FAIRness and Usability for Open-access Omics Data Systems

Daniel C. Berrios, MD MPH PhD[1,4], Afshin Beheshti, PhD[2,4], Sylvain V. Costes, PhD[3,4]
[1]Universities Space Research Association; [2]Wyle Labs; [3]NASA
[4]Ames Research Center, Moffett Field, CA

Abstract

Omics data sharing is especially crucial to the biological research community, and the last decade or two has seen a huge rise in collaborative analysis systems, databases, and knowledge bases for omics and other systems biology data. We assessed the "FAIRness" of NASA's GeneLab Data Systems (GLDS) along with four similar kinds of systems in the research omics data domain, using 14 FAIRness metrics. 14 metrics. The range of Pass ratings was 29-79% of the 14 metrics, Partial Pass 0-21%, and Fail 7-50%. The range of overall FAIRness scores was 5-12 (out of 14). The systems we evaluated performed the best in the areas of data findability and accessibility, and worst in the area of data interoperability. We propose two new principles that Big Data systems, in particular, should consider for increasing data accessibility. We relate our experiences implementing semantic integration of omics data from several systems for the federated querying and retrieval functions of the GLDS, given the shortcomings in data interoperability of these systems.

Introduction

Over the last decade the vast increase in outputs from "omics" (genomics, transcriptomics, proteomics, and metabolomics) assays, together with decreasing assay and data storage costs, have resulted in an exponentially increasing volume of these "big" data. The high feature-number aspect of these big data has highlighted the importance of data sharing (reuse) functions of these systems, as data sets often have relevance for investigating a very large number of hypotheses. Thus, developers of big data systems increasingly are focusing on the needs of users to discover, annotate, share, and analyze data. Omics data sharing is especially crucial to the biological research community, and the last decade or two has seen a huge rise in collaborative analysis systems[1-3], databases[4-10], and knowledge bases[11-13] for omics and other systems biology data.

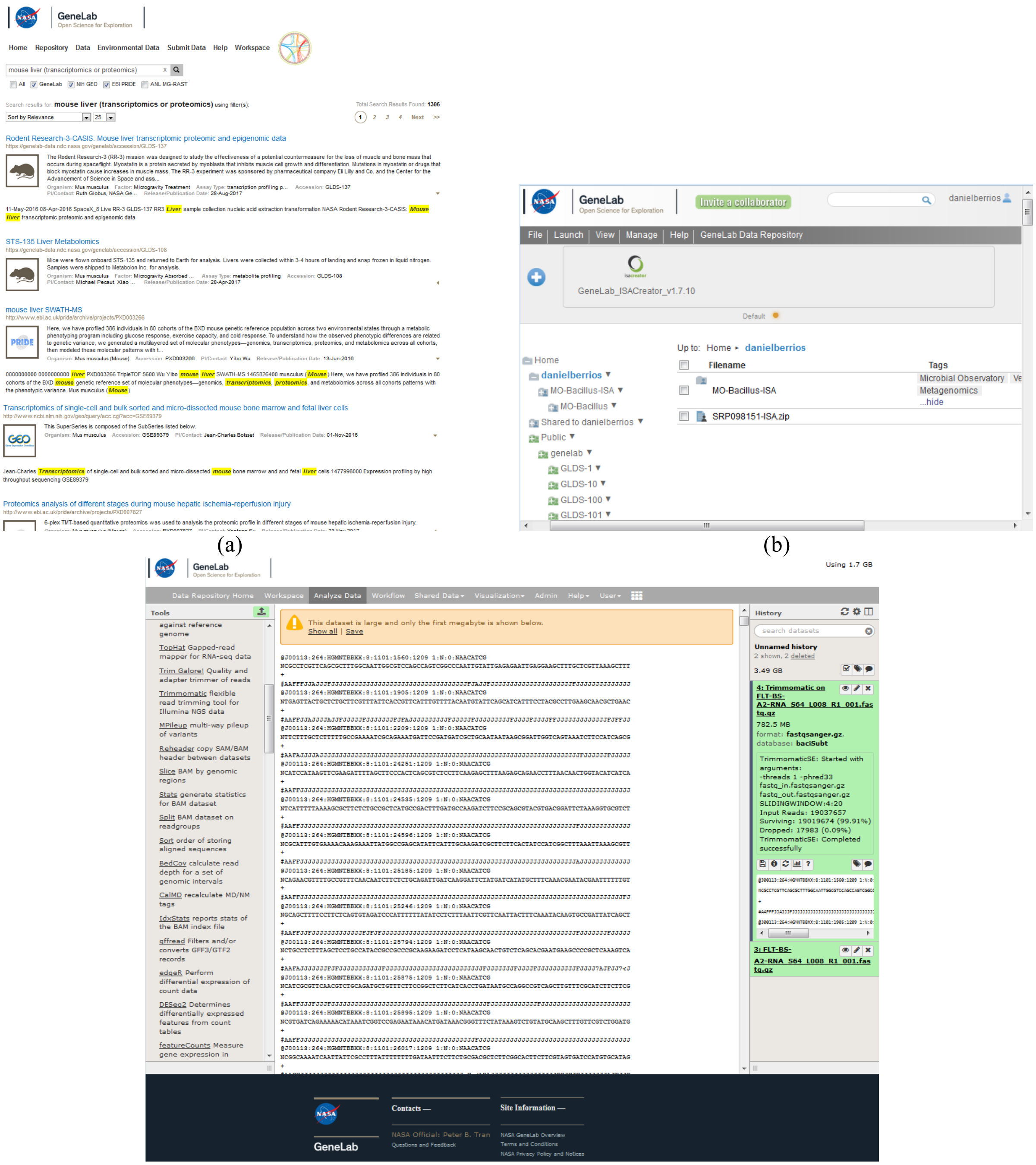
The FAIR[14] principles are a set of guiding principles that are being advocated as being foundational in the broad areas of data findability, accessibility, interoperability, and reusability. However, the challenges involved in using current technologies and real-world infrastructures within organizational guidelines to develop data repositories and/or data analytical systems that are "FAIR" are substantial. The FAIR principles deal with aspects of data handling such as identification, representation, and attribution. Metrics to evaluate the "FAIRness" of a data system or its design are currently being developed[15]. System developers should consider familiarizing themselves with these metrics, as many of the FAIR principles have significant implications for big-data management and analysis systems in terms of feasible design options and operations resource planning. Below, we assess the "FAIRness" of NASA's GeneLab Data Systems (GLDS) along with four similar kinds of systems in the research omics data domain, using 14 FAIRness metrics. We discuss FAIRness concepts for the design of Big Data omics systems in particular, relating some of the challenges we faced developing our system.

