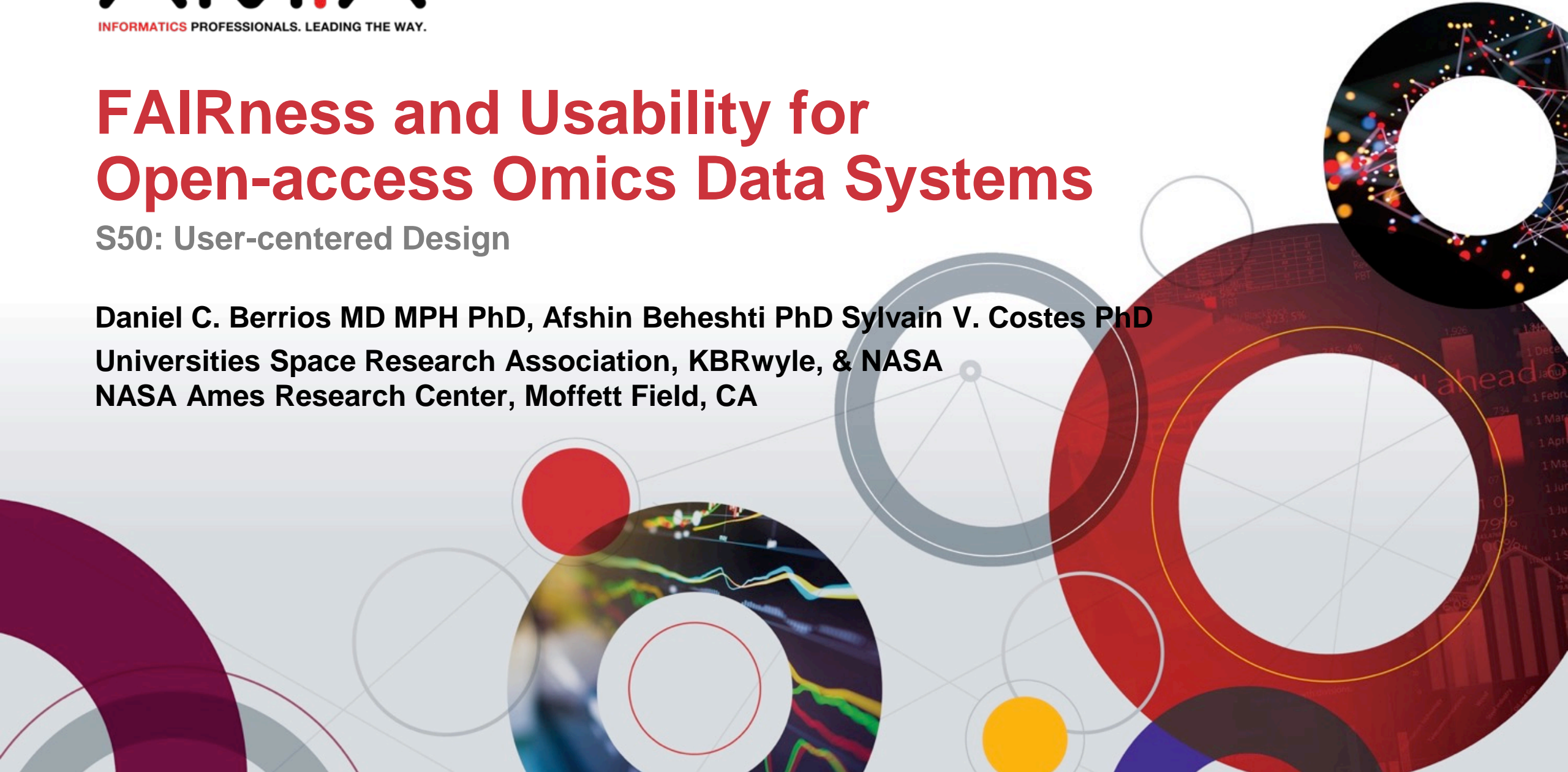


FAIRness and Usability for Open-access Omics Data Systems

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

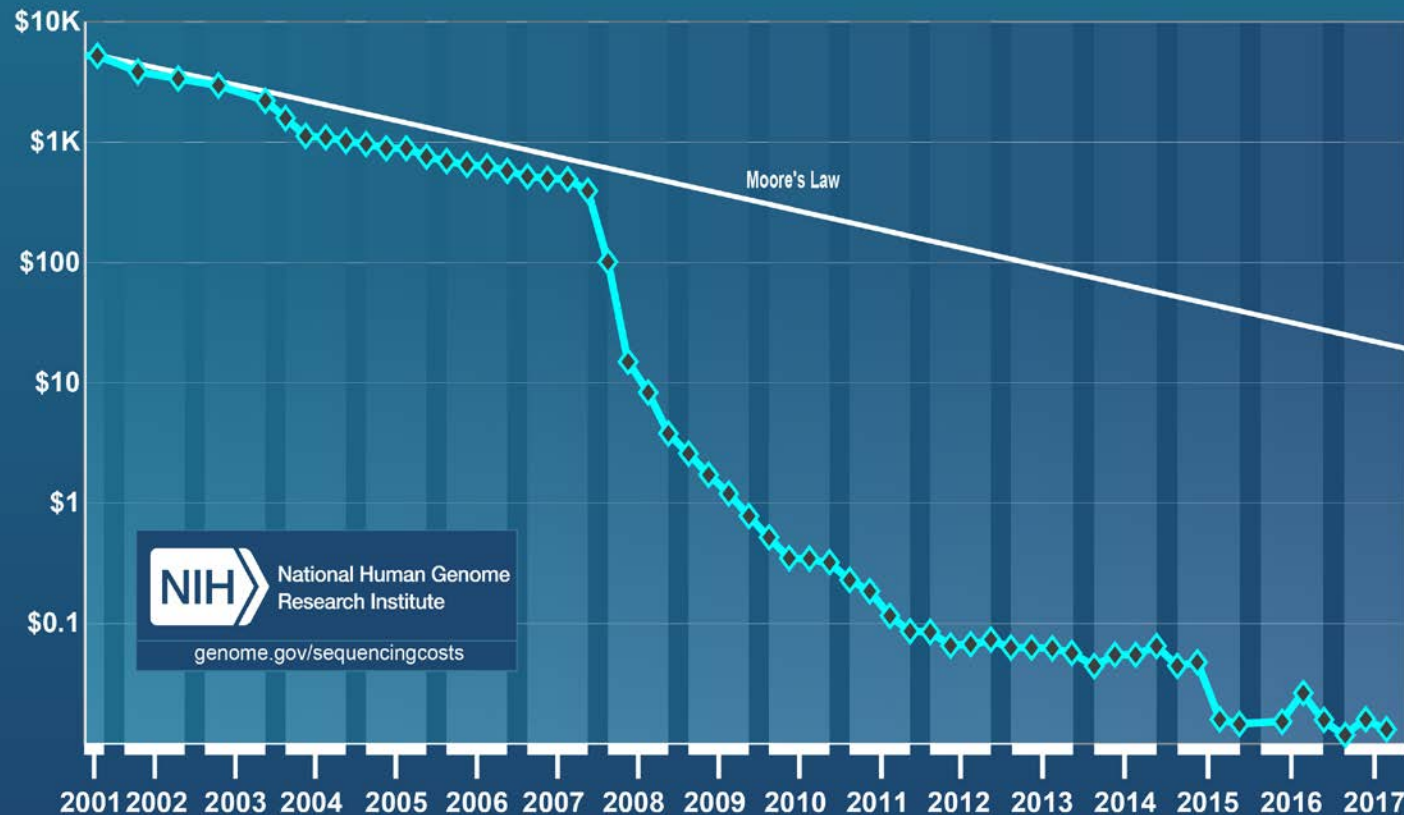


- Introduction: Omics Data & Data Systems, FAIRness
- Methods: Metrics, Raters, Systems Rated
- Results
 - Findability
 - Accessibility
 - Interoperability
 - Reusability
- Conclusions

* No relevant financial disclosures

Trajectory of Sequence Data Generation

Cost per Raw Megabase of DNA Sequence



<https://www.genome.gov/27541954/dna-sequencing-costs-data/>

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

Data Identification Schemes

- 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

ACCESSIBILITY

Integrated Credentialing

Metadata Normalization

Inferring Data Relationships

Data Federation

Processed Data Metadata Schemas

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

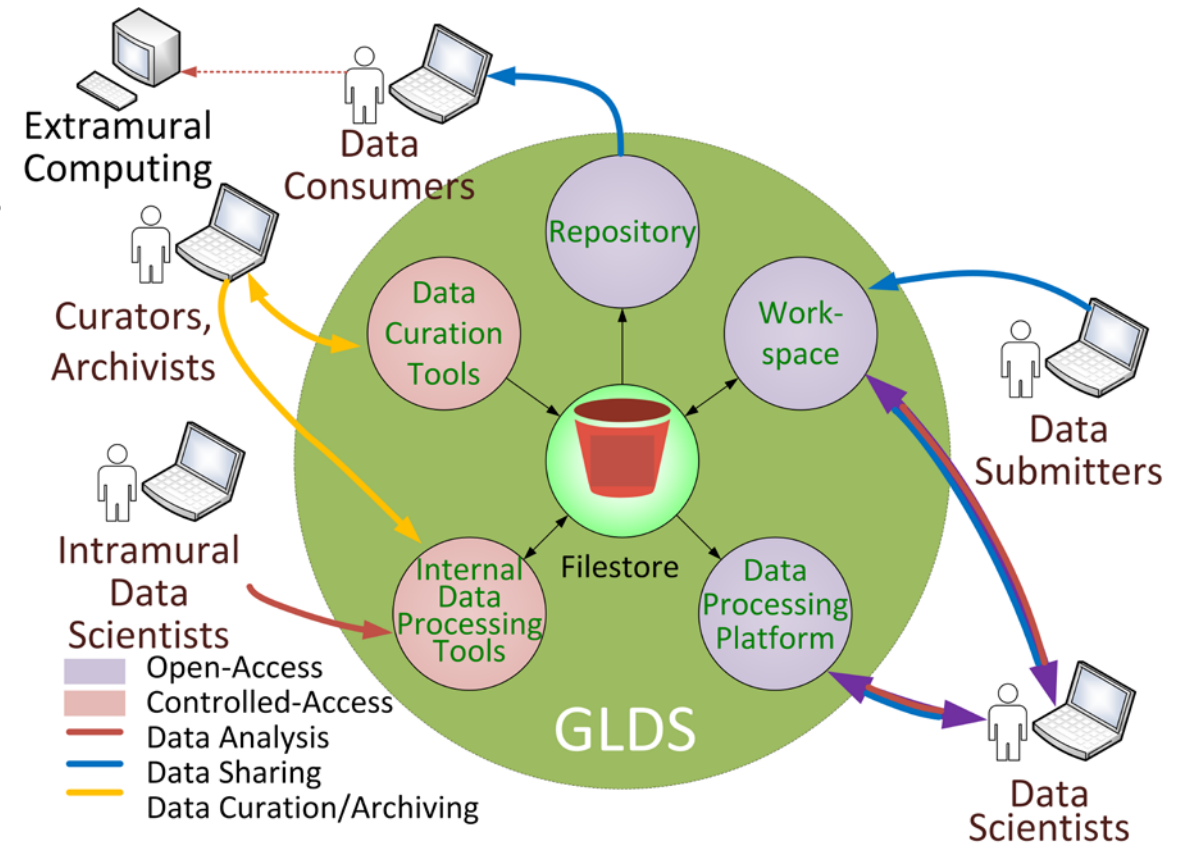
Cost-efficient Metadata Acquisition

- 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

REUSABILITY

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

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



MG-RAST

Consensus

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

NCBI GEO

ENA European
Nucleotide Archive

 MetaboLights

Results

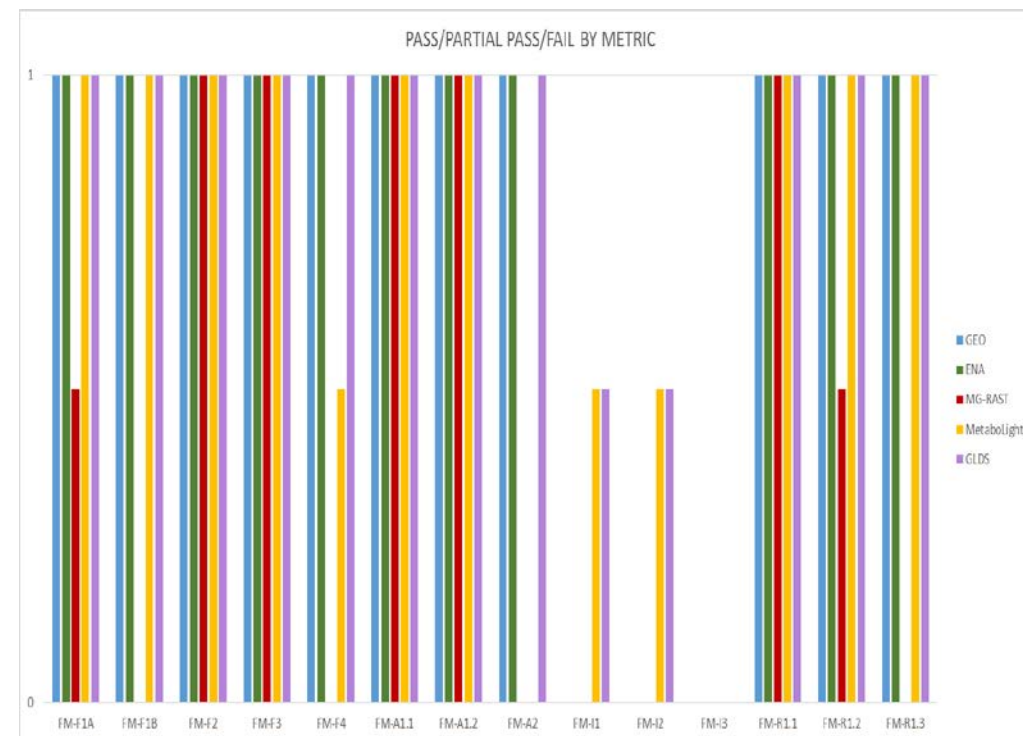
FAIR Principle	Metric	GEO ⁴	ENA ⁷	MG-RAST ⁶	Metabolights ⁸	GLDS
<div style="border: 1px solid black; padding: 5px; width: fit-content;"> Pass Partial pass Fail </div>						
F1. (meta)data are assigned globally unique and persistent identifier	FM-F1A	●	●	○	●	●
F1. (meta)data are assigned globally unique 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)

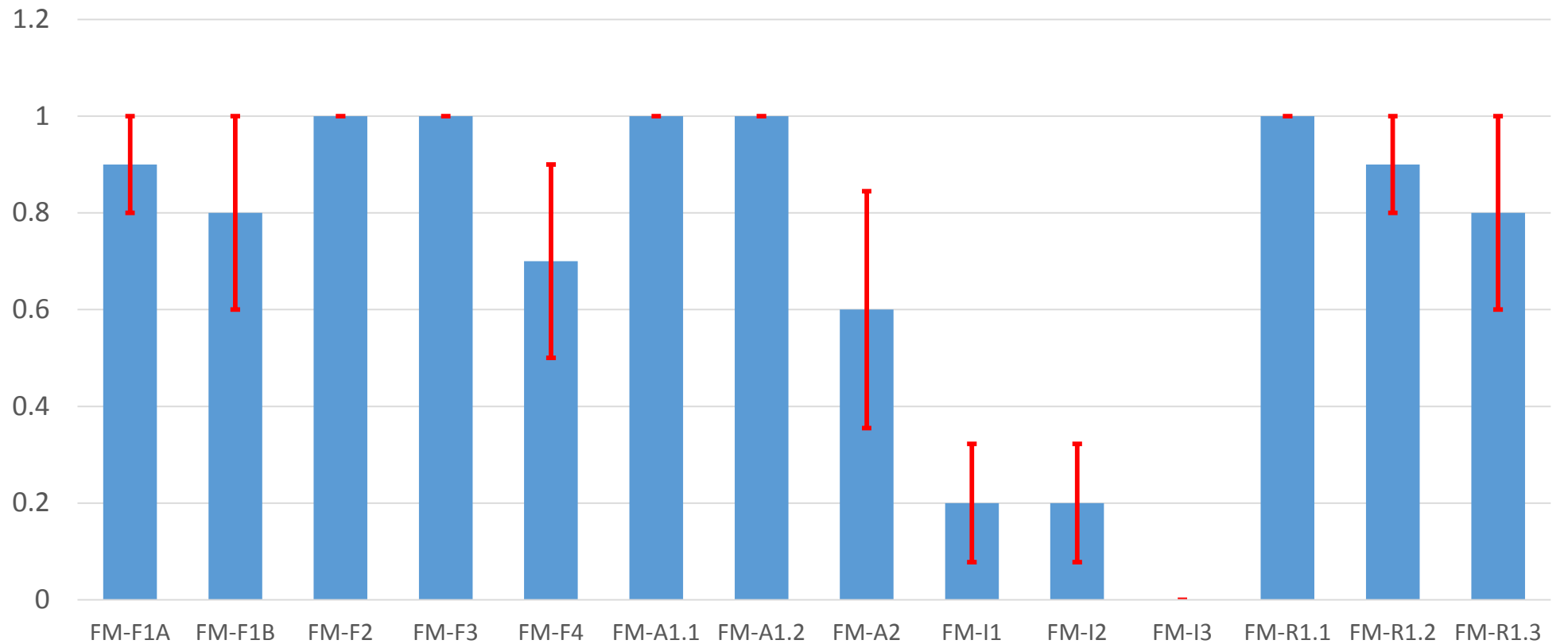
PASS: 29-79%

PARTIAL PASS: 0-21%

FAIL: 7%-50%



Mean and S.E. for Each FAIR Metric Assessed



FM-F1A: Data identifier uniqueness

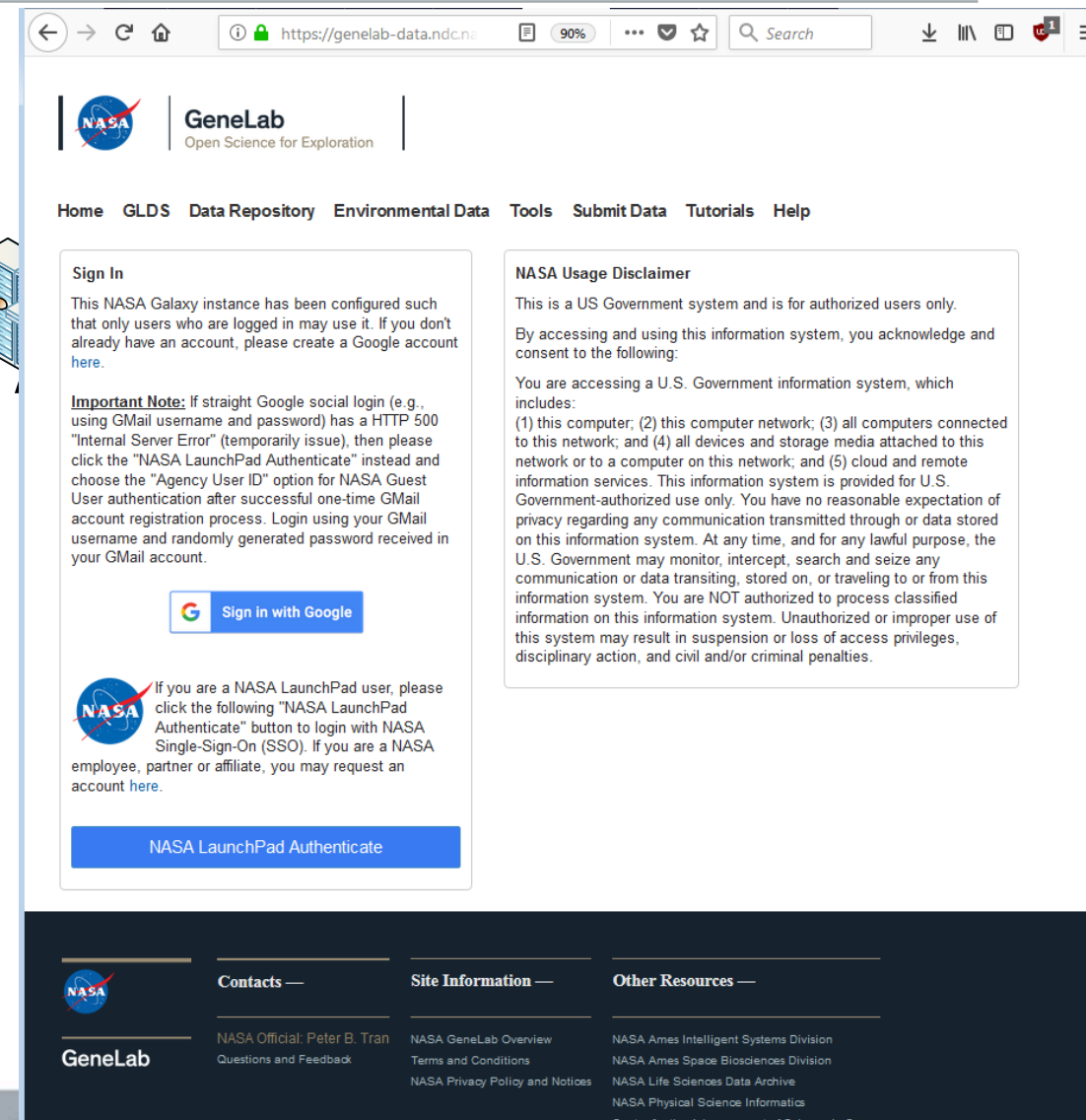
FM-F1B: Data identifier persistence

- 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




GeneLab
Open Science for Exploration

Home GLDS Data Repository Environmental Data Tools Submit Data Tutorials Help

Sign In
This NASA Galaxy instance has been configured such that only users who are logged in may use it. If you don't already have an account, please create a Google account [here](#).

Important Note: If straight Google social login (e.g., using Gmail username and password) has a HTTP 500 "Internal Server Error" (temporarily issue), then please click the "NASA LaunchPad Authenticate" instead and choose the "Agency User ID" option for NASA Guest User authentication after successful one-time Gmail account registration process. Login using your Gmail username and randomly generated password received in your Gmail account.

[Sign in with Google](#)

 If you are a NASA LaunchPad user, please click the following "NASA LaunchPad Authenticate" button to login with NASA Single-Sign-On (SSO). If you are a NASA employee, partner or affiliate, you may request an account [here](#).

[NASA LaunchPad Authenticate](#)

NASA Usage Disclaimer
This is a US Government system and is for authorized users only. By accessing and using this information system, you acknowledge and consent to the following:
You are accessing a U.S. Government information system, which includes:
(1) this computer; (2) this computer network; (3) all computers connected to this network; and (4) all devices and storage media attached to this network or to a computer on this network; and (5) cloud and remote information services. This information system is provided for U.S. Government authorized use only. You have no reasonable expectation of privacy regarding any communication transmitted through or data stored on this information system. At any time, and for any lawful purpose, the U.S. Government may monitor, intercept, search and seize any communication or data transiting, stored on, or traveling to or from this information system. You are NOT authorized to process classified information on this information system. Unauthorized or improper use of this system may result in suspension or loss of access privileges, disciplinary action, and civil and/or criminal penalties.

GeneLab

Contacts —
NASA Official: Peter B. Tran
Questions and Feedback

Site Information —
NASA GeneLab Overview
Terms and Conditions
NASA Privacy Policy and Notices

Other Resources —
NASA Ames Intelligent Systems Division
NASA Ames Space Biosciences Division
NASA Life Sciences Data Archive
NASA Physical Science Informatics

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

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

- 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