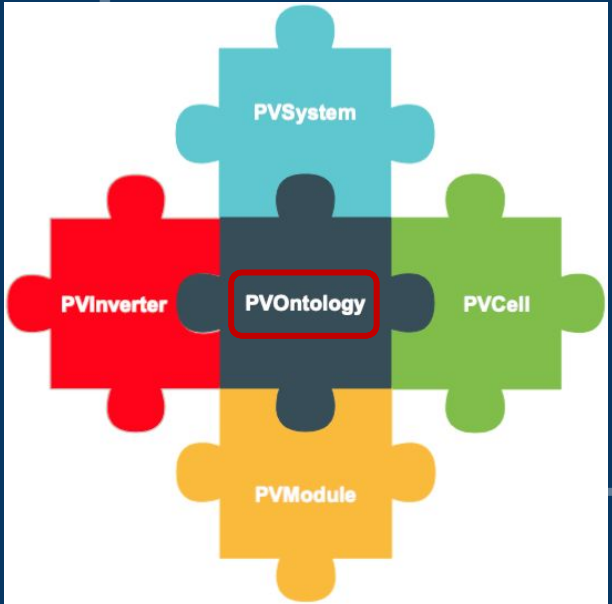
FAIRification And Materials Domain Ontologies: Enabling Open Data & Models for Improved Knowledge Management and Learning



Roger H. French[1,2,3,4,5], Mingjian Lu[2,3,4], Liangyi Huang[2,3,4], Will C. Oltjen[2,4,5], Xuanji Yu[1,4,5], Arafath Nihar[2,3,4], Thomas Ciardi[2,3,4], Erika Barcelos[1,4,5], Pawan Tripathi[1,4,5], Abhishek Daundkar[1,4,5], Deepa Bhuvanagiri[2,3,4], Hope Omodolor[1,4,5], Olatunde Akanbi[1,4,5], Hein Htet Aung[1,4,5], Kristen J. Hernandez[1,4,5], Mirra M. Rasmussen[1,4,5], Raymond J. Wieser[1,4,5], Sameera N. Venkat[1,4,5], Tian Wang[2,3,4], Weiqi Yue[2,3,4], Yangxin Fan[2,3,4], Rounak Chawla[2,3,4], Leean Jo[1,4,5], Zelin Li[1,4,5], Jiqi Liu[1,4,5], Justin P. Glynn[1,4,5], Kehley A. Coleman[1,4,5], Jeffrey M. Yarus[1,4,5], Mengjie Li[1,6], Kristopher O. Davis[1,6], Laura S. Bruckman[1,4,5], Yinghui Wu[2,3,4]

1. Department of Materials Science and Engineering, 2. Department of Macromolecular Science and Engineering, 3. Department of Computer and Data Sciences, 4. SDLE Research Center, 5. Case Western Reserve University, Cleveland OH 44106, 6. University of Central Florida, Orlando Florida, 32816

Materials Data Science for Stockpile Stewardship (MDS^3) Center of Excellence
SDLE Research Center
Materials Science & Eng. Dept., Computer & Data Sciences Dept.
Case Western Reserve University, Cleveland OH 44106 USA

A selection of Lifetime Extension Articles

roger.french@case.edu
http://dmseg5.case.edu/people/faculty.php?id=rxf131

FAIRification Using Materials Domain Ontologies: Enabling Open Data & Models for Improved Knowledge Management and Learning

Roger H. French[1,2,3,4,5], Mingjian Lu[2,3,4], Liangyi Huang[2,3,4], Will Oltjen[2,4,5], Xuanji Yu[1,4,5], Arafath Nihar[2,3,4], Thomas Ciardi[2,3,4], Erika Barcelos[1,4,5], Pawan Tripathi[1,4,5], Abhishek Daundkar[1,4,5], Deepa Bhuvanagiri[2,3,4], Hope Omodolor[1,4,5], Olatunde Akanbi[1,4,5], Hein Htet Aung[1,4,5], Kristen Hernandez[1,4,5], Mirra Rasmussen[1,4,5], Raymond Wieser[1,4,5], Sameera Nalin Venkat[1,4,5], Tian Wang[2,3,4], Weiqi Yue[2,3,4], Yangxin Fan[2,3,4], Rounak Chawla[2,3,4], Leean Jo[1,4,5], Zelin Li[1,4,5], Jiqi Liu[1,4,5], Justin Glynn[1,4,5], Kehley Coleman[1,4,5], Jeffrey Yarus[1,4,5], Mengjie Li[1,6], Kristopher Davis[1,6], Laura Bruckman[1,4,5], Yinghui Wu[2,3,4]

1. Department of Materials Science and Engineering
2. Department of Macromolecular Science and Engineering
3. Department of Computer and Data Sciences
4. SDLE Research Center
5. Case Western Reserve University, Cleveland OH 44106
6. University of Central Florida, Orlando Florida, 32816

The 2016 paper by Wilkinson[1] described making data Findable, Accessible, Interoperable, and Reusable as is needed to achieve W3C's goals for WEB 3.0, the Semantic Web: make internet data machine readable. One can incorporate metadata, relationships between entities and categories of things, and enable automated reasoning over data (and models) across the web by encoding semantics using web ontology language (OWL) and Resource Data Framework (RDF). The value of achieving this goal has been globally recognized, and medical research has used ontologies to define terms, relationships and variable names for a long time. The challenge for materials science and energy sciences is that our communities (or scientific domains) are new to this semantic, conceptual approach.

To FAIRify scientific domains, such as materials science, requires two major steps; 1) an agreed upon standardized schema for naming data and metadata variables, and 2) an accepted form for sharing these variables and their values in a human and machine readable form. Implementation of FAIR principles has been a serious challenge because there are no established domain ontologies for Materials Research, and there are inconsistencies between the older XML formats vs. the newer JSON-LD for linked data.

FAIRification in our fields requires addressing these issues, and has advanced with the new ISO standards for Basic Formal Ontology and JSON. In our research to date,, we have initiated an open-source collaborative effort developing 18 materials domain ontologies and JSON-LD templates for use with the FAIRmaterials packages,. FAIRmaterials is our first "bi-lingual" code package for both Python (on PyPI) and R (on CRAN), it contains JSON-LD templates with defined schema for key:value pair naming and notation.

1. M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Sci Data, vol. 3, no. 1, Art. no. 1, Mar. 2016.
2. P. Hitzler, "A Review of the Semantic Web Field," Commun. ACM, vol. 64, no. 2, pp. 76–83, Jan. 2021.
3. W3C, "OWL - Semantic Web Standards," World Wide Web Consortium (W3C), Dec. 2012.
4. W3C, "RDF 1.1 Primer," World Wide Web Consortium (W3C), Jun. 2014.
5. Manu Sporny, et al., "JSON-LD 1.1: A JSON-based Serialization for Linked Data," Jul. 2020.
6. ISO/IEC, "21838-2 Ed. 1: Top-level ontologies - Part 2: Basic Formal Ontology," ISO/IEC, Geneva, Nov. 2021.
7. ISO/IEC, "21778 Ed. 1 : The JSON data interchange syntax," ISO/IEC, Geneva, Nov. 2017.
8. A. Nihar, et al., "Toward Findable, Accessible, Interoperable & Reusable PV System Time Series Data," 48th IEEE PVSC, 2021, pp. 1701–1706.
9. W. C. Oltjen, et al., "FAIRification, Quality Assessment, and Missingness Pattern Discovery for Spatiotemporal Photovoltaic Data," 49th IEEE PVSC, 2022, pp. 0796–0801.
10. M. Lu, et al., "FAIRmaterials: Generate JSON-LD files based on FAIRification standard." Python, v0.2, Jan-2023.
11. W C. Oltjen, et al., "FAIRmaterials: Make Materials Data FAIR," R, v0.2, January, 2023.

FAIRmaterials: Funding Acknowledgements

A diverse FAIRification effort across a number of our projects and centers

- **DOE-EERE-SETO Project: PV-stGNN Project:** This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number **DE-EE0009353**.
- **DOE-EERE-SETO Project: Multi-scale Project:** This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number **DE-EE0009347**.
- **ARPA-E ESTCP Project:** This work was supported by the U.S. Department of Energy's Advanced Research Projects Agency–Energy (ARPA-E) award **DE-AR0001250**.
- **DOE-NNSA Project: Materials Data Science for Stockpile Stewardship Center of Excellence (MDS3-COE):** This material is based upon research in the Materials Data Science for Stockpile Stewardship Center of Excellence (MDS3-COE), and supported by the Department of Energy's National Nuclear Security Administration under Award Number **DE-NA0004104**.
- **NSF Engineering Research Center for Advancing Sustainable and Distributed Fertilizer Production (CASFER) Phase I:** This material is based upon work supported by the U.S. National Science Foundation, Div Of Engineering Education and Centers, under Award Number **2133576**

SDLE Research Center: Acknowledgements

CWRU Faculty
- Roger French, Laura Bruckman, Jeffrey Yarus, Yinghui Wu, Alp Sehirlioglu, Ina Martin

Post-doctoral Research Associates
- Pawan Tripathi, Xuanji Yu, Erika Barcelos

Graduate Students
- PhD: Arafath Nihar, Deepa Bhuvanagiri, Emmanuel Olatunde, Haiping Ai, Hein Htet Aung, Hope Omodolor, Kristen Hernandez, Mingjian Lu, Raymond Wieser, Redad Mehdi, Sameera Nalin Venkat, Thomas Ciardi, Tian Wang, Weiqi Yue, Yangxin Fan
- MS: Dominique Yao, Leean Jo, David Meshnick, Mirra Rasmussen, Olatunde Akanbi, Rounak Chawla, Scott Maurer, Shanthan Reddy, Sophia Hall, Will Oltjen, Zelin Li,

Undergraduates
- Adam Goodman, Ethan Fang, Gabriel Ponon, Jose Diego Zubieta Sempertegui, Jonathan Libenson, Jube Augustino, Kemal Ozdemirli, Lam Nguyen, Mason Jester, Neha Panduri, Noah Butler, Peter Schlueter, Ravi Lin, Ray Le, Summer Sun, Vibha Mandayam, Zhuldyz Ualikhankyzy,

SDLE Staff: Jonathan Steirer, Tariq Shabazz, Rich Tomazin

A BHEF CASE STUDY

CREATING A MINOR IN

APPLIED DATA SCIENCE

Case Western Reserve University Engages Business Leaders to Produce T-Shaped Professionals

THROUGH THE COLLABORATION of its business and higher education members, the Business-Higher Education Forum (BHEF) launched the National Higher Education and Workforce Initiative (HEWI) to create new undergraduate pathways in high-skill, high-demand fields such as data science and analytics. Data science and analytics must be integrated with T-shaped skills, such as critical thinking, collaboration, and effective communication, which are critical for all graduates entering the 21st century workforce. Knowledge of data science and analytics in recent years has become as fundamental as any other skill for graduates' career readiness. BHEF's Strategic Business Engagement Model with higher education addresses this demand by moving the two sectors from transactional relationships to strategic partnerships through five strategies:

1. **ENGAGE** corporate leadership;
2. **FOCUS** corporate philanthropy on undergraduate education;
3. **IDENTIFY** and tap core competencies and expertise;

undergraduate education.

This case study examines how BHEF member Case Western Reserve University (Case Western Reserve) is integrating T-shaped skills into a minor in applied data science.

PROGRAM OVERVIEW

THE APPLIED DATA SCIENCE (ADS) MINOR AT CASE WESTERN RESERVE serves as a national model for undergraduate education in data science. Available to every undergraduate student across all schools at the university, this program of study requires experiential learning opportunities, embeds T-shaped skills, and allows students to master fundamental ADS concepts in their chosen domain area. From strong leadership engagement to funded undergraduate research opportunities, Case Western Reserve applied BHEF's Strategic Business Engagement Model to create a minor that responds to the fundamental need for data science in today's global business community.

AY 2014-15	AY 2015-16	AY 2016-17	AY 2017-18	AY 2018-19	AY 2019-20	AY 2020-21	AY 2021-22	AY 2022-23	Total
9	36	49	57	100	106	92	159	220	828

Medical Mutual of Ohio
Medtronic
Philips Healthcare
Sherwin-Williams Company
Siemens
Teradata Corporation
Timken Company
University Hospitals

5

http://www.bhef.com/publications/creating-minor-applied-data-science

BHEF BUSINESS HIGHER EDUCATION FORUM
Creating Solutions. Inspiring Action.