In addition to designing a data system to be FAIR, developers of Big Data systems, in particular, should consider how system design can maximize data usability for its users. We propose a few new design principles for FAIR systems that handle Big Data like omics data from our experiences with the GLDS. By complying with these usability principles and the FAIR principles together, systems that manage Big Data can maximize both the usability and reusability of their data.

Methods

NASA GeneLab

GeneLab (http://genelab.nasa.gov) is a NASA initiative designed to accelerate "open science" biomedical research in support of the human exploration of space and the improvement of life on earth.[16] GeneLab data have been used to research impacts of space on mice, humans, and plants. Phase I of GeneLab Data Systems (GLDS) project implemented an "omics" (genomics, transcriptomics proteomics, and metabolomics) data repository for biomedical research data conducted in or relevant to space environments (Figure 1). This initial phase developed processes and systems for data submission, curation, indexing, search, and retrieval. In Phase II GeneLab implemented federated data search and retrieval capabilities from GeneLab and three other open-access data systems, in order to facilitate data integration and biological meta-investigation (Figure 1a).[17] Such meta-investigations are key to corroborating



mouse liver (transcriptomics or proteomics) x

□ All ☑ GeneLab ☑ NIH GEO ☑ EBI PRIDE □ ANL MG-RAST

Search results for: **mouse liver (transcriptomics or proteomics)** using filter(s):

Total Search Results Found: **1306**

Rodent Research-3-CASIS: Mouse liver transcriptomic proteomic and epigenomic data
https://genelab-data.ndc.nasa.gov/genelab/accession/GLDS-137

The Rodent Research-3 (RR-3) mission was designed to study the effectiveness of a potential countermeasure for the loss of muscle and bone mass that occurs during spaceflight. Myostatin is a protein secreted by myoblasts that inhibits muscle cell growth and differentiation. Mutations in myostatin or drugs that block myostatin cause increases in muscle mass. The RR-3 experiment was sponsored by pharmaceutical company Eli Lilly and Co. and the Center for the Advancement of Science in Space and ass...
Organism: Mus musculus Factor: Microgravity Treatment Assay Type: transcription profiling p... Accession: GLDS-137
PI/Contact: Ruth Globus, NASA Ge... Release/Publication Date: 28-Aug-2017

11-May-2016 08-Apr-2016 SpaceX_8 Live RR-3 GLDS-137 RR3 ***Liver*** sample collection nucleic acid extraction transformation NASA Rodent Research-3-CASIS: ***Mouse liver*** transcriptomic proteomic and epigenomic data

STS-135 Liver Metabolomics
https://genelab-data.ndc.nasa.gov/genelab/accession/GLDS-108

Mice were flown onboard STS-135 and returned to Earth for analysis. Livers were collected within 3-4 hours of landing and snap frozen in liquid nitrogen. Samples were shipped to Metabolon Inc. for analysis.
Organism: Mus musculus Factor: Microgravity Absorbed ... Assay Type: metabolite profiling Accession: GLDS-108
PI/Contact: Michael Pecaut, Xiao ... Release/Publication Date: 28-Apr-2017

mouse liver SWATH-MS
http://www.ebi.ac.uk/pride/archive/projects/PXD003266

Here, we have profiled 386 individuals in 80 cohorts of the BXD mouse genetic reference population across two environmental states through a metabolic phenotyping program including glucose response, exercise capacity, and cold response. To understand how the observed phenotypic differences are related to genetic variance, we generated a multilayered set of molecular phenotypes—genomics, transcriptomics, proteomics, and metabolomics across all cohorts, then modeled these molecular patterns with t...
Organism: Mus musculus (Mouse) Accession: PXD003266 PI/Contact: Yibo Wu Release/Publication Date: 13-Jun-2016

0000000000 0000000000 ***liver*** PXD003266 TripleTOF 5600 Wu Yibo ***mouse liver*** SWATH-MS 1465826400 musculus (***Mouse***) Here, we have profiled 386 individuals in 80 cohorts of the BXD ***mouse*** genetic reference set of molecular phenotypes—genomics, ***transcriptomics***, ***proteomics***, and metabolomics across all cohorts patterns with the phenotypic variance. Mus musculus (***Mouse***)

Transcriptomics of single-cell and bulk sorted and micro-dissected mouse bone marrow and fetal liver cells
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89379

This SuperSeries is composed of the SubSeries listed below.
Organism: Mus musculus Accession: GSE89379 PI/Contact: Jean-Charles Boisset Release/Publication Date: 01-Nov-2016

Jean-Charles ***Transcriptomics*** of single-cell and bulk sorted and micro-dissected ***mouse*** bone marrow and and fetal ***liver*** cells 1477998000 Expression profiling by high throughput sequencing GSE89379

Proteomics analysis of different stages during mouse hepatic ischemia-reperfusion injury
http://www.ebi.ac.uk/pride/archive/projects/PXD007827

6-plex TMT-based quantitative proteomics was used to analysis the proteomic profile in different stages of mouse hepatic ischemia-reperfusion injury.

(a)

GeneLab_ISACreator_v1.7.10

Up to: Home ▸ danielberrios

□	Filename	Tags
□	MO-Bacillus-ISA	Microbial Observatory Ve Metagenomics ...hide
□	SRP098151-ISA.zip	

(b)

