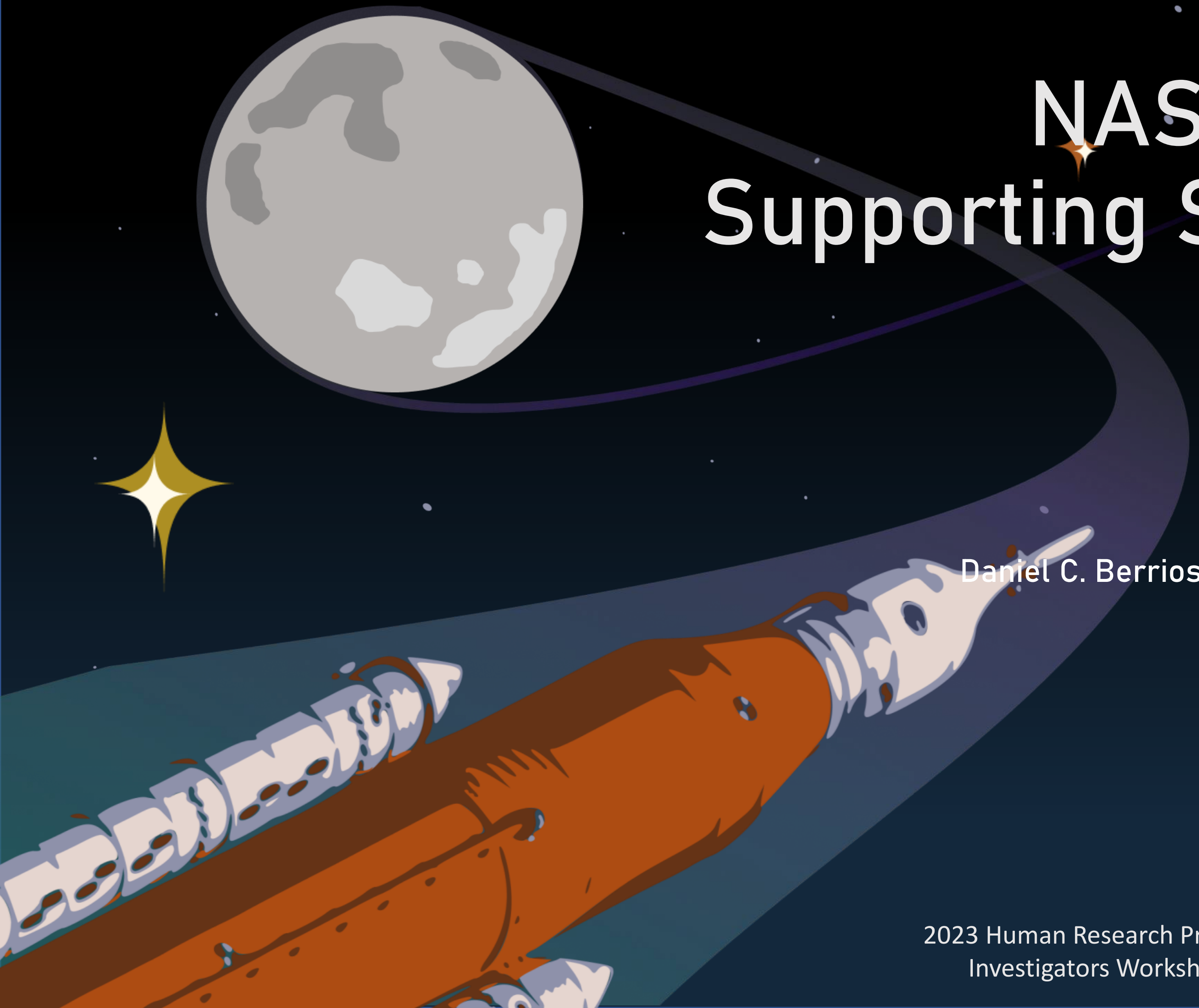


NASA Life Sciences Portal: Supporting Scientific Transparency and Reproducibility

