.

[6] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.