against reference genome

TopHat Gapped-read mapper for RNA-seq data

Trim Galore! Quality and adapter trimmer of reads

Trimmomatic flexible read trimming tool for Illumina NGS data

MPileup multi-way pileup of variants

Reheader copy SAM/BAM header between datasets

Slice BAM by genomic regions

Stats generate statistics for BAM dataset

Split BAM dataset on readgroups

Sort order of storing aligned sequences

BedCov calculate read depth for a set of genomic intervals

CalMD recalculate MD/NM tags

IdxStats reports stats of the BAM index file

gffread Filters and/or converts GFF3/GTF2 records

edgeR Perform differential expression of count data

DESeq2 Determines differentially expressed features from count tables

featureCounts Measure gene expression in

This dataset is large and only the first megabyte is shown below.
Show all | Save

```
@J00113:264:HGMNTBBXX:8:1101:1560:1209 1:N:0:NAACATCG
NCGCCTCGTTCAGCGCTTTGGCAATTGGCGTCCAGCCAGTCGGCCCAATTGTATTGAGAGAATTGAGGAAGCTTTGCTCGTTAAAGCTTT
+
#AAFFFJJAJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJAJJFJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:1905:1209 1:N:0:NAACATCG
NTGAGTTACTGCTCTGCTTCGTTTATTCACCGTTCATTTGTTTTACAATGTATTCAGCATCATTTCCTACGCCTTGAAGCAACGCTGAAC
+
#AAFFJJAJJJJAJJFJJJJJFJJJJJJJFJFAJJJJJJJJJJJFJJJJJFJJJJJJJJJFJJJJFJJJJFFJJJJJJJJJJJJJJFJFFJJ
@J00113:264:HGMNTBBXX:8:1101:2209:1209 1:N:0:NAACATCG
NTTCTTTGCTCTTTTTGCCGAAAATCGCAGAAATGATTCCGATGATCGCTGCAATAATAAGCGGATTGGTCAGTAAATCTTCCATCAGCG
+
#AAFAJJJJAJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJFJJJJJ
@J00113:264:HGMNTBBXX:8:1101:24251:1209 1:N:0:NAACATCG
NCATCCATAAGTTCGAAGATTTTAGCTTCCCACTCAGCGTCTCCTTCAAGAGCTTTAAGAGCAGAACCTTTAACAACTGGTACATCATCA
+
#AAFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:24535:1209 1:N:0:NAACATCG
NTCATTTTTAAAAGCGCTTCTCTGCCGCTCATGCCGACTTTGATGCCAAGATCTTCCGCAGCGTACGTGACGGATTCTAAAGGTGCGTCT
+
#AAFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:24596:1209 1:N:0:NAACATCG
NCGCATTTGTGAAAACAAAGAAATTATGGCCGAGCATATTCATTTGCAAGATCGCTTCTTCACTATCCATCGGCTTTAAATTAAAGCGTT
+
#AAFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJAJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:25185:1209 1:N:0:NAACATCG
NCAGAACGTTTTGCCGTTTCAACAATCTTCTCTGCAGATTGATCAAGGATTCTATGATCATATGCTTTCAAACGAATACGAATTTTTTGT
+
#AAFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJ
@J00113:264:HGMNTBBXX:8:1101:25246:1209 1:N:0:NAACATCG
NGCAGCTTTTCCTTCTCAGTGTAGATCCCATTTTTTATATCCTCTTTAATTCGTTCAATTACTTTCAAATACAAGTGCCGATTATCAGCT
+
#AAFFJJFJJJJJJJJJJJJJJFJFJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJFJJJJJ
@J00113:264:HGMNTBBXX:8:1101:25794:1209 1:N:0:NAACATCG
NCTGCCTCTTTAGCTCCTGCCATACCGCCGCCCGCAAGAAGATCCTCATAAGCAACTGTCTCAGCACGAATGAAGCCCCGCTCAAAGTCA
+
#AAFAJJJJJJJFJFJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJFJJJJFJJJJJJJJJJFJJJJ7AJFJJ7<J
@J00113:264:HGMNTBBXX:8:1101:25875:1209 1:N:0:NAACATCG
NCATCGCGTTCAACGTCTGCAGATGCTGTTTCTTCCGGCTCTTCATCACCTGATAATGCCAGGCCGTCAGCTTGTTTCGCATCTTCTTCG
+
#AAFFJJJFJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:25895:1209 1:N:0:NAACATCG
NCGTGATCAGAAAAACATAAATCGGTCCGAGAATAAACATGATAAACGGGTTTCTATAAAGTCTGTATGCAAGCTTTGTTCGTCTGGATG
+
#AAFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:26017:1209 1:N:0:NAACATCG
NCGGCAAAATCAATTATTCGCCTTTATTTTTTTTGATAATTTCTTCTGCGACGCTCTTCGGCACTTCTTCGTAGTGATCCATGTGCATAG
+
```

Unnamed history
2 shown, 2 deleted

3.49 GB

4: Trimmomatic on FLT-BS-A2-RNA_S64_L008_R1_001.fastq.gz

782.5 MB
format: **fastqsanger.gz**, database: **baciSubt**

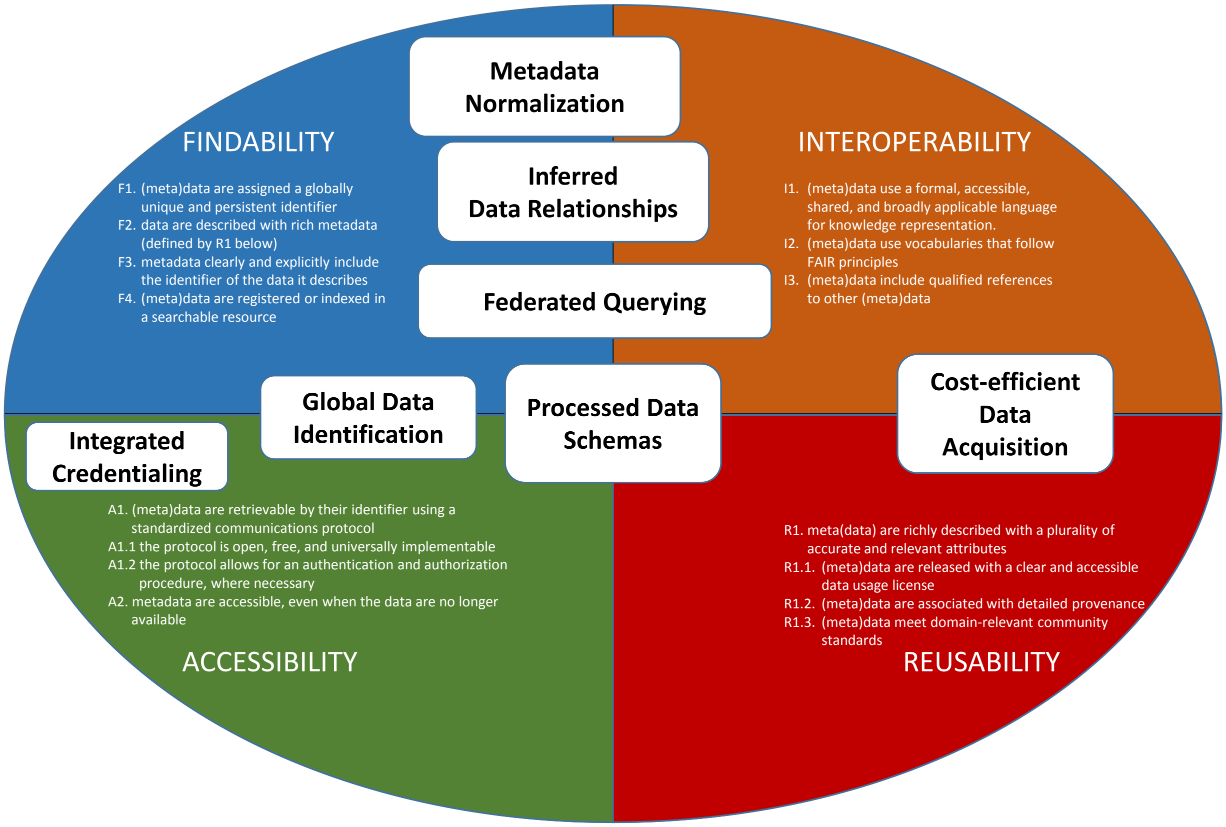
```
TrimmomaticSE: Started with
arguments:
-threads 1 -phred33
fastq_in.fastqsanger.gz
fastq_out.fastqsanger.gz
SLIDINGWINDOW:4:20
Input Reads: 19037657
Surviving: 19019674 (99.91%)
Dropped: 17983 (0.09%)
TrimmomaticSE: Completed
successfully
```

```
@J00113:264:HGMNTBBXX:8:1101:1560:1209 1:N:0:
NCGCCTCGTTCAGCGCTTTGGCAATTGGCGTCCAGCCAGTCGGCC
+
#AAFFFJJAJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@J00113:264:HGMNTBBXX:8:1101:1905:1209 1:N:0:
```