Daniel C. Berrios[1], Macresia Alibaruho[2], Truong Le[3], John Dunn[1,4], Sandeep Shetye[1]

[1]NASA Ames Research Center, Mountain View, CA

[2]NASA Glenn Research Center, Cleveland, OH

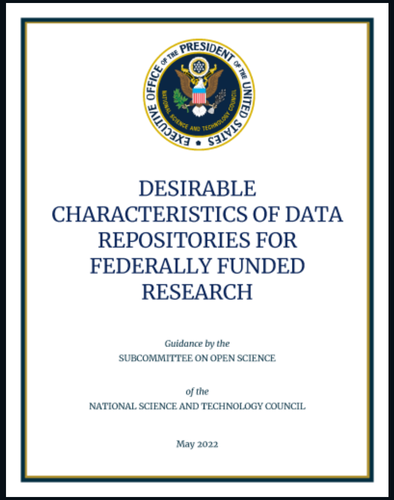
[3]NASA Johnson Space Center, Houston, TX

[4]Universities Space Research Association, Mountain View, CA

2023 Human Research Program
Investigators Workshop

NASA and Open Science

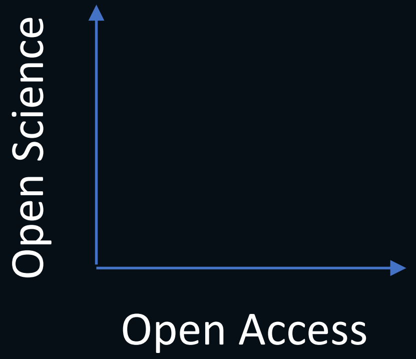
- Science cannot be termed “open” unless its conduct is transparent
  - Metadata transparency means conveying what was done clearly and uniformly
    - Unambiguous and richly annotated attributes, values
    - Community-developed and –maintained (open-source) terminology models
  - Data transparency means using open standards for data whenever possible
- Transparency enables scientific reproducibility
  - Data cannot be reproduced if the context in which it is generated is not well understood
- Open Science ≠ Open Access
  - Open Science can be conducted and supported when:
    - Access to data and/or metadata is controlled
    - Subjects/samples are not identified/identifiable
    - Protocols, personnel, assay instruments, etc. are not (fully) revealed



DESIRABLE CHARACTERISTICS OF DATA REPOSITORIES FOR FEDERALLY FUNDED RESEARCH

Guidance by the
SUBCOMMITTEE ON OPEN SCIENCE
of the
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

May 2022



Open Science and FAIR Systems

- Critical features of FAIR systems
  - Metadata standardization and harmonization
  - Linked data
- Foundational components for Open Science,
  - Enhance transparency of investigations
  - Facilitate scientific reproducibility.
- NASA biomedical repositories could improve their FAIR scores through:
  - The increased use of community-based standards for metadata
  - Ensuring more uniformity of metadata values within and across biomedical data systems
  - Capturing more correspondences between metadata (linked data)
    - “This specimen in this experiment is a sample of that organism in that experiment”
    - “This instrument used in this experiment is the same as that instrument used in that experiment”
    - Etc.

NLSP Plan for Increasing FAIR/Open Science Compliance

Low **FAIR** Compliance	Improved **FAIR** Compliance
Lack of Standard Metadata Metamodel Lack of Standard Metadata Model	Implement **ISA-tab** Metadata Metamodel
	Develop and Deploy Open-source Metadata Model (**Ontologies**)
Lack of Standard Metadata Format	Implement the **ISA-tab** format standard
Lack of Data Identifiers	Implement DOI for Data Objects
Lack of Data Licenses	Implement Licenses for Data Objects

