2022 Data Ark Survey Comments and Response
Scientific Computing and Data & GGS
March 9, 2022

The Data Ark survey —distributed in January 2022—solicited feedback from 662 active Minerva users. Of these, 20 users responded (3% response rate)

Survey Questions:

1. **Have you used Data Ark: A Mount Sinai Data Commons?**
   Yes: 16%
   No: 84%

2. **[Of those who answered "No" to Question 1,] Do you plan to use Data Ark?**
   Yes: 32%
   No: 5%
   Maybe: 42%
   No reply: 21%

3. **What data sets would you like to see added?**

   - Epidemiology sets
   - TCGA
   - COVID-19 biobank
   - Mount Sinai Brain Bank
   - RNAseq and eQTL datasets
   - Human Genomes Diversity Project
   - MassIVE (UCSD)
   - Gabriella Miller Kids First Pediatric Research Program
   - WGS
   - PsychENCODE
   - ROSMAP
   - UK Biobank
   - Sleep Studies
   - CommonMind
   - gnomAD
   - Molecular dynamics trajectories;DataBases of Molecular Dynamics (MD) Simulations

   - 1000 Genomes Project and UK Biobank are already included within Data Ark
   - We are adding the heavily-used portions of TCGA to Data Ark. All the open-access RNA-seq counts and WXS maf (Mutation Annotation Format) files will be added, as well as the cBioPortal database associated with the TCGA data set.
   - We are reviewing the rest datasets listed above and will discuss with the Data Ark committee

4. **What features would you like to see added to the Data Ark?**
   - Effective search and retrieval of data
     - We agree that this is a good suggestion. This feature is currently outside of the scope of Data Ark, but it is an excellent suggestion for future development.

- Summary stats and resources, quality metrics, cross-sectional vs longitudinal information, limitations of the datasets
  - Excellent suggestion. We will provide a detailed description of the data set, basic stats, experiment type if applicable, and a summary of the data set. Additionally, once more people use the data, we can gather additional information/feedback and add it to the website.

- An Instance of a HPC or AWS that allows to analyze the data
  - We have the resources to run your pipeline with access to those data under the Data Ark on our HPC. Let us know more details if this is not sufficient and what you need at data-ark-team@lists.mssm.edu

- A codebook to understand what variables are stored
  - Data Ark has module files that can be loaded to see all the environmental variables that have been defined for each dataset. Additionally, we will add the README file into each data set to provide information and instructions on navigating the data set.

- Neuroimaging data
  - We are actively working on getting IRW (imagine research warehouse) on Minerva and will plan and announce on sharing. If you have specific neuroimaging data that you think would benefit the Mount Sinai research community and would like to share it through Data Ark, please contact us at data-ark-team@lists.mssm.edu

- Streamlined pipelines for processing most commonly used data types (WGS, RNAseq, etc.)
  - We are exploring the resources to address the computational needs and will discuss with the Data Ark committee on a centralized data analysis pipeline for WGS, RNAseq, etc. At the same time, we will also be looking into community curated pipeline resources, such as nf-co.re/, and provide more support for users who would like to use those pipelines.

- Better folder organization, with dataset release/versions/formats clearly separated.
  - Currently, we only have one version. For future releases, we will separate the new versions. The Data Ark group will discuss the organization of the folders inside the data set to seek a clear and better presentation.

5. **What barriers exist preventing your usage of Data Ark?**
   - Did not know how to access and use resources, Instruction of how to use it, Lack of knowledge. It sounds like a great resource that needs better advertisement in the Sinai community.

        - Thanks for the suggestion. There will be more promotions/marketing surrounding Data Ark. Additionally, we will hold workshops, training and seminars to get users informed.

    - Unclear about permissions, IRB rules, prices or fees associated
        - Right now, there are no fees associated with the Data Ark Data Commons.

    - Not clear that it is useful/faster than just using versions of these datasets I already have downloaded and have been using for years
        - One of the purposes of Data Ark is to reduce the repetitive downloads of data sets and promote cooperation among researchers using the same data sets. We would love to hear your experiences utilizing the data sets. Please email us at data-ark-team@lists.mssm.edu

    - Interactions with Data Ark team
        - We have the email list where any questions can be asked, and any issues can potentially be resolved. This email address is data-ark-team@lists.mssm.edu. Also we have the Data Ark slack channel that you can join at https://join.slack.com/t/data-ark/signup

All comments and responses from users:

- We produce extensive MD simulations of proteins, membrane proteins, nucleic acid, and more. We could contribute these data to the Data Ark.
    - Please email us at data-ark-team@lists.mssm.edu with more details or complete survey form here. We will set up meeting with you accordingly.

- Would be great if some more descriptive EHR information could be uniformly processed (like medication information, physician notes)
    - We already have included some of the EHR data under Data Ark on Minerva, such as Mount Sinai Data Warehouse COVID-19 Electronic Health Record (EHR) Data Set. For the other EHR data hosted in MSDW/OMOP database, you can access the OMOPdatabase directly on Minerva login nodes (li03c02/3/4) with sqlcmd as below (you need to request access to MSDW database via sailpoint if you haven't)

      ```
      $kinit <your ID>@MSSMCAMPUS.MSSM.EDU
      $/opt/mssql-tools/bin/sqlcmd -S
      msdw2-mssql-prd.msnyuhealth.org -d omop
      ```

      Please email us at data-ark-team@lists.mssm.edu for more details

- There are data analysts in varying positions across departments and research groups, it would be an excellent think tank to review what the most common data problems are and universal MSSM data sets that would be helpful
    - This is precisely the mission that Data Ark is trying to tackle. In the beginning, we worked on the problem of users' difficulties in accessing data. After this is completed, we will better tackle the other standard data problems while users utilize those data sets. We are also getting guidance from the Data Ark

committee on those common data sets and problems.

- Instructions to navigate types of data and how to use the data to answer certain research questions with examples would be really helpful
  - We have a description of each data set on the Data Ark website. The user will have to go to our website to get the details of each particular data set and then request access to that data set. To answer certain research questions with examples and assist in the research process is presently not a feature of the Data Ark.

- A webapp providing a searchable metadata index of the various datasets through which to access information about each dataset and dataset releases/versions/formats would be highly desirable (similar comments to the previous one)
  - Good suggestion. This feature is currently outside of the scope of Data Ark, but it is an excellent suggestion for future development.