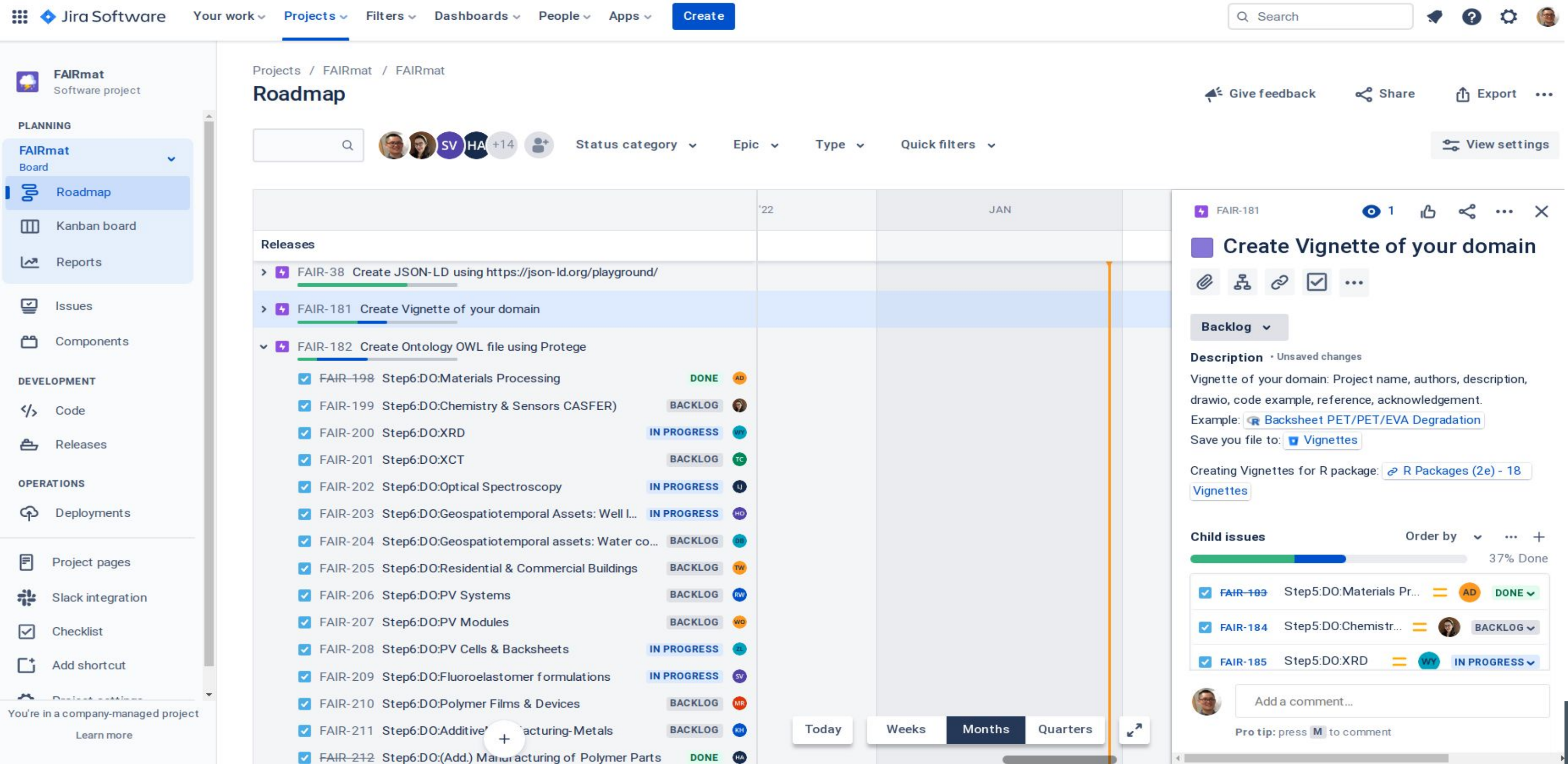
lle.case.edu

MDS³ COE

Agile Software Development: Jira to coordinate 32 people's work

FAIRmaterails: v0.0.2 in Sept. 2021.

- Sprints from Nov. 22 to Jan. 23: v0.1 on 12/10/22, v0.2 on 1/10/23



CASE WESTERN RESERVE UNIVERSITY

Common Research Analytics & Data Lifecycle Environment

CRADLE Analytics & Compute Cluster

Distributed & High Performance Computing[1]

FAIRification of Datasets

1. A. Khalilnejad, et al., Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data, PLOS ONE. 15 (2020) e0240461. https://doi.org/10.1371/journal.pone.0240461.

CRADLE Compute Environment

Running in CWRU HPC
- **Rider (RHEL7)**
- **Pioneer (RHEL8 OS)**
- **Markov (Teaching Cluster)**

OnDemand Containerized Apps
- Using Ubuntu 20.04 OS

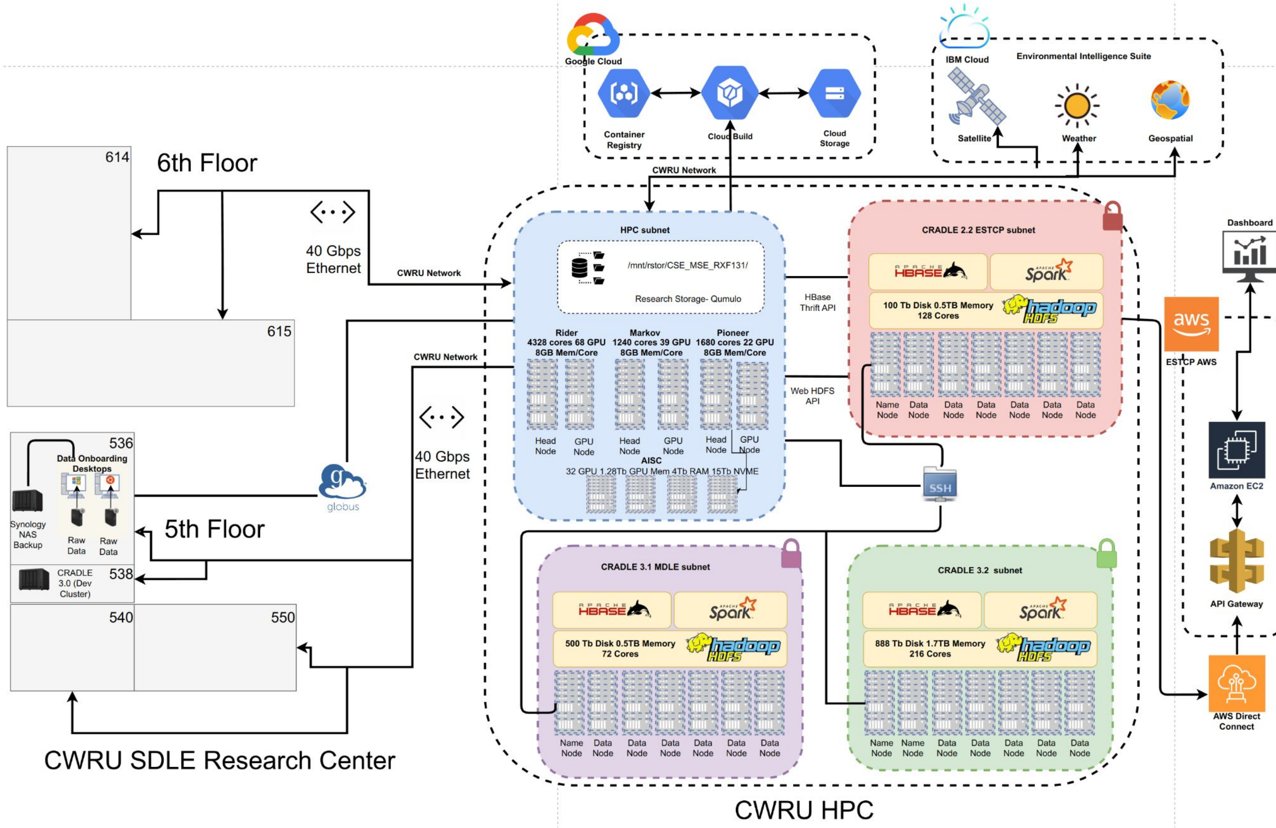
Cloudera Data Platform
- Comm. supported distribution
- Of Apache Hadoop/Hbase/Spark/….

2 Petabytes of Distrib. Comp. Storage
- **Meeting NIST 800-171 level 1**
- **For Controlled Unclassified Data**

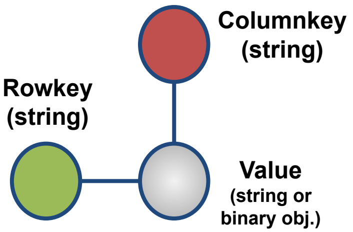
Nvidia AISC
- 32 A100 GPU (40 Gb vRAM each)
  - = 1.28 Tb of GPU vRAM
- With 15 Tb of NVME Fast Storage

Able to train 100s of Deep Learning Models

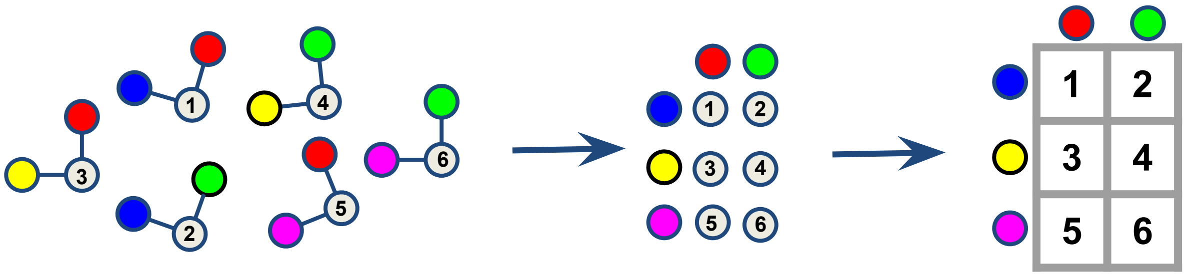
Common Research Analytics & Data Processing Environment



The "NoSQL" Database Abstraction of Hadoop/Hbase: RDF Triples



HBase RDF 'Triple'



Combines Lab data (Spectra, Images, Videos etc.)

With Geospatiotemporal Data (PV Power Plant Data)

Distributed & High Performance Computing:

Petabyte Data Lake In A Petaflop HPC Environment

- In-place Analytics: Distributed Spark Analytics in Hadoop/HDFS/Hbase
- In-memory Data Extraction: To Separate HPC Compute Nodes

A non-relational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites

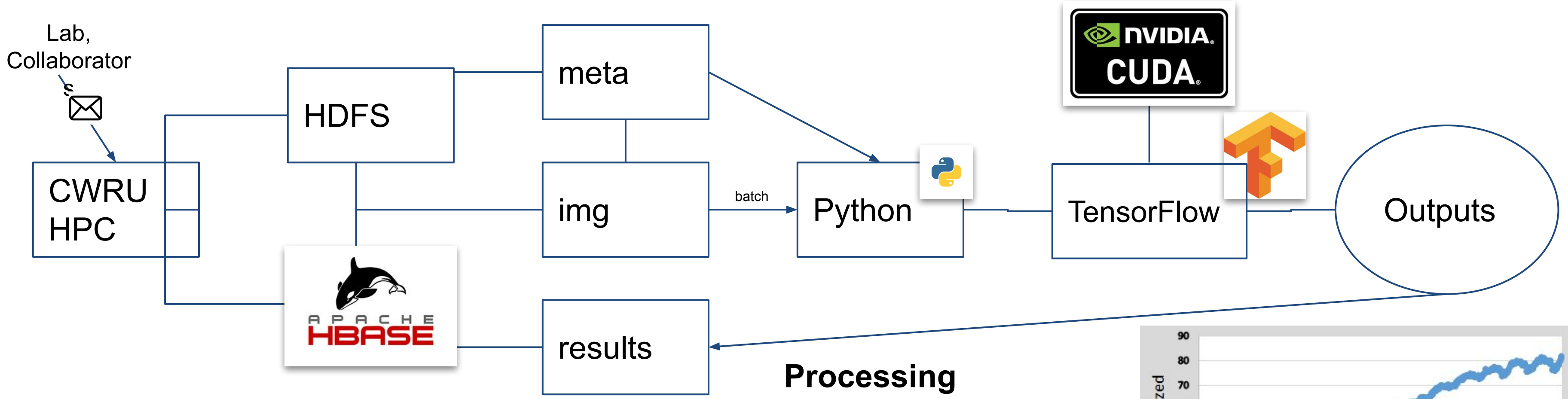
Yang Hu, *Member, IEEE*, Venkat Yashwanth Gunapati, Pei Zhao, Devin Gordon, Nicholas R. Wheeler, Mohammad A. Hossain, *Member, IEEE*, Timothy J. Peshek, *Member, IEEE*, Laura S. Bruckman, Guo-Qiang Zhang, *Member, IEEE*, and Roger H. French, *Member, IEEE*

Automated pipeline framework for processing of large-scale building energy time series data

Arash Khalilnejad[1,5], Ahmad M. Karimi[2,5], Shreyas Kamath[1,5], Rojiar Haddadian[2,5], Roger H. French[2,4,5]*, Alexis R. Abramson[3,6¤]

Hu, Y., et al., "A Nonrelational Data Warehouse for the Analysis of Field & Lab Data From Multiple Heterogeneous Photovoltaic Test Sites." IEEE JPV, 7, 1, 2017, 230–36.
A. Khalilnejad, et al., Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data, PLOS ONE. 15 (2020) e0240461.

