AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

FAIRness and Usability for Open-access Omics Data Systems

S50: User-centered Design

Daniel C. Berrios MD MPH PhD, Afshin Beheshti PhD Sylvain V. Costes PhD

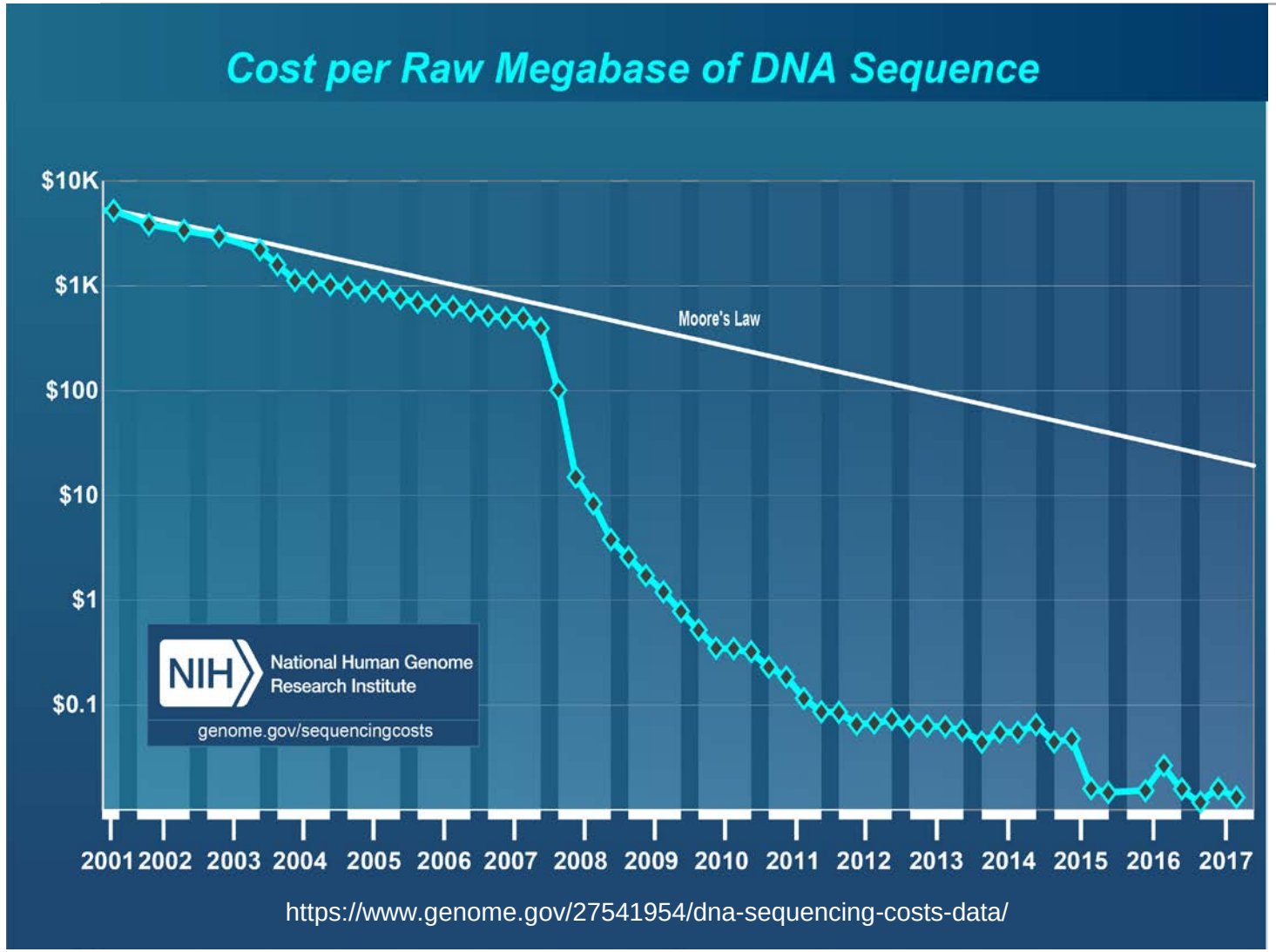
Universities Space Research Association, KBRwyle, & NASA
NASA Ames Research Center, Moffett Field, CA

Outline

- Introduction: Omics Data & Data Systems, FAIRness
- Methods: Metrics, Raters, Systems Rated
- Results
  - Findability
  - Accessibility
  - Interoperability
  - Reusability
- Conclusions

* No relevant financial disclosures

Trajectory of Sequence Data Generation



Cost of sequencing data acquisition falling precipitously

Continued development of even more high-throughput sequencing applications and techniques

- Single-cell omics (e.g., scRNA-Seq)
- Spatial transcriptomics
- Epigenetic applications: Epigenomics, Epitranscriptomics, Proteogenomics

Inundation of the biomedical community by omics data unlikely to abate

Developers are increasingly focused on the needs of users to discover, annotate, share, and analyze omics

Omics Data Systems Multiplying

BaseSpace
Arvados
GenomeSpace
INSDC Databases
MG-RAST
Metabolights
BlueBee
Many, many, more…

The "FAIR" Principles

FINDABILITY

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

INTEROPERABILITY

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

ACCESSIBILITY

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 if data are no longer available

REUSABILITY

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

Metadata Normalization

Inferring Data Relationships

Data Federation

Processed Data Metadata Schemas

Cost-efficient Metadata Acquisition

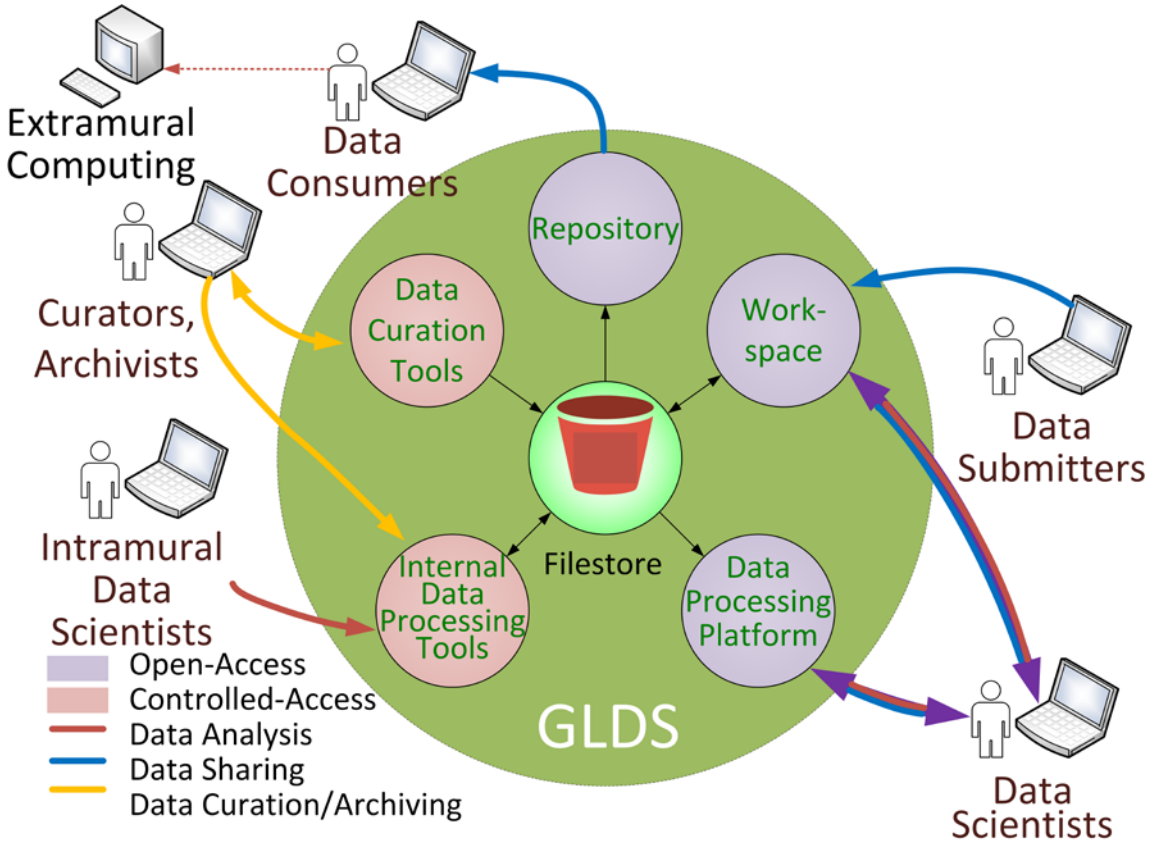
Data Identification Schemes

Integrated Credentialing

NASA GeneLab Data Systems (GLDS)

- Single platform for Omics data archiving, sharing, and analysis
  - Diversity of users: citizen scientists, students, educators, NASA-funded PIs, other investigators
  - Diversity of usage: data browsing to data analysis, data submission, data set curation
  - Open access, with private data option (for pre-publication)

- Seeking best practices for complying with FAIR principles
  - Identify costs of implementing, particularly “trail-blazing” system features
  - Identify widely vs. rarely compliant principles



Goals

- Assess the “FAIRness” of 4 systems in the research omics data domain
- Compare FAIRness with that of the GLDS
- Gather knowledge of technical and cost challenges for FAIR compliance
- Incorporate this knowledge in GLDS designs

Methods

5 Omics Data Systems, similar to GLDS

- Data archiving capabilities
- Open access
- Government-operated or government-funded, research lab-developed

14 candidate FAIRness metrics

- 5 Findability, 3 Accessibility, 3 Interoperability, 3 Reusability
- Developed by the FAIR Metrics Group (http://fairmetrics.org/)

MG-RAST

NCBI GEO

ENA European Nucleotide Archive

MetaboLights

Methods Cont'd

3 Raters

- PASS (1): No evidence of failure of any test, for any input
- PARTIAL PASS (0.5): Failure of some, but not all tests, or test steps/components
- FAIL (0): No evidence of compliance to the principle, for any inputs tested

Consensus

- Individually-assessed ratings combined through dialogue among the raters until consensus reached

MG-RAST

NCBI GEO

ENA European Nucleotide Archive

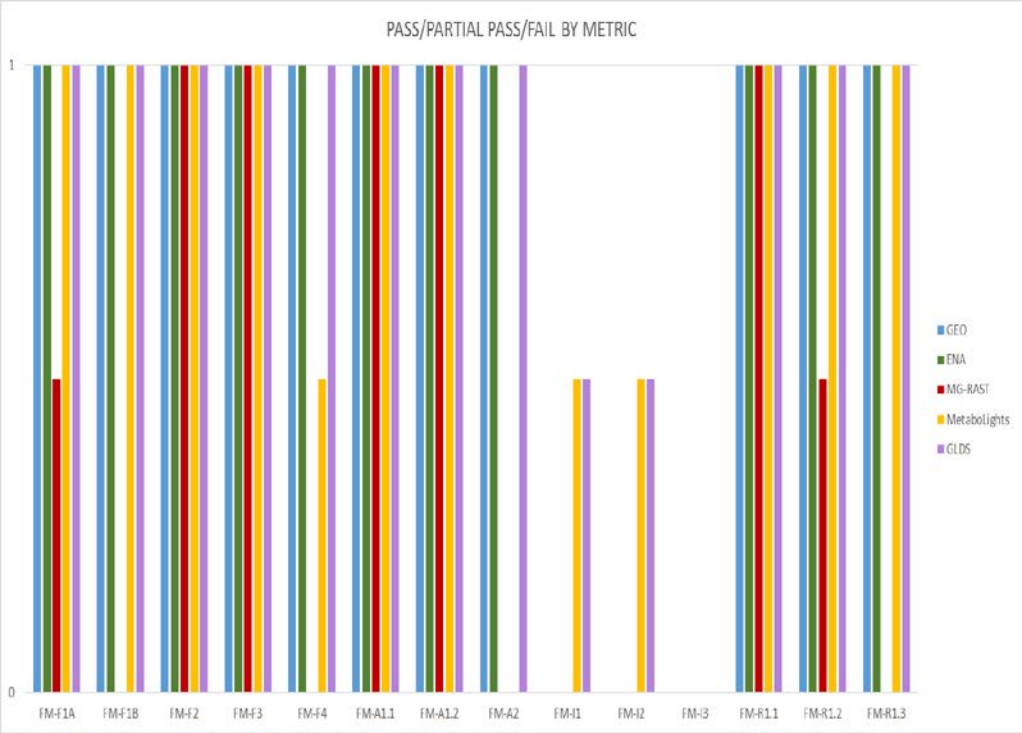
MetaboLights

Results

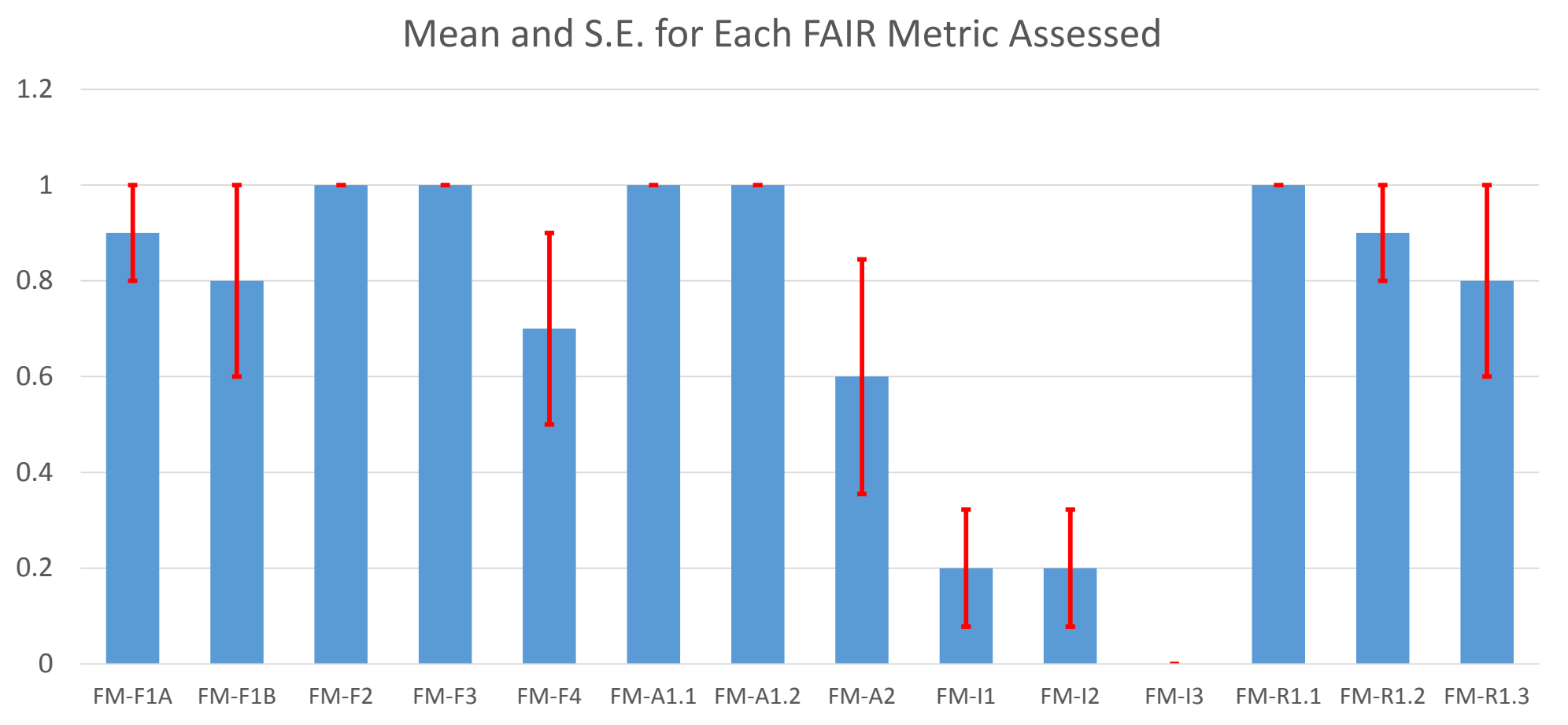
● Pass
◉ Partial pass
○ Fail

FAIR Principle	Metric	GEO[4]	ENA[7]	MG-RAST[6]	Metabolights[8]	GLDS
F1. (meta)data are assigned globally unique and persistent identifier	FM-F1A	●	●	◉	●	●
F1. (meta)data are assigned globally unique and persistent identifier	FM-F1B	●	●	○	●	●
F2. data are described with rich metadata (defined by R1 below)	FM-F2	●	●	●	●	●
F3. metadata clearly/explicitly include identifier of data it describes	FM-F3	●	●	●	●	●
F4. (meta)data are registered or indexed in a searchable resource	FM-F4	●	●	○	◉	●
A1. (meta)data are retrievable by 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 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 released with clear, 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 FAIRness Score		11	11	6	10.5	12

Range of FAIRness Scores (all systems)
PASS: 29-79%
PARTIAL PASS: 0-21%
FAIL: 7%-50%



Results


Mean and S.E. for Each FAIR Metric Assessed
1.2
1
0.8
0.6
0.4
0.2
0
FM-F1A
FM-F1B
FM-F2
FM-F3
FM-F4
FM-A1.1
FM-A1.2
FM-A2
FM-I1
FM-I2
FM-I3
FM-R1.1
FM-R1.2
FM-R1.3

FINDABILITY

FM-F1A: Data identifier uniqueness

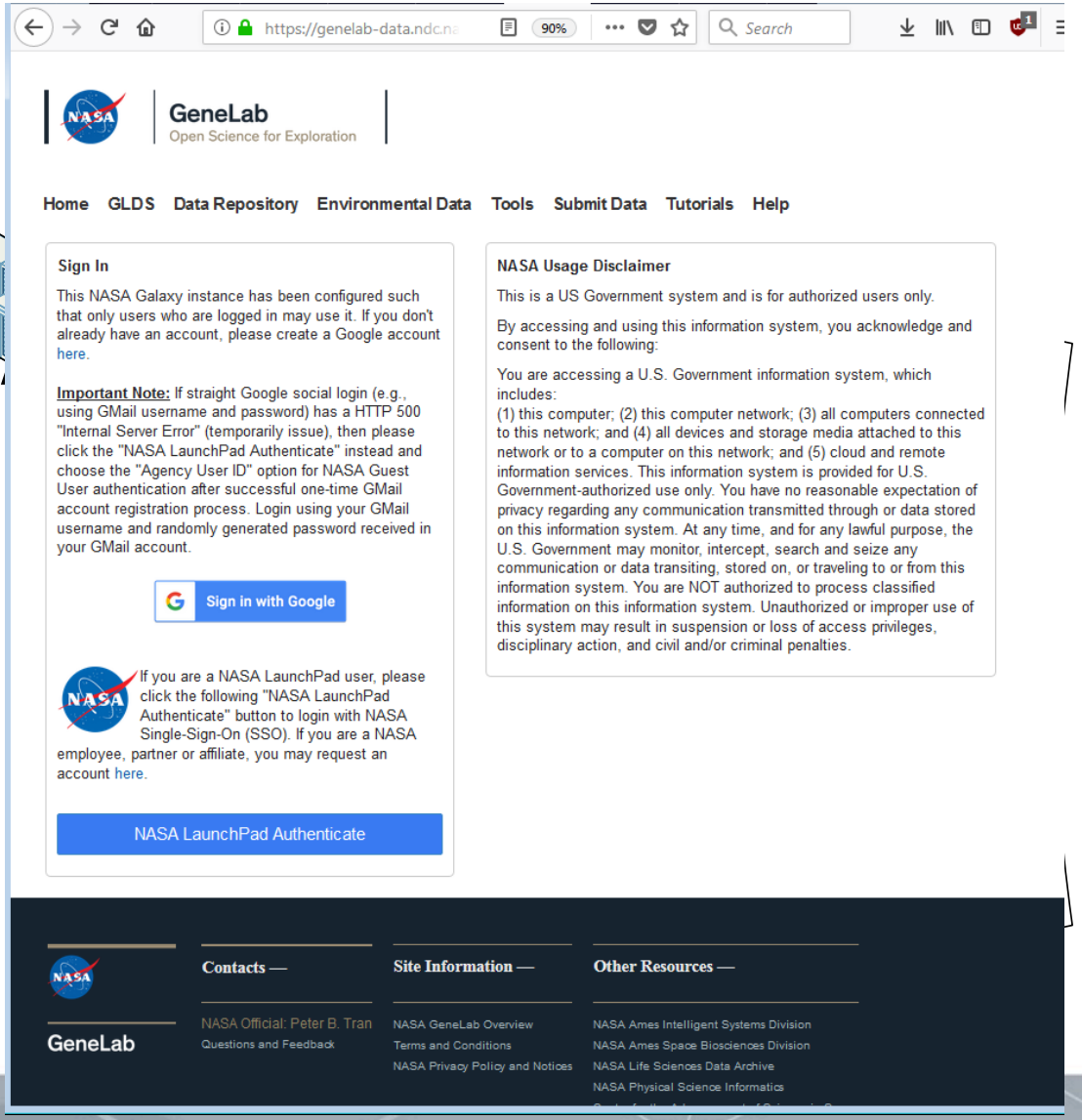
FM-F1B: Data identifier persistence

- URIs/IRIs were not designed to be persistent, but are being used as if they were
- DOIs were, yet as still not widely used for Data Citation
- GeneLab Data Systems will employ DOIs using in-house DOI services

FM-F4: Submit metadata to data search engines

ACCESSIBILITY

- Integrated credentialing and authentication ("single sign-on") are valuable for users, yet not common
- GeneLab Data Systems employing Google or Agency credentials for logins



ACCESSIBILITY

Data transport is frequently required by omics data analysis systems

- Can require minutes to hours to move data into analysis environment
- Would be of value to have a FAIR principles for data location/transport

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

Interoperability & Reusability

All systems lacking interoperable metadata (and data)

- Lack of metadata schemes grounded in formal semantics (DL languages like RDF/OWL/etc.)
- Some use ISA-Tab, which has at least been modeled in RDF, although they do not represent metadata in RDF
- Community-derived, controlled vocabularies usage uncommon
- Semantic linkage of data through metadata also uncommon

Conclusions

- All the omics systems assessed had some FAIRness shortcomings
- Un-FAIRness in areas that support data interoperability is common
  - Lack of metadata representations with formalized *semantics*
  - Lack of use of FAIR vocabularies
  - Low prevalence of semantic normalization, perhaps due in part to
    - High costs of manual semantic normalization
    - Lack of automated semantic normalization resources
  - Impedes the development of functions relying on system-system interoperability

Acknowledgements

NASA GeneLab Team

Jon Galazka
Sigrid Reinsch
Homer Fogle
Sam Gebre

To see a system demonstration of NASA GeneLab:

NASA's GeneLab: An Integrated Omics Data Commons and Workbench
S64: Ontology Driven Health Information Systems Architectures
Tuesday, November 6, 2018 9:00 AM Yosemite A/B

FAIR Metrics Group