3: FLT-BS-A2-RNA_S64_L008_R1_001.fastq.gz

(c)

Figure 1. The GeneLab Data Systems. (a) The GeneLab Data Repository. The search interface shown includes not only search of GLDS repository data, but of the extramural data sources Gene Expression Omnibus (NIH GEO), the European Bioinformatics Institute's PRoteomics IDEntifications repository (EBI PRIDE), and the Argonne National Laboratory's Metagenomics Analysis server (ANL MG-RAST). (b) The GLDS User Workspace. (c) One of the tools for omics data analysis hosted within the GLDS (an instance of a Galaxy server[1]).



F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Figure 2**.** Impacts of FAIR concepts on the design of the GeneLab Data Systems.

findings from different studies and/or kinds of bioassay data, and translating them into systems biology knowledge and, eventually, therapeutics[18]. In Phase III, GeneLab has begun to deploy an omics data analysis platform as part of the GLDS. Using this platform, investigators can design, execute and share in silico experimentation of omics data sets. At the end of Phase III, users with results of these kinds of experiments will be able to store them directly into the GLDS repository, linked to originating data and to the analysis workflows that generated them (Figure 1b and Figure 1c).

Even though the inception of the GLDS predated publication of the FAIR principles, many of the concepts behind the principles were discussed extensively prior to its design (**Figure 2**). These discussions included substantial input from the National Academies of Sciences and from a steering committee of outside experts in omics biological data, and had significant impacts on required resources for development and operations, on selection of technological approaches and system architectures, and on the policies developed for operating the GLDS. These concepts included metadata normalization, inferred data relationships, and federation of data querying and retrieval, all of which have impacts on data findability and interoperability. Issues of integrated credentialing and authentication, and global data identification impact findability and accessibility of data. Costs were a big factor in designs for data acquisition and curation that adequately support data interoperability and reusability. Finally, a lack of available schemas for metadata associated with processed data has impacts on our ability to design for compliance with each of the FAIR principles.

FAIRness of Omics Data

The FAIR principles were designed to guide the management of data so that it is maximally findable, accessible, interoperable, and reusable. They are not precise prescriptions for FAIR system design specifications; rather they are aspirational guidelines with some references to exemplar standards and technologies. Among the findability principles are requirements that data be assigned a persistent, global unique identifier (GUID), be described by rich metadata, and be indexed by a system for retrieving the data. Accessibility requires the use of standardized, open (available for implementation to all) communications protocols for retrieval of data by GUID. Interoperability is supported through

Table 1. The FAIR principles, corresponding draft FAIRness metrics, and semi-quantitative FAIRness ratings for select omics data systems.

● Pass ⭘ Partial pass ○ Fail **FAIR Principle**	**Metric**	**NIH GEO**	**EBI ENA**	**ANL MG-RAST**	**Metabolights**	**GLDS**
F1. (meta)data are assigned a globally unique and persistent identifier	FM-F1A	●	●	⭘	●	●
F1. (meta)data are assigned a globally unique and persistent identifier	FM-F1B	●	●	○	●	●
F2. data are described with rich metadata (defined by R1 below)	FM-F2	●	●	●	●	●
F3. metadata clearly and explicitly include the identifier of the data it describes	FM-F3	●	●		⭘	●
F4. (meta)data are registered or indexed in a searchable resource	FM-F4	●	●	○	○	●
A1. (meta)data are retrievable by their identifier using a standardized communications protocol	N/A	-	-	-	-	-
A1.1 the protocol is open, free, and universally implementable	FM-A1.1	●	●	●	●	●
A1.2 the protocol allows for an authentication and authorization procedure, where necessary	FM-A1.2	●	●	●	●	●
A2. metadata are accessible, even when the data are no longer available	FM-A2	●	●	○	○	●
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	FM-I1	○	○	○	⭘	⭘
I2. (meta)data use vocabularies that follow FAIR principles	FM-I2	○	○	○	⭘	⭘
I3. (meta)data include qualified references to other (meta)data	FM-I3	○	○	○	○	○
R1. meta(data) are richly described with a plurality of accurate and relevant attributes	N/A	-	-	-	-	-
R1.1. (meta)data are released with a clear and accessible data usage license	FM-R1.1	●	●	●	●	●
R1.2. (meta)data are associated with detailed provenance	FM-R1.2	●	●	⭘	●	●
R1.3. (meta)data meet domain-relevant community standards	FM-R1.3	●	●	○	●	●
Overall Score		11	11	5	9.5	12

the use of machine-actionable metadata and data. Reusability requires systems to be compliant in the management of their data objects with the first three principles, specifically so that data and metadata can be linked and integrated across FAIR systems.

We assessed compliance of five open-access, omics data systems and the GLDS for FAIRness using the current working draft of the FAIRness metrics developed by the GO FAIR Metrics group[19] (**Table 1**). We chose the systems for their similarity to the GLDS; all are open-access, government or government-funded research lab-developed omics data systems. As the FAIR principles themselves are only aspirational, the assessment of the systems can only be at most semi-quantitative, and they would likely be affected by further clarifications or details in the principles and/or metrics which are almost certain to come. We rated systems with Pass when we had no evidence of any failure of a test by the metric. Systems were rated as Partial Pass, if the metric had multiple steps or components, and the system passed some, but not all; or if we had evidence that some data inputs for the metric yielded a pass, while others did not. We chose Fail only if we could find no evidence the system was compliant with any part of the metric, for any inputs we tried. We developed an overall FAIRness Score by assigning a score of 1 for each Pass, 0.5 for each Partial Pass, and 0 for each Fail.

