

Breakout Session 6: Track B

Metadata for the Masses: Making CEDAR Portable and Cloud-Based

Dr. Mark Musen (Moderator)
Professor, Stanford University

Metadata for the Masses: Making CEDAR Portable and Cloud-Based

Mark A. Musen, M.D., Ph.D
Stanford University
musen@stanford.edu

Supported in part by grant 3R01 LM013498-02S1:

“Improved metadata authoring to enhance AI/ML readiness of associated datasets”



SCIENTIFIC DATA



Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measurable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016











Workspace

Shared with Me

FILTER RESET

TYPE



	Title	Created	Modified
	GEO	9/5/17 9:48 AM	9/5/17 10:24 AM
	BioCADDIE	9/5/17 9:48 AM	9/5/17 10:24 AM
	BioSample Human	9/5/17 9:49 AM	9/5/17 11:28 AM
	Optional Attribute	9/5/17 10:38 AM	9/5/17 10:38 AM
	ImmPort Investigation	9/5/17 9:49 AM	9/5/17 10:21 AM
	LINCS Cell Line	9/5/17 9:49 AM	9/5/17 9:49 AM
	LINCS Antibody	9/5/17 9:49 AM	9/5/17 9:49 AM
	ImmPort Study	9/5/17 9:49 AM	9/5/17 9:49 AM





All / Users / Mark A. Musen











Workspace

Shared with Me

FILTER RESET

TYPE



	Title	Created	Modified
	GEO	9/5/17 9:48 AM	9/5/17 10:24 AM
	BioCADDIE	9/5/17 9:48 AM	9/5/17 10:24 AM
	BioSample Human	9/5/17 9:49 AM	9/5/17 11:28 AM
	Optional Attribute	9/5/17 10:38 AM	9/5/17 10:38 AM
	ImmPort Investigation	9/5/17 9:49 AM	9/5/17 10:21 AM
	LINCS Cell Line	9/5/17 9:49 AM	9/5/17 9:49 AM
	LINCS Antibody	9/5/17 9:49 AM	9/5/17 9:49 AM
	ImmPort Study	9/5/17 9:49 AM	9/5/17 9:49 AM

Open

Populate

Share...

Copy to...

Move to...

Rename...

Delete





▼ BioSample Human

- * Sample Name
- * Organism
- * Tissue
- * Sex
- * Isolate
- * Age
- * Biomaterial Provider
- ▼ **Attribute**
 - Name
 - Value

CANCEL

VALIDATE

SAVE

▼ BioSample Human

* Sample Name 056

* Organism Homo sapiens

* Tissue

* Sex

* Isolate

* Age

* Biomaterial Provider

▼ Attribute

Name

Value

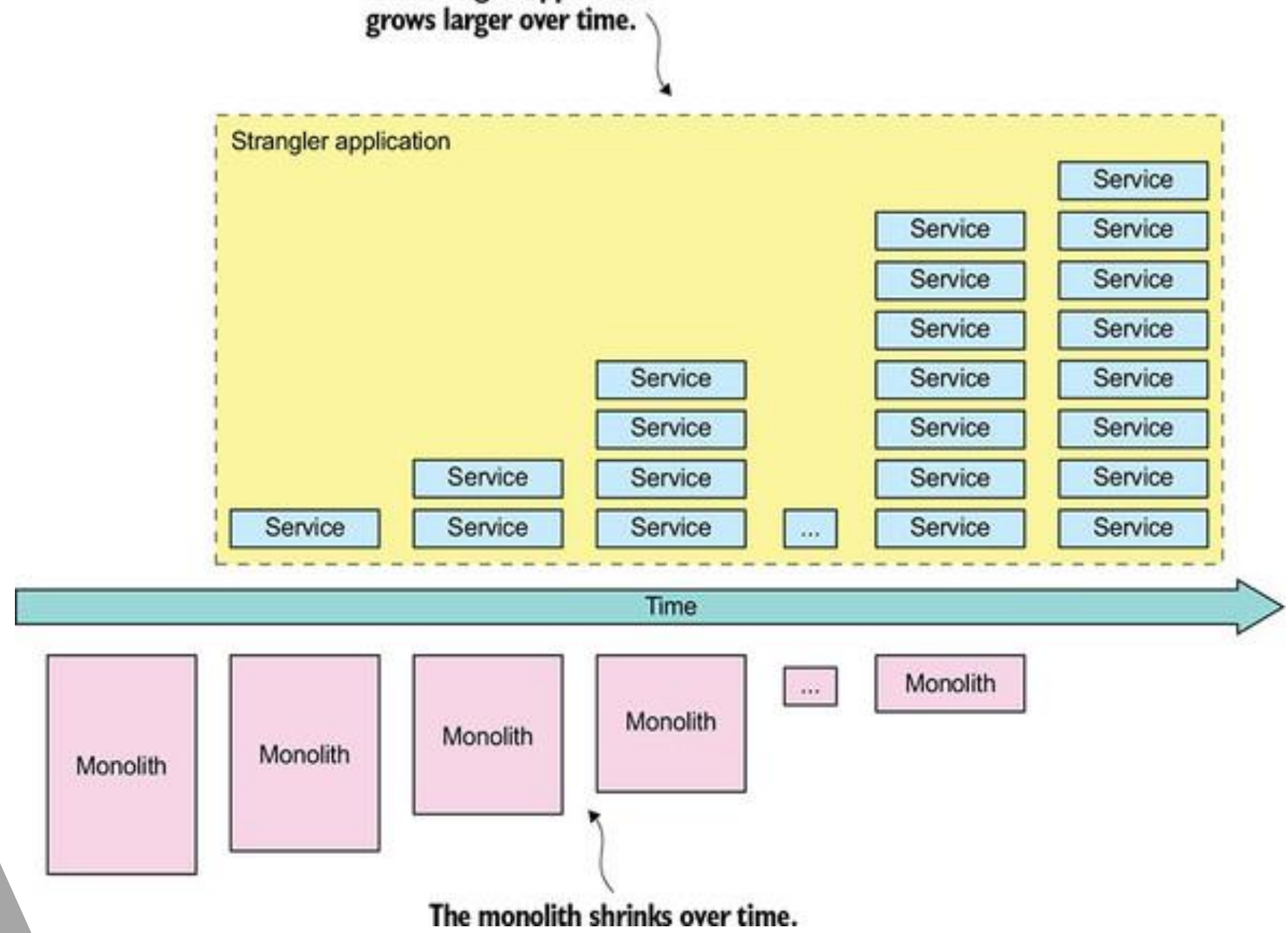
Goals of our project:

- Dockerize CEDAR to support cloud-based deployment
- Provide reusable components to acquire and view metadata
- Use these components to develop standalone, reusable tools for metadata management



Componentizing CEDAR and Moving It to the Cloud Using the “Strangler Vine” Pattern

Strangling the Monolith





Putting CEDAR
Components to
use in the
RADx Data Hub

RADx Data Hub

- Archives and harmonizes data from hundreds of studies related to COVID-19
- Designed to support secondary analysis of disparate data sets



Study Explorer

Variables Catalog

Helpful Information 

Visualize Metadata

COMMIT (0.4 MB)

dbGaP Link: [phs003081](https://dbgap.ncbi.nlm.nih.gov/oa/GET.cgi?acc=phs003081)

[Home](#) > [Study Explorer](#) > [Study Overview](#)

Study Information

NIH Institute/Center: NIMHD

RADx Data Program: RADx-UP

Study Description: In the United States underserved and socially vulnerable populations experience higher rates of COVID-19 infection, morbidity, and mortality. This disproportionate burden has shown that systemic racial bias in health care delivery, discrimination, and poor social determinants of health such as asthma, diabetes, hypertension, and obesity, all of which are root causes, academic and other research institutions and health care systems behaviors among underserved and vulnerable populations. Behaviors among structural barriers to trust, testing, treatment, and prevention of COVID-19. Primary researchers, the focus should be on radical institutional transformation to address issues (SEBI) influencing access acceptability and uptake of COVID-19 testing and existing community-academic partnerships. The distinction between trust and trust. Our proposed study will employ a continuous engagement approach to existing community-engaged research (CEnR) partnership. In collaboration with to co-design a sustainable model for trustworthy CEnR partnerships to address

Principal Investigator: C. Daniel Mullins

Has Data Files: Yes



RADx Metadata Specification

Expand All Collapse All

Data File Titles

Title 

Community Mistrust and Measures of Institutional Trustworthiness (COMMIT)

Language 

en

Data File Identity

Identifier 

Identifier Type 

Start typing to filter

File Name 

20230131_project101_DATA_transformcopy.csv

Version 

3

SHA256 digest 

b2f91603895f28326c91267725f43a53d66714df26dfbbdb2ccdf6359559b032

Metadata Viewer for RADx Data Hub



Putting CEDAR
Components to
use for
HuBMAP

Human BioMolecular Atlas Program

An open, global atlas of the human body at the cellular level

The HuBMAP Data Portal is the central resource for discovery, visualization, and download of single-cell tissue data generated by the consortium. A standardized data curation and processing workflow ensure that only high quality is released.

Navigate healthy human cells with the Common Coordinate Framework

Interact with the human body data with the Anatomical Structures, Cell Types and Biomarkers (ASCT+B) Tables and CCF Ontology. Also explore two user interfaces: the Registration User Interface (RUI) for tissue data registration and Exploration User Interface (EUI) for semantic and spatial data.

Get Started



Screenshot

The screenshot displays the HuBMAP Data Portal interface. At the top, the HuBMAP logo is on the left, and navigation links for 'Atlas & Tools', 'Resources', and 'User Profile' are on the right. Below the header, a search bar is labeled 'Search ontology terms ...'. The main content area features a 3D anatomical model of the human body with internal organs highlighted in red. To the left of the model is a navigation menu with a search bar and a list of ontology terms, including 'body', 'heart', 'lung', 'kidney', 'right kidney', 'left kidney', 'kidney capsule', 'cortex of kidney', 'renal medulla', 'renal column', 'renal pyramid', 'hilum of kidney', 'kidney interstitium', 'kidney calyx', 'renal pelvis', 'ureter', 'renal papilla', 'renal fat pad', 'nephron', 'spleen', 'spleen capsule', 'trabecula of spleen', 'spleen pulp', 'marginal zone of spleen', 'spleen perifollicular zone', 'hilum of spleen', and 'colon'. To the right of the model is a list of samples with columns for 'body', 'Centers', 'Donors', and 'Samples'. The list includes samples from KPMP, CODEX, and LC, with details such as patient ID, age, BMI, and collection date.

	A	B	C	D	E	F	G	I
1	sample_ID	source_storage_ti	source_storage_ti	preparation_medium	preparation_cond	processing_tim	processing_tim	storage_meth
2	Visium_90LC_A4_S1	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embed
3	Visium_90LC_A4_S2	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embed
4	Visium_90LC_I4_S1	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embed
5	Visium_90LC_I4_S2	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embed
6		86 days	days	Formalin		10 minutes	minutes	Paraffin em
7		86 days	days	Formalin		10 minutes	minutes	Paraffin em
8		86 days	days	Formalin		10 minutes	minutes	Paraffin em
9		86 days	days	Formalin		10 minutes	minutes	Paraffin em
10		86 days	days	Formalin		10 minutes	minutes	Paraffin em
11	Visium_40AZ_Q9_S1	100	d	Agar-agar		5	min	OCT embed
12	Visium_40AZ_Q9_S2	100	d	Agar-agar		5	min	OCT embed
13	Visium_40AZ_Q9_S3	100	d	Agar-agar		5	min	OCT embed
14	Visium_40AZ_Q9_S4	100	d	Agar-agar		5	min	OCT embed
15	Visium_90LC_W3_S1	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
16	Visium_90LC_W3_S2	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
17	Visium_90LC_W3_S3	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
18	Visium_90LC_W3_S4	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
19	Visium_90LC_W3_S5	208	day	Methanol (100%)	-20 celsius	4	minute	Unknown
20	Visium_90LC_W3_S6	208	day	Methanol (100%)	-20 celsius	4	minute	Unknown
21	Visium_90LC_W3_S7	208	day	Methanol (100%)	-20 celsius	4	minute	Unknown

HuBMAP Metadata Spreadsheet Validator



Upload and submit your spreadsheet file to validate the metadata records

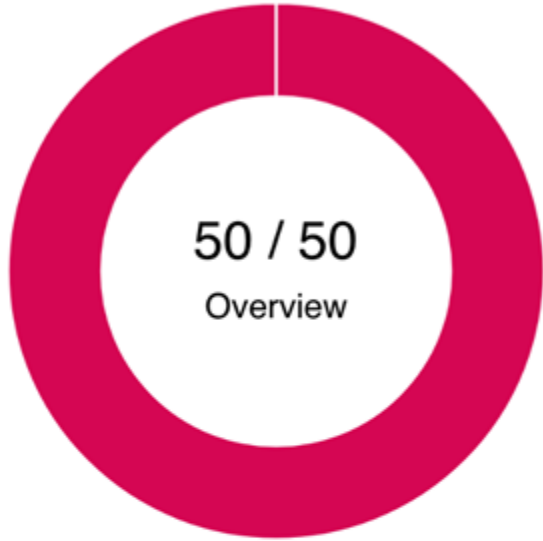
Drag & drop your spreadsheet file here or [Browse](#)

RNA_eq_latest.xlsx

START VALIDATING

Validation Result

Found 50 metadata records in the spreadsheet



Invalid metadata Valid metadata

Validation Summary of [RNAseq](#) Metadata

 maseq.xlsx

The validity of a metadata record is measured by two metrics: *completeness* and *adherence*.

Completeness measures the presence of all required values in the metadata record defined by the metadata specification.

Adherence measures the conformance of the stated value in the metadata field to the data type defined by the metadata specification.

A metadata record is called invalid when the system detects errors using these two metrics. Use the button below to start the repair action.

REPAIR COMPLETENESS ERRORS

REPAIR ADHERENCE ERRORS

Completeness Error Analysis

Evaluating 50 metadata records for detecting missing values in the spreadsheet.

Field name

of invalid metadata records

analyte_class

10

40



Adoption of CEDAR Components within GREI

Research Project

[Overview](#)[Metadata](#)[Files](#)[Wiki](#)[Analytics](#)[Registrations](#)[Contributors](#)[Add-ons](#)[Settings](#)

Select a Metadata Template

OSF has partnered with CEDAR <https://metadatacenter.org> to provide more ways to annotate your research with domain or community-specific metadata records. If you would like to request the addition of a new metadata template, contact us at .

Available Templates from CEDAR

Psych-DS Official Template

Psych-DS metadata template

Human Cognitive Neuroscience Data (v1)Human cognitive neuroscience data (v1)
template schema generated by the CEDAR
Template Editor 2.6.49**Generic Dataset Metadata Template (GDMT)**Generic dataset metadata template (gdmt)
template schema generated by the CEDAR
Template Editor 2.6.0**Testing Record**

unique demo template for testing on OSF



Psych-DS Official Template

Generic.ExpandAll

Generic.CollapseAll

Name * ?

0

Description * ?

0

VariableMeasured * (1 .. ∞) ?

1



Generic.AllValues: 1 null

0

Author (1 .. ∞) ?

1



Generic.AllValues: 1 null

0

CEDAR Metadata Editor in the Open Science Framework Web Platform

factchecking_factcheckers.pdf

[Return to factchecking_factcheckers.pdf](#)



Psych-DS Official Template

Expand All

Collapse All

Name * ?

0

Description * ?

0

VariableMeasured * (1 .. ∞) ?

1



All Values: 1 null

0

Author (1 .. ∞) ?

1



CEDAR Metadata Editor in the **Open Science Framework** App

Standardized metadata

Fill out a standardized me

+ Add metadata form: H

Related works

Are there any preprints, artic
Publication?

Work type

Supplemental information

Work type

Supplemental information

Work type

Data management plan

Work type

Supplemental information

Work type

Software

Work type

Supplemental information

+ Add another related work

◀ Back to My datasets

[Privacy](#) [Accessibility](#) [Term](#)

Copyright (c) 2024 Dryad

Preprocessing

Preprocessing status [?]

- Preprocessed
- Raw

Information about the preprocess used to produce the dataset. Please provide the link to the documentation or publication describing the analysis process, using DOI when possible. (e.g. [Brainlife](#) workflow publication).

Leave the field blank if not applicable.

Preprocessing Pipeline (1 .. ∞) [?]

1   


Provide a link to the location where the preprocessing code is hosted, i.e. GitHub repository.

To ensure the accessibility and compatibility of the code, consider depositing a copy of the code together with the dataset following the [Dryad submission process](#). Leave the field blank if not applicable.

Preprocessing Script (1 .. ∞) [?]

1   

Standard 

Source dataset 

Experiment 

Analysis 



related to this Data

remove

remove

remove

remove

remove

remove

All progress saved

Proceed to README 

[news](#)  [Jobs & opportunities](#)

Version: v3.0.3;

CEDAR Metadata Editor in the Dryad Platform



The Vine Has
Been Strangled!

Componentization of CEDAR is paying off for multiple projects

- Acquisition and management of specialist data
 - RADx Data Hub for COVID study data
 - HuBMAP 'omics biomarker data
- Generalist data repositories
 - Open Science Framework
 - Dryad

