

FAIRification for HPC Datasets and AI models

DOE Data Days Workshop

June 3rd, 2022



Pei-Hung Lin
Computer Scientist



FAIR Guiding Principles

- F: Findable
 - F1: (Meta) data are assigned globally unique and persistent identifiers
 - F2: Data are described with rich metadata
 - F3: Metadata clearly and explicitly include the identifier of the data they describe
 - F4: (Meta)data are registered or indexed in a searchable resource
- A: Accessible
 - A1: (Meta)data are retrievable by their identifier using a standardised communication protocol
 - A1.1: The protocol is open, free and universally implementable
 - A1.2: The protocol allows for an authentication and authorisation where necessary
 - A2: Metadata should be accessible even when the data is no longer available
- I: Interoperable
 - I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
 - I2: (Meta)data use vocabularies that follow the FAIR principles
 - I3: (Meta)data include qualified references to other (meta)data
- R: Reusable
 - 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



Existing FAIR Assessment Strategies:

Questionary-based (manual):

- Answering a checklist or list of single-selection questions
- Straight-forward process but no answer validation
- Assessment result can be biased according to answers

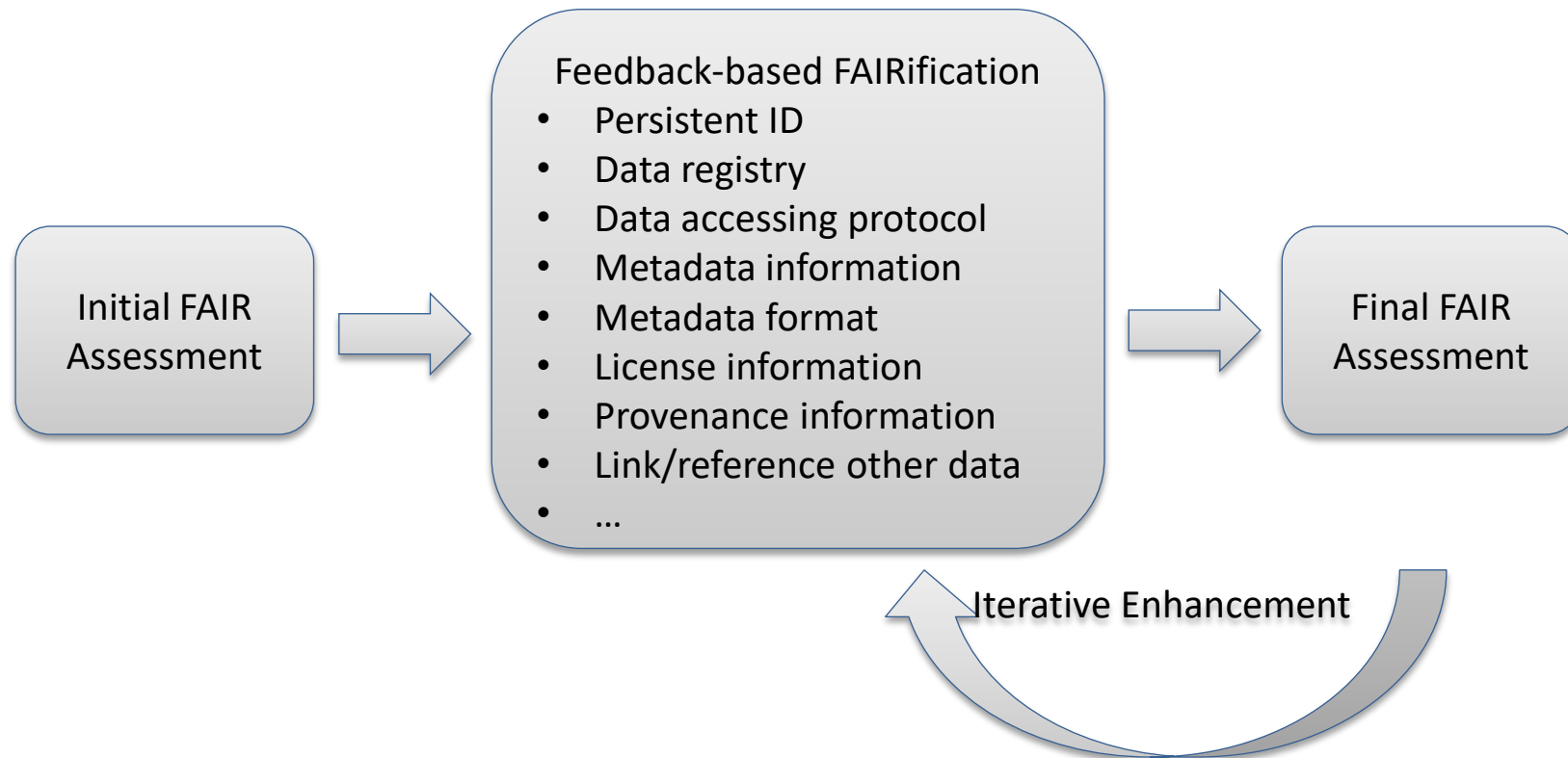
Automatic/semi-automatic:

- Assessment requires GUID input
- Evaluated with predefined metrics and associated tests (executed automatically); evaluation feedback and recommendation provided for improvement.
- Capability limited by software support and metadata providers

Criteria	Description	Manual	Automated
Metadata	Metadata information that can be evaluated	Flexible	Fixed
Productivity	Human effort required for evaluation	Low	High
Completeness	FAIR Guiding Principles coverage	High	Low
Granularity	Data granularity that can be evaluated	Yes	No



Proposed FAIRness Improvement Process



Proposing Hybrid FAIRness Assessment



- Automatic assessment is preferred but inadequate for full coverage
 - A2 (Metadata are accessible, even when the data are no longer available): Automatic approach might not be able to check due to the need to make data no longer available
 - I2: (Meta)data use vocabularies that follow FAIR principles: Automatic approach might not be feasible to check all the vocabularies
- Automatic assessment assesses at a coarse granularity:
 - Checking metadata for the whole dataset, and might not check the details such as headers of CSV data, names used in JSON.
- Propose a hybrid approach involving both manual and automated assessments
 - Automatic assessment: FAIRness assessment service by F-UJI
 - Manual assessment: The self-assessment tool by Research Data Alliance (RDA) FAIR data maturity model
- Design a scoring system to combined results from both assessments
 - A total of 47 points
 - A point is given if its result is determined by the RDA maturity indicator as fully implemented, or by the F-UJI metric as a fully passed test

Example actions for commonly seen FAIRness inadequacies

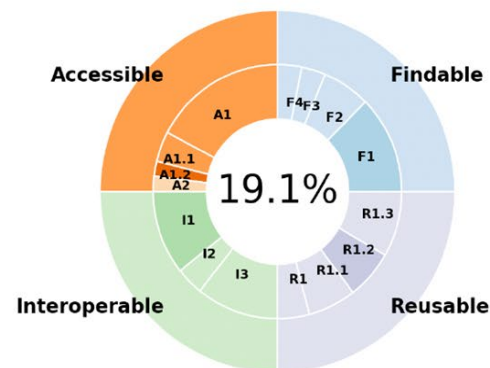
- Getting persistent identifier
- Providing coarse-grain metadata information
- Generating rich attributes for different granularity of data
- Automatic annotating data elements
- Provenance information
- License information



Case Study for Improving Data FAIRness

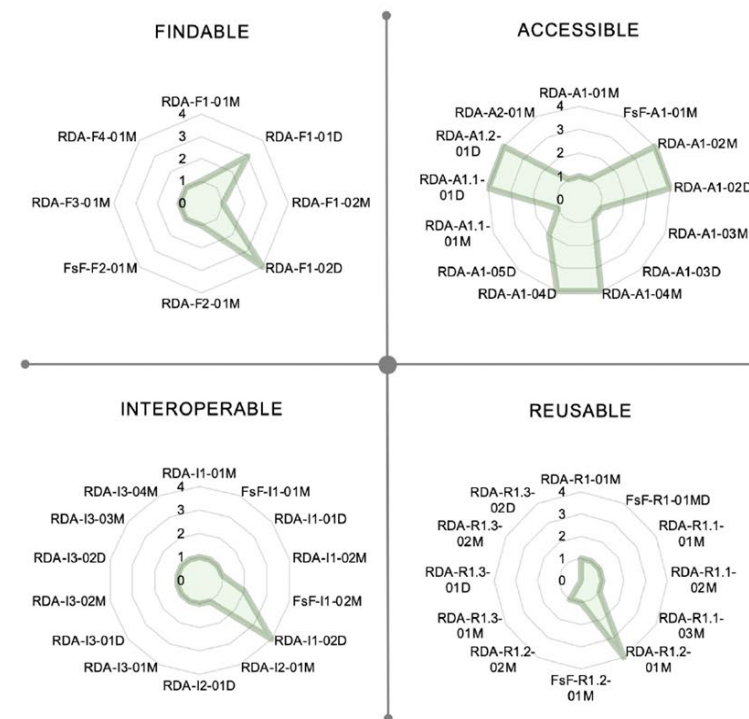
Example Dataset: XPlacer profiling data with labels-used to train machine learning models predicting best NVIDIA GPU memory APIs for arrays.

Initial Assessment:
Raw CSV data hosted at Github



	Score	%
Findable:	1 of 8	12.5%
Accessible:	6 of 13	46.2%
Interoperable:	1 of 14	7.1%
Reusable:	1 of 12	8.3%
Total:	9 of 47	19.1%

(a) Initial FAIRness score



(b) Initial FAIRness maturity levels

Enhancement applied for FAIRification

- Register DOI at Zenodo.org
- Obtained metadata support from Zenodo.org
- Revising metadata information
- Updating data provenance information for Zenodo.org
- Applying Creative Commons 4.0 license
- Applying HPC ontology for fine-grain metadata support
- Exploiting Tarql to automatically convert the corresponding CSV file into JSON-LD

```
{
  "@id": "http://example.org/test.csv#L1",
  "@type": "hpc:TableRow",
  "hpc:codeVariant": "111100",
  "hpc:allocatedDataSize": 8000000,
  "hpc:arrayID": "0",
  "hpc:commandLineOption": "graph1MW.6",
  "hpc:gpuPageFault": 5,
  "hpc:hostToDeviceTransferSize": {
    "@id": "_:Nbdd222a0d12a483d8f1a4cef274f18fc"
  }
},
{
  "@id": "_:Nbdd222a0d12a483d8f1a4cef274f18fc",
  "@type": "http://qudt.org/schema/qudt/QuantityValue",
  "http://qudt.org/schema/qudt/unit": {
    "@id": "http://qudt.org/vocab/unit/KiloBYTE"
  },
  "http://qudt.org/schema/qudt/value": {
    "@type": "http://www.w3.org/2001/XMLSchema#decimal",
    "@value": "7872.0"
  }
}
```

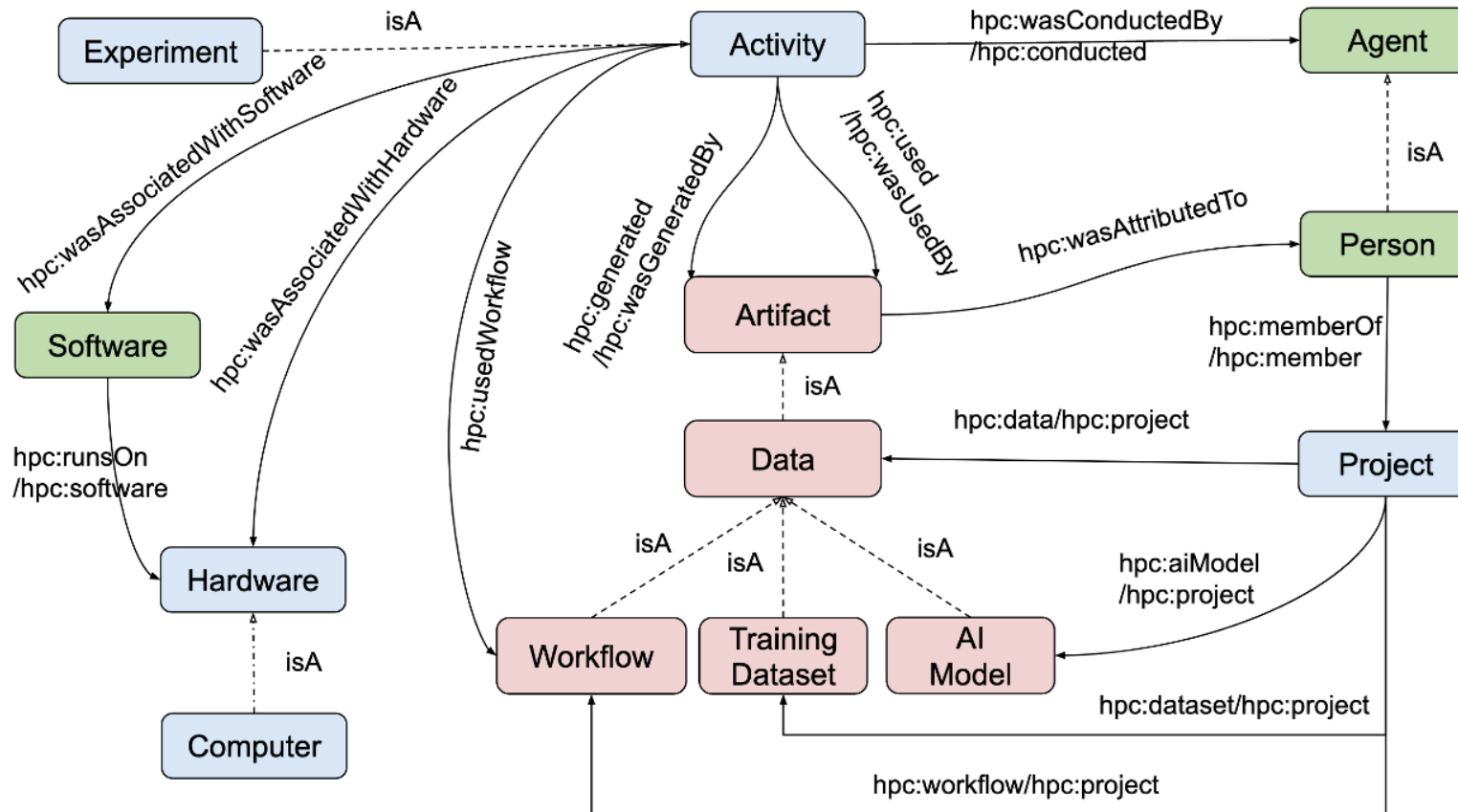


HPC Ontology: First Dedicated Ontology for HPC



- Domain: training datasets and AI models used for HPC software analyses and optimizations
- Design principles: a two-level modular design
 - Manually defined to have a curated class hierarchy and leverage domain knowledge
 - Incremental and use-case driven process
- High-level core ontology: describing entire datasets, AI models, and related software, hardware, administrative information
 - Keeping core concepts into a single namespace (the hpc: prefix)
 - Link to mainstream ontologies: e.g. Dublin Core metadata and Schema.org
- Low-level supplemental components: representing fine-grain, internal information of various subdomains:
 - Program constructs, hardware features, performance metrics, etc.
 - QUDT (Quantities, Units, Dimensions and Types)

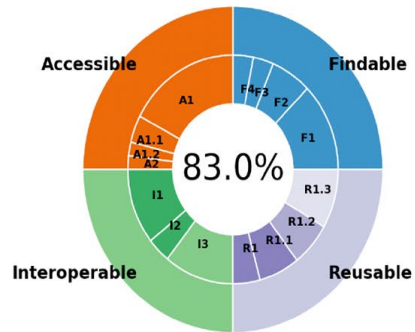
High-Level Core Ontology



Liao et al. HPC Ontology: Towards a Unified Ontology for Managing Training Datasets and AI Models for High-Performance Computing, 2021 IEEE/ACM Workshop on Machine Learning in High Performance Computing Environments (MLHPC)

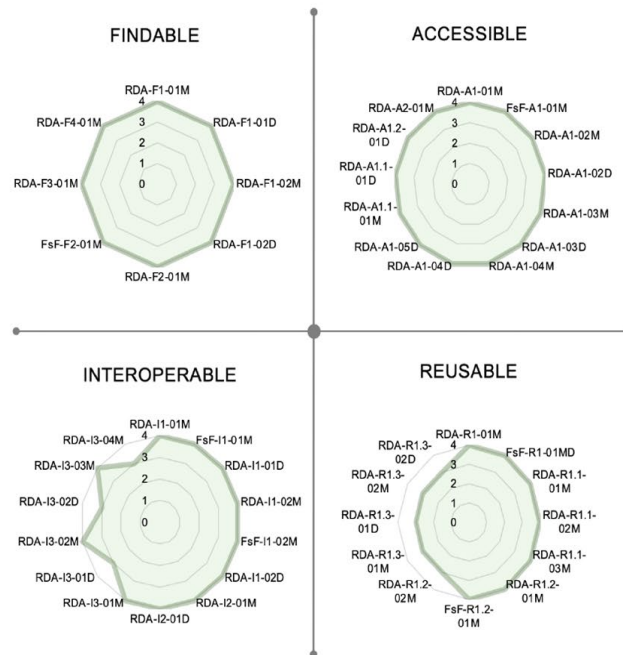


Final Assessment



	Score	%
Findable:	8 of 8	100%
Accessible:	13 of 13	100%
Interoperable:	11 of 14	78.6%
Reusable:	7 of 12	58.3%
Total:	39 of 47	83.0%

(a) Final FAIRness score



(b) Final FAIRness maturity levels

List of unfulfilled metrics:

- Full and qualified references to other data standards
 - RDA-I3-01D (Data includes references to other data)
 - RDA-I3-02D (Data includes qualified references to other data)
 - RDA-I3-04M (Metadata include qualified references to other data)
- PROV-O for cross-community provenance info:
 - RDA-R1.2- 02M (Metadata includes provenance information according to a cross-community languages)
- Missing community standards:
 - RDA-R1.3-01M
 - RDA-R1.3-01D
 - RDA-R1.3-02M
 - RDA-R1.3-02D



Conclusion & Acknowledgement

- Propose a concrete methodology to FAIRify HPC datasets and AI models
- Design a hybrid FAIRness assessment to have full coverage for FAIR principles
- Demonstrate the FAIRness improvement with existing HPC dataset
- Prepared by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-826494). This work is also supported by the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Program under Award Number DE-SC0021293.
- Project website: <http://hpcfair.org/>