Findability

The first findability principle, F1 (**Table 1**), states that systems should identify data using globally unique and persistent identifiers. The objective is to make data findable, but its implementation also supports data accessibility

(FAIR principle A1). The qualifier “global” should be interpreted as global within the domain of conceivable use of the data. Thus, it is not sufficient to use locally (system-specific) controlled processes to manage data identifiers[20], as this approach can never absolutely guarantee against overlap with other systems' local identifier schemes. In addition, F1 specifies the persistence of issued identifiers; identification systems that employ Uniform Resource Locators (which rely on alterable domain name resolution) cannot guarantee persistence of identifiers.

All but one of the systems we assessed passed the metric for F1 metric by employing persistent, globally unique identifiers. (ANL MG-RAST only offers global identifiers as an option for data submission, and so was assessed as Partial Pass.) However, among the schemes for providing such identifiers, the most reliable over time are likely to be those that use third-party resolution (versus those that rely on in-house redirection). Third-party resolution identifiers include Digital Object Identifiers (DOIs) and those provided by identifers.org, among others[21]. However, in the biomedical research and clinical practice realms, publishers have now widely selected the DOI scheme for identification and citation of literature; more than 80% of 2015 articles indexed by PubMed have assigned DOIs[22]. As a consequence, the biomedical community is already (and continues to become more and more) familiar with using these identifiers to retrieve and cite literature. Yet to date only a few omics data repositories have similarly implemented DOIs for data publications. Consequently, actual findability and accessibility of data in these repositories is less than optimal, even though they are technically in compliance with F1, and will continue to be so, unless and until the biomedical community is trained on how to use other identifying schemes for citing data. Furthermore, metadata from these repositories are not currently indexed in DOI meta-indexes (such as the one produced by DataCite[23]). In order to be compliant with F1 and A1, the GLDS is implementing methods to issue DOIs for all data publications. Additionally, curated metadata records of the publications are transmitted to DataCite[23] for meta-indexing, increasing the chances for reuse of GLDS-hosted data.

FAIR data also need to be described “with rich metadata” in order to be findable. The draft metric to assess F2 compliance merely requires a metadata description document that specifies a format for machine-readable metadata. The definition of the term “machine-readable” varies somewhat, but it is generally agreed that it cannot be natural language. (Although F2 refers to reusability principles requiring data “have a plurality of accurate and relevant attributes” (R1), the metric for F2 does not attempt to assess compliance with principle R1 [perhaps it should]). All the systems we assessed passed this metric. Furthermore, all systems passed F3, requiring association of metadata with data it describes (although it should be noted that FM-F3 refers only to IRIs and not broadly to any other kinds of acceptable data identifiers; we assume that ultimately FM-F3 will be broadened to allow other kinds of F1-compliant data identifiers. Also, it is difficult to determine if any of the systems truly support IRIs, given that none of them currently appear to use non-ASCII character encodings in their data identifying schemes).

The FAIRness metric FM-F4 is rather ill-defined currently. It requires submission of an identifier of FAIR data and its metadata to search engines (which precisely, it does not yet specify, but some common search engines are mentioned as examples in the comments for the metric). A Pass should be given if the returned search results include links to the published data (again, it does not yet specify where in the returned results, although among both the top 10 and top 40 are mentioned in metric comments). Some of the systems we assessed failed this metric, and it seems their data records are likely not indexed by at least some search engines. For example, search of the ANL MG-RAST data record identified by “http://metagenomics.anl.gov/linkin.cgi?metagenome=mgm4447971.3” or any of its tabular metadata yielded no search results referring to the repository using Google. More of the systems would fail this metric if it required data to be submitted specifically to DOI-based search systems (such as DataCite).

Accessibility

All of the systems for which we assessed compliance with FAIR principles A1 and A1.1 passed; all use open source communication protocols to access data, which are retrieved by their identifiers. FAIR Principle A1.2 stipulates that any protocol used to retrieve data “allows for an authentication and authorization procedure, where necessary.” All the systems we assessed describe which functions of the system require credentials, and how to obtain the credentials. It should be noted that A1.2 can be applied to the design of all types of systems, from those with highly restricted access controls, to the kinds of “open access” systems we assessed, that typically require minimal authentication and grant the widest possible access to all users. Furthermore, the data in question could any kind of data, including raw instrument output, shared data analysis workflows or data products, or scientifically-relevant user comments and opinions. Even for open access systems, some systems functions such as user-workspace file management, user commenting and messaging, and user attribution of data always require authentication, because these functions must necessarily leverage user-specific information.

A specific challenge that GeneLab faced when implementing A1.2 for the GLDS was a diversity of credentialing services available to the system's target users, with no single service common to all users. Furthermore, the goal of GeneLab is to maximize access to GLDS data across all kinds of users, but within the bounds of authentication and credentialing policies of NASA – another challenge. Aside from policy restrictions, the development of any new credentialing and authentication functions or services can require considerable resources.

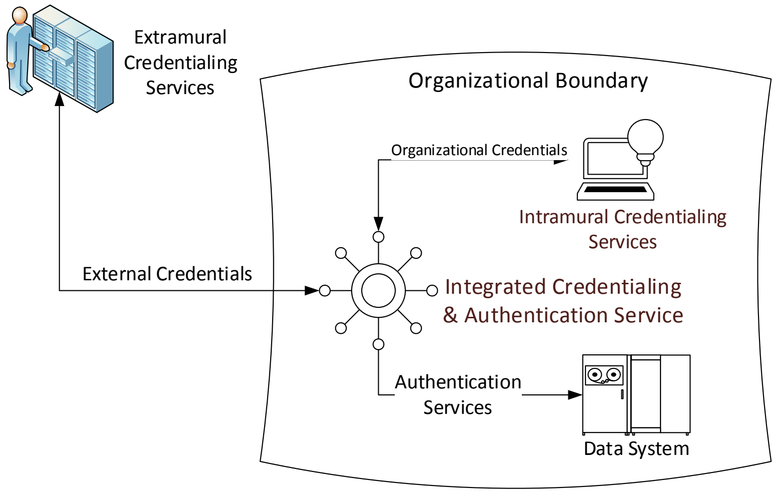

Extramural Credentialing Services
Organizational Boundary
Organizational Credentials
Intramural Credentialing Services
External Credentials
Integrated Credentialing & Authentication Service
Authentication Services
Data System

Figure 3. Use of an Integrated Credentialing and Authentication Architecture (ICAA) to comply with FAIR principle A.2. The ICAA provides services and policies that use information from intra- and extra-mural credentialing services to authenticate users of a data system.

A solution for organizations that are seeking to develop open-access systems for is to consider investments in developing an organization-wide integrated credentialing and authentication architecture (ICAA, Figure 3), and policies for allowing users to leverage extramural credentials, with appropriate access limitations. ICAAs and attendant policies can provide flexibility to develop and/or deploy data systems with available resources that can meet policy requirements of the organization that support extramural user access. It should be noted that using an ICAA to leverage credentialing systems developed independently from the system requires 1) technical compatibility between the authentication services of credentialing providers and the ICAA; and 2) compatibility of the credentialing policies of credentialing organizations and those of the organization developing the data system. Technical compatibility is facilitated through the use of authentication protocols and standards. Organizational credentialing policy incompatibility is more problematic to resolve. For example, it may be that the organization deploying the data system requires credentials be updated at an interval that is different from that required by one or more extramural credential provider. If these differences cannot be resolved, the only option may be to limit access for some extramural users data system functions requiring authentication.

NASA is developing an ICAA and adopting policies that permit certain intramural data systems to leverage it for user credentials and authentication. These policies stipulates that users within the organization must use the organization's own credentialing services, while extramural users may authenticate using one from a defined set of externally-managed credentialing systems (currently including authentication by selected "social media" systems). At NASA (and at most U.S. federal agencies), intramural credentialing is quite a long process involving corroboration of much identifying information. The use of the ICAA, users without NASA credentials can specify a minimum of information (e.g., only a verifiable email address), which they can provide easily through a web browser, and gain almost immediate access to the GLDS. Through use of the ICAA, the GLDS is able to recognize a lower level of confidence when authenticating extramural users, and implement appropriate access controls to some functions for these users (e.g., they may not join system administrator groups). It seems reasonable to suggest that an additional principle be considered when designing FAIR data systems that have at least some functions requiring authentication or authorization (A1.2.1 in **Table 2**). This principle, would stipulate that an integrated credentialing architecture, like that shown in Figure 3 be leveraged, in order to offer access to data to the widest number of users possible, when and where data policies permit such access.

Most of the systems we assessed for FAIR compliance to A2 have some policy to retain at least some metadata when data are removed from the system. It is frequently unclear, however, which metadata were to be retained when data

Table 2. Additional principles for optimizing data accessibility

A1.2.1	Authentication protocols should support multiple credential providers
A3	Data transport should be minimized

records are removed from the system. Data systems should ensure that policies for retaining metadata are clearly stated, including which metadata and for what period of time.

Another factor that specifically affects accessibility to Big Data is data transport. Data to be used as inputs in Big Data systems, but which are not located in the same environment as executable code of the system, commonly must first be transported (duplicated or moved) to the environment first. This environment typically is a user-specific workspace. The transport process for Big Data can require minutes, hours or even days in some Big Data systems, and represents a significant barrier to efficient access. Designers of Big Data would benefit from a principle that stipulates the collocation within Big Data systems of data and executable code (A3 in **Table 2**).

Interoperability

I1 stipulates the "use a formal, accessible, etc. language for knowledge representation. The draft metric for this principle requires that the language for data or metadata be available by URL, and have a BNF representation. The comments on the metric suggest additional required language qualities, including that the language must have semantics ("['vanilla'] xml and json ..should fail"). While current knowledge representation languages such as RDF and OWL[24] are available to generate metadata specifications with inherent semantics, tools that leverage such languages require some sophistication in their use. All the repositories we assessed have metadata specifications that do not have inherent semantics, and thus would not yield a Pass for this metric. Metabolights and the GLDS both use the ISATab metadata specification, which can be converted to RDF, and so were rated with Partial Pass.

The metric for principle I2 has a related requirement: that data systems use open, community-developed vocabularies for data and metadata, and that such vocabularies be more than simple keywords (i.e., include semantics in some form). As the ISATab specification supported ontological lookup from such vocabularies, we again only rated Metabolights and the GLDS with Partial Pass, and the rest of the systems as FAIL.

In addition to lack of experience with knowledge representation technologies like RDF and OWL, another likely reason for lack of compliance of the systems with I1 and I2 is a general lack of available semantic resources developed by the community to generate metadata specifications and vocabularies, particularly in more recent domains like genomics, proteomics and metabolomics. We have ourselves sought such resources for use in the GLDS development, where we are seeking to provide users with omics data analysis capabilities, an omics "workbench" consisting of user-specific workspaces and data access controls, together with access to a suite of omics analysis tools (see Figure 1b and Figure 1c). Users of these tools will create data processing (analysis) workflows to transform input data into "higher-order" data products, which they will likely want to share with others and/or associate with the workflows and data inputs. One of the challenges we have observed for making these kinds of processed information FAIR include a lack of semantic resources for creating metadata specifications for such processed data and the processing workflows that generate them. While schemas like the Common Workflow Language[25] (CWL) specification provide a method to represent and reproduce the mechanics of data processing, they lack domain-specific semantics regarding the nature of these data transformations (e.g., what are the goals of the transformation from a biological perspective? What are the biological data concepts produced in output files? how do these relate to input data elements? etc.). Without these descriptors, systems may not adequately index these processed data and workflows for discovery and retrieval. While sophisticated users (e.g., trained bioinformaticians) who do encounter processed data and workflows may be able to understand and reuse them, others without domain-specific experience and knowledge of omics data processing may not. A challenge for the GLDS will be to help spur the creation of community resources so that it can implement rich metadata for processed omics data and workflows in a timely fashion.

The objective of the metric for principle I3 is to measure how much data in one system are linked semantically to data in other systems. It is not surprising that none of the data sources link data "out" to data in external systems in semantically qualified ways; this is a challenging objective for any data system. (We ignored links to records that were merely duplicated between systems) Manual linking of data requires subject matter expertise and can be time-consuming for systems with many data records. Automated linking requires that source and linked data both be characterized semantically so that links may be inferred. As mentioned above, semantic resources for characterizing omics data are by and large still forthcoming.

Reusability

R1 is referred to by the findability FAIR principle F2 (see above); at this time there is no metric specifically for R1. R1.1 requires data systems to provide licenses for data usage (and is curiously intended to facilitate "legal interoperability" rather than reusability[26]. All the data repositories we assessed are governmental entities, and as such, licensing is traditionally prohibited.

R1.2 and R1.3 require that data are published together with provenance metadata, and that data and metadata conform to community-derived standards, respectively. Such standards include taxonomies, ontologies and other vocabularies for metadata, and data and file formats. The metric for R1.2 involves demonstration that a system can provide FAIR-compliant citational and contextual provenance metadata. While all the repositories provide such metadata, some do not use FAIR-compliant schemes for these metadata. Similarly, most of the repositories assessed validate submitted metadata and data files against FAIR-schemes and formats.

