NIST Materials Data Informatics Efforts

C. E. Campbell, U.R. Kattner

Material Measurement Laboratory, Material Science and Engineering Division

A.A. Dima and S. Youssef

Information Technology Laboratory, Software and Systems Division

L. Bartolo

Kent State University

April 29, 2014

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

NIST Collaborators

Information Systems Group

Mary Brady, Alden Dima, Y.-S. Li-Baboud, Benjamin Long, , Philippe Dessauw, Pierre Savonitto , Sharief Youssef

Thermodynamics and Kinetics Group

Carrie Campbell, Ursula Kattner, Ben Burton, Chandler Becker, Francesca Tavazza, Zach. Trautt , Dan Wheeler, Andrew. Reid. Shenyen Li

Thermodynamic Research Group

Stephanie Hooker, Robert. Chirico, Michel Frenkel, Vladimir Diky, Chirs Munzy.

Cell Systems Science Group TN Bhat and John Elliott

Kent State, Laura Bartolo

NIST Data Efforts

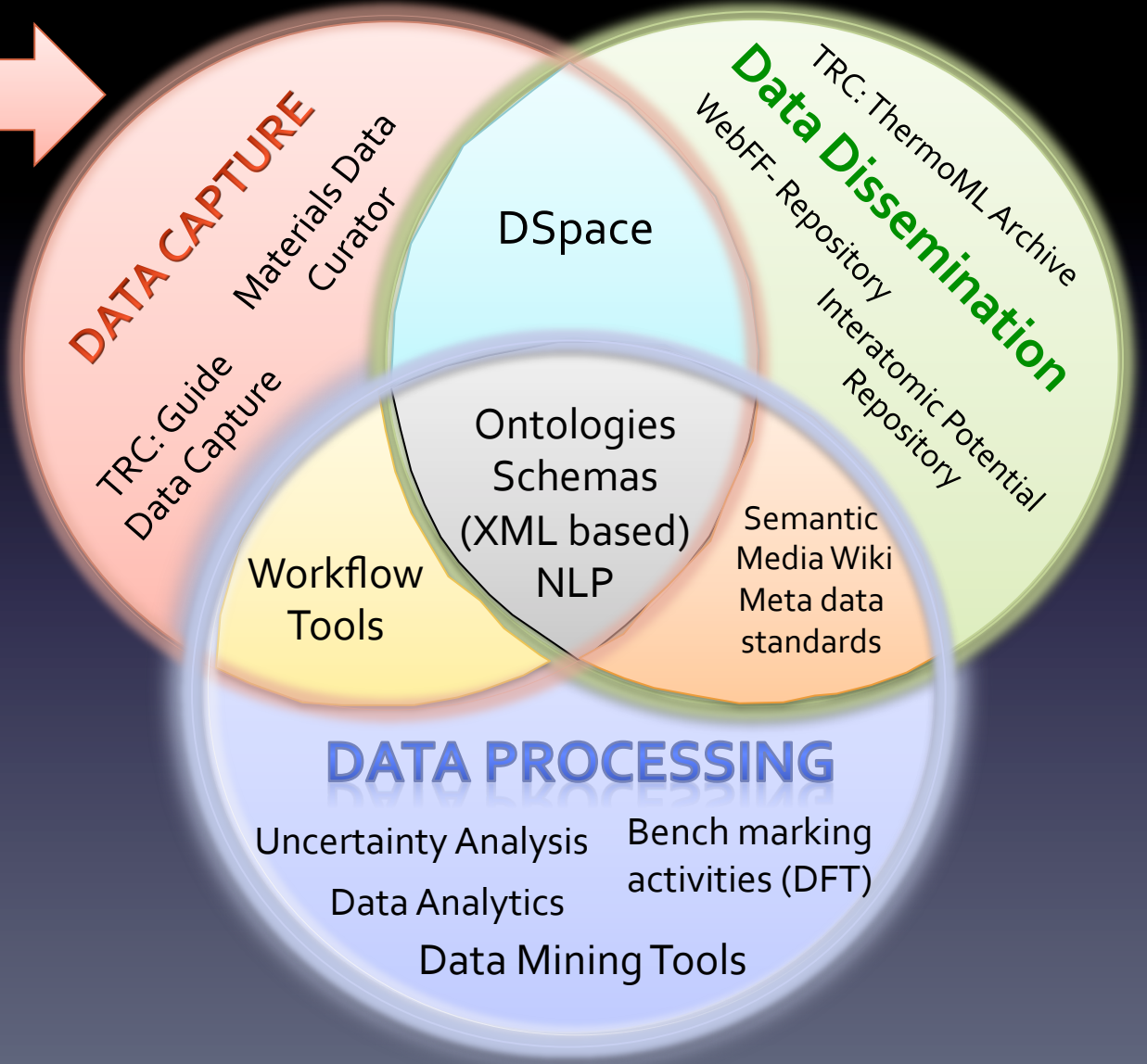
Collaborations

ASM International: Structural Data Demonstration Project

DOE/EERE Kinetics of Cast Mg Alloys

Journals collaboration

- IMMI
- *Others under discussion*



Phase-Based Property Database

- Material Property Database Exist

GRANTA MATERIAL INTELLIGENCE

MatWeb

MATERIALS PROJECT

ASM INTERNATIONAL Everything Material. The Materials Information Society

MatNavi

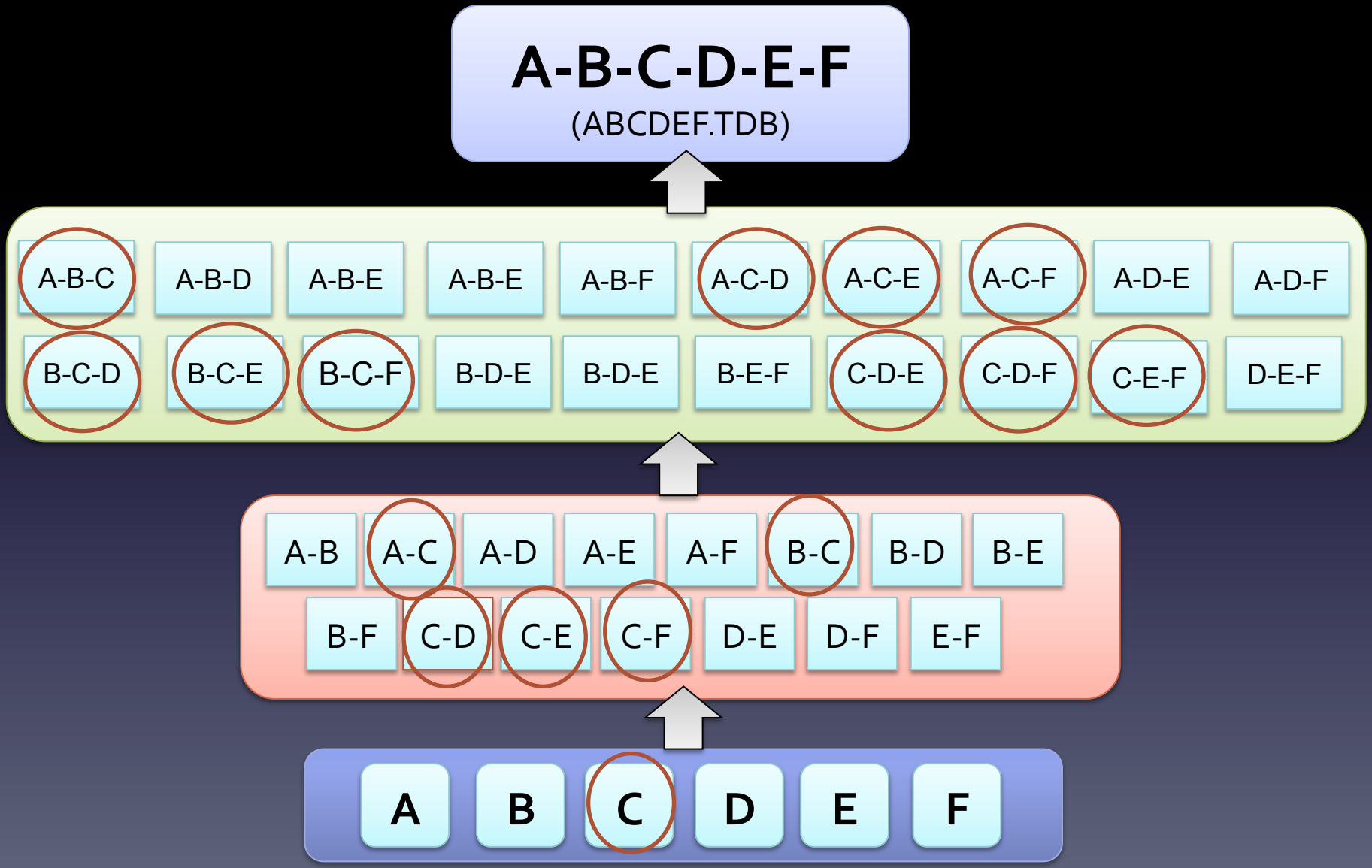
AFLOWLIB.ORG

OQMD:
An Open Quantum Materials Database

Generally, focused on engineering/design specs or first-principle calculations results.

- **Focus on phase-based properties that are needed to describe the composition, temperature, and pressure functions of a phase.**
- **Unary, binary and ternary data are primary focus.**
- **Multicomponent data are needed for validation**

Data Dependencies



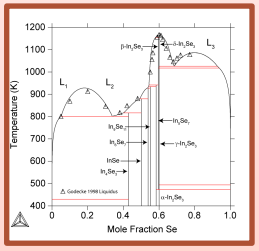
Examples of CALPHAD Data

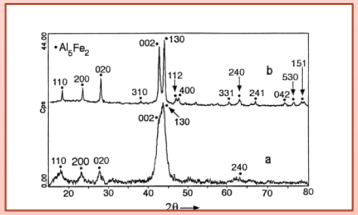
For each assessment: Evaluated data file (e.g. POP, DOP)
Functional descriptions for phase quantity (e.g. TDB)

- Emphasis on binary and ternary data to predict multicomponent properties
- Data can be experimental or computational.

1-D (Points)

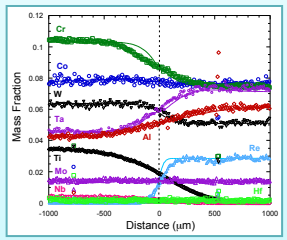
- Melting Temperatures
- Critical Temperatures (Phase Changes)
- Lattice Parameters
- Heat of Formations
- Phase fractions and compositions
- Tracer Diffusivities
- Activation energies

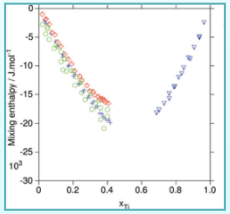


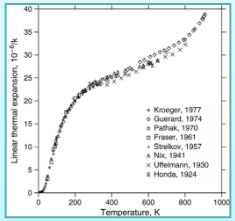


2-D (Lines)

- Composition Profiles
- Heat Capacities
- Enthalpies of mixing

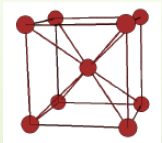


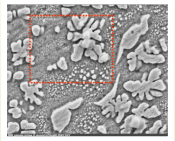


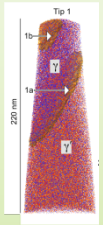


3-D

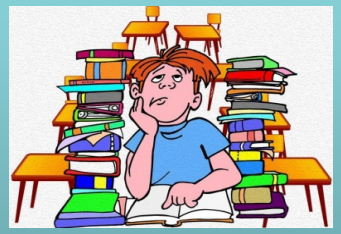
- Crystal structures
- Micrographs/Morphologies
- 3-D Atom probe Tomography





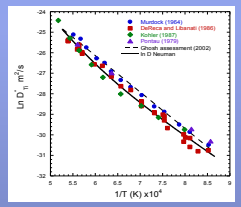


Current Assessment Process

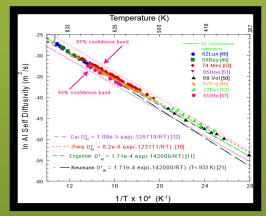


Search data

Digitize Data



Plot data



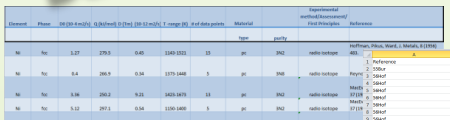
Reference values

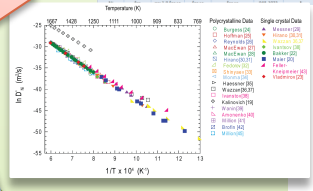
Future

Search infrastructure

CALPHAD user needs data for A-B-C system,

Computational and Experimental Databases and Repositories





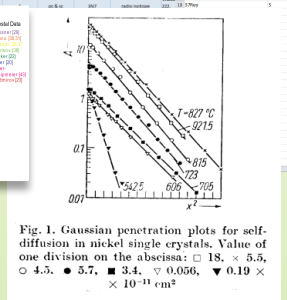
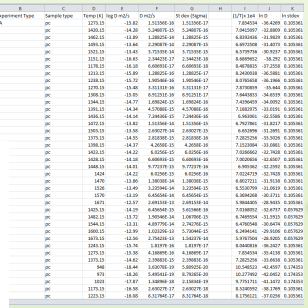


Fig. 1. Gaussian penetration plots for self-diffusion in nickel single crystals. Value of one division on the abscissa: □ 18, × 5.5, ○ 4.5, ● 5.7, ■ 3.4, ▽ 0.056, ▼ 0.19 × $\times 10^{-11}$ cm²



Federated System

Interfaces for input of computational and experiment data

Search returns available data And points out missing data

EXAMPLE OF DIFFERENT TYPES OF DATA USERS: DIFFUSION DATA

Industry user just wants to the know the diffusion coefficient at given temperature for a given material.

Calphad user who wants to complete a new assessment

Advanced expert wants to understand the diffusion mechanism

Materials Data Infrastructure

D = 2.16e-15 m2/s
$D = D_0 \exp(-Q/RT)$

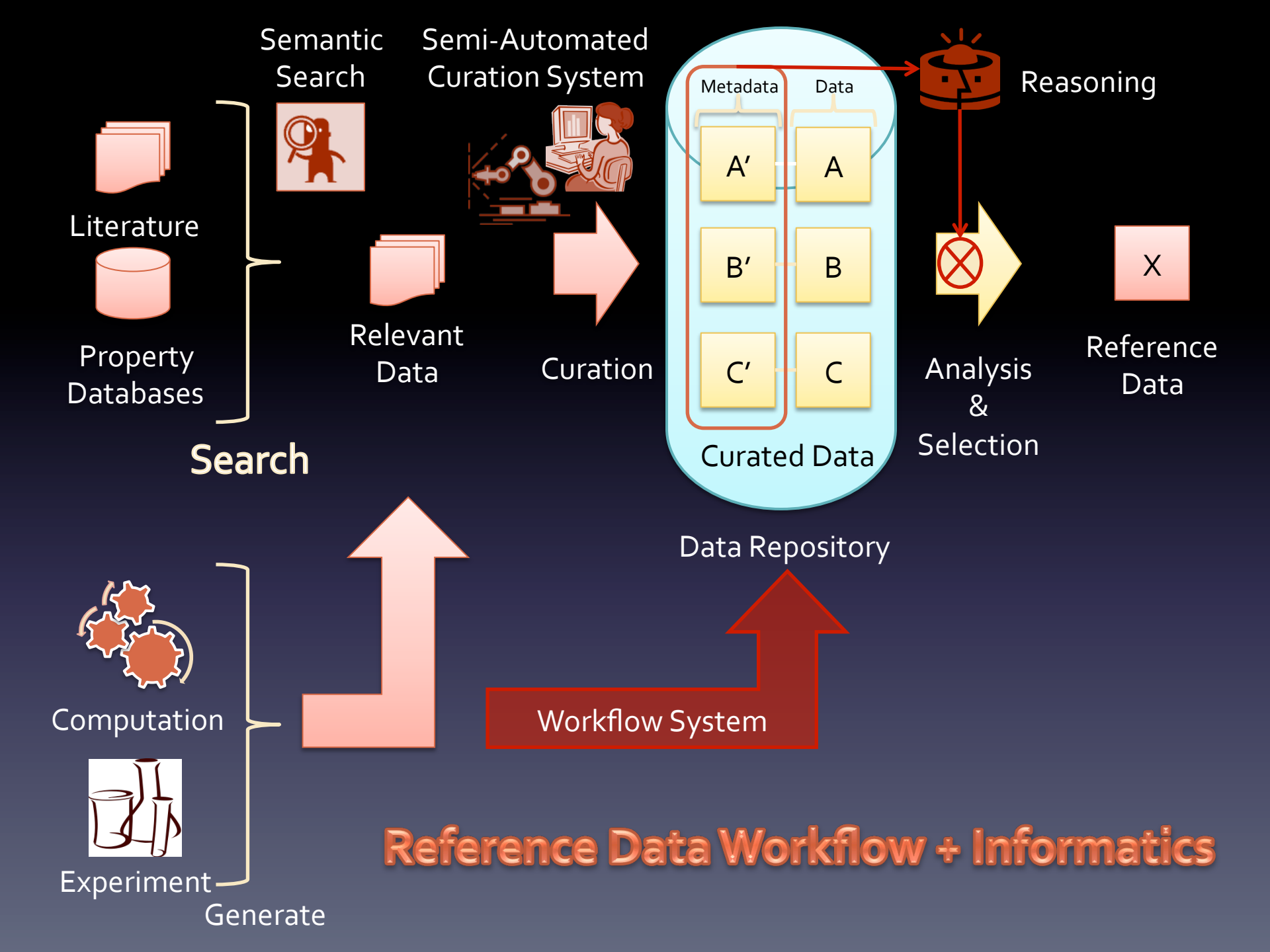
Diffusion Coefficients
Diffusion Couple
Composition Profiles

Composition Profiles
Raw Data (i.e Intensity vs Distance)

Data are diverse

Data are semi-structured

Need complete data sets


Semantic Search
Semi-Automated Curation System
Literature
Property Databases
Search
Relevant Data
Curation
Metadata
Data
A′
A
B′
B
C′
C
Curated Data
Data Repository
Reasoning
Analysis & Selection
X
Reference Data
Computation
Experiment
Generate
Workflow System
Reference Data Workflow + Informatics

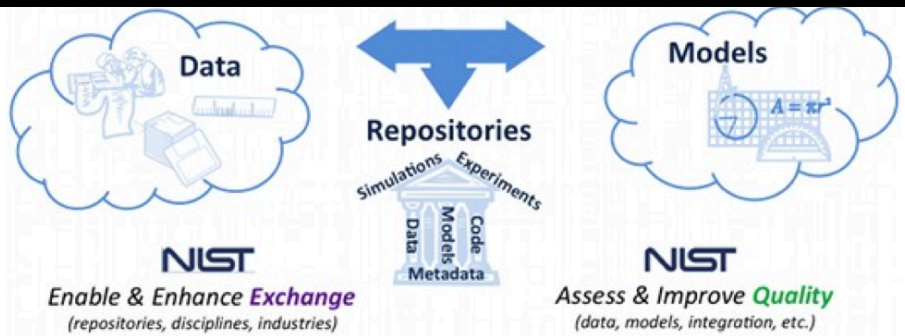
Two Biggest Problems

- **Need to collect data and organize**
  - Data are diverse
  - Data are semi-structured
  - Data are incomplete

- **Need to make the data useful to a wide variety of users.**

D-SPACE CUSTOMIZED FILE REPOSITORY

nist.matdl.org



Login

NIST Repositories → Community List

NIST Repositories

The National Institute of Standards and Technology is establishing essential data exchange protocols and mechanisms for widespread adoption to ensure quality materials data and models and to foster data sharing and reuse.

- **NIST Computational File Repository**
  - CALPHAD Assessments
  - First Principles Phase Stability (FPPS) Files
  - Interatomic Potentials
- **NIST Experimental Data Repository**
  - Diffusion Data
  - Phase Equilibria
  - Texture Data
- **NIST Structural Materials Data Demonstration Project ASM**
  - 6061 Aluminum Alloys: Diffusion Data
  - 6061 Aluminum Alloys: Handbooks, Journal Articles, and Technical Papers
  - 6061 Aluminum Alloys: Mechanical Properties
  - 6061 Aluminum Alloys: Phases, Phase Diagrams
- **TMS Springer Integrating Materials and Manufacturing Innovation (IMMI)**
  - General Articles
  - Thematic Series on Three-Dimensional Materials Science

Search NIST Repositories

Computational
CALPHAD
First Principles

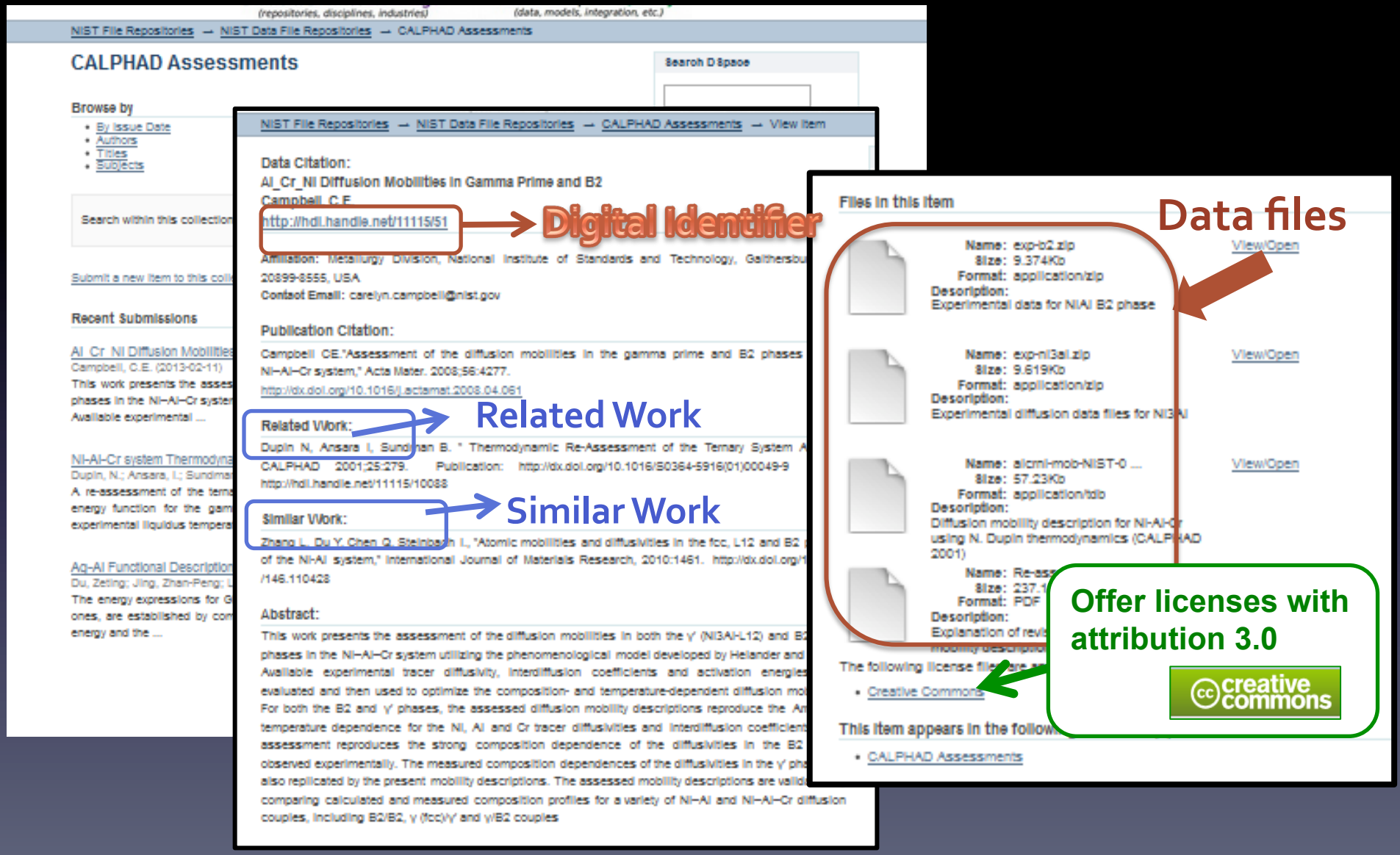
Browse

All of NIST Repositories

Experimental
Phase Equilibria
Diffusion

Login
Register

Sample Entry



https://materialsdata.nist.gov/dspace/xmlui/

NIST Repositories → Community List

NIST Repositories

The National Institute of Standards and Technology is establishing essential data exchange protocols and mechanisms for widespread adoption to ensure quality materials data and models and to foster data sharing and reuse.

- **Computational File Repository**
  - Atomistics Simulations
  - CALPHAD Assessments
  - First Principles Simulations
  - Other Computational Methods
- **Experimental Data Repository**
  - Diffusion Data
  - Mechanical Properties
  - Other Experimental Data
  - Phase Equilibria and Thermodynamic Data
- **NIST/DOE-EERE Advanced Automotive Cast Magnesium Alloys**
  - A systematic multiscale modeling and experimental approach to protect grain boundaries in magnesium alloys from corrosion
  - Corrosivity and Passivity of Metastable Mg Alloys
  - Dealloying, Microstructure and the Corrosion/Protection of Cast Magnesium Alloys
  - High-Throughput Study of Diffusion and Phase Transformation Kinetics of Mg-Based Systems
  - In-situ Investigation of Microstructural Evolution During Solidification and Heat-Treatment in a Die-Cast Magnesium Alloy
  - Phase Transformation Kinetics and Alloy Microsegregation in High Pressure Die Cast Magnesium Alloys
- **NIST Thermodynamics and Kinetics Test Space**

Search NIST Repositories

Go

Advanced Search

Browse

All of NIST Repositories
Communities & Collections
Subjects
Titles
Authors

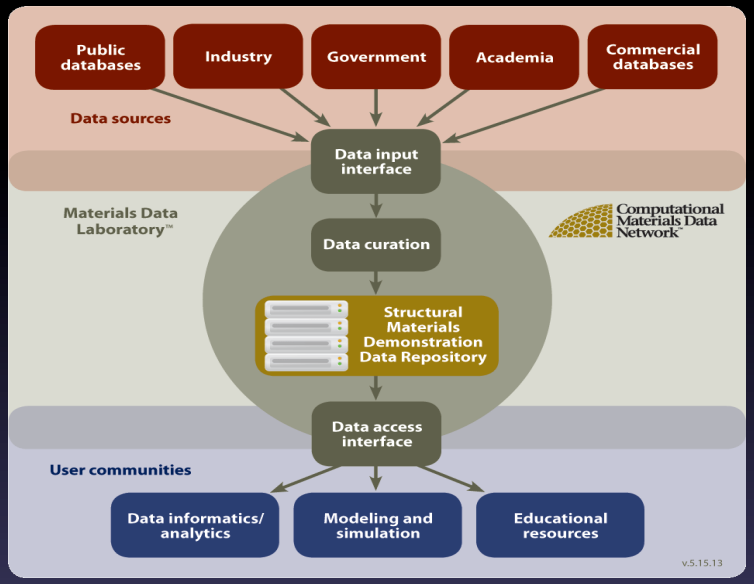
My Account

Login

To be added soon:
Huesler First Principle Simulations (DFT Magnetic Properties)

Theme by NIST MatDL

Structural Data Demonstration Project: Al6061



March 2014: Phase 1 release.
June 2014: Phase 2 release.
Dec 2014: Project Completion

Goal: Establish well-pedigreed and curated demonstration datasets for non-proprietary metallic structural materials data over all length scales.

NIST's role

- Provide data schemas and meta-data formats for diffusion and phase equilibria data.
- Provide sample diffusion and phase equilibria data for the Al-Mg-Si system.
- Use expanded TRC Guided Data Capture program with available binary and ternary phase equilibria literature
- Expand use and implementation of DSpace Repository
- Link with developing ontology and semantic web tools

Start of process

1. Experiment Planning (Article Authors)

2. Article Preparation and Submission (Article Authors)

ThermoLit

Journal Support Websites

NIST

A

NIST Literature Report

End

3. Journals (Editors)

4. Traditional Peer Review

End

5. Decision

Approve (not "Accept")

6a. In-House Data Capture (Student Associates)

6b. Guided Data Capture

6c. ThermoData Engine

NIST

B

NIST/TRC SOURCE Database

NIST Data Report

7. Journals (Editors)

7a. Revisions (Authors)

End

8. Final Decision

Accept

Publish

After publication

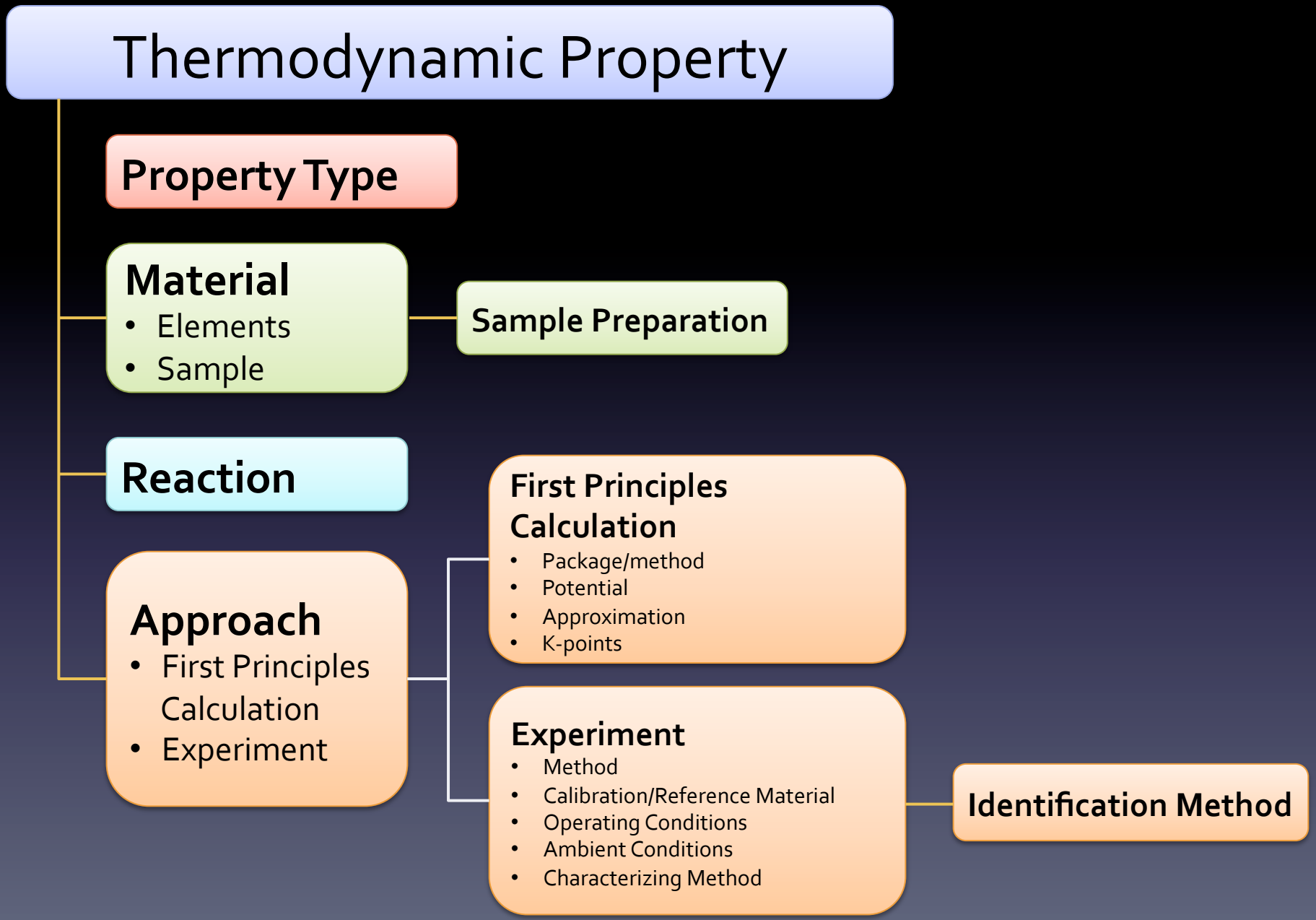
9. ThermoML Archive of published experimental data

NIST

C

10. Data Users

End of process


Thermodynamic Property
Property Type
Material
• Elements
• Sample
Sample Preparation
Reaction
Approach
• First Principles Calculation
• Experiment
First Principles Calculation
• Package/method
• Potential
• Approximation
• K-points
Experiment
• Method
• Calibration/Reference Material
• Operating Conditions
• Ambient Conditions
• Characterizing Method
Identification Method

(D) Sample Description			
1	Sample series Composition Composition unit Composition error	☒ Series ☐ Single 0.95Al-xMg-(0.05-x)Si ☒ Mass fraction ☐ Mole fraction not reported	xMg in Table
2	Form Size/Dimension	☒ Chunk ☐ Foil ☐ Powder 0.3 g	
3	Sample preparation	☒ Casting ☐ Sintering ☐ Mechanical Alloying ☐ Rapid cooling	
4	Mechanical history	☐ Rolling ☐ Extrusion ☐ Milling ☒ Forming	
5	Thermal history	☒ Annealing	
6	Sample analysis	☐ EPMA ☐ EDS ☐ WDS ☐ AES ☐ XRF ☐ Chemical	

Casting

1. Melted in high frequency furnace under argon gas
2. Casted into 8 mm diameter 100 mm length copper mold

Forming

1. Hammered to 6.5 mm diameter

Sample Preparation Description

Annealing	
Container	not reported
Atmosphere	not reported
Temperature, Time	550 °C, 5 days
Quench	not reported

Facilitating Materials Ontology Development via NLP and Machine Learning

Starting with a corpus of 5893 PDFs

MS&T

TMS 2009, 2010, 2011, 2012

NIST Diffusion Data Center

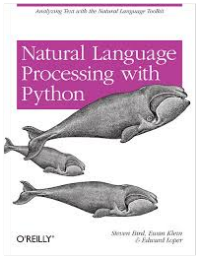
www.matscitech.org www.tms.org patapsco.nist.gov/diffusion

1 → 2 → 3 → 4

OCROPUS

thoonk.py

redis



gensim

python

radimrehurek.com/gensim

www.nltk.org

Focus of current development

5893 PDFs - 5.8 GB

63K pages

3.8M lines

53GB

3.4M lines extracted

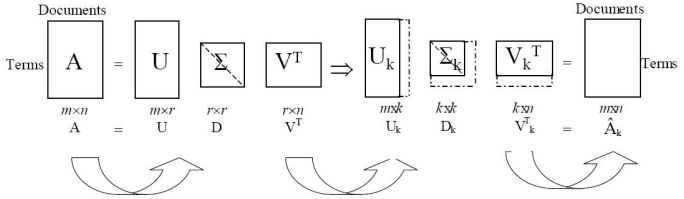
20M tokens

Extract text[1,2]

temperatur	69079
materi	53673
diffus	49743
process	48594
result	43713
surfac	33899

Identify key stems by frequency[3]

Use Latent Semantic Indexing to group similar concordance entries[4]



http://liqiangguo.wordpress.com/2011/06/09/latent-semantic-analysis/

Generate concordances for key stems[3]

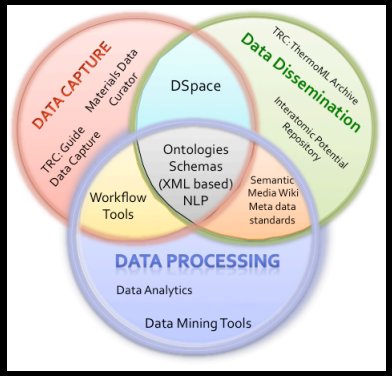
```
tivity method. Lattice diffusion coefficients and grain boundary diffusion coe
nts and grain boundary diffusion coefficients were measured. Plots of the data
r all the alloys , the diffusion coefficients at tempe- ratures 1150 'C and hi
, were only accepted. From these coefficients , the parameters of temperature
 those alloys the self-diffusion coefficients of which were also determined at
ting intercrystallite diffu sion coefficients , the following rel ation coeffi
 place. Values of bulk diTfusion coefficients at temperatures below 1j50 ' C
```

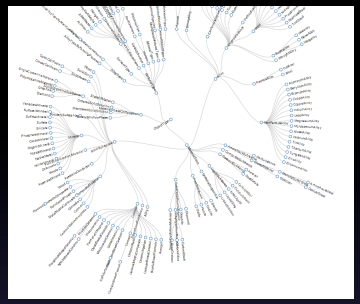
Alden Dima - 2/24/2014

STOP

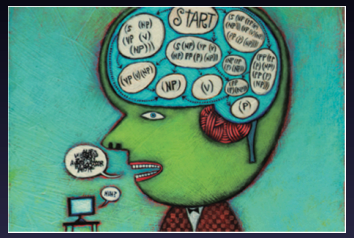
Go to Alden and Youssef Presentations

Integration Tools:
How to Gain Knowledge from Unstructured Data





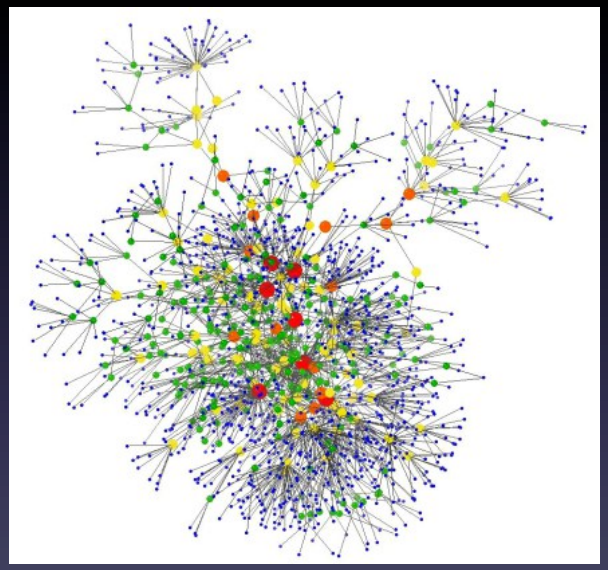
- Ontologies
- Natural Language Processing
- Semantic Web/ Semantic Media Wiki



Semantic Web

What is an Ontology and Why Build One?

- An ontology defines concepts and the relationships between concepts for a given community.



Purpose:

- To share a common understanding of the structure of information among people or software tools.
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

http://www.ksl.stanford.edu/Ontology 101

An ontology model is dynamic. The domain knowledge changes.

Informatics for Ontology Acceleration

Significant Automation Possible

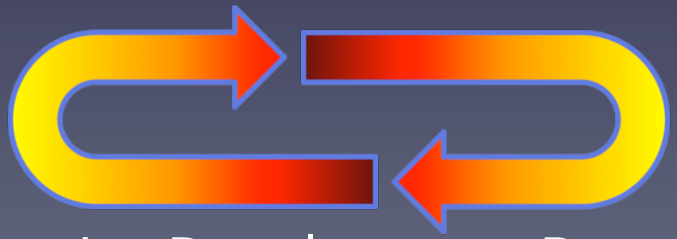
Corpus Generation
Page Extraction
Text Extraction
Natural Language Processing

Literature

Ontology

Search
Visualization
Modeling

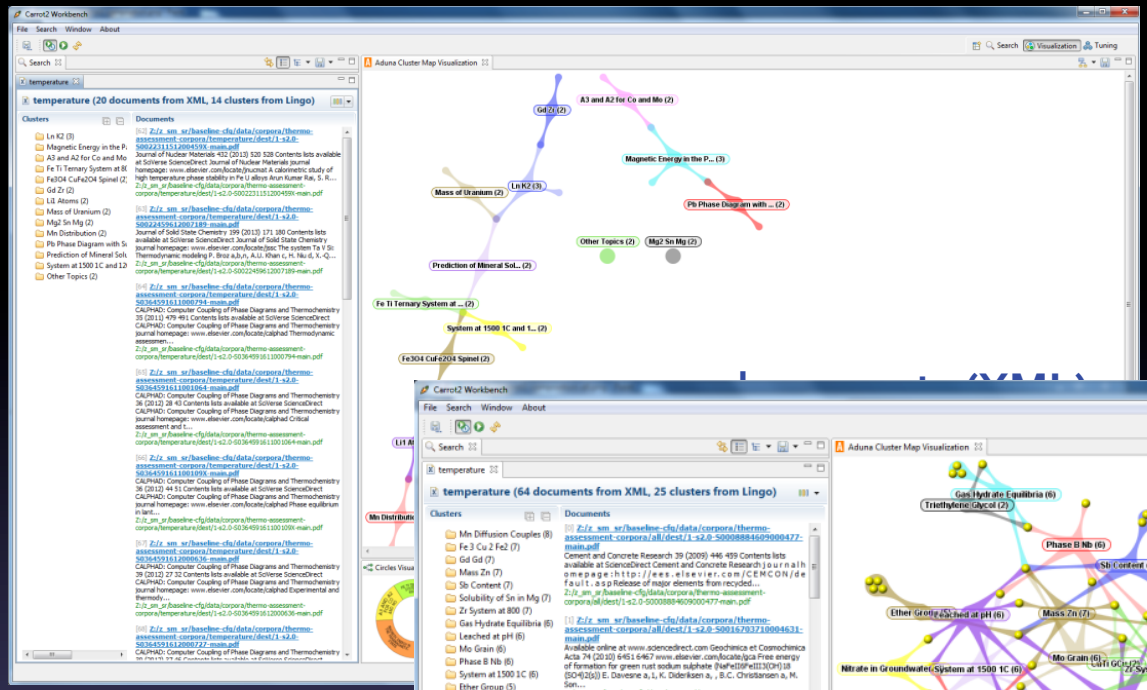
Requires Human Interaction

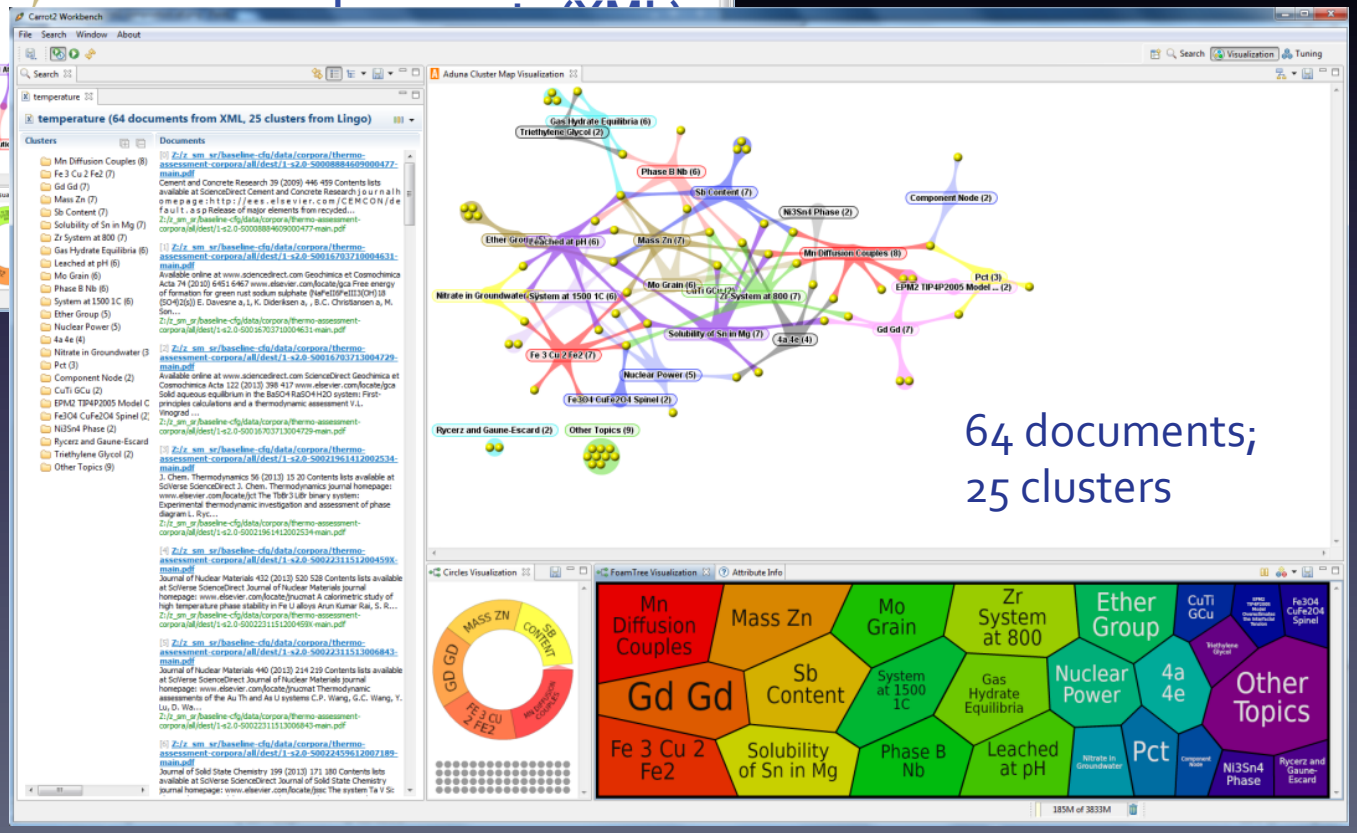


Iterative Development Process

NLP identifies the concepts needed for the ontology

Ontology Acceleration – Clustering



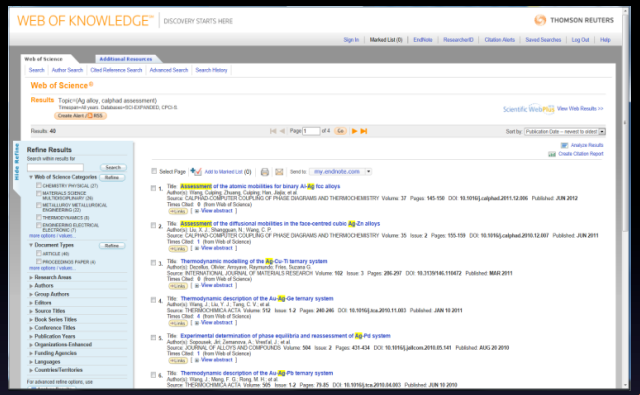


Search term:
"Temperature"

Carrot²

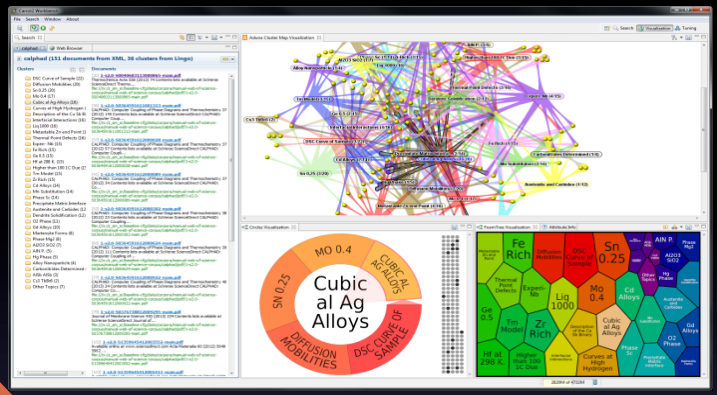
64 documents;
25 clusters

Informatics Task (research): Research Ag Alloys for CALPHAD Assessment



Informatics Task

Research Ag Alloys + CALPHAD Assessment



Traditional Approach

Web of Science Search (Ag alloy, CALPHAD assessment)

- Term-only criteria
- Relies on pre-processed, structured corpora

Our Approach

Carrot2 Search (Ag alloy, CALPHAD assessment)

- Domain-specific representation
- Tunable clustering
- Can process less structured, noisy corpora
- Thematic visualization & browsing

Semantic Medline (SM)

1. What is it?

1. SM is a semantic layer on top of PubMed
2. SM supports semantic applications of semantic search, semantic visualization, and automatic summarization

2. Collaborators: NIH/NLM

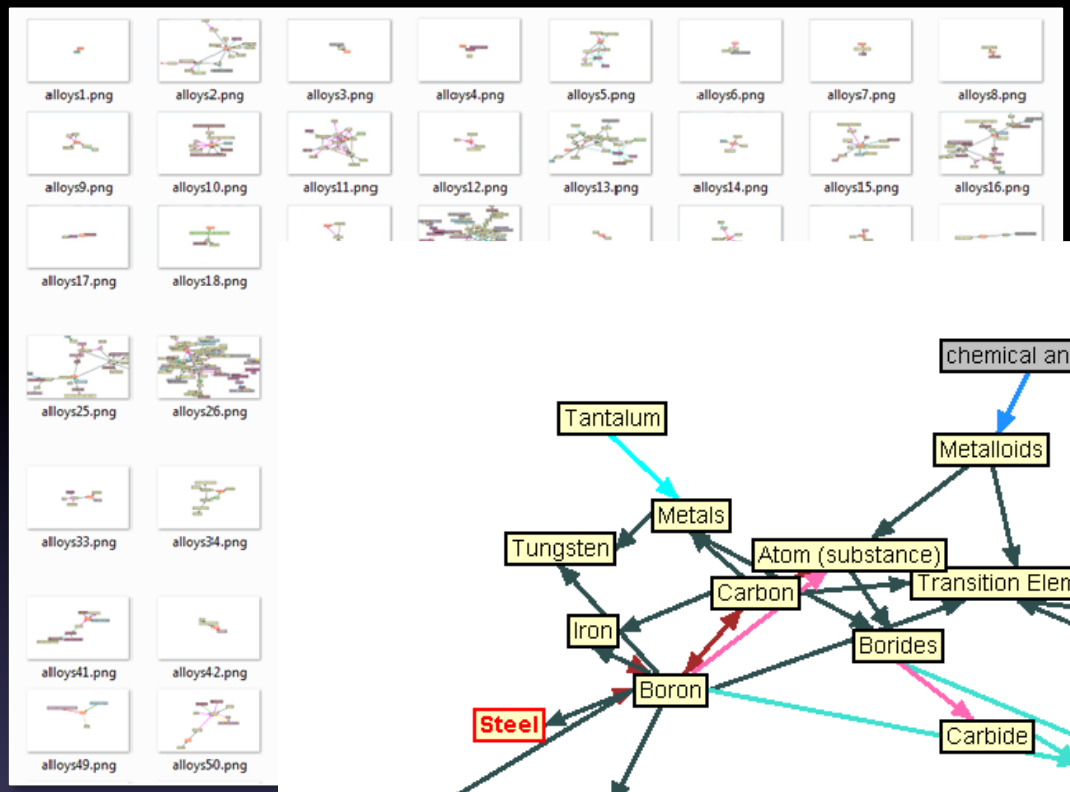
3. How does it work?

1. Supported by DB of corpus-based semantic information
2. Uses search to access semantic-info which supports semantic application operations

4. GOAL: *Retarget* SM from **biomedical** domain to **MS** domain

"Semantic Matline"

Ontology Acceleration – Semantic Medline

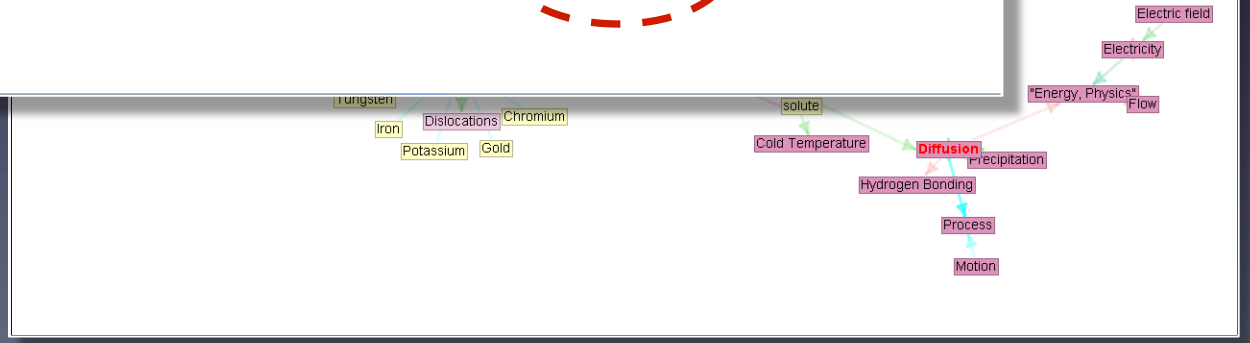


Iterative Process for Ontology Evolution

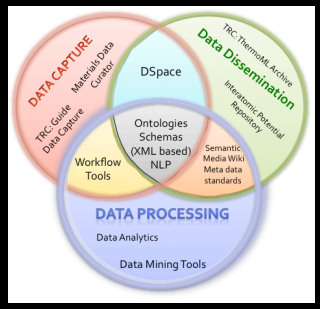
1. Start with Existing Ontology

rpus

y Coverage

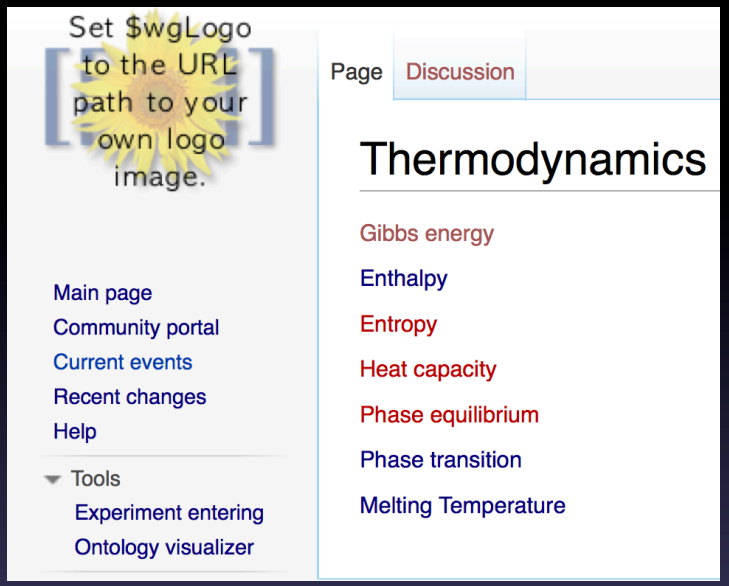


Semantic Web

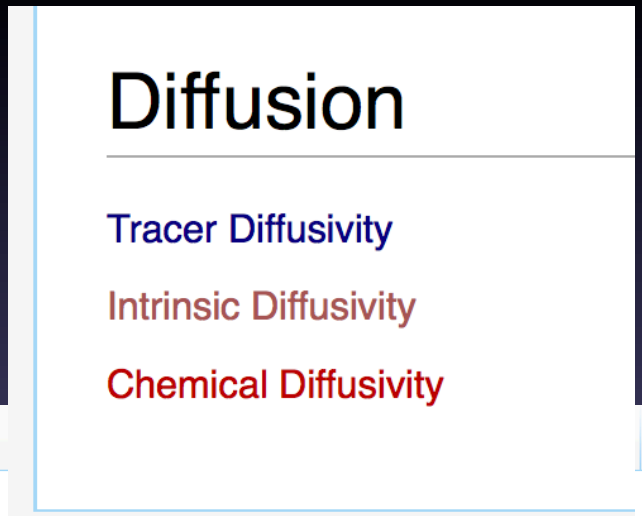


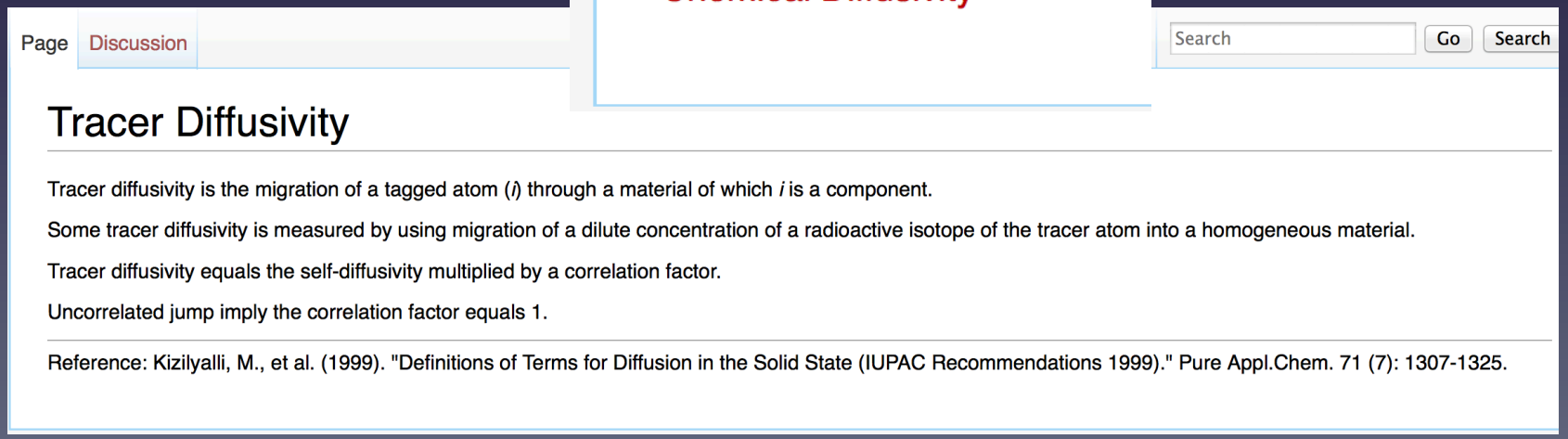
- Creates a "Web of Data" that can be machined processed
- Provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.
- Elements of the Semantic Web
  - XML: elemental syntax for content structure within documents
    - Examples (ThermoML, MatML, UnitsML)
  - XML Based- Schemas : a language for providing and restricting the structure and content of elements contained within XML documents.
  - RDF: a simple language for expressing data models

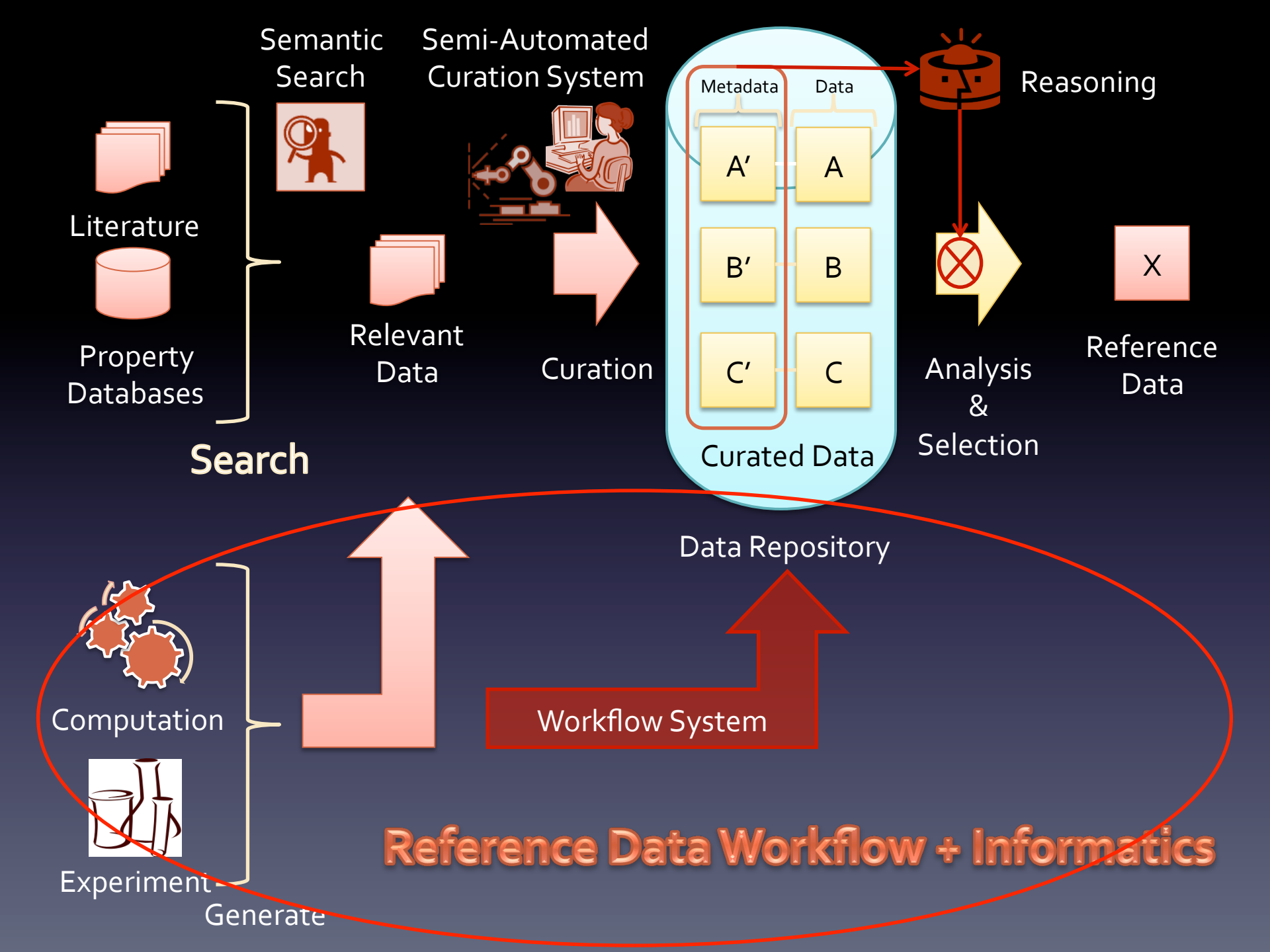
Linking Sematic Media Wiki to D-Space: Better Searching



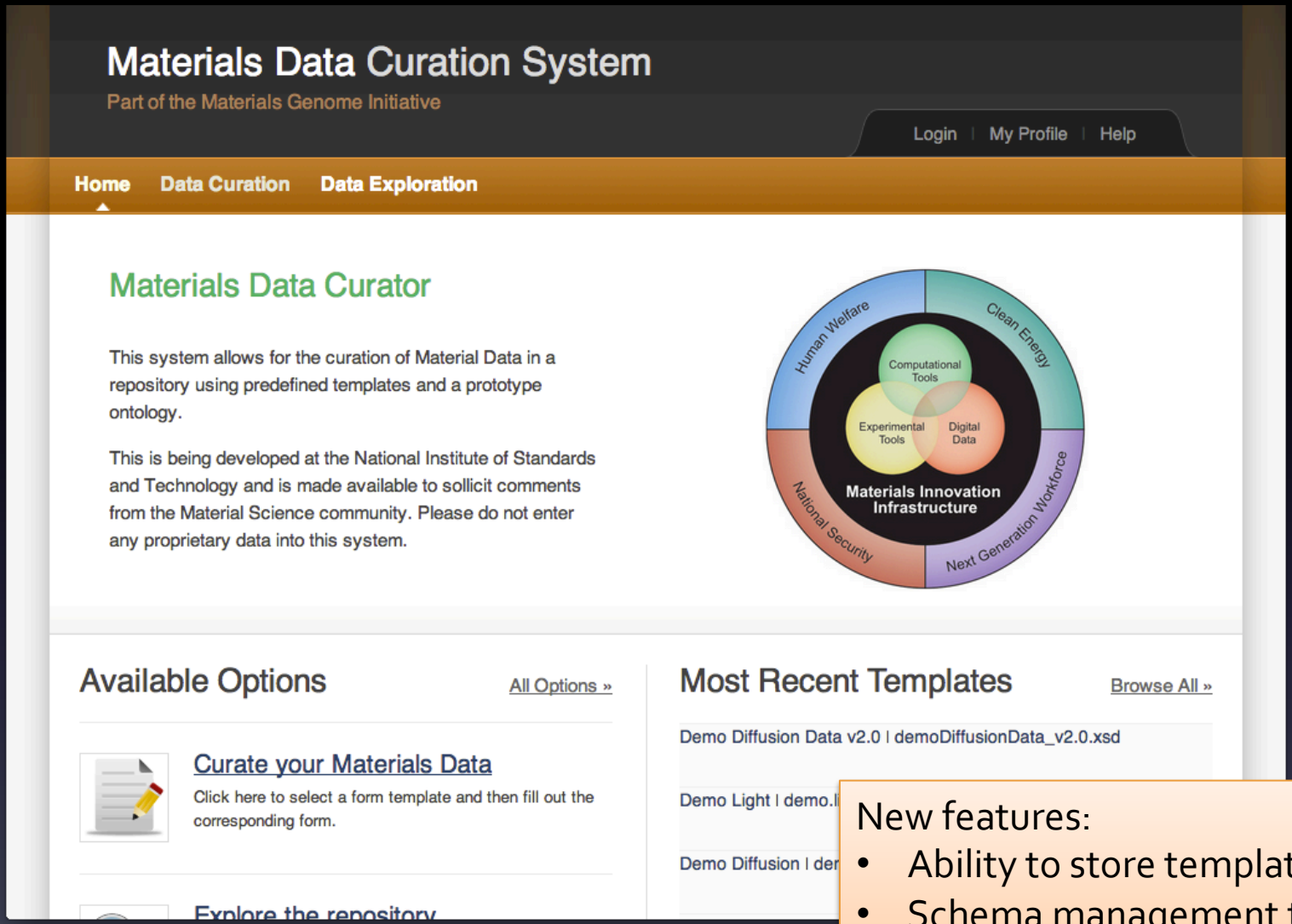
Beginning an effort to define critical concepts using controlled natural language






Semantic Search
Semi-Automated Curation System
Literature
Property Databases
Search
Relevant Data
Curation
Metadata
Data
A′
A
B′
B
C′
C
Curated Data
Reasoning
Analysis & Selection
X
Reference Data
Data Repository
Computation
Workflow System
Experiment
Generate
Reference Data Workflow + Informatics

Data Curation

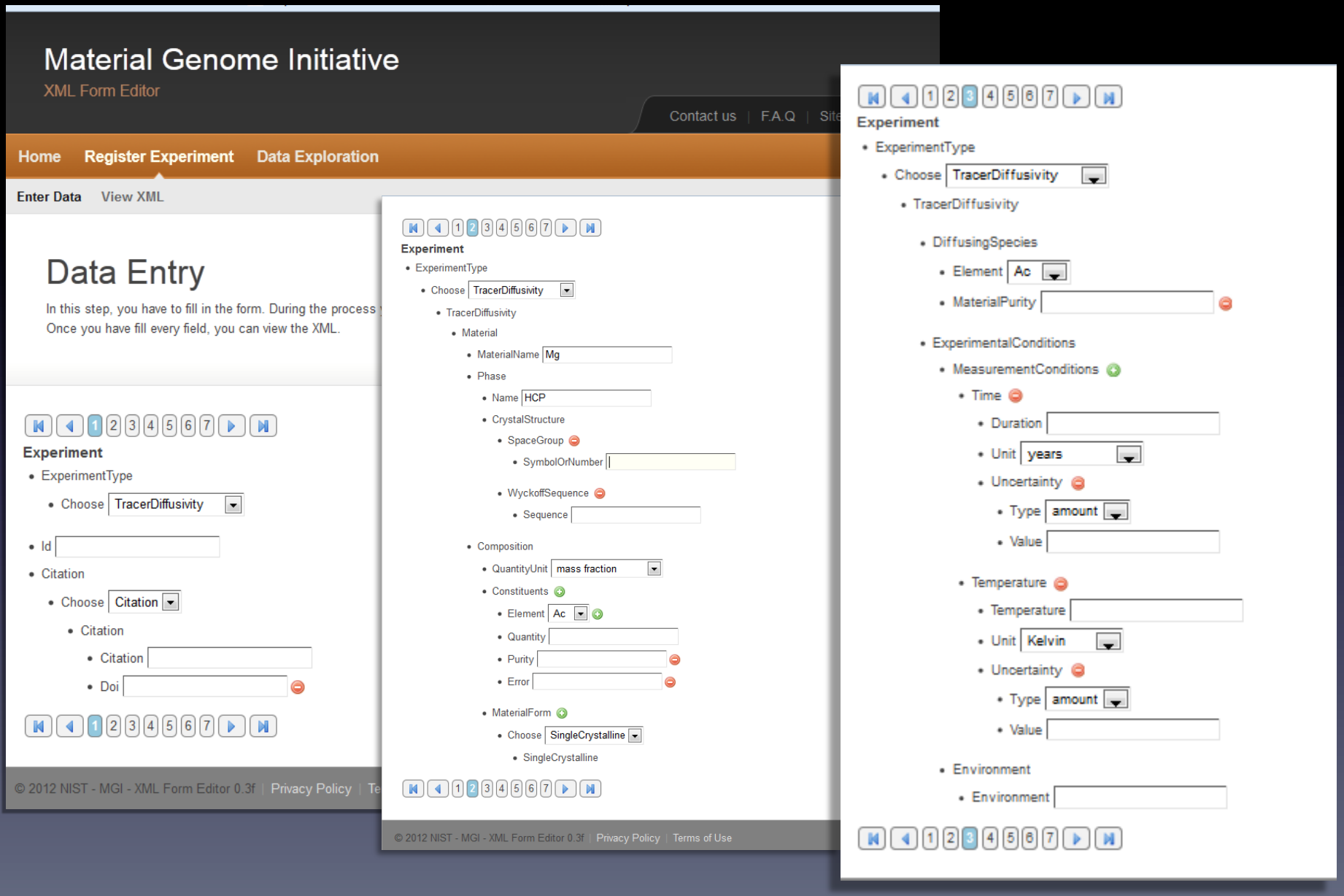


- Re-written in python
- Backed by MongoDB

New features:

