Data Infrastructure

Carelyn Campbell, Ben Blaiszik, Laura Bartolo

November 1, 2016

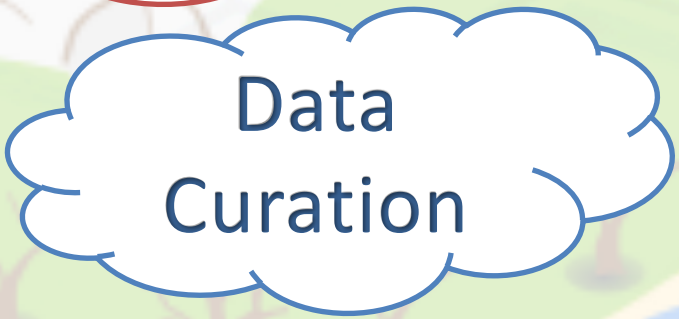
NIST National Institute of Standards and Technology U.S. Department of Commerce

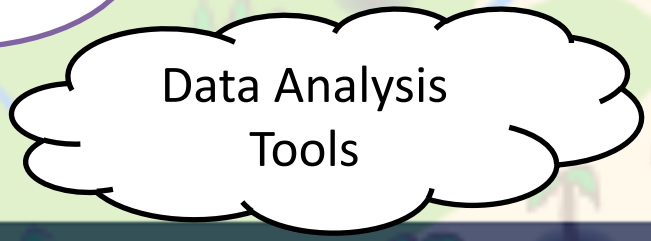
CHiMaD

THE UNIVERSITY OF CHICAGO 1890

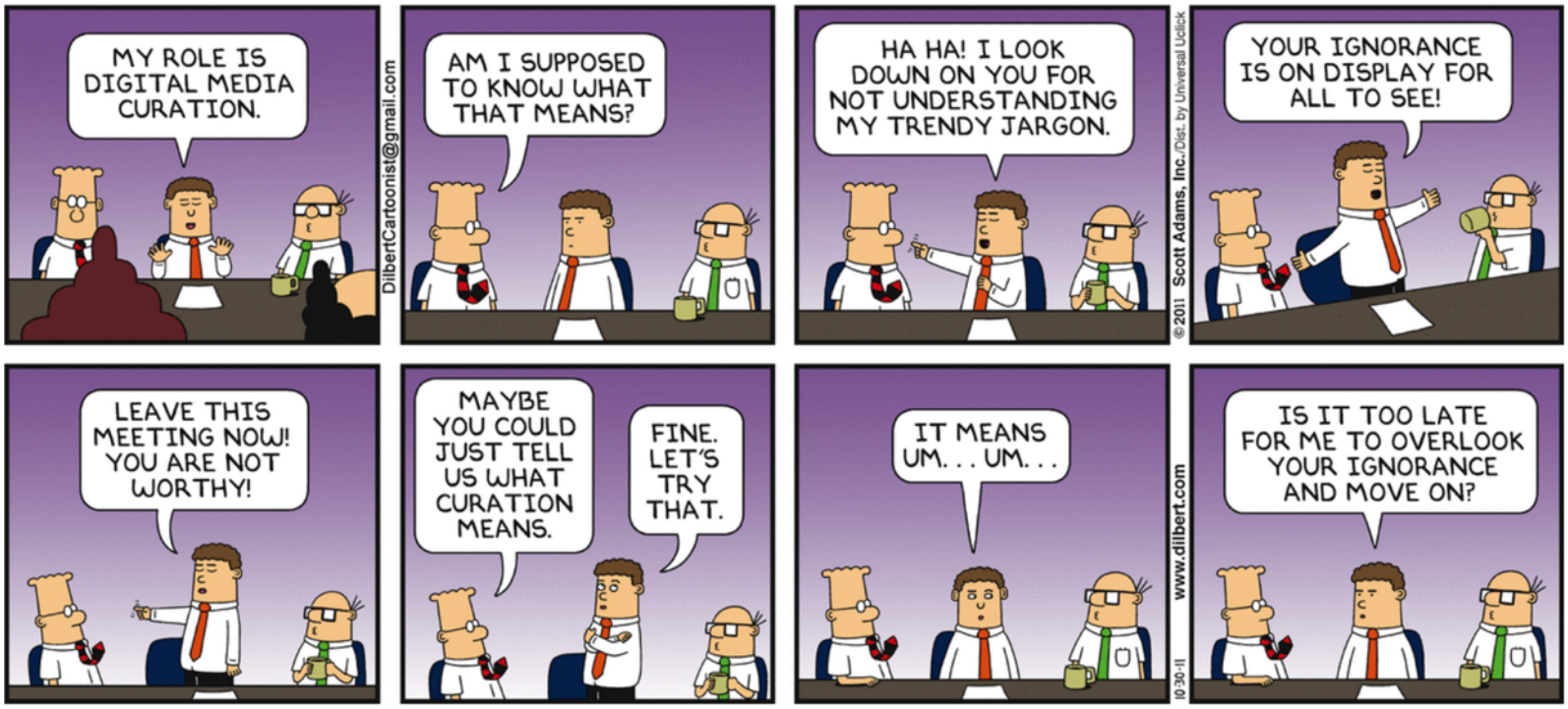
MATERIAL MEASUREMENT LABORATORY

Data Landscape





What is Data Curation?



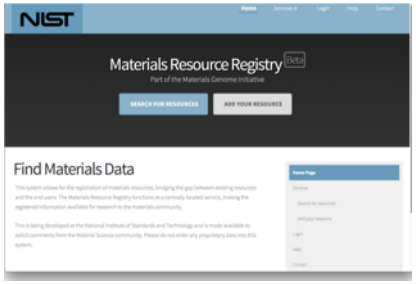
Scott Adams, October 30, 2011

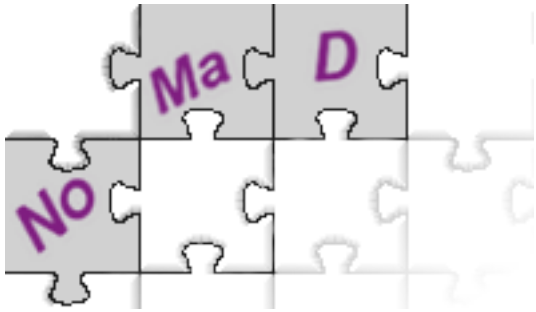
Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education.

http://ischool.illinois.edu/academics/degrees/specializations/data_curation

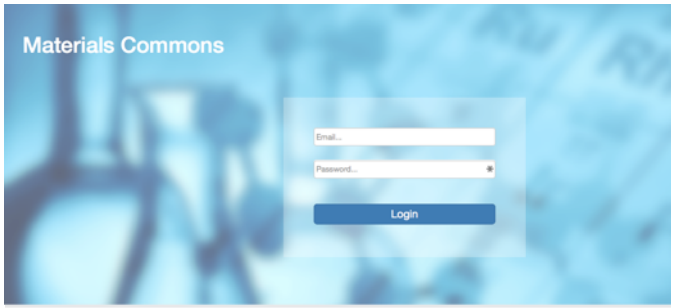
Materials Data Curation Tools

GRANTA MATERIAL INTELLIGENCE





ICE INTEGRATED COLLABORATIVE ENVIRONMENT



CoRR

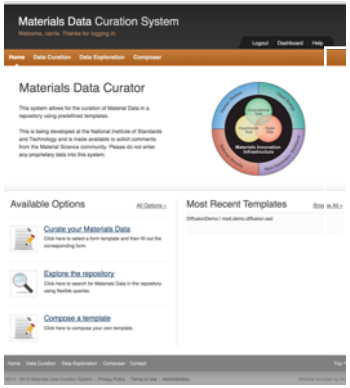
AiiDA

T2C2

MATERIALS PROJECT

Georgia Tech Institute for Materials | MATIN

Citrine



MATERIALS DATA FACILITY

TMS WEBINARS
LEARN. NETWORK. ADVANCE.

Representative list of tools focused on materials data, but not comprehensive

Data Model Definition

Defines the structure of metadata and data

Measurement Data Model

Metadata e.g.

- Sample owner
- Date of measurment Kα1
- Sample stage position
- Apparatus temperature

Data e.g.

- As XML
- Raw data (text, ASCII, binary)
- Imported table
- *Link to image or raw data

Example APS Data (Large Data Example)

GitHub

docker

Generate Data at APS → Analysis Tools → Analyzed Data → Share Data And Metadata

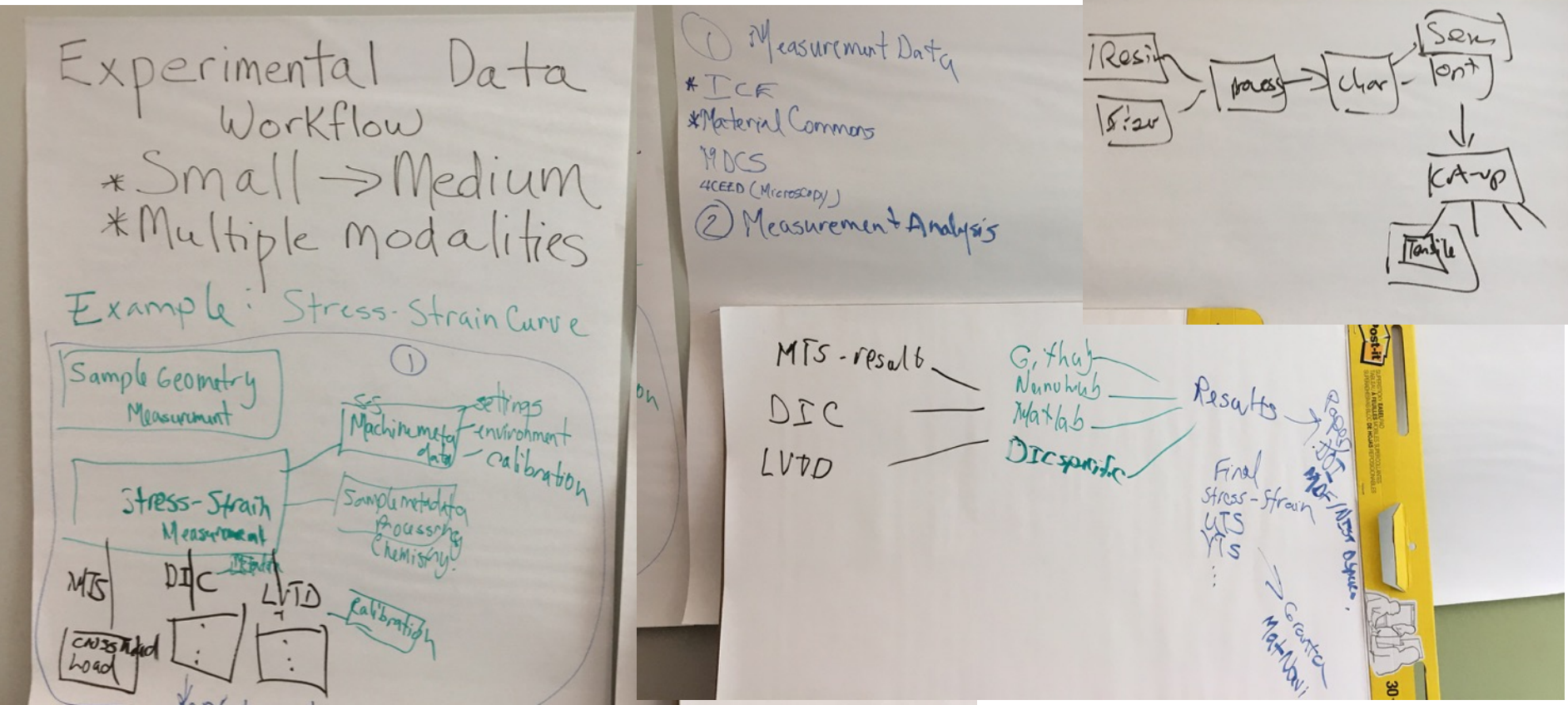
MATERIALS DATA FACILITY

Workflows

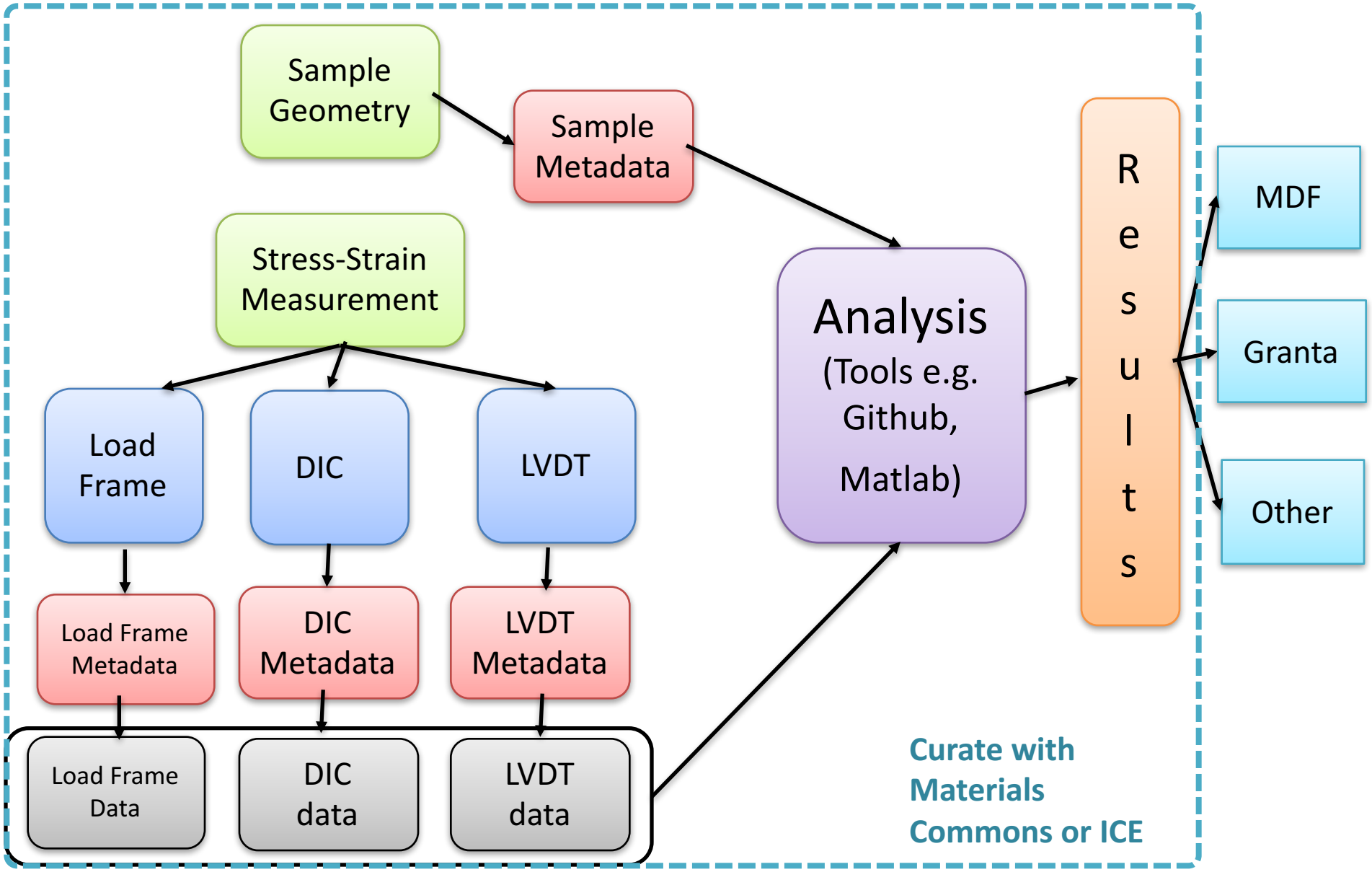
- Large Data sets: Single Point Source (e.g. APS)
- Experimental data (small to medium size), multiple source generation
- Computational Data

- Infrastructure Selection Tool

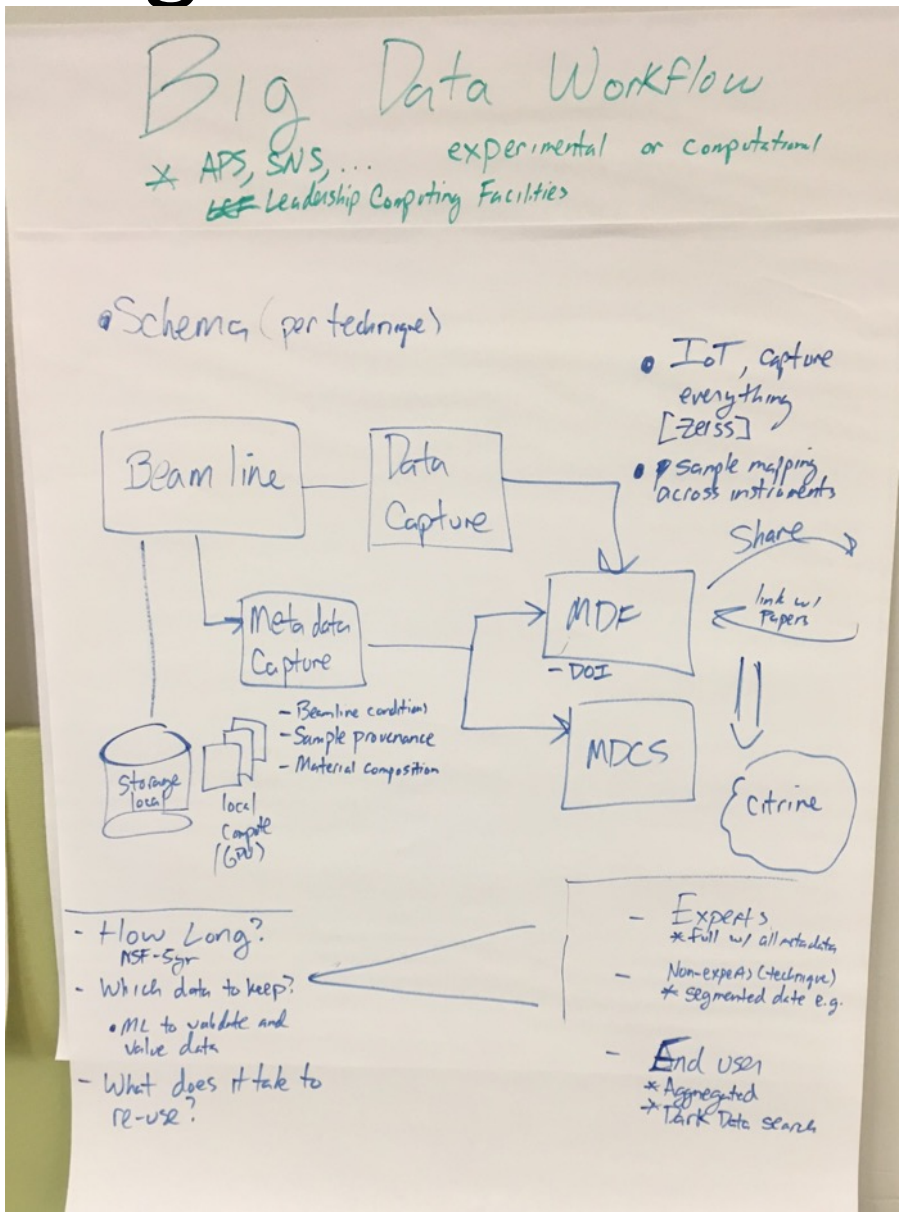
Stress-Strain Measurement



Experimental Workflow: Stress-strain Measurement



Big Data Workflow



Computational Data Workflow

- Lots of different techniques
- PhaseField modeling: no standards.
- Community standards needed

- Codes changes quickly
- Social change needed.
- FEM - more benchmark. -- more standardize

How do I select a Materials Data Infrastructure Tool?

Example: Workflow Tool Selection

Journal of Grid Computing (2006) 3: 171–200
DOI: 10.1007/s10723-005-9010-8

A Taxonomy of Workflow Management Systems for Grid Computing

Jia Yu and Rajkumar Buyya*
Grid Computing and Distributed Systems (GRIDS) Laboratory, Department of Computer Science and Software Engineering, The University of Melbourne, Melbourne, Australia
E-mail: raj@cs.mu.oz.au

Received 28 May 2005; accepted in revised form 6 December 2005

Key words: Grid computing, resource management, scheduling, taxonomy, workflow management

Abstract

With the complex resources Therefor computin building develope onomy n workflow

1. Intro

Grids [5 structure cations heteroge such as h ics, geop utilizing sets. In ments, d devices, need to b tion wor [92].

Work procedur tween pa

* Corres

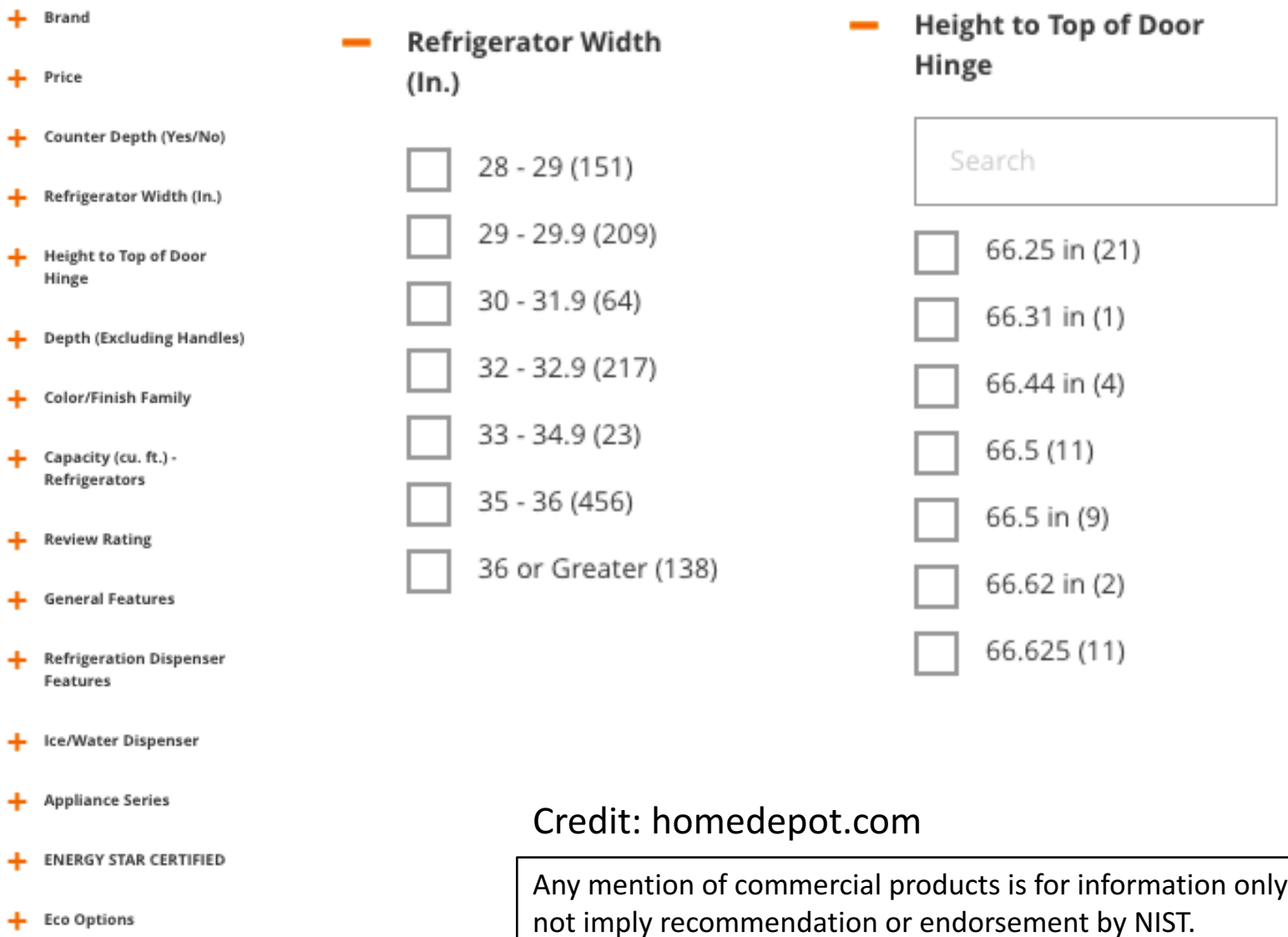
Table 2. Workflow design taxonomy mapping.

Project name	Structure	Model	Composition systems	QoS constraints
DAGMan	DAG	Abstract	User-directed • Language-based	User specified rank expression for desired resources
Pegasus	DAG	Abstract	User-directed • Language-based Automatic	N/A
Triana	Non-DAG	Abstract	User-directed • Graph-based	N/A
ICENI	Non-DAG	Abstract	User-directed • Language-based • Graph-based	Metrics specified by users
[illegible]	[illegible]	[illegible]	[illegible] • Language-based • Graph-based	N/A
GridAnt	Non-DAG	Concrete	User-directed • Language-based	N/A
GrADS	DAG	Abstract	User-directed • Language-based	Estimated application execution time

Table 2. Workflow design taxonomy mapping.

Project name	Structure	Model	Composition systems
DAGMan	DAG	Abstract	User-directed • Language-based
Pegasus	DAG	Abstract	User-directed • Language-based Automatic
Triana	Non-DAG	Abstract	User-directed • Graph-based
ICENI	Non-DAG	Abstract	User-directed • Language-based • Graph-based

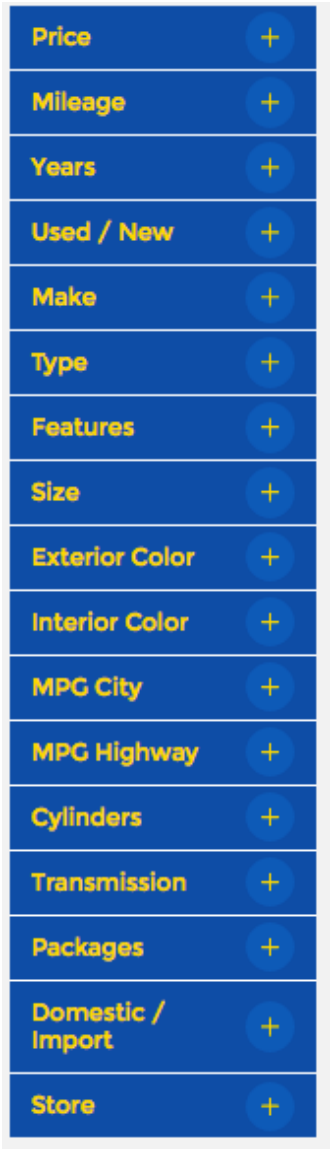
Example: Hardware Store Website

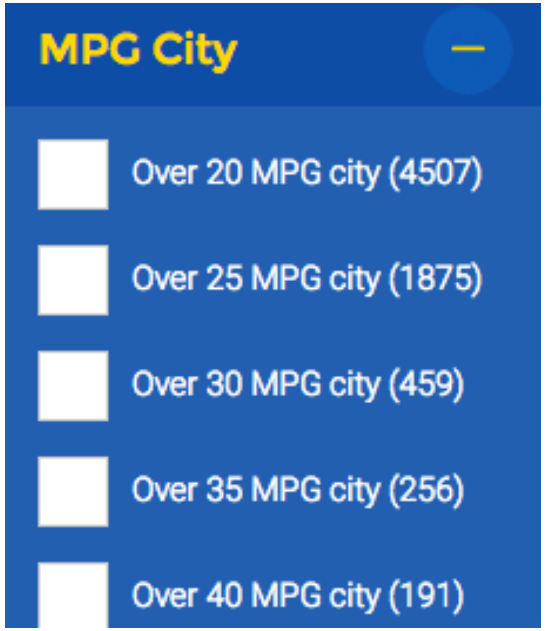


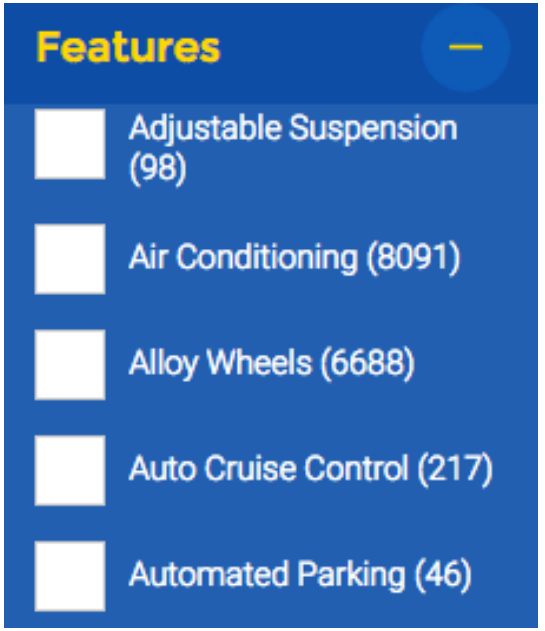
Credit: homedepot.com

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

Example: Used Car Website







Credit: carmax.com

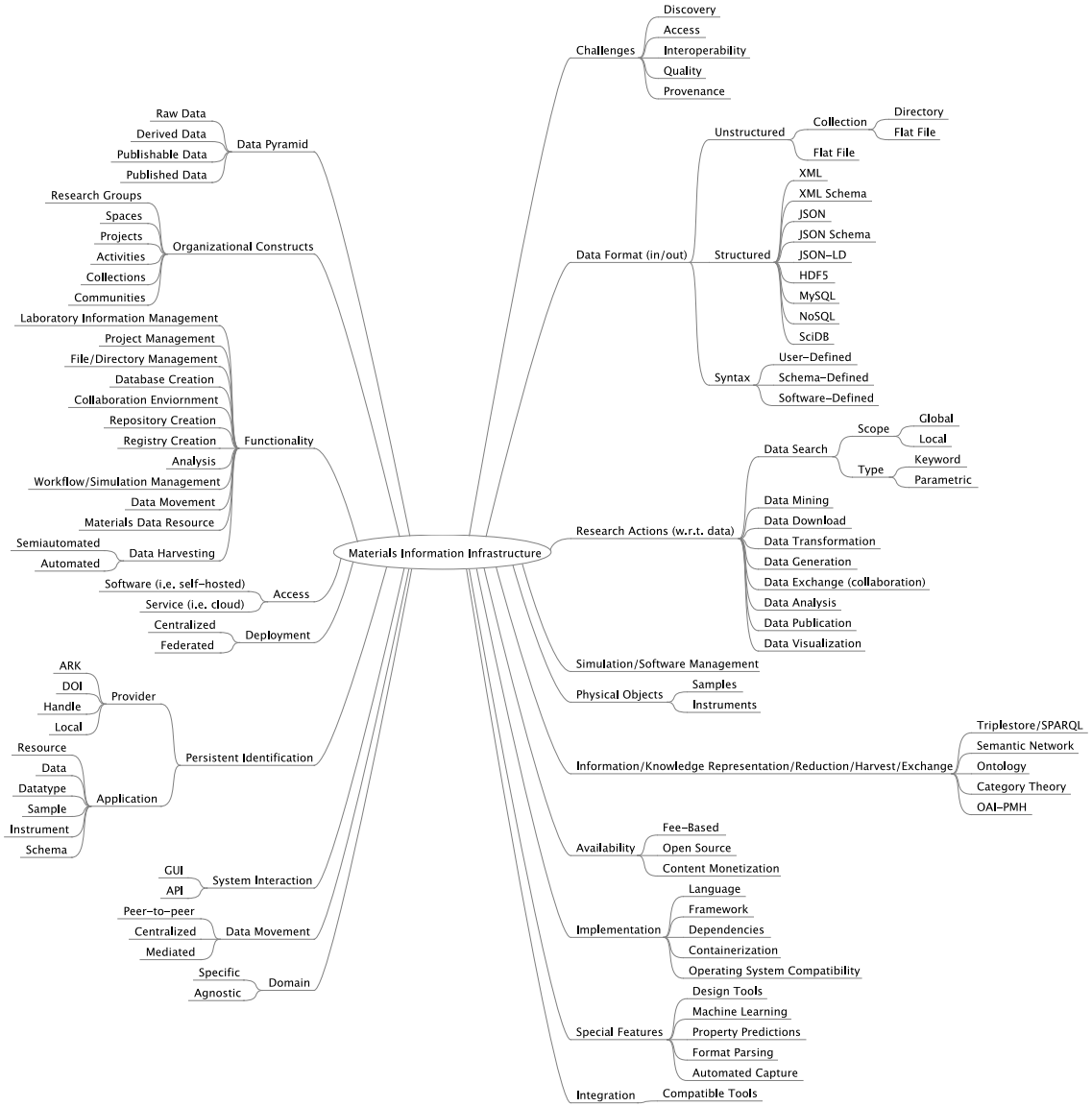
Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

Registry: Materials Data Infrastructure Tools

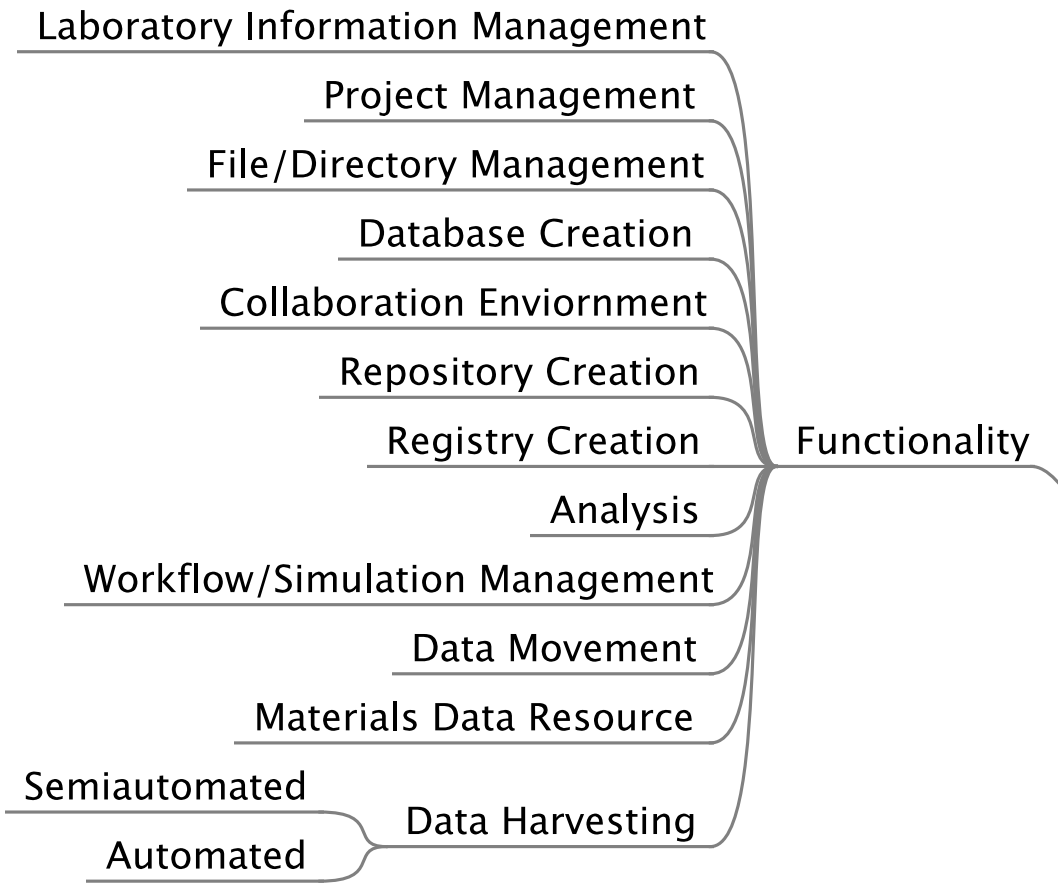
Material Types	☐ Metal ☐ Semiconductor ☐ Ceramic ☐ Polymer ☐ Biomaterial	☐ Organic ☐ Inorganic ☐ Oxide ☐ Composite ☐ Nanomaterials	☐ Superconductor ☐ Non-Specific ☐ Other	ⓘ	(recommended)
Morphology/Structures	☐ Crystalline ☐ Amorphous ☐ Fluid ☐ Quasi-periodic ☐ Bulk ☐ 2-Dimentional	☐ 1-Dimentional ☐ Film ☐ Nanotube ☐ Fiber ☐ Composite ☐ Interfacial	☐ Interphase ☐ Line Defect ☐ Point Defect ☐ Non-Specific ☐ Other	ⓘ	
Material Property Classes	☐ Optical ☐ Mechanical ☐ Thermodynamic	☐ Structural ☐ Simul		ⓘ	(recommended)
Experimental Data Aquisition Methods	☐ Electr	☐ Microscopy ectroscopy ☐ Optical Microscopy ☐ Impact Testing	☐ Indentation ☐ Dilatometry ☐ Other	ⓘ	(recommended)
al Data Aquisition ods	☐ Density Functional Theory ☐ Molecular Dynamics Simulation ☐ Numerical Simulations ☐ Multiscale ☐ Finite Element Analysis ☐ Computational Thermodynamics	☐ Statistical Mechanics ☐ Dislocation Dynamics ☐ Phase Field ☐ Crystal Plasticity ☐ Other		ⓘ	(recommended)
Sample Processing Methods	☐ Casting ☐ Annealing ☐ Vapor Deposition ☐ Milling	☐ Extrusion ☐ Pressing ☐ Exfoliation ☐ Melt Blending	☐ Polymerization ☐ Curing ☐ Evaporation ☐ Other	ⓘ	(recommended)

Need "checkboxes" for Materials Data Infrastructure Tools

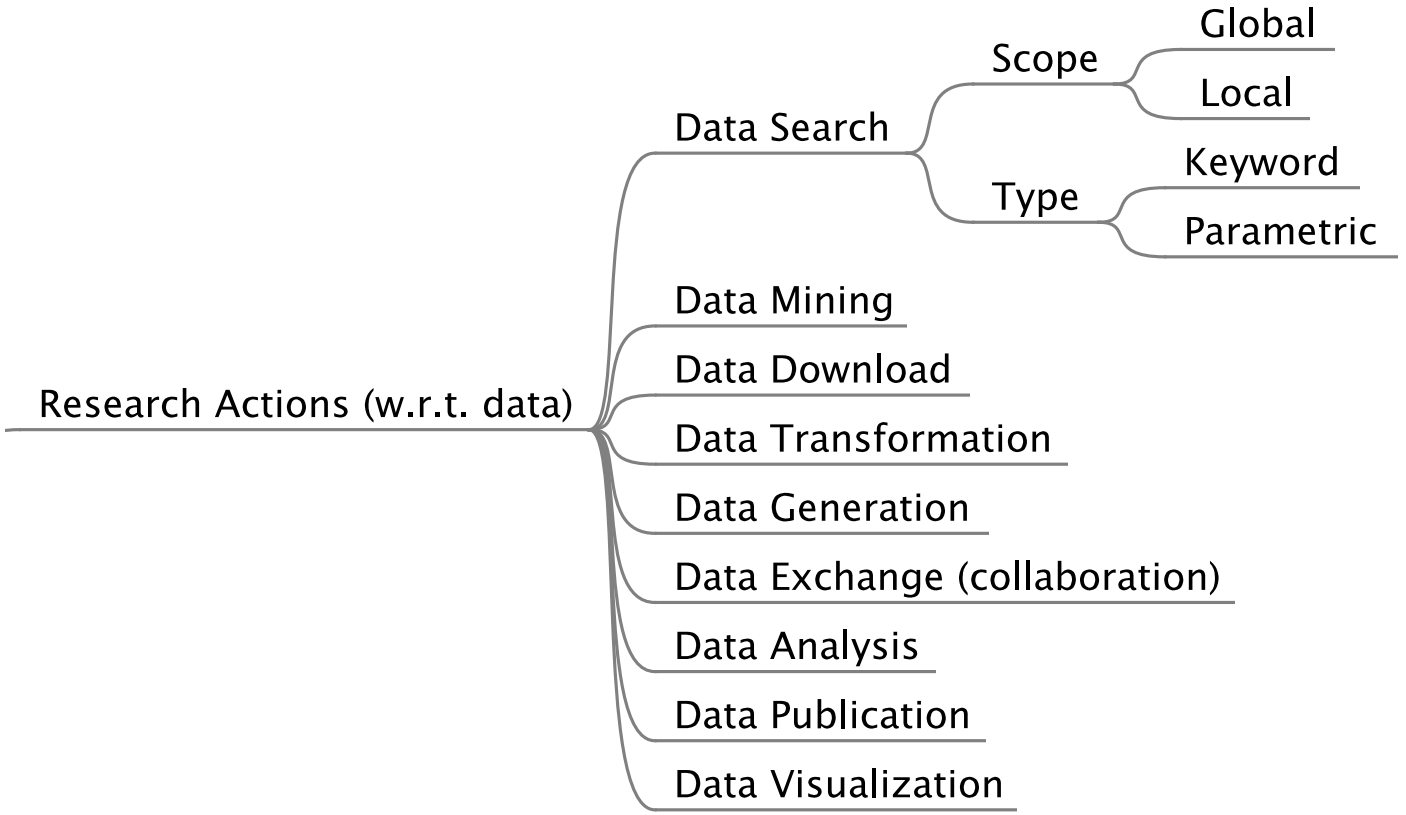
First Rough Draft Taxonomy



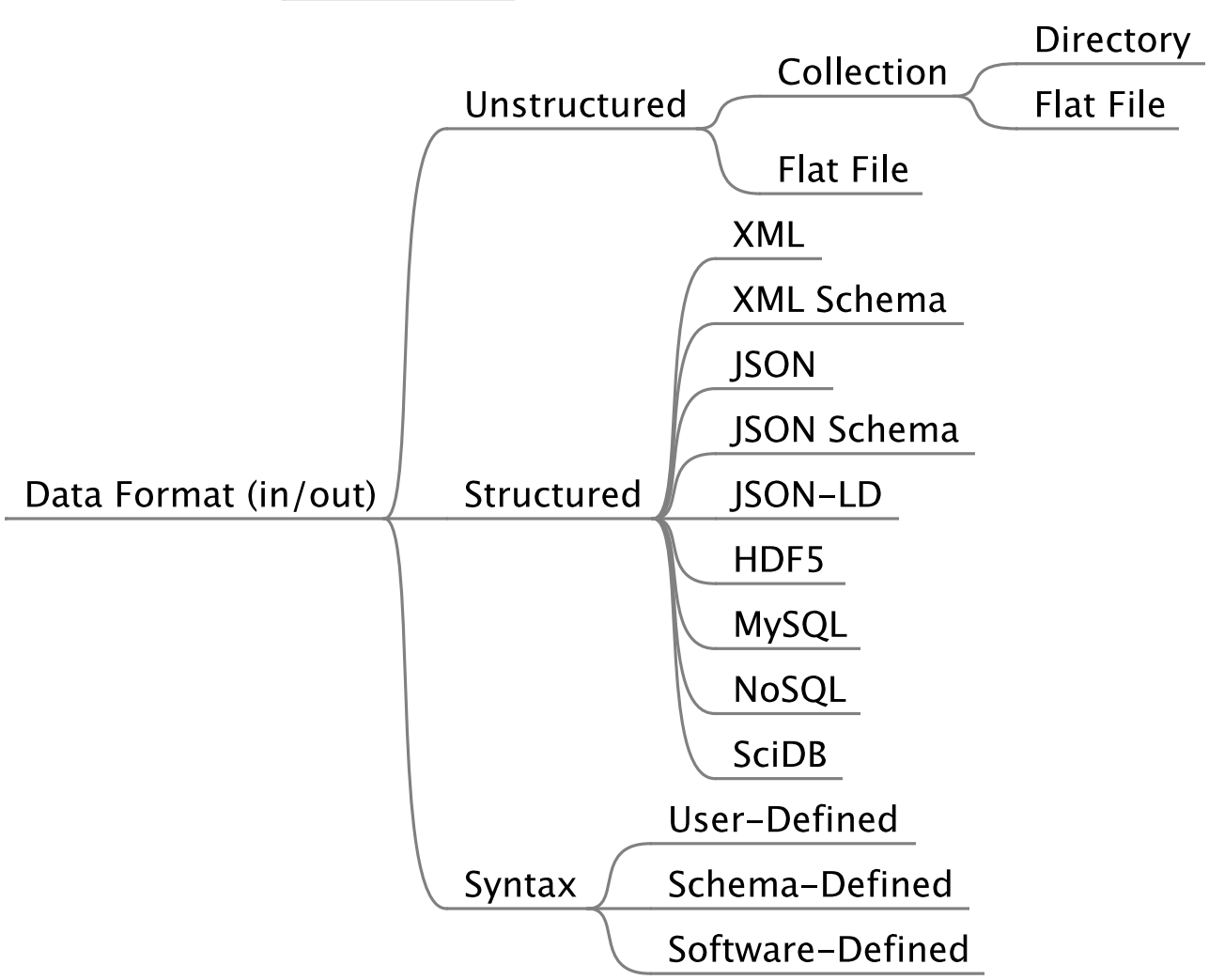
First Rough Draft Taxonomy



First Rough Draft Taxonomy



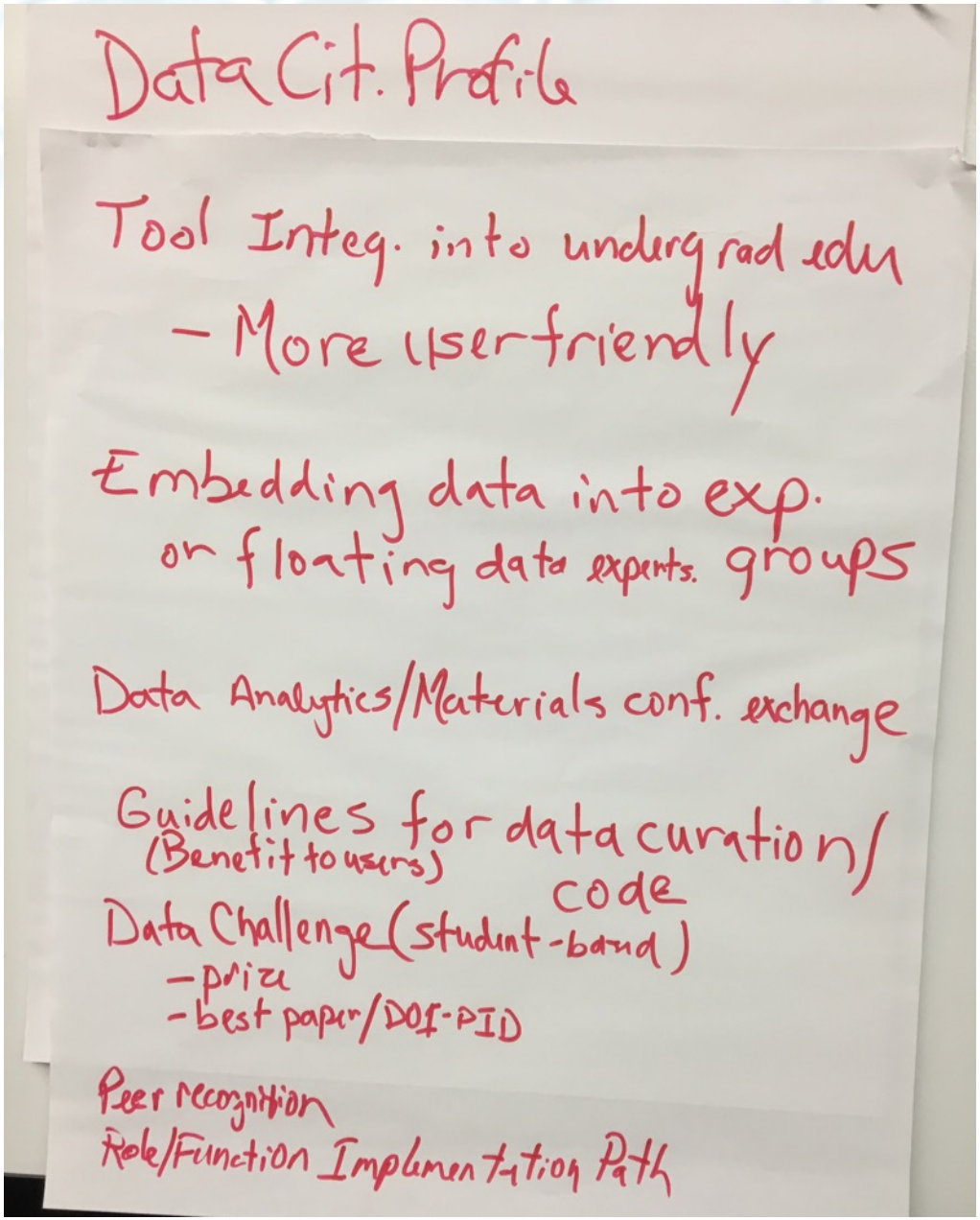
First Rough Draft Taxonomy



Notes from Summit Wrap-up Session

- Integrate tools into undergraduate education
  - Tools need to more user friendly
- Embed data experts into experimental groups
  - Alternate: floating data experts available for experimental groups.
  - Need to define skills needed for these data experts
- Encourage more conference exchanges at Data Analytics and Materials communities
- Define data curation guidelines/code
  - Benefit to users
- Data Challenge (Student)
  - Prize for data set
  - Best paper/DOI/PID
- Develop implementation path
- Improve peer recognition
- Develop data cite profile

Interest in following up with small working groups on specific issues.


Data Cit. Profile
Tool Integ. into undergrad edu
- More user-friendly
Embedding data into exp.
or floating data experts. groups
Data Analytics/Materials conf. exchange
Guidelines for data curation/
(Benefit to users)
code
Data Challenge (student-band)
- prize
- best paper/DOI-PID
Peer recognition
Role/Function Implementation Path