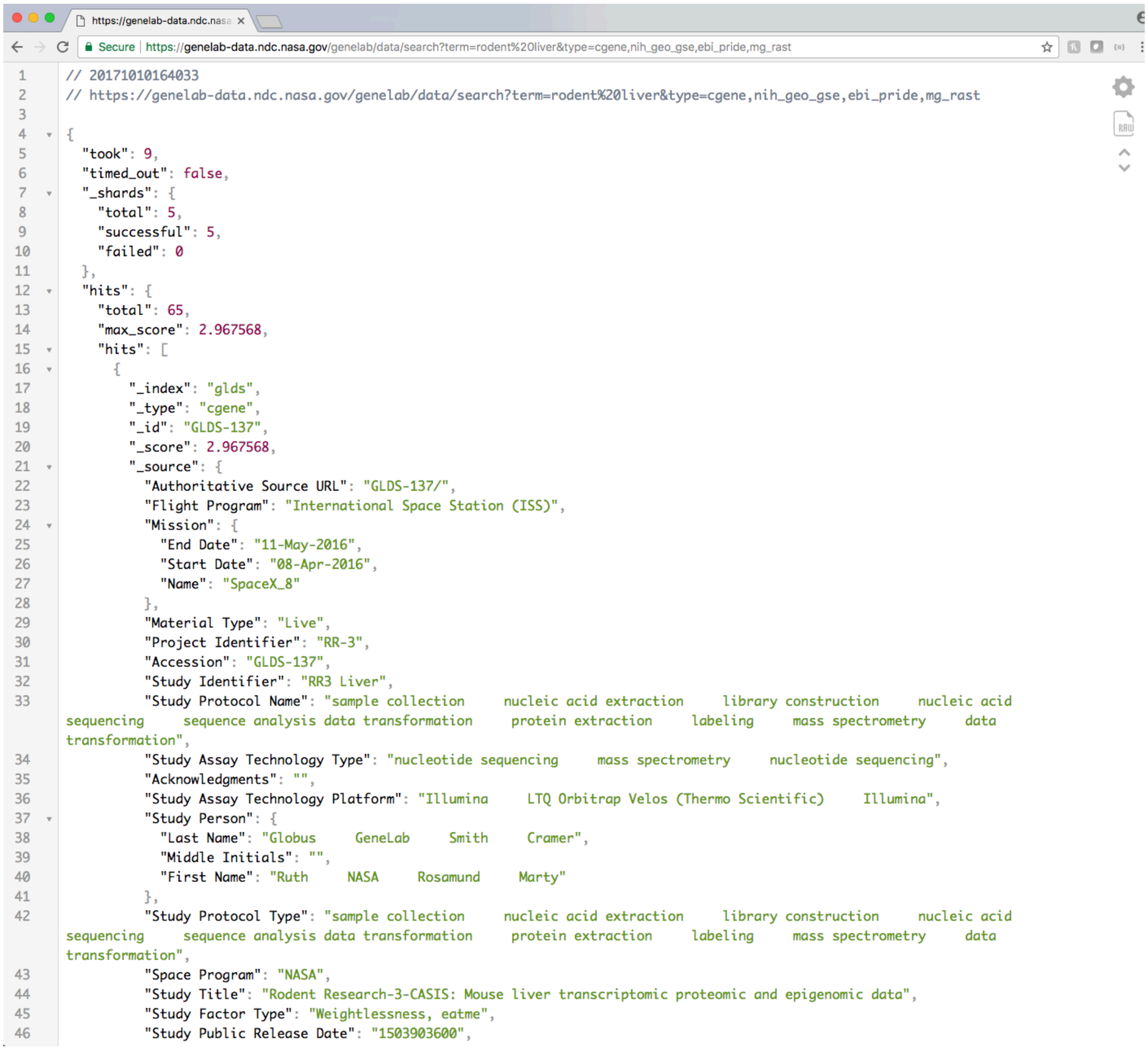
Summary of FAIRness Assessments

There is quite some variability in FAIRness among the omics data systems we assessed using the 14 metrics. The range of Pass ratings was 29-79% of metrics, Partial Pass 0-21%, and Fail 7-50%. The range of overall FAIRness scores was 5-12. The systems evaluated performed the best in the areas of data findability and accessibility, and worst in the area of data interoperability.

Discussion

The systems we assessed for FAIRness clearly have shortcomings, particularly in the area of supporting data interoperability. None of the systems currently use a metadata representation language with sufficient semantics. In addition, none fully leverage FAIR vocabularies for metadata. There are perhaps several causes at work for these observations. There are often myriad vocabularies (ontologies) for metadata (each of which is a *de facto* community standard) even within a given biological sub-domain, and often much overlap in semantic content between them, creating the need for systems to be able to "normalize" input metadata across them somehow. For example, the concept of zero or very low-gravity experimental conditions appears in at least 5 different, widely-used biomedical vocabularies: SNOMED CT, MeSH, Read Codes Clinical Terms, Psychology Ontology, PLOS Thesaurus, and FAST Topical Facet, with varying language for the concept ("weightlessness" vs. "zero gravity"). True semantic interoperability and data reusability that would permit a user to find all data in tagged with this concept require either that all these concept instances be individually associated with the data, or use of a system that reliably maps all the concepts from these ever-changing, ever-expanding set of ontologies. Even if such metadata normalizing mechanisms are developed, their performance and value in generating "rich" metadata is dependent on the quality of input concepts supplied either by original data submitters or biodata curators (typically employed by the system). The expertise that submitters have is critical to shaping metadata that have value for many different kinds of audiences, and that allow associated data to be found and reused. In the biomedical field, submitters to shared data repositories have relatively low motivation to go to any great lengths to provide "rich metadata" (or assist biocurators in generating them). This is likely due, at least in part, to the fact that only a small fraction of shared data appears to be cited in publication references when they are reused.[27] Without their input, niche repositories and data systems, like the GLDS, with limited resources for acquiring (and maintaining) rich metadata that meet all the criteria above must constantly trade the use of resources to acquire or generate rich metadata with those that could be used for other worthy goals.

