NASA's GeneLab Phase II: Federated Search and Data Discovery

Daniel C. Berrios, MD MPH PhD[1,2], Sylvain V. Costes, PhD[2], Peter B. Tran, PhD[2]
[1]Universities Space Research Association; [2]NASA Ames Res. Ctr., Moffett Field, CA

Purpose

GeneLab is currently being developed by NASA to accelerate "open science" biomedical research in support of the human exploration of space and the improvement of life on earth.[1,2] Phase I of the four-phase GeneLab Data Systems (GLDS) project emphasized capabilities for submission, curation, search, and retrieval of genomics, transcriptomics and proteomics ("omics") data from biomedical research of space environments. The focus of development of the GLDS for Phase II has been federated data search for and retrieval of these kinds of data across other open-access systems, so that users are able to conduct biological meta-investigations using data from a variety of sources. Such meta-investigations are key to corroborating findings from many kinds of assays and translating them into systems biology knowledge and, eventually, therapeutics.

System Design

The GLDS currently serves over 100 omics investigations to the biomedical community via open access. In order to expand the scope of metadata record searches via the GLDS, we designed a metadata warehouse that collects and updates metadata records from external systems housing similar data. To demonstrate the capabilities of federated search and retrieval of these data, we imported metadata records from three open-access data systems into the GLDS metadata warehouse: NCBI's Gene Expression Omnibus (GEO), EBI's PRoteomics IDEntifications (PRIDE) repository, and the Metagenomics Analysis server (MG-RAST). Each of these systems defines metadata for omics data sets differently. One solution to bridge such differences is to employ a common object model (COM) to which each systems' representation of metadata can be mapped. Warehoused metadata records are then transformed at ETL to this single, common representation. Queries generated via the GLDS are then executed against the warehouse, and matching records are shown in the COM representation (Fig. 1). While this approach is relatively straightforward to implement, the volume of the data in the omics domain presents challenges in dealing with latency and currency of records. Furthermore, the lack of a coordinated, universal registry of these kinds of data creates the issue of data duplication in federated search systems.

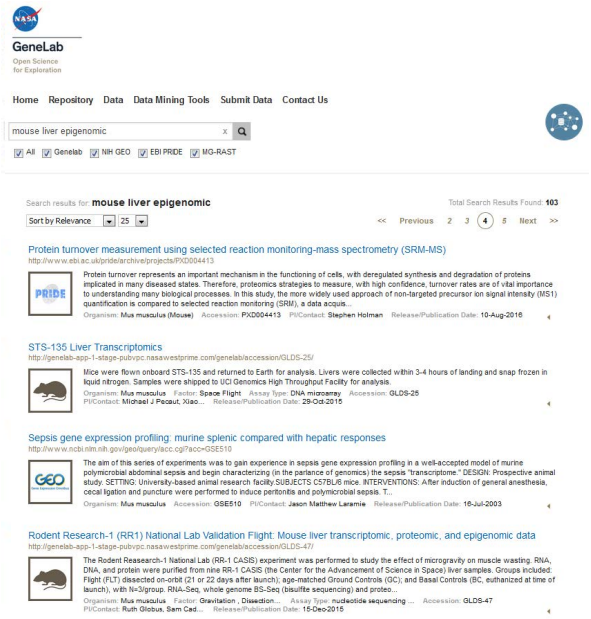

Figure 1. Federated Search of Omics Data

Results

Prototype federated data search capabilities are currently accessible to internal (NASA) users, with open access to these capabilities anticipated in the GLDS no later than Sep. 2017. We will demonstrate the flexibility, performance, and power of our metadata representation mapping approach. The execution of these kinds of studies will be furthered in Phases III and IV through the development of collaborative omics meta-analysis workspaces.

References

1. Berrios DC, Thompson TG, Fogle HW, Rask JC, Coughlan JC. GeneLab: NASA's Open Access, Collaborative Platform for Systems Biology and Space Medicine. AMIA Annual Symposium Proceedings 2015.
2. Berrios DC, Welch, JD, Fogle HW, Skidmore, M, Marcu O. GeneLab: NASA's GeneLab: Phase I Results and Plans. AMIA Annual Symposium Proceedings 2016.