Increasing FAIR Compliance: Rich Metadata

- Use of Reference Vocabularies obviate need for retrospective metadata harmonization
  - SMEs develop and maintain the vocabularies
  - Re-use existing where appropriate
  - Both data producers and data consumers have access to browse, search

- Use of "Object-oriented" Vocabularies supports data linking
  - XML/RDF/OWL ontologies can be used as highly-annotated and well-organized vocabularies
  - Ontologies have classes, instances, relations, and relationships (relations between instances)

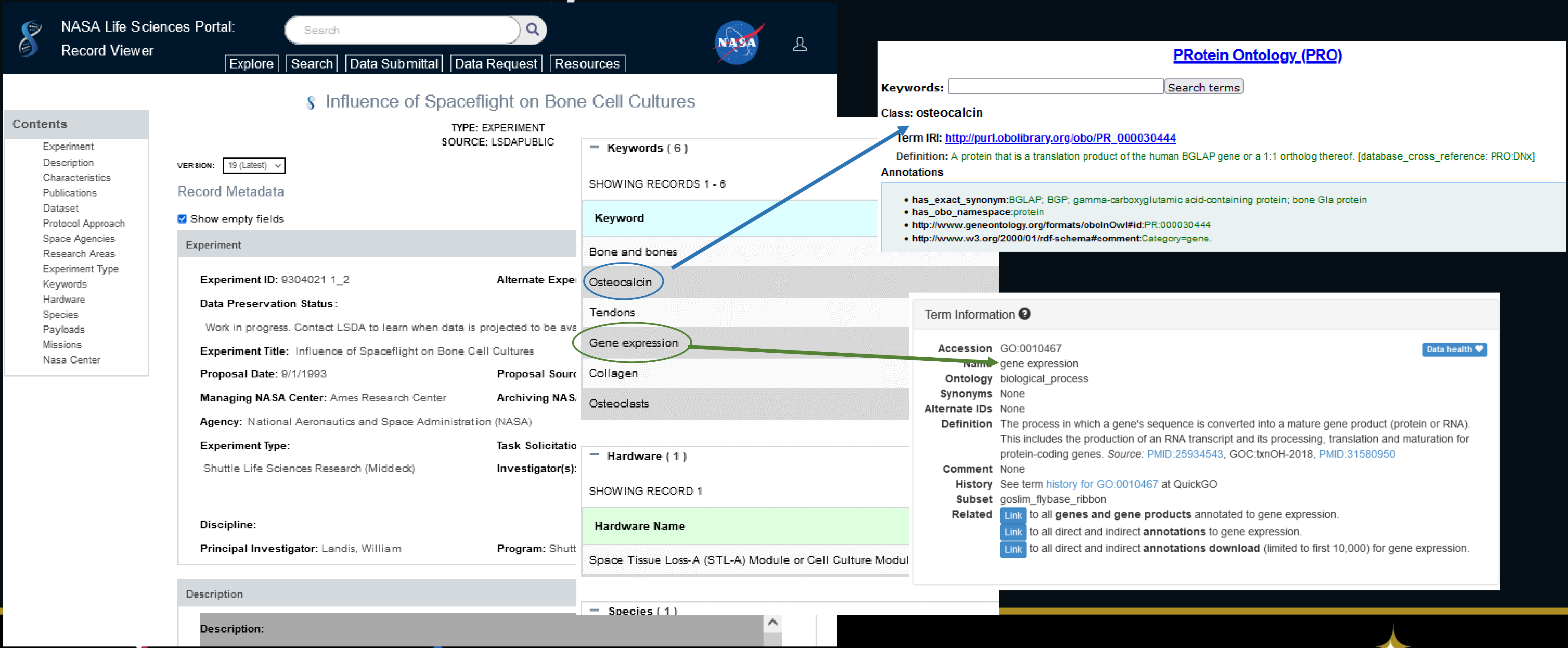
Biomedical Investigation Ontologies

- OBO Foundry (~ 200 ontologies)
  - OBI Ontology for Biomedical Investigations
  - GO Gene Ontology
  - ENVO Environment Ontology
  - RBO Radiation Biology Ontology
- W3C
  - SOSA/SSN (Semantic Sensor Network)
  - TO Time Ontology
- NIH / NCBO (National Center for Biomedical Ontology) (1136 Ontologies, and counting)
  - NCBO Taxon: Ontological transformation of NCBI Taxonomy