The findability and interoperability principles, when implemented in data systems, support data integration functions like multi-source data querying and retrieval. Principles F2, I1, I2, and R1.3 require the use of community-developed metadata standards, formal languages for metadata representation, FAIR vocabularies, and qualified references to other data. These principles are unfortunately not yet widely or fully implemented across biomedical data systems, making it difficult for systems to develop functions that integrate data semantically. To illustrate this, we examine our own experience developing the GLDS to provide federated search and retrieval of omics data from several data sources. We sought to import metadata records from three open-access data systems into our own metadata warehouse: the National Center for Biotechnology Information's Gene Expression Omnibus (GEO), the European Bioinformatics Institute's PRoteomics IDEntifications (PRIDE) repository, and the Argonne National Laboratory's Metagenomics Analysis server (MG-RAST). Each of these systems defines metadata for omics data sets differently, but none of them make these schemas available in a formal semantic language format (such as RDF or OWL). Furthermore, there are no existing maps of any of the metadata schemas to each other. As such, we resorted to manually creating semantic


https://genelab-data.ndc.nasa ×
Secure
https://genelab-data.ndc.nasa.gov/genelab/data/search?term=rodent%20liver&type=cgene,nih_geo_gse,ebi_pride,mg_rast

```
// 20171010164033
// https://genelab-data.ndc.nasa.gov/genelab/data/search?term=rodent%20liver&type=cgene,nih_geo_gse,ebi_pride,mg_rast

{
  "took": 9,
  "timed_out": false,
  "_shards": {
    "total": 5,
    "successful": 5,
    "failed": 0
  },
  "hits": {
    "total": 65,
    "max_score": 2.967568,
    "hits": [
      {
        "_index": "glds",
        "_type": "cgene",
        "_id": "GLDS-137",
        "_score": 2.967568,
        "_source": {
          "Authoritative Source URL": "GLDS-137/",
          "Flight Program": "International Space Station (ISS)",
          "Mission": {
            "End Date": "11-May-2016",
            "Start Date": "08-Apr-2016",
            "Name": "SpaceX_8"
          },
          "Material Type": "Live",
          "Project Identifier": "RR-3",
          "Accession": "GLDS-137",
          "Study Identifier": "RR3 Liver",
          "Study Protocol Name": "sample collection     nucleic acid extraction     library construction     nucleic acid
sequencing     sequence analysis data transformation     protein extraction     labeling     mass spectrometry     data
transformation",
          "Study Assay Technology Type": "nucleotide sequencing     mass spectrometry     nucleotide sequencing",
          "Acknowledgments": "",
          "Study Assay Technology Platform": "Illumina     LTQ Orbitrap Velos (Thermo Scientific)     Illumina",
          "Study Person": {
            "Last Name": "Globus     GeneLab     Smith     Cramer",
            "Middle Initials": "",
            "First Name": "Ruth     NASA     Rosamund     Marty"
          },
          "Study Protocol Type": "sample collection     nucleic acid extraction     library construction     nucleic acid
sequencing     sequence analysis data transformation     protein extraction     labeling     mass spectrometry     data
transformation",
          "Space Program": "NASA",
          "Study Title": "Rodent Research-3-CASIS: Mouse liver transcriptomic proteomic and epigenomic data",
          "Study Factor Type": "Weightlessness, eatme",
          "Study Public Release Date": "1503903600",
```

Figure 4. Federated Search of Omics Data in the GLDS. The RESTful search API result shows 64 records from the four data sources searched (GeneLab, NIH GEO, EBI PRIDE, MG-RAST). Each records metadata is mapped using a single schema for display in the GLDS (see **Figure 1**a).

maps of these schemas for use in our system; queries generated via the GLDS or its RESTful search API are executed against the mapped schema concepts, and matching records are shown in a common object model representation (**Figure 4**). While this approach met our immediate goals of allowing users to query several data sources using our system, it required a good deal of resources and domain expertise to develop a schema mapping. Furthermore, maintaining this mapping over time as each of the data sources inevitably change will require more of our limited resources. The research community should recognize the value of federated system functions like the querying function that the GLDS offers, and urge all open access repositories and data systems to make their metadata schemas available to all according to the aforementioned FAIR principles.

References

1. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome research. 2005;15(10):1451-5.

2. Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, Dehal P, et al. The DOE Systems Biology Knowledgebase (KBase). bioRxiv. 2016.
3. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, et al. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. Frontiers in plant science. 2011;2:34.
4. Motenko H, Neuhauser SB, O'Keefe M, Richardson JE. MouseMine: a new data warehouse for MGI. Mammalian genome : official journal of the International Mammalian Genome Society. 2015;26(7-8):325-30.
5. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic acids research. 2001;29(1):82-6.
6. Arend D, Junker A, Scholz U, Schuler D, Wylie J, Lange M. PGP repository: a plant phenomics and genomics data publication infrastructure. Database : the journal of biological databases and curation. 2016;2016.
7. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic acids research. 2011;39(Database issue):D19-21.
8. Jones P, Cote RG, Martens L, Quinn AF, Taylor CF, Derache W, et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucleic acids research. 2006;34(Database issue):D659-63.
9. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, et al. The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic acids research. 2001;29(1):102-5.
10. FlyBase: the Drosophila database. Nucleic acids research. 1996;24(1):53-6.
11. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic acids research. 2005;33(Database issue):D428-32.
12. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research. 1999;27(1):29-34.
13. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-9.
14. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3:160018.
15. Wilkinson MD, Sansone S-A, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. bioRxiv. 2017.
16. Berrios D, Thompson T, Fogle H, Rask J, Coughlan J. GeneLab: NASA's Open Access, Collaborative Platform for Systems Biology and Space Medicine. Poster presented at: 2015 AMIA Annual Symposium; November 14-18; San Francisco, CA2015.
17. Berrios D, Welch J, Fogle H, Skidmore M, Marcu O. GeneLab: NASA’s GeneLab: Phase I Results and Plans. Poster presented at: 2016 AMIA Annual Symposium; Chicago, IL2016.
18. Beheshti A, Cekanaviciute E, Smith DJ, Costes SV. Global transcriptomic analysis suggests carbon dioxide as an environmental stressor in spaceflight: a systems biology GeneLab case study. Sci Rep. Forthcoming 2018.
19. FAIR Metrics ALL: GO FAIR Metrics Group; 2018 [updated January 11, 2018. Available from: https://doi.org/10.5281/zenodo.1065973.
20. McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology. 2017;15(6):e2001414.
21. Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, et al. Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Comput Sci. 2015;1.
22. Boudry C, Chartron G. Availability of digital object identifiers in publications archived by PubMed. Scientometrics. 2017;110(3):1453-69.
23. Neumann J, Brase J. DataCite and DOI names for research data. Journal of computer-aided molecular design. 2014;28(10):1035-41.
24. Patel-Schneider PF, Hayes P, Horrocks I. OWL Web Ontology Language: Semantics and Abstract Syntax. W3C; 2004.
25. Peter A, Michael R. C, Nebojša T, Brad C, John C, Michael H, et al. Common Workflow Language, v1.02016.
26. R1.1: (Meta)data are released with a clear and accessible data usage license: GO FAIR Initiative; [Available from: https://www.go-fair.org/fair-principles/r1-1-metadata-released-clear-accessible-data-usage-license/.
27. Kratz JE, Strasser C. Researcher perspectives on publication and peer review of data. PloS one. 2015;10(2):e0117619.