PO BOX 400249
31 Bonnycastle Drive
Charlottesville VA 22904

January 4, 2019

University Librarian and Dean of Libraries
University of Virginia Library
P.O. Box - 400109
160 McCormick Road
Charlottesville, VA 22904

Dear Dean Unsworth:

I am writing to you to provide an update on the status of collaboration between the Data Science Institute and the Library with regards to the Open Data Lab. A formal report for 2018 is forthcoming and will be delivered upon completion (expected Q1 2019). There are three main areas of overlap outlined below.

1. **Data Working Group**: This group is led by Esther Onega and uses computational resources allocated from the Open Data Lab. The first project developed an algorithm to optimize volume location (Clemons or Ivy) on circulation costs while Alderman Renewal is in progress. This algorithm was developed with input from the liaisons. The working group Library team was Esther Onega, Tim Morton, and Anthony Lindsay. Pete Alonzi represented the Data Science Institute. The materials pertaining to this project are stored on GitHub at: "https://github.com/alonzi/alderman".
2. **Research Data Services Workshop Program**: This program includes workshops produced by the Open Data Lab group in the Data Science institute. In 2018 a workshop was given about the scale data solution Spark. The workshop was powered by the Open Data Lab computational resources to provide an interactive experience for the attendees. The materials are stored on GitHub at: "https://github.com/alonzi/spark-intro". This workshop was produced by Pete Alonzi and the series is managed by Michele Claibourn. There are three workshops scheduled for the Spring 2019 semester focusing on Python and Machine Learning.
3. **Dataverse and Libra** The Open Data Lab is exploring Dataverse as a discovery tool. The model is to put the metadata for data resources into a Dataverse entry including a pointer to the data in the Open Data Lab cloud resources (AWS S3 buckets,. GitHub repos, etc.). The pilot dataset is 13 TB and is best stored centrally so that researchers may execute on that data with a colocated computational resource. Additionally downloading 13 TB is prohibitive and would stifle research. This model fits with Libra and ongoing work with Sherry Lake has been very helpful in understanding how to proceed. In particular given the limited storage nature of Libra the metadata pointer paradigm with remote storage is ideal.

Thank you for your time and consideration. Please pass along any follow up questions you may have to me via datascientist@virginia.edu.

Sincerely,

Loreto Peter Alonzi