Clinical Ontologies

- SNOMED CT OWL
- ICD 9, 10 OWL
  - and other WHO ontologies
  - See Bioportal.bioontology.org for more
- RxNORM
- LOINC
- ENVO
  - To characterize environments/exposures

Use of Ontologies for “Rich” Metadata



Influence of Spaceflight on Bone Cell Cultures

TYPE: EXPERIMENT
SOURCE: LSDAPUBLIC

VERSION: 19 (Latest)

Record Metadata

☑ Show empty fields

Experiment

Experiment ID: 9304021 1_2

Alternate Expe

Data Preservation Status:

Work in progress. Contact LSDA to learn when data is projected to be ava

Experiment Title: Influence of Spaceflight on Bone Cell Cultures

Proposal Date: 9/1/1993

Proposal Sourc

Managing NASA Center: Ames Research Center

Archiving NAS

Agency: National Aeronautics and Space Administration (NASA)

Experiment Type:

Shuttle Life Sciences Research (Middeck)

Task Solicitatio

Investigator(s):

Discipline:

Principal Investigator: Landis, William

Program: Shutt

Description

Description:

− Keywords (6)

SHOWING RECORDS 1 - 6

Keyword
Bone and bones
Osteocalcin
Tendons
Gene expression
Collagen
Osteoclasts

− Hardware (1)

SHOWING RECORD 1

Hardware Name
Space Tissue Loss-A (STL-A) Module or Cell Culture Modul

− Species (1)

PRotein Ontology (PRO)

Keywords: Search terms

Class: osteocalcin

Term IRI: http://purl.obolibrary.org/obo/PR_000030444

Definition: A protein that is a translation product of the human BGLAP gene or a 1:1 ortholog thereof. [database_cross_reference: PRO:DNx]

Annotations

- has_exact_synonym:BGLAP; BGP; gamma-carboxyglutamic acid-containing protein; bone Gla protein
- has_obo_namespace:protein
- http://www.geneontology.org/formats/oboInOwl#id:PR:000030444
- http://www.w3.org/2000/01/rdf-schema#comment:Category=gene.

Term Information

Data health

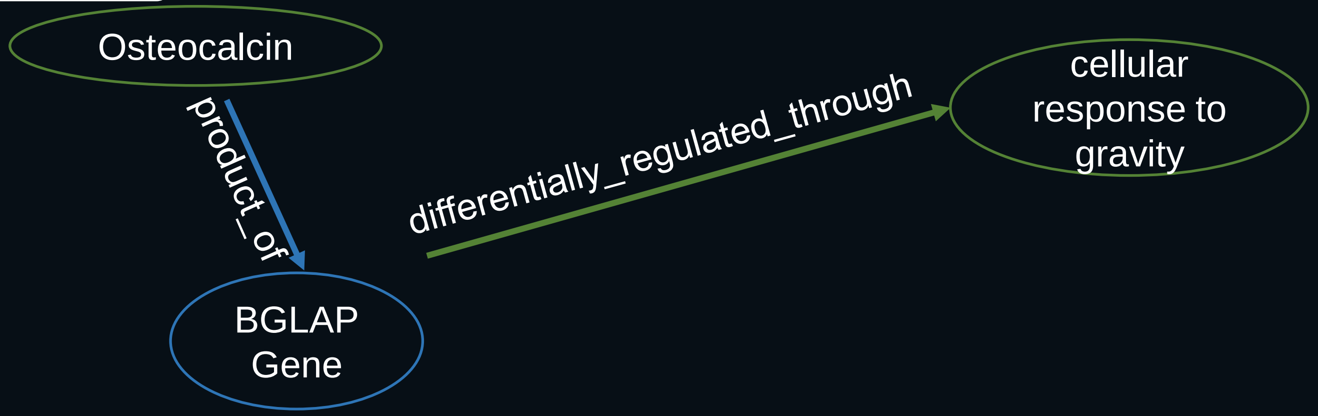
Accession	GO:0010467
Name	gene expression
Ontology	biological_process
Synonyms	None
Alternate IDs	None
Definition	The process in which a gene's sequence is converted into a mature gene product (protein or RNA). This includes the production of an RNA transcript and its processing, translation and maturation for protein-coding genes. *Source:* PMID:25934543, GOC:txnOH-2018, PMID:31580950
Comment	None
History	See term history for GO:0010467 at QuickGO
Subset	goslim_flybase_ribbon
Related	Link to all **genes and gene products** annotated to gene expression. Link to all direct and indirect **annotations** to gene expression. Link to all direct and indirect **annotations download** (limited to first 10,000) for gene expression.

From Ontologies to Linked Data (Knowledge Graphs)

OWL Graph constraints

RDF Graph structures (nodes, vertices)

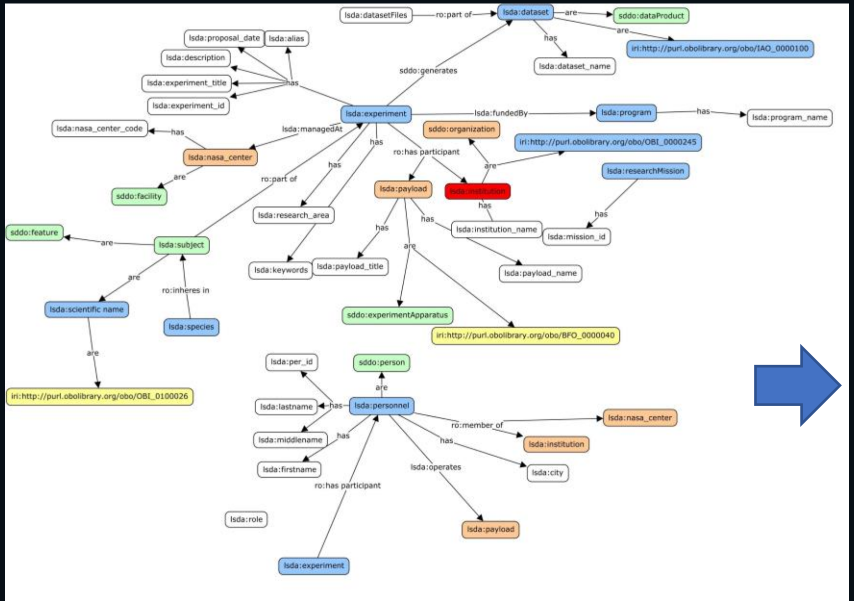
XML Hierarchical containment, namespaces,


Osteocalcin
product_of
BGLAP Gene
differentially_regulated_through
cellular response to gravity

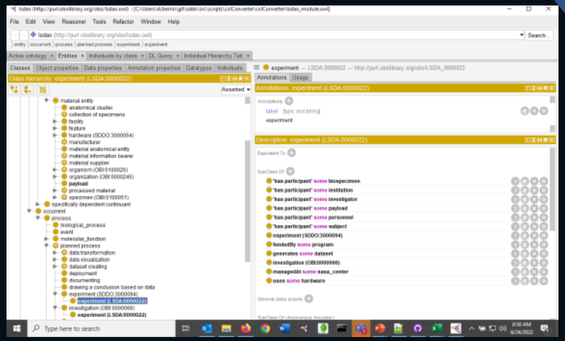
- RDF/OWL natively support logical property assertions for classes that connect *instances* through meaningful *links* to form graphs of knowledge

Life Sciences Data Archive Ontology

- An *application* ontology
- Contains
  - Classes
  - Properties/relationships
  - Inferred from the legacy LSDA
  - Contextualized within the Science Data Discovery Ontology
- Currently being enhanced with critical annotations and relationships not captured by the SDDO



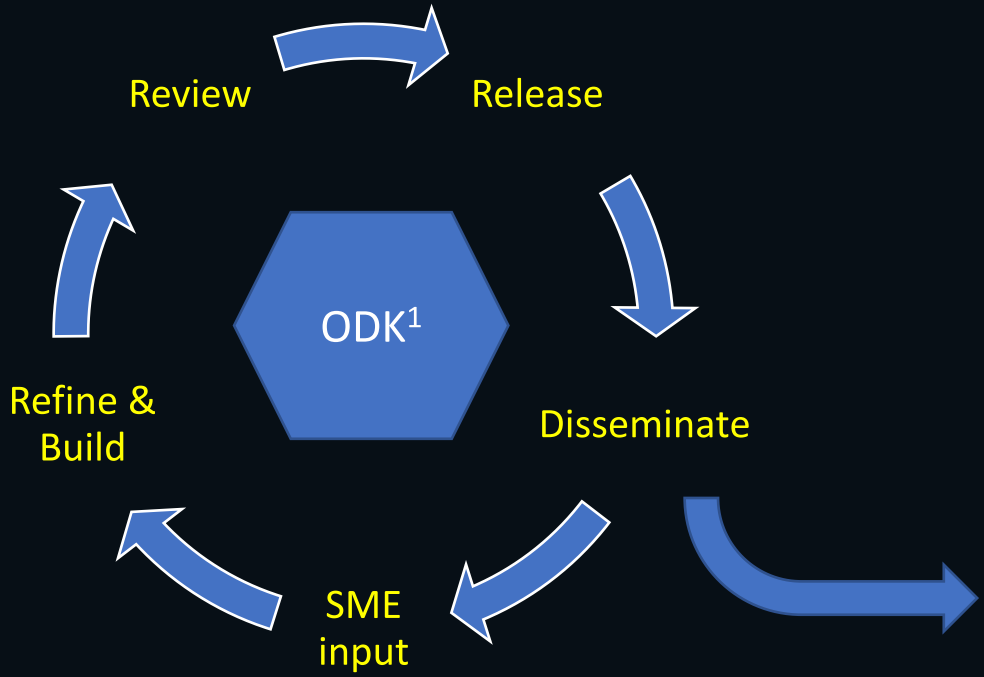
LSDA Ontology



LSDA Ontology

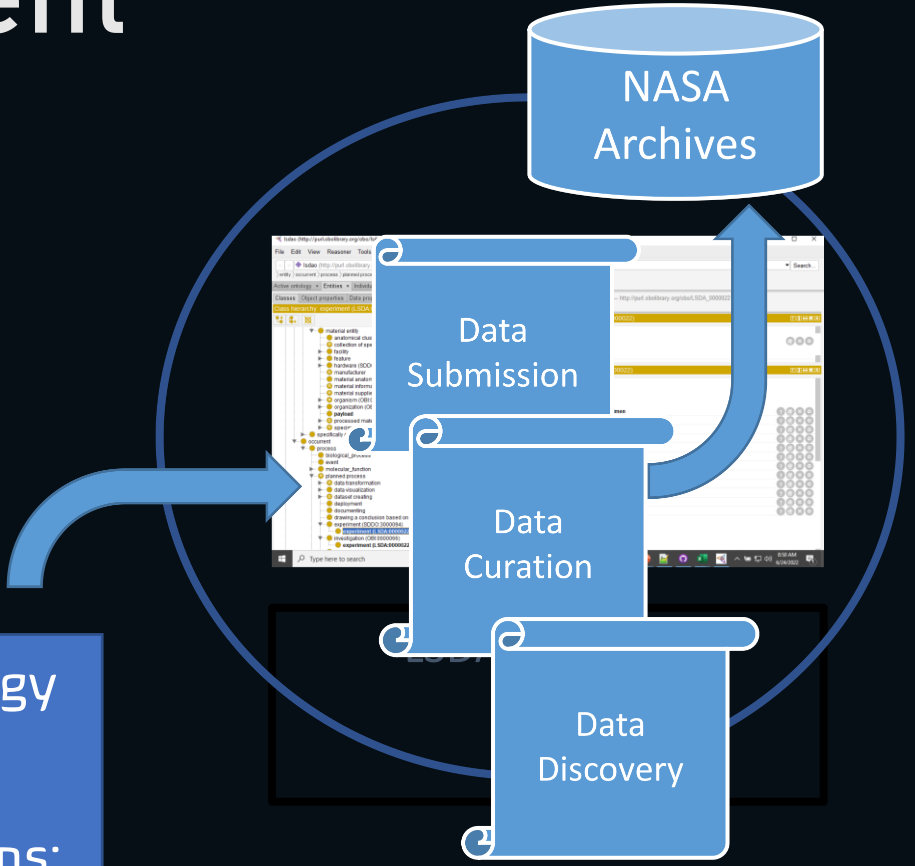
Hosted at: https://github.com/nasa/LSDAO

LSDA Ontology Development



Internal Ontology Server

External Systems:
Ontobee[2]
BioPortal[3]



[1] Ontology Development Kit
https://github.com/INCATools/ontology-development-kit

[2] https://ontobee.org/
[3] https://bioportal.bioontology.org/

Conclusions

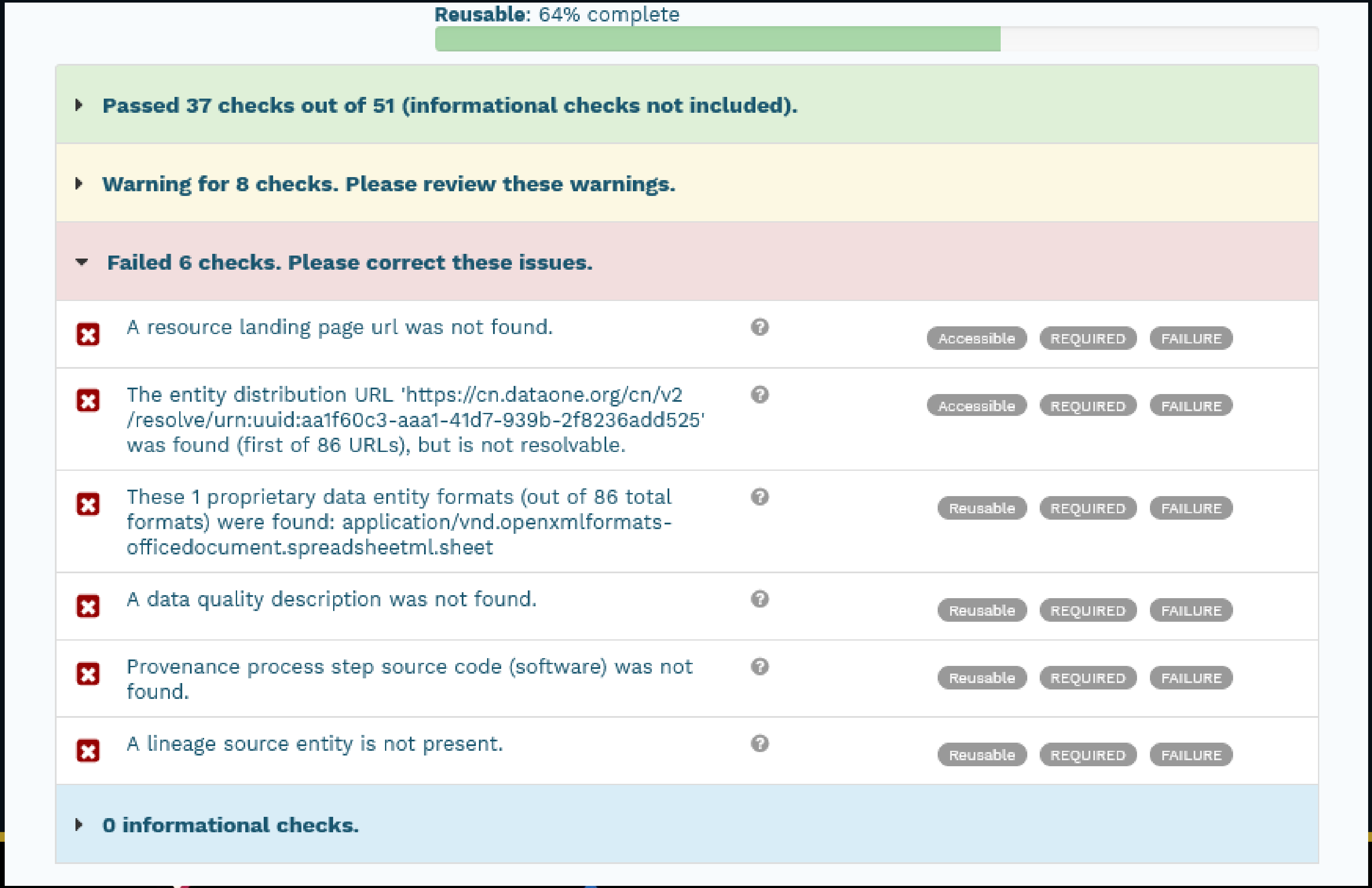
- Efforts towards frameworks that support semantic harmonization and data linkage increase transparency of science and FAIR compliance
- The NASA life sciences repositories are working with the scientific research communities to develop and use knowledge resources such as
  - Metadata frameworks/models (e.g., ISATab)
  - Standard Vocabularies (like those that are part of OBO Foundry ontologies)
  - Citation and Licensing standards and services
- Future Work: NASA will develop FAIR compliance assessment and monitoring tools for these systems

Backup

FAIR Dashboard Development

- Requirements for a FAIR Dashboard are in work
- Dashboard should give broad overview of all data holdings and their range of FAIR Compliance
  - How many objects have DOIs? Of what types? What are the DOI management metrics? What are current DOI mgmt. issues?
  - How many Data objects have DOIs? Of what types? What are the DOI management metrics? What are current DOI mgmt. issues?
  - What % of Experiments have metadata issues wrt FAIR Metrics? What % of public-access Experiments?
  - What % of metadata values are “free text” vs. ontological references?

FAIR Workbench



Reusable: 64% complete

▸ **Passed 37 checks out of 51 (informational checks not included).**

▸ **Warning for 8 checks. Please review these warnings.**

▾ **Failed 6 checks. Please correct these issues.**

☒	A resource landing page url was not found.	Accessible	REQUIRED	FAILURE
☒	The entity distribution URL 'https://cn.dataone.org/cn/v2/resolve/urn:uuid:aa1f60c3-aaa1-41d7-939b-2f8236add525' was found (first of 86 URLs), but is not resolvable.	Accessible	REQUIRED	FAILURE
☒	These 1 proprietary data entity formats (out of 86 total formats) were found: application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	Reusable	REQUIRED	FAILURE
☒	A data quality description was not found.	Reusable	REQUIRED	FAILURE
☒	Provenance process step source code (software) was not found.	Reusable	REQUIRED	FAILURE
☒	A lineage source entity is not present.	Reusable	REQUIRED	FAILURE

▸ **0 informational checks.**

- This dataset failed on 2 Accessibility and 4 Reusability Checks