Data Processing Infrastructure: A Data Analysis Pipeline (Python or R)

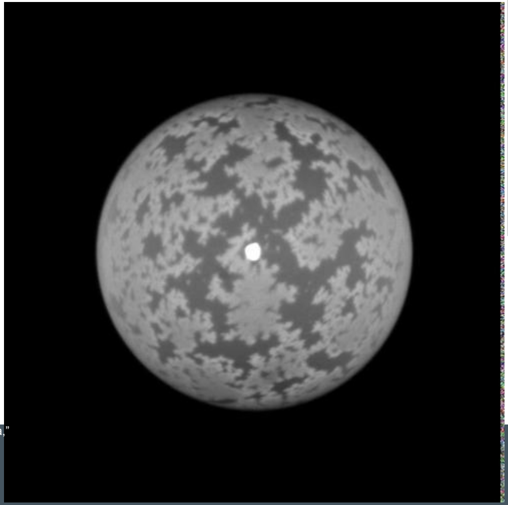


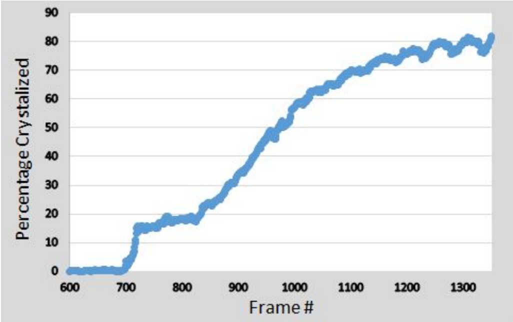
Processing

Storage **Tables**

Nucleation & Growth of AlN Crystals

- From Al/Ni Melt
- 1 Million Images





CRADLE infrastructure

NoSQL database

- Apache HBase

HPC environment

- Nvidia GPU acceleration for deep learning

Python/TensorFlow language

M. Adachi, S. Hamaya, D. Morikawa, B. G. Pierce, A. M. Karimi, Y. Yamagata, K. Tsuda, R. H. French, H Fukuyama, Temperature dependence of crystal growth behavior of AlN on Ni–Al using electromagnetic levitation and computer vision technique", Mat. Sci. in Semicon. Proc., 153, 2023, 107167, ISSN 1369-8001, https://doi.org/10.1016/j.mssp.2022.107167
.

[1] A. Khalilnejad, et al., "Automated Pipeline Framework for Processing of Large-Scale Building Energy Time Series Data," PLOS ONE, p. e0240461, Dec. 2020, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240461 .
[2] Masayoshi Adachi et al., "In-situ observation of AlN formation from Ni-Al solution using an electromagnetic levitation technique," J Am Ceram Soc, p. jace.16960, Jan. 2020, https://onlinelibrary.wiley.com/doi/abs/10.1111/jace.16960 .

Stream Processing Example: Tesla, ML for “Autonomous Driving”

Automated Data Analysis Pipelines

- Enable Terabyte Dataset Analysis
  - In-situ Manu. Datastreams (7 Tb)
  - Beamline XRD (12 Tb)

Write-back All Models & Results

- Future Analysis Builds On Priors
- Datasets & AI/ML Models Get Smarter

Minimize Large Data Transfer

- Prefer In-place Analytics (Hadoop/Spark)

Focus on Fast/Efficient Modeling

- Such as YOLO CNN
  - for Autonomous Driving

Knowledge Graphs

Spatiotemporal Graphs

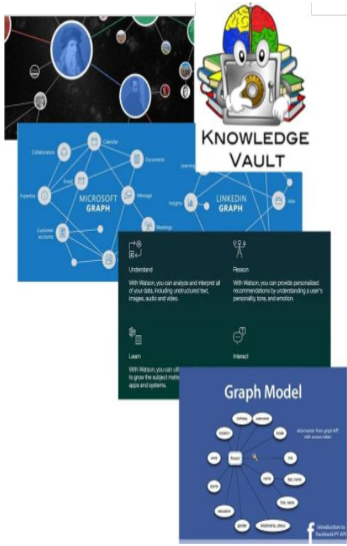
And Their Role in

Deep Representation Learning

Knowledge graphs: Nodes & Edges Define Relationships

What is a knowledge Graph?

- **Nodes:** entities w. types & attributes;
- **Edges:** relations
- capture (factual) knowledge as graphs



Where do KGs come from?

- Structured data: sensors, tables, Wiki infoboxes,
  databases, social nets, …
- Unstructured data: text, images, videos

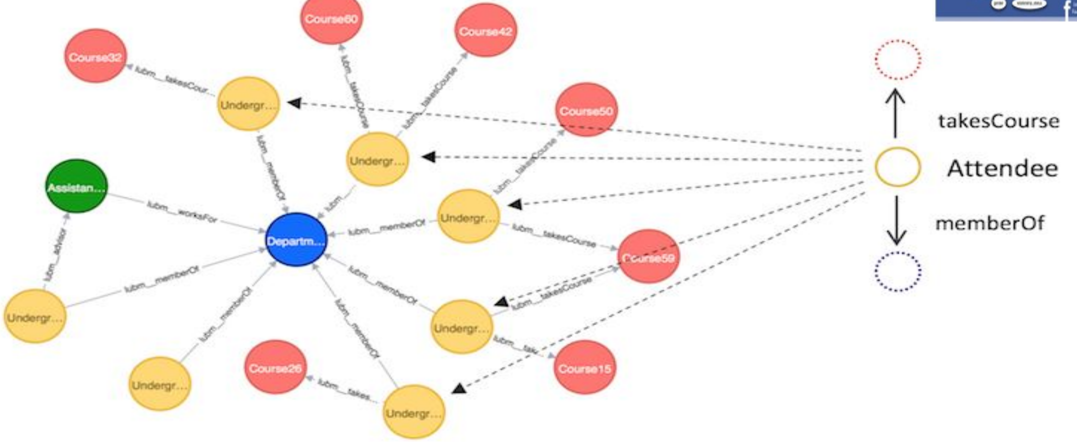
Why (Knowledge) Graphs?

Humans:

- Explore data via intuitive/processible structure
- Combat information overload
- Tool for supporting knowledge-driven tasks

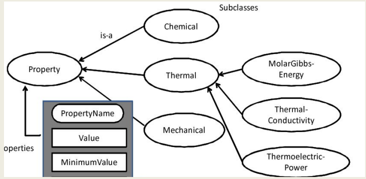
Also:

- Key ingredient for many AI tasks
- Bridge from data to human semantics
- Use decades of work on graph analysis



A Knowledge Graph Learning Framework

Metadata, Ontologies
RDF triples, JSON
(abstraction/semantics/constraints)



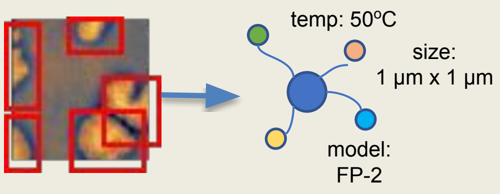
... to data standardization & knowledge sharing

1 *FAIRification*

validate ⟷ *curate*

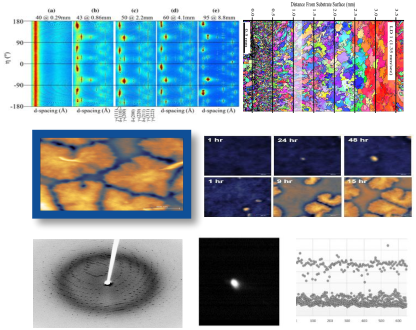
2 *Featurization*

Objects, Observations & Properties
(instances)



... to create AI/ML ready data resources

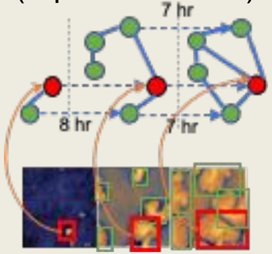
Raw datasets
Image/videos, design data, in-situ/ex-situ data..



ML/AI models · *physical con-straints*

linked entities · *enriched features*

Scenes or st-Graphs
(representations)



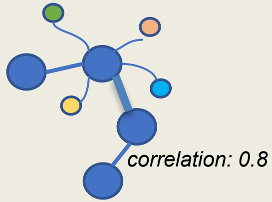
Deep st-graph representation learning
... to inferential & predictive models

4 *Deep Learning*

cost-effective learning ⟷ *efficient access & interpretation*

3 *Summarization*

Summary Graphs
(patterns)

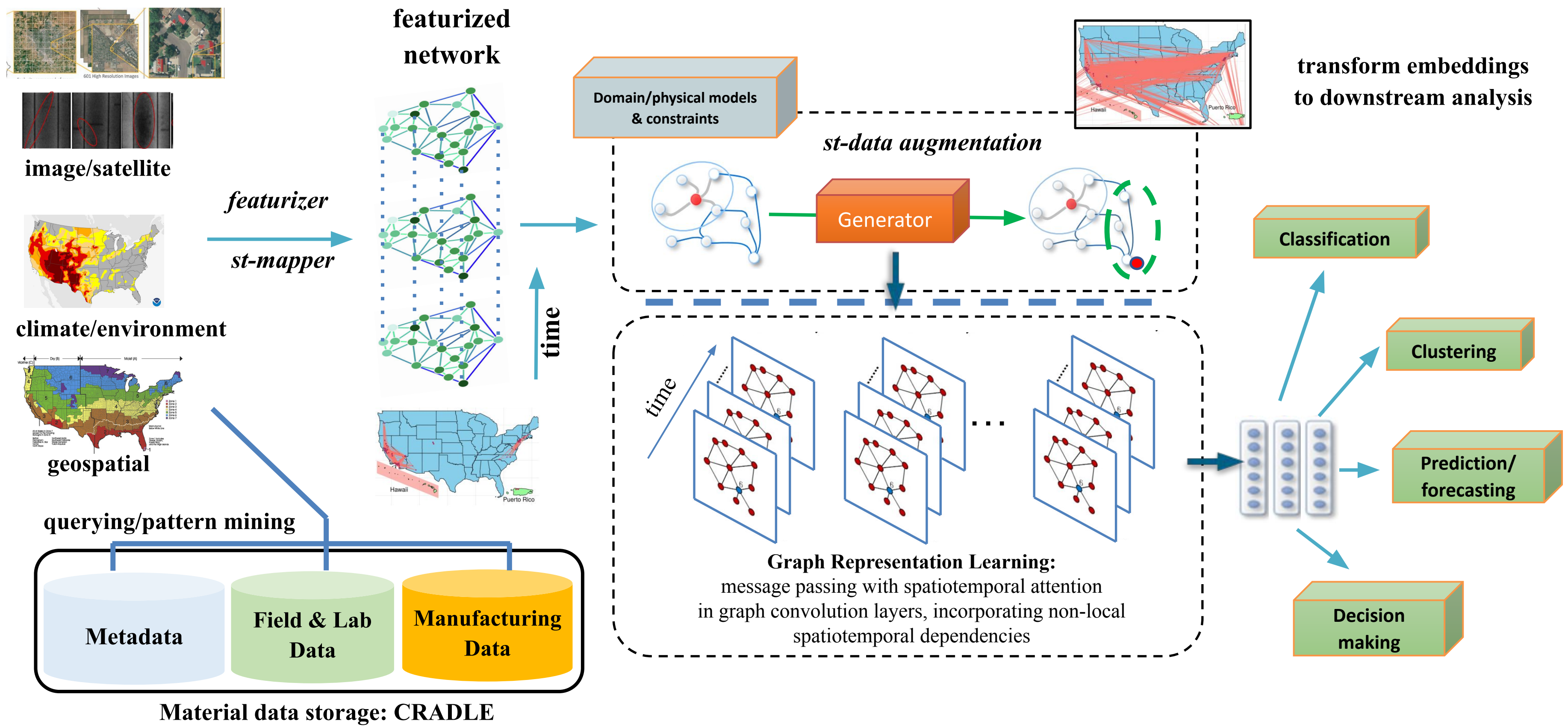


... to cost-effective data access, analysis & interactive exploration

[1] Roger H. French et al., "Fairmaterials." The Python Package Index (PyPI), Oct. 08, 2021, https://pypi.org/project/fairmaterials/ .
[2] Willam C. Oltjen, et al., "FAIRmaterials: Make Materials Data FAIR." CRAN, Sep. 14, 2021, https://CRAN.R-project.org/package=FAIRmaterials .
[3] A. M. Karimi, et al., "Spatiotemporal Graph Neural Network for Performance Prediction of Photovoltaic Power Systems," in Proc. of AAAI,, 2021, https://ojs.aaai.org/index.php/AAAI/article/view/17799 .

PVplr-stGNN: Degradation Science with Spatiotemporal-Graph Models

DOE-EERE-SETO AI for PV, PV-stGNN Project: DE-EE0009353, French, Wu, Bruckman, CWRU



A. M. Karimi, Y. Wu, M. Koyuturk, R. H. French, "Spatiotemporal Graph Neural Network for Performance Prediction of Photovoltaic Power Systems," in Proc. AAAI Conf. on Artificial Intelligence, May 2021, vol. 35, p. 8.

FAIRification:
Scientific Domain Ontologies And JSON-LDs
for the FAIR Datasets & Models
And the Semantic Web

FAIRification[1] of Datasets & Models, Enables AI learning

Making Datasets & Models FAIR

- By “FAIRification”

Enables Models to find Data

- And Data to find Models

So that they can advance

- Without human intervention: AI Learning

This is an aspect of the Semantic Web[2]

- And Resource Description Framework[3]
- CRADLE’s Hbase triples are an example of RDF

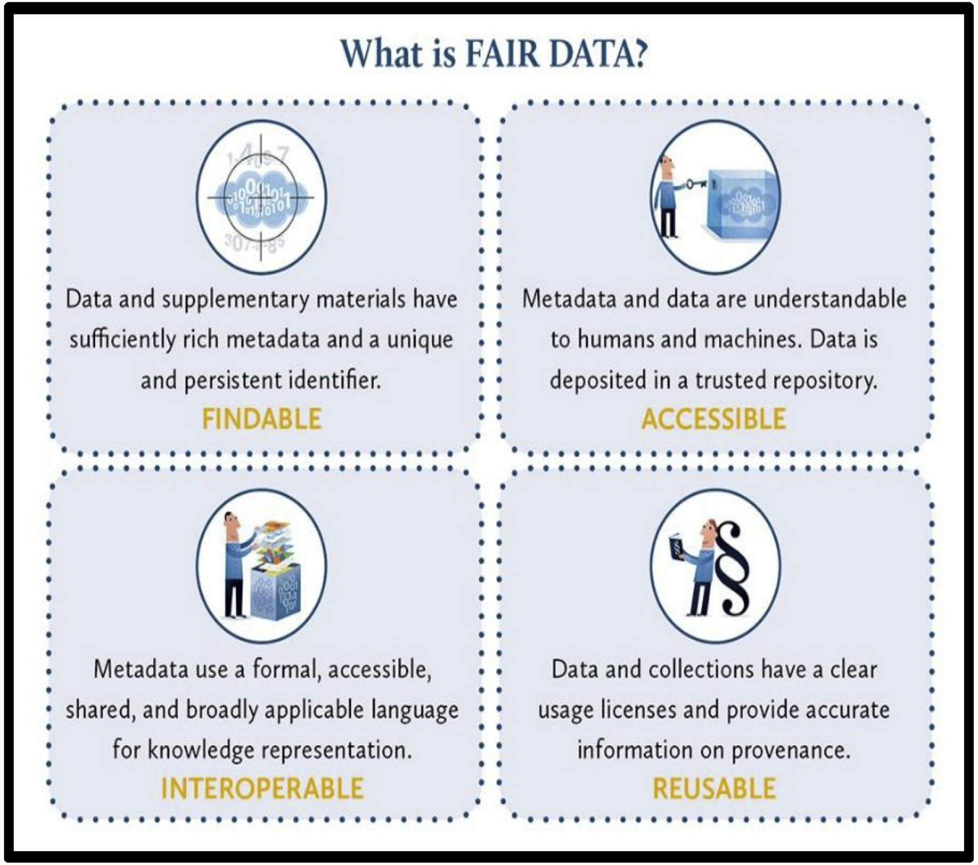
FAIRification essential for DOE SETO AI awards

- For PV st-GNN (9353-CWRU), we include FAIRification
- And PV Multiscale (9347-UCF)

And for other projects & centers

- PV System Owners in IEA-PVPS Task 13
- DOE-NNSA: MDS[3]-Center of Excellence
- CASFER NSF ERC

For the petabytes of data SDLE ingests!



1.M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Sci. Data, vol. 3, no. 1, pp. 1–9, Mar. 2016.
2. P. Hitzler, “A Review of the Semantic Web Field,” Commun. ACM, vol. 64, no. 2, pp. 76–83, Jan. 2021. 3. RDF - Semantic Web Standards

FAIRification: make datasets FAIR using JSON-LDs

Project data, metadata FAIRified to

- **"key-value" pairs for variables**
  - key is a standardized variable name
- **& triples: Column key, row key & value**
  - Columns are variables
  - Rows are observations

FAIRify by defining "key-value" pairs

- Each "key" is a defined word
- Defined either by schema.org of W3C.org
- Using JSON-LD
  - JSON for Linked Data
  - Previously XML was used

Ex.: Split metadata of PV Power Plants Into human readable "chunks"

- DOE RTC Baseline System:
  - *c10hov6.json*
- Inverter:
  - *c10hov6-inverter-metadata.json*
- Modules:
  - *c10hov6-module-metadata.json*
- Timeseries Power data:
  - *c10hov6.csv*

FAIRmaterails first published in 2021: but not Semantic Web Consistent

We used JSON-LD, and our local "keys"

- Not human readable!

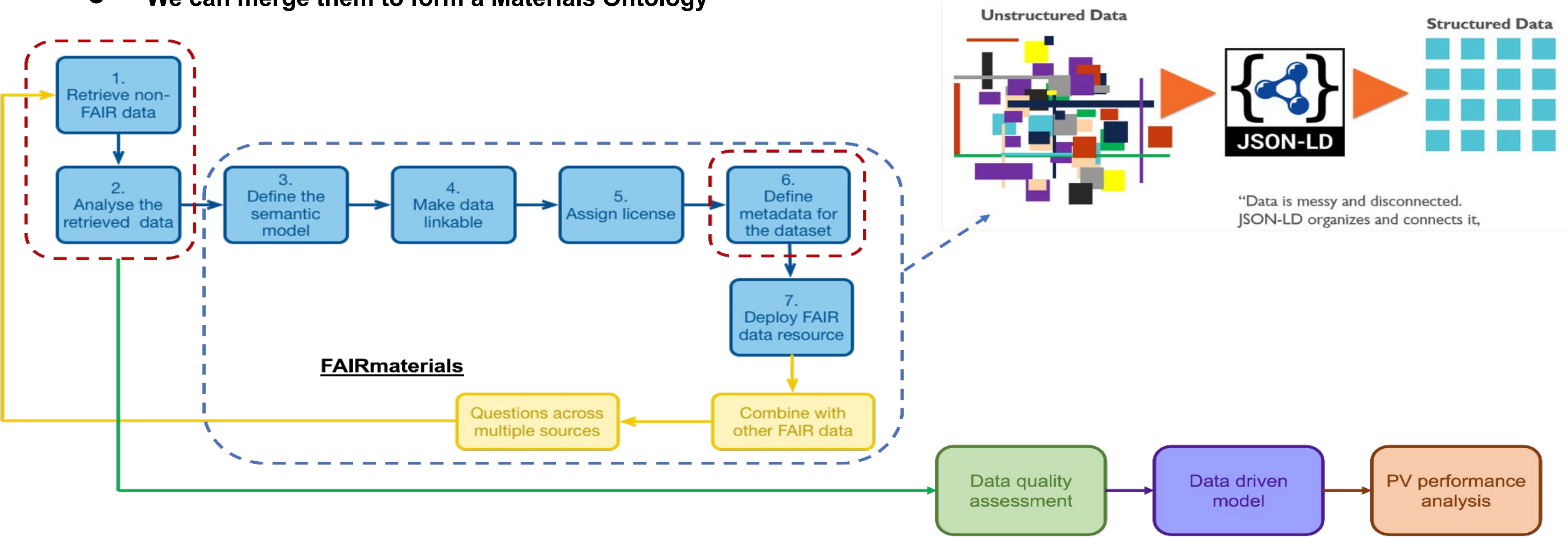
We needed to define our Semantic Model & Domains

So begin with 20 Scientific Domains & Their Defined Keys

- And Define their keys & their Ontology Graphs
- We can merge them to form a Materials Ontology

Develop an evolving Materials Ontology

- By definition requires an Open-Source Community
- Invite Community to submit new domains and keys
- Currently email to fairmaterials@gmail.com



How: Agile Software Development: FAIRmaterails for R & Python

Agile Software Engineering

- Define basic components
  - Inputs
  - Outputs
  - Functions etc.
- Assign components to people
- Unit Tests
- Git Code Management

Create User "Stories"

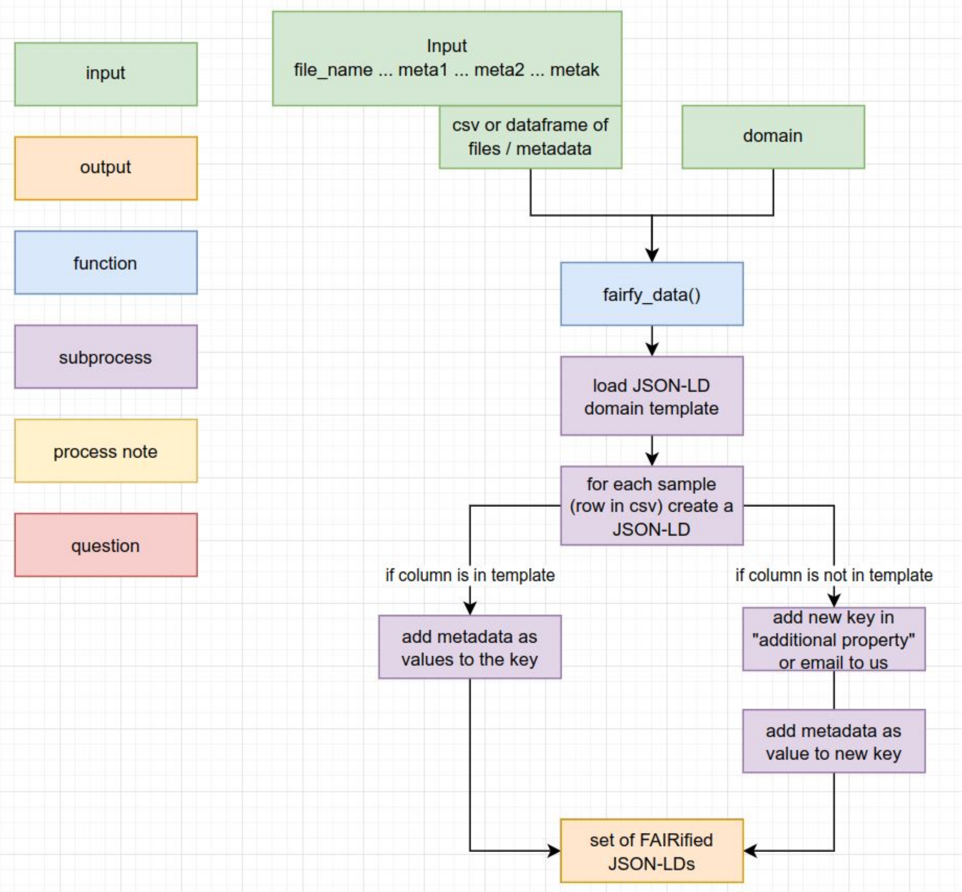
- To define the capabilities needed

Create "Epics"

- R package
- Python package
- 6 Steps to define Domain Ontologies

For each of the 20 Domains

- Diagram the structure of datasets
- This defines the required information



Appreciating the Semantic Web and Basic Formal Ontology: Schema.org

Major advances since 2016 for FAIR

- Orange Button PV Taxonomy, since 2016
- JSON ISO Standard: ISO/IEC 21778 Ed. 1: 2017

& 2020: JSON-LD v1.1 & Basic Formal Ontology

- **JSON-LD v1.1: Specification released 2020**

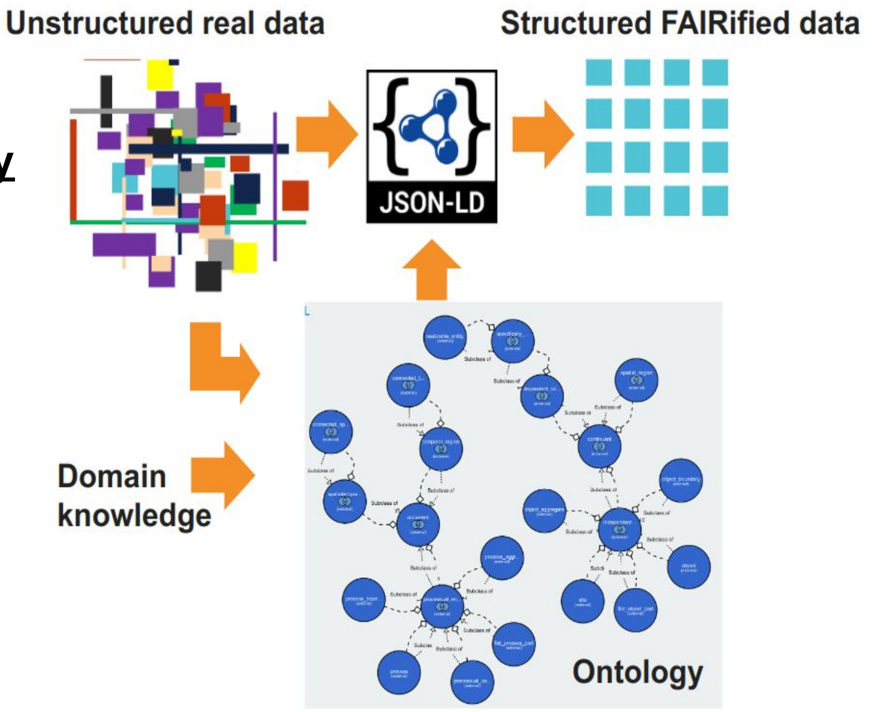
  A JSON-based Serialization for Linked Data

  a. Specifically for Linked Data
  b. Distinct from JSON-OpenAPI
     i. PV Taxonomy uses JSON-OpenAPI

- **Top-level ontologies (TLO) ISO/IEC 21838**

  Released in 2021 as ISO Standards

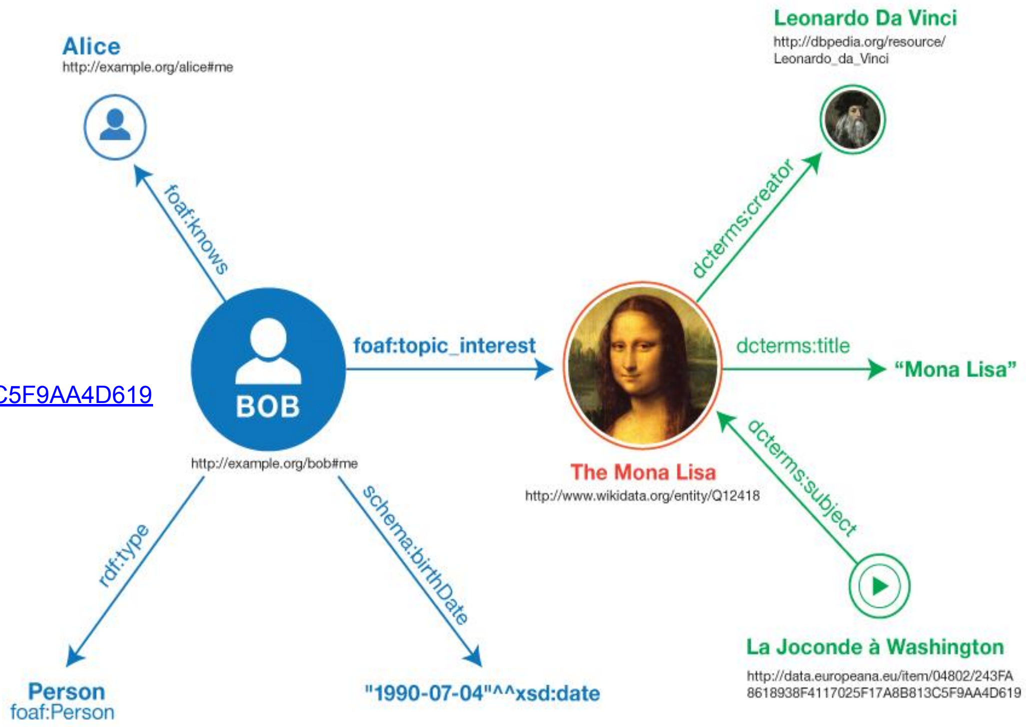
  a. Part 1 Requirements
     i. ISO/IEC 21838-1 Ed. 1: 2021
  b. Part 2: **Basic Formal Ontology (BFO)**
     i. ISO/IEC 21838-2 Ed. 1: 2021



Incorrect Example of JSON-LD representation of a linked data Graph

JSON-LD

```
{
 "@context": "example-context.json ",
 "@id": "http://example.org/bob#me",
 "@type": "Person",
 "birthdate": "1990-07-04",
 "knows": "http://example.org/alice#me",
 "interest": {
  "@id": "http://www.wikidata.org/entity/Q12418",
  "title": "Mona Lisa",
  "subject_of":
"http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619
",
  "creator": "http://dbpedia.org/resource/Leonardo_da_Vinci"
  }
 }
```

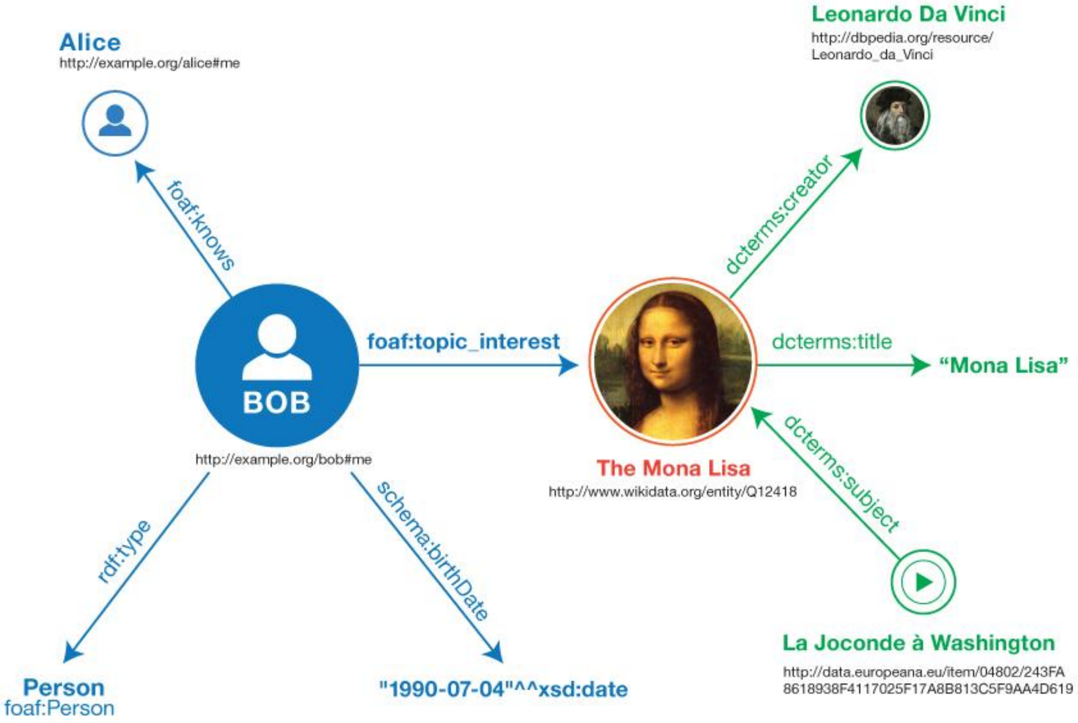


This is wrong

- **i.e. it doesn't Validate in JSON-LD Playground**

Good Example of JSON-LD representation of a linked data Graph

```
{
  "@context": { "foaf": "http://xmlns.com/foaf/0.1/",
        "Person": "foaf:Person",
        "interest": "foaf:topic_interest",
        "knows": {
          "@id": "foaf:knows",
          "@type": "@id"
        },
        "birthdate": {
          "@id": "http://schema.org/birthDate",
          "@type": "http://www.w3.org/2001/XMLSchema#date"
        },
        "dcterms": "http://purl.org/dc/terms/",
        "title": "dcterms:title",
        "creator": {
          "@id": "dcterms:creator",
          "@type": "@id"
        },
        "subject_of": {
          "@reverse": "dcterms:subject",
          "@type": "@id"
        }
      },
  "@id": "http://example.org/bob#me",
  "@type": "Person",
  "birthdate": "1990-07-04",
  "knows": "http://example.org/alice#me",
  "interest": {
    "@id": "http://www.wikidata.org/entity/Q12418",
    "title": "Mona Lisa",
    "subject_of": "http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619",
    "creator": "http://dbpedia.org/resource/Leonardo_da_Vinci"
  }
 }
```



Bi-lingual R & Python Package: With common stack JSON-LD Domain Templates

FAIRmaterials Package website

- https://cwrusdle.bitbucket.io/

20 Scientific Domain Ontologies

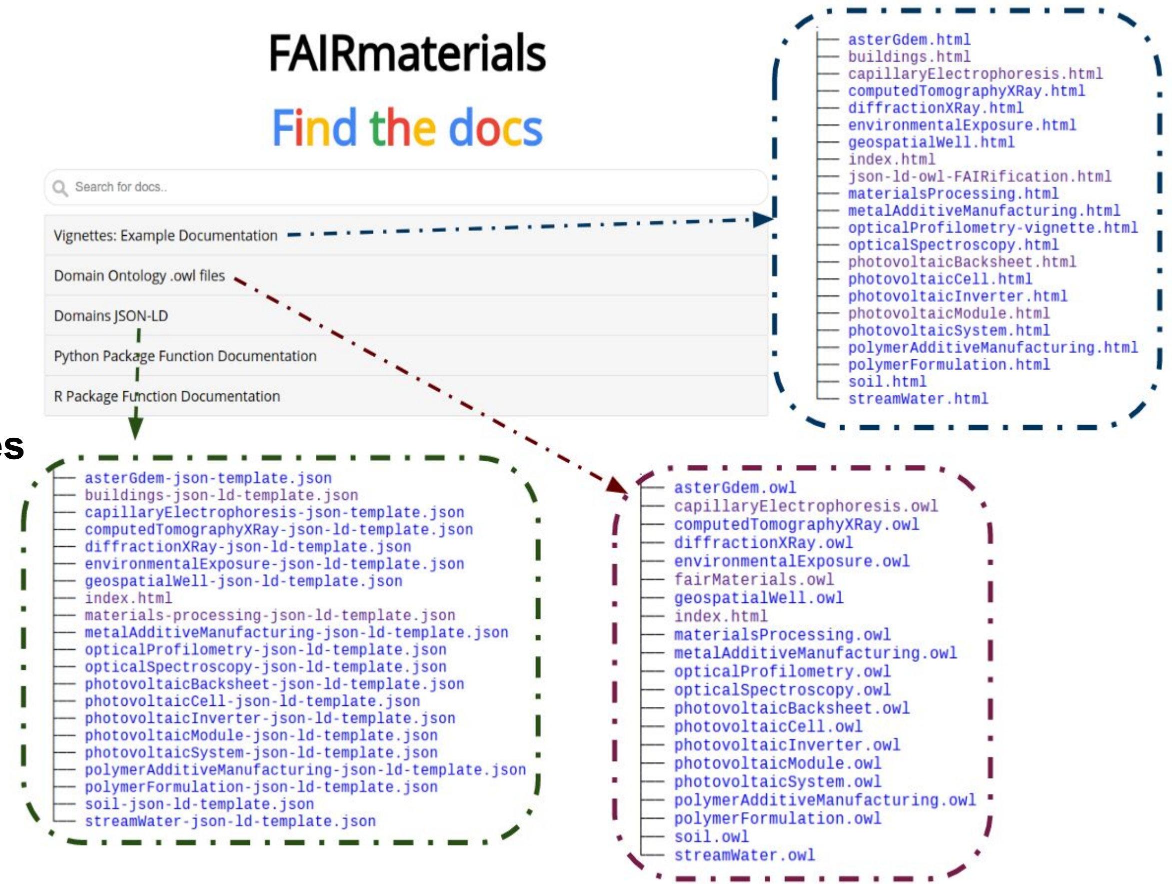
- Defined by OWL Files
- And 1 Combined OWL file

20 json-ld templates

- **for these domains**

21 domain documentation vignettes

- How to FAIRify for that domain



FAIRmaterials

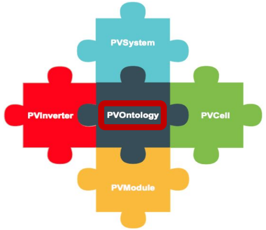
Package Documentat

Schema Diagram of the Data

JSON-LD Description

Code examples to FAIRify

- From R
- From Python



Photovoltaic Module JSON-LD

William C. Oltjen, Laura S. Bruckman, Roger H. French

2023-01-10

Photovoltaic Module JSON-LD Description

This json-ld template is used to store metadata information commonly used to describe photovoltaic (PV) Modules. There is information included on the cells, encapsulant, backsheet, electrical properties, and module dimensions. Note that a lot of the material included in this json pulls from the OrangeButton Taxonomy.

Creating JSON-LD for PV Modules in R

```r
library(FAIRmaterials)

# An example data frame for PV Modules
module_example <- data.frame('CellTechnologyType' = c('PERC', 'Al-BSF'),
                             'CellMaterial' = c('mono-crystalline', 'poly-crystalline'),
                             'EncapsulantMaterial' = c('EVA', 'POE'),
                             'BacksheetMaterial' = c('PVF', 'PET'),
                             'BacksheetColor' = c('White', 'Black'),
                             'ProdMfr' = c('CSI', 'FirstSolar'),
                             'ProdCode' = c('BiHiKu7', 'BiHiKu7255'),
                             'CellCount' = c(60, 72),
                             'ModuleEfficiency' = c(20.1, 22),
                             'GlassThickness' = c(2, 1.5),  # GT and J_Box are not included in
                             'J_Box' = c('IP68', 'IP70')) # base template, but they still have

# This will generate JSON-LD file for the example data
output <- fairify_data(module_example, domain = 'PVModule', saveLocal = TRUE)
```

Creating JSON-LD for PV Modules in Python

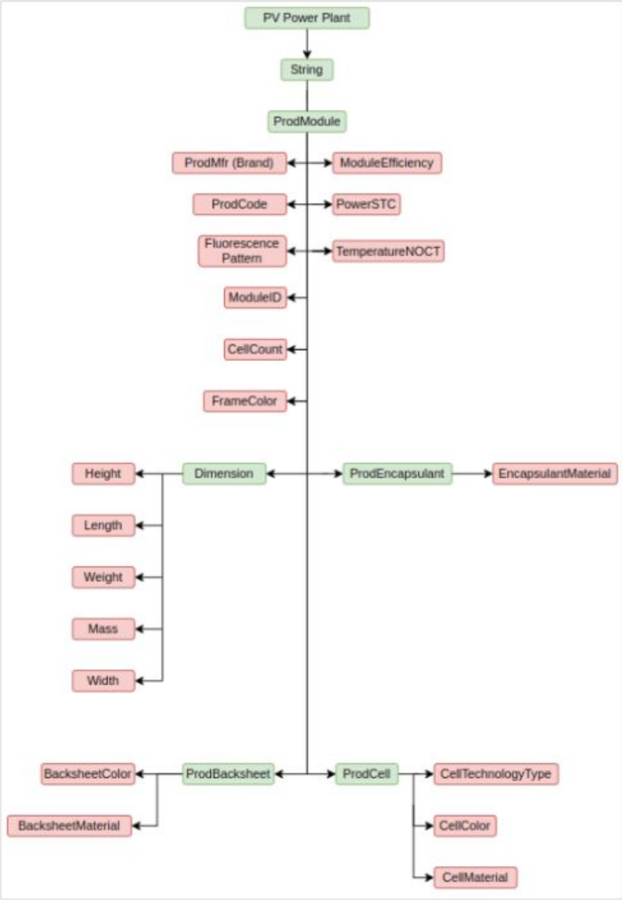
```python
from fairmaterials.fairify_data import *
import pandas as pd

# An example data frame for PV Modules
module_example = {'CellTechnologyType' : ['PERC', 'Al-BSF'],
                  'CellMaterial' : ['mono-crystalline', 'poly-crystalline'],
                  'EncapsulantMaterial' : ['EVA', 'POE'],
                  'BacksheetMaterial' : ['PVF', 'PET'],
                  'BacksheetColor' : ['White', 'Black'],
                  'ProdMfr' : ['CSI', 'FirstSolar'],
                  'ProdCode' : ['BiHiKu7', 'BiHiKu7255'],
                  'CellCount' : [60, 72],
                  'ModuleEfficiency' : [20.1, 22]}

module_example = pd.DataFrame(module_example)

# This will generate JSON-LD file for the example data
output <- fairify_data(module_example, domain = 'PVModule')
```

PV Module schema diagram



PV Module schema diagram

Acknowledgment

This material is based upon work supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE) under Solar Energy Technologies Office (SETO) Agreement Number DE-EE0009347.

PV System Ontology Graph

For a PV System, or PV Power Plant

- We define System
- And have sub-domains of
  - PV Inverter, aka ProdInverter
  - PV Module, aka ProdModule
  - PV Cell

We Orange Button terms

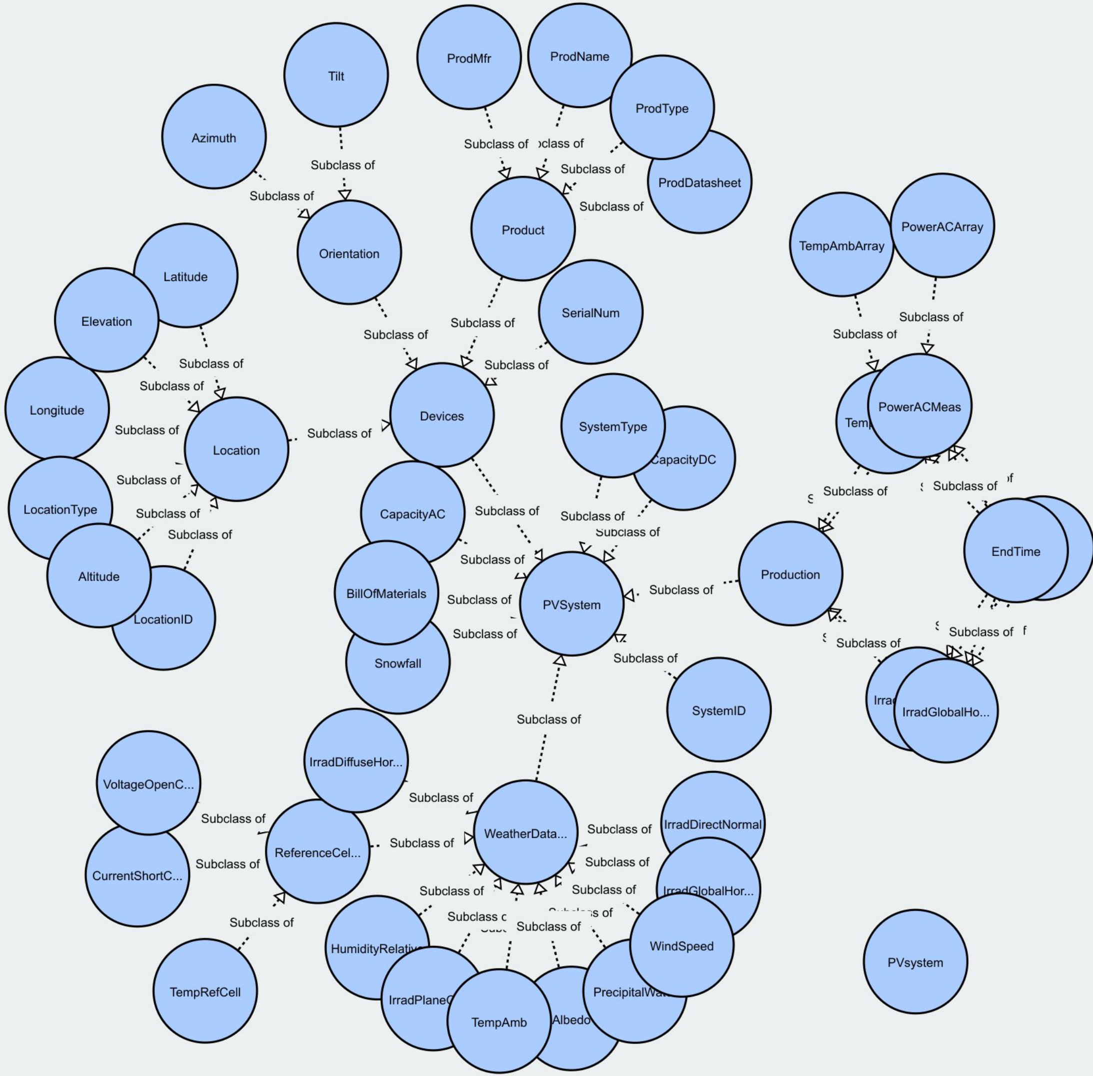
- "Orange Button is an open data exchange standard for the solar+energy storage industry. "
- As defined in their Taxonomy

Their Taxonomy is for Contractual Relations

- For the PV Value Chain
- And uses OpenAPI JSON Syntax
- So not JSON-LD

For JSON-LD & Ontologies

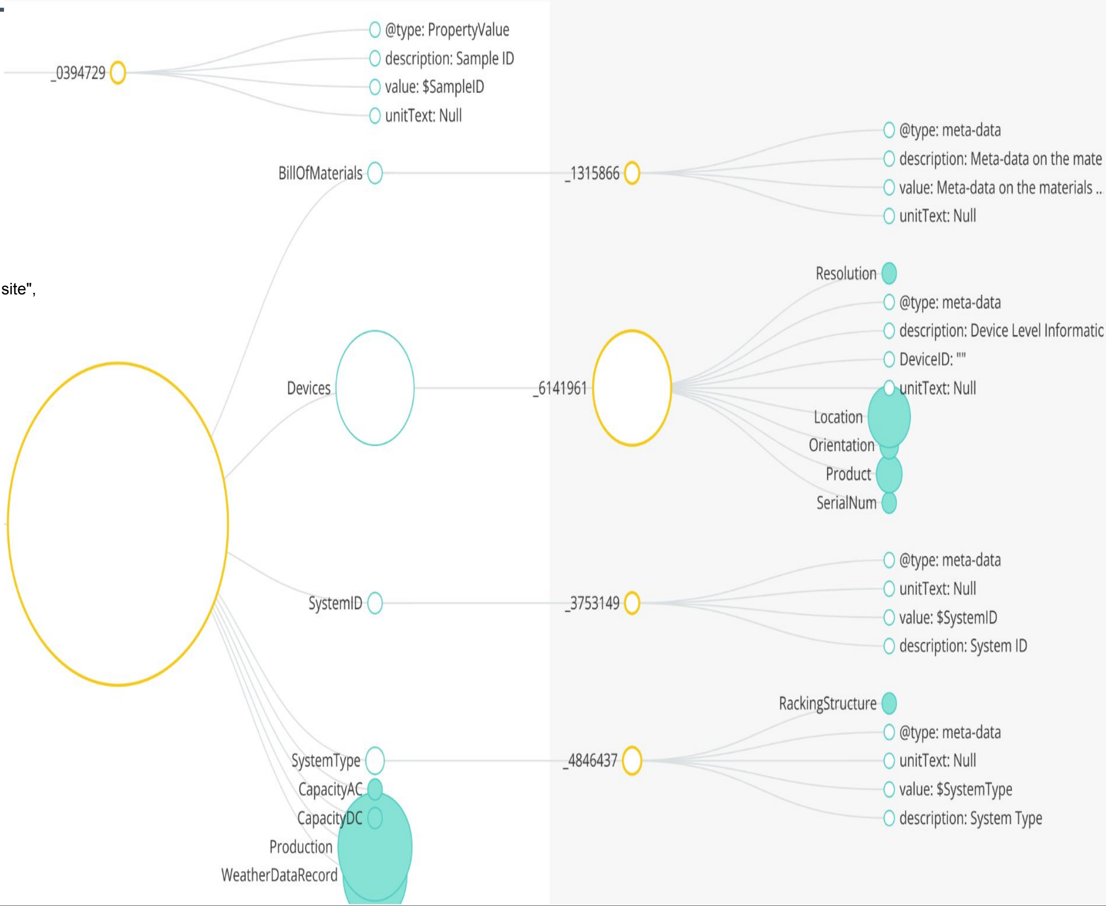
- All keys must point to a web defined location
- Using Internationalized Resource Identifiers (IRIs)
- **We point our IRIs to OrangeButton**



PV System JSON-LD visualization


```
{
  "@context": {
    "SampleID": "https://schema.org/",
    "@vocab": "http://www.w3.org/2001/XMLSchema#",
    "ProdModule": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "PVSystem": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "BillOfMaterials": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "Devices": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "DeviceID": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "Location": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "Altitude": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "Elevation": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "Latitude": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "LocationID": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "LocationType": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "Longitude": "https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
}
```

```
{
 "SampleID": {
  "@type": "PropertyValue",
  "description": "Sample ID",
  "value": "$SampleID",
  "unitText": "Null"
 },
"PVSystem": {
  "BillOfMaterials": {
   "@type": "meta-data",
   "description": "Meta-data on the materials in the site",
   "value": "Meta-data on the materials at the site",
   "unitText": "Null"
  },
  "Devices": {
   "Resolution": {
    "@type": "meta-data",
    "unitText": "Null",
    "value": "$Resolution",
    "description": "Resolution of Data Record"
   },
   "@type": "meta-data",
   "description": "Device Level Information",
   "DeviceID": "\"\"",
   "unitText": "Null",
   "Location": {
    "ClimateZone": {
     "@type": "meta-data",
     "unitText": "Null",
     "value": "$ClimateZone",
     "description": "Koppen Geiger Climate Zone"
    },
    "Altitude": {
     "@type": "meta-data",
     "unitText": "ft",
     "value": "$Altitude",
     "description": "Distance above ground"
```

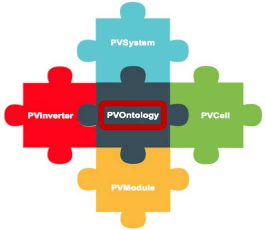
PV Module Ontology Graph

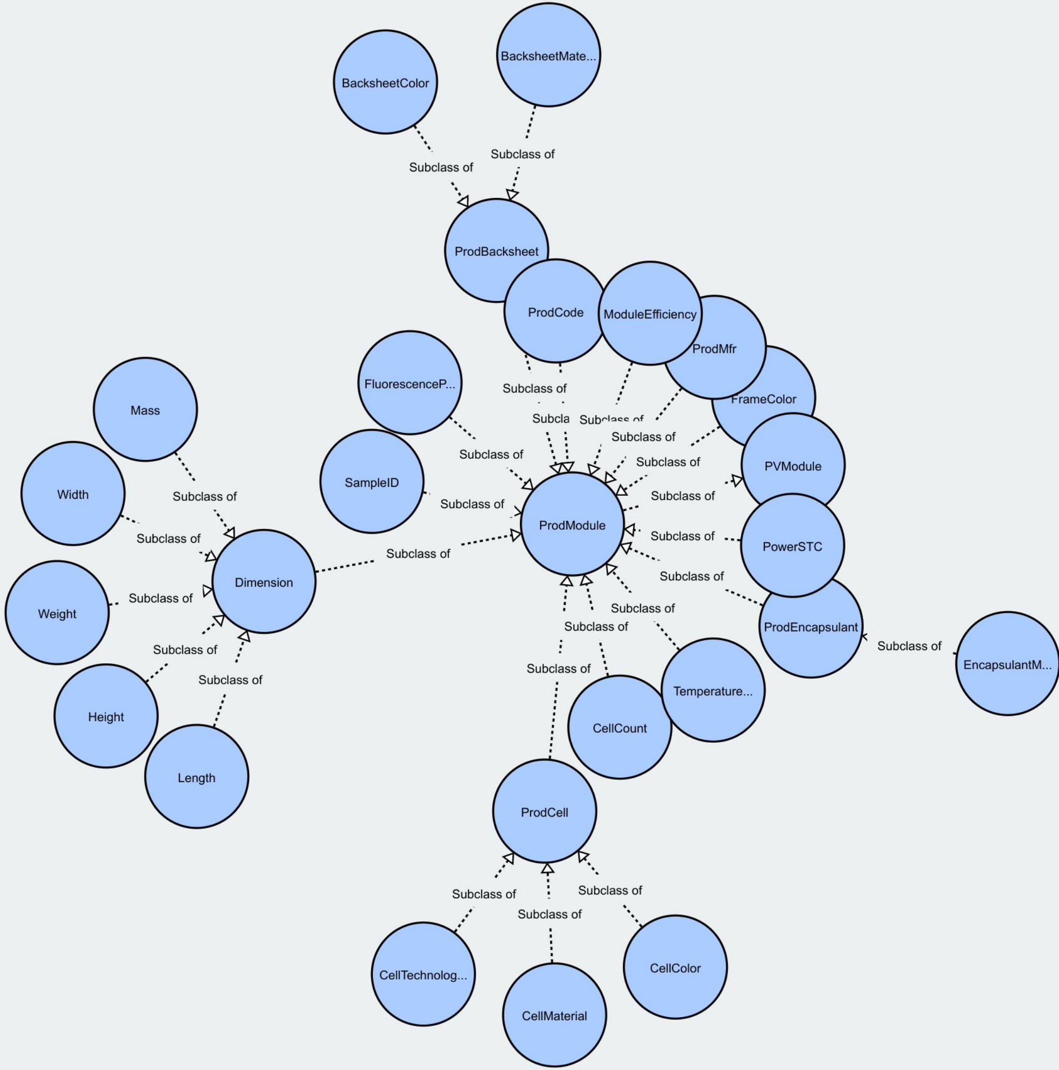
Here we using the Orange Button Key

- **ProdModule**

We also will write out Photovoltaic

- **Instead of using PV**
- **Since PV is not human-readable**
- **Its an acronym that is not universal**

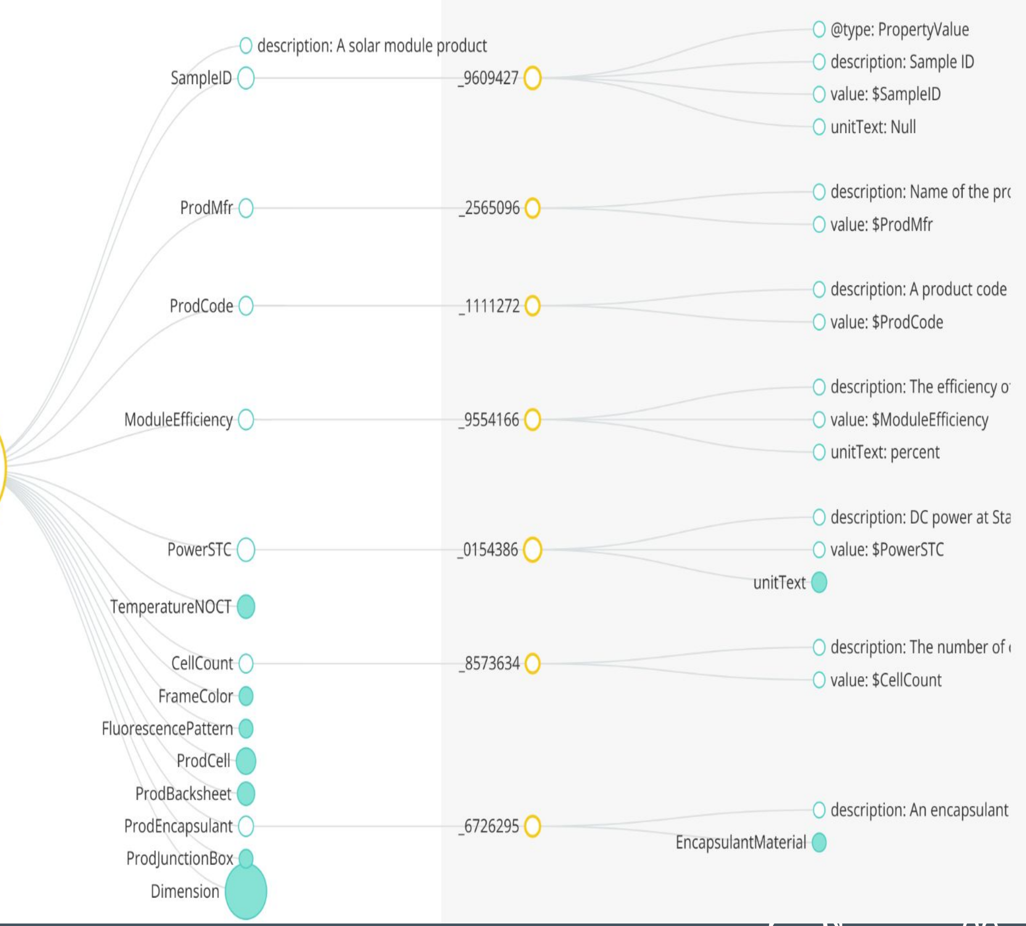




ProdModule JSON-LD

```
{
  "@context": {
    "SampleID": "https://schema.org/",
    "@vocab": "http://www.w3.org/2001/XMLSchema#",
    "ProdModule":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "ProdMfr":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "ProdCode":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "ModuleEfficiency":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "PowerSTC":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "TemperatureNOCT":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "CellCount":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "FrameColor":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "FluorescencePattern":
"https://cwrusdle.bitbucket.io/Ontology/photovoltaicModule.owl",
    "ProdCell":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "CellTechnologyType":
"https://github.com/Open-Orange-Button/Orange-Button-Taxonomy",
    "CellMaterial":
"https://cwrusdle.bitbucket.io/Ontology/photovoltaicModule.owl",
  }
```

```
{
  "ProdModule": {
    "description": "A solar module product",
    "SampleID": {
      "@type": "PropertyValue",
      "description": "Sample ID",
      "value": "$SampleID",
      "unitText": "Null"
    },
    "ProdMfr": {
      "description": "Name of the product or equipment manufacturer",
      "value": "$ProdMfr",
      "unitText": null
    },
    "ProdCode": {
      "description": "A product code is a standardized, unique human-readable identifier that is compact,and can be easily parsed",
      "value": "$ProdCode",
      "unitText": null
    },
    "ModuleEfficiency": {
      "description": "The efficiency of the solar module at standard test condition (STC)",
      "value": "$ModuleEfficiency",
      "unitText": "percent"
    },
    "PowerSTC": {
      "description": "DC power at Standard Test Condition (STC). STC is defined as 1000 W/m2 irradiance, 25C and ASTM G173-03 standard solar spectrum",
      "value": "$PowerSTC",
      "unitText": ["W", "kW", "MW"]
    },
    "Temperature NOCT": {
      "description": "The nominal operating cell temperature (NOCT) of a solar module",
      "value": "$Temperature NOCT",
      "unitText": ["C", "K", "F"]
```



Open Science Framework is a great platform for web linked data

OSF.io

- Accelerates open scholarship

CWRU is an institutional member

- SDLE Res. Center has our site
- We share public datasets

Every dataset gets its own DOI

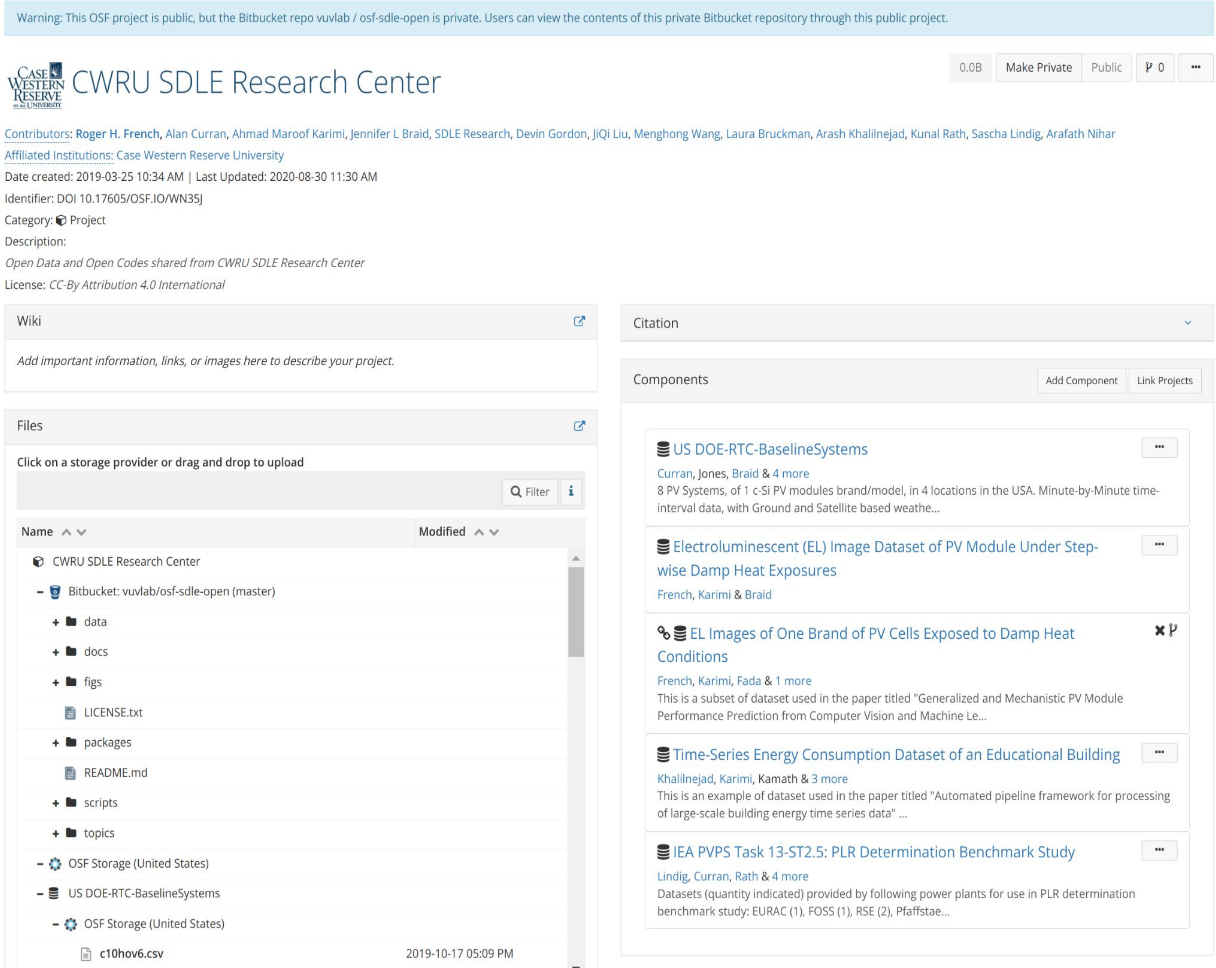
- Permanent

We can share data from

- Many storage systems

We get noticeable Citation Traffic

- From our OSF.io site



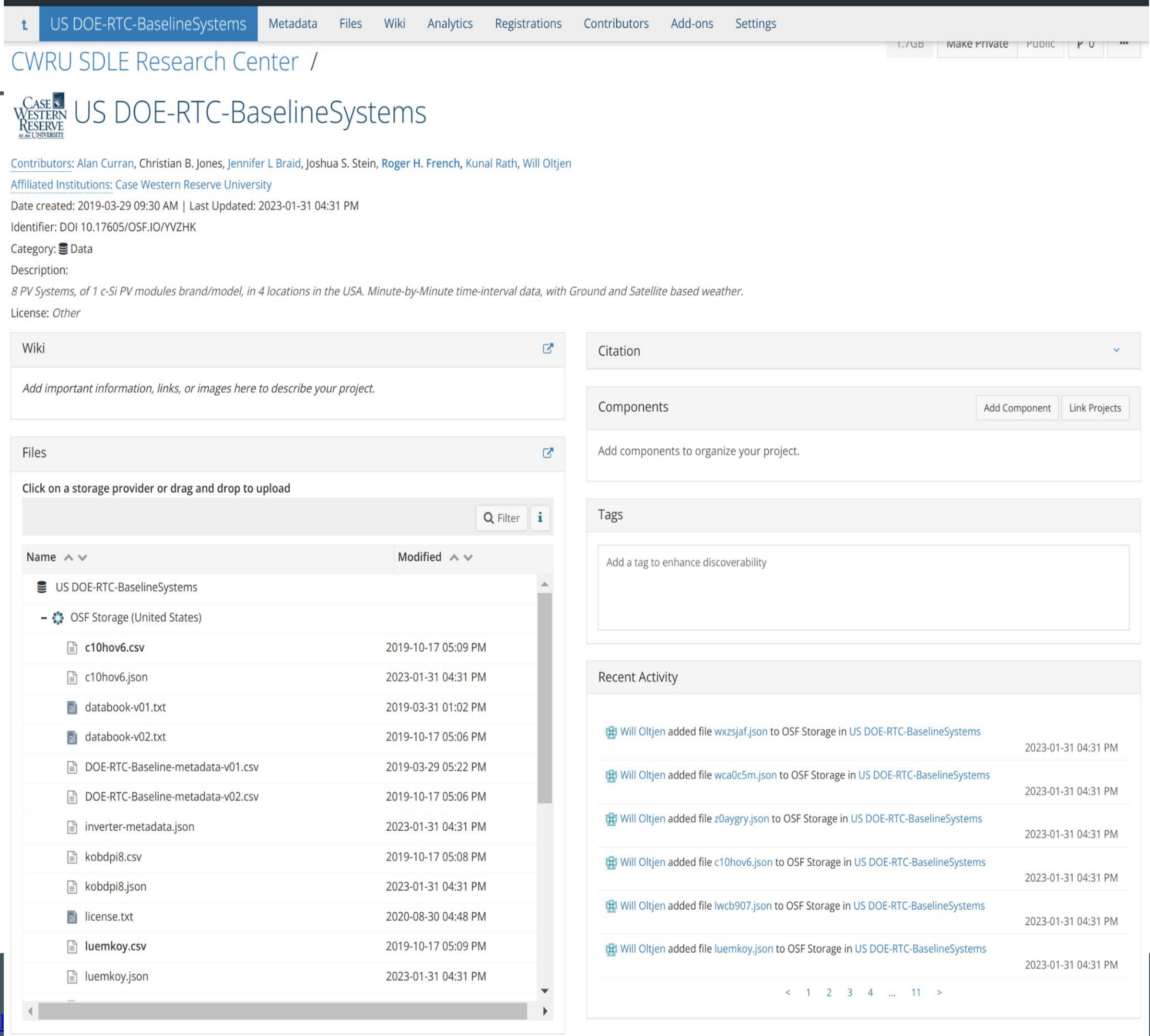
DOE RTC Baseline Systems

The DOE Regional Test Centers[1,2,3]

- Have 4 locations
- With 2 identical systems at each

These RTC baseline systems

- Are distributed as open data
- And are now FAIRified
  - Using FAIRmaterials Package
- On the SDLE OSF.io page
  - https://osf.io/yvzhk/



1. L. Burnham, et al., "The US DOE Regional Test Center Program: Driving Innovation Quality and Reliability," Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 1279686, 2015.

2. M. S. Lave, et al., "Performance Comparison of Stion CIGS Modules to Baseline Monocrystalline Modules at the New Mexico Florida and Vermont Regional Test Centers: January 2015-December 2016.," Sandia National Lab. (SNL-CA), Livermore, CA (United States); Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), Jan. 2017.

3. J. Stein, "US DOE Regional Test Centers Program - 2016 Annual Report.," Sandia National Lab. (SNL-NM), Albuquerque, NM (United States), SAND-2017-4853R, May 2017.

Conclusions & Future Directions

Remember: FAIRification applies not just to datasets

- **Models can also be FAIRified**

We plan the next release of FAIRmaterails R & Python in Feb. 2023

Materials Ontologies introduce one to Graphs & Knowledge Graphs (KGs)

- Wikidata is a … multilingual knowledge graph"
- ICIJ: International Consortium of Investigative Journalists
  - Provides Open Access to all their resources as a Graph Database

We are developing Spatiotemporal Graphs (stGraphs)

- Such as for Fleet Performance Analysis and Power Prediction

So after OWL & JSON-LD, we move to RDF, Graph Databases and SPARQL

- We are implementing JanusGraph on our 2Pb CRADLE Hadoop Clusters
- So as to enable native, distributed Graph Computation

We are developing Knowledge Graphs as repositories of data science models

We are also developing Scientific Workflows

- For Data Analyses and their Automation
- And for Materials Processing

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY

[1] Chelsey Bryant, Nicholas R. Wheeler, Franz Rubel, and Roger H. French, “kgc: Koeppen-Geiger Climatic Zones.” The Comprehensive R Archive Network, Dec. 20, 2017 [Online]. Available: https://cran.r-project.org/package=kgc . [Accessed: May 24, 2021]

[2] W.-H. Huang et al., “netSEM: Network Structural Equation Modeling.” The Comprehensive R Archive Network, Nov. 28, 2018 [Online]. Available: https://CRAN.R-project.org/package=netSEM . [Accessed: Dec. 02, 2018]

[3] A. M. Karimi, B. G. Pierce, J. S. Fada, N. A. Parrilla, R. H. French, and J. L. Braid, “PVimage: Package for PV Image Analysis and Machine Learning Modeling.” May 08, 2020 [Online]. Available: https://pypi.org/project/pvimage/ . [Accessed: Feb. 28, 2020]

[4] Alan J. Curran, Tyler Burleyson, Sascha Lindig, David Moser, and Roger H. French, “PVplr: Performance Loss Rate Analysis Pipeline.” The Comprehensive R Archive Network, Oct. 07, 2020 [Online]. Available: https://CRAN.R-project.org/package=PVplr . [Accessed: Oct. 18, 2020]

[5] Wei-Heng Huang et al., “ddiv: Data Driven I-V Feature Extraction.” The Comprehensive R Archive Network, Apr. 14, 2021 [Online]. Available: https://CRAN.R-project.org/package=ddiv . [Accessed: Jul. 30, 2019]

[6] Menghong Wang et al., “SunsVoc: Constructing Suns-Voc from Outdoor Time-series I-V Curves.” The Comprehensive R Archive Network, Apr. 30, 2021 [Online]. Available: https://CRAN.R-project.org/package=SunsVoc

[7] Willam C. Oltjen, Liangyi Huang, Roger H. French, and Liangyi Huang, “FAIRmaterials: Make Materials Data FAIR.” The Comprehensive R Archive Network, Sep. 14, 2021 [Online]. Available: https://CRAN.R-project.org/package=FAIRmaterials

[8] Roger H. French et al., “Fairmaterials.” The Python Package Index (PyPI), Oct. 08, 2021 [Online]. Available: https://pypi.org/project/fairmaterials/

[9] Kris Zhao and Roger H. French, “hbspark: Package to pipe data from HBase to Spark (2+).” 2021 [Online]. Available: https://github.com/kxz167/hbspark . [Accessed: 29-Jan-2022]

[10] R. F.-0002-6162-0532) Huang(ORCID:0000-0003-0845-3293) Liangyi, “pointextract: Extract points information from 2d images to build a 3D object.” Jun-2022 [Online]. Available: https://pypi.org/project/pointextract/ . [Accessed: 22-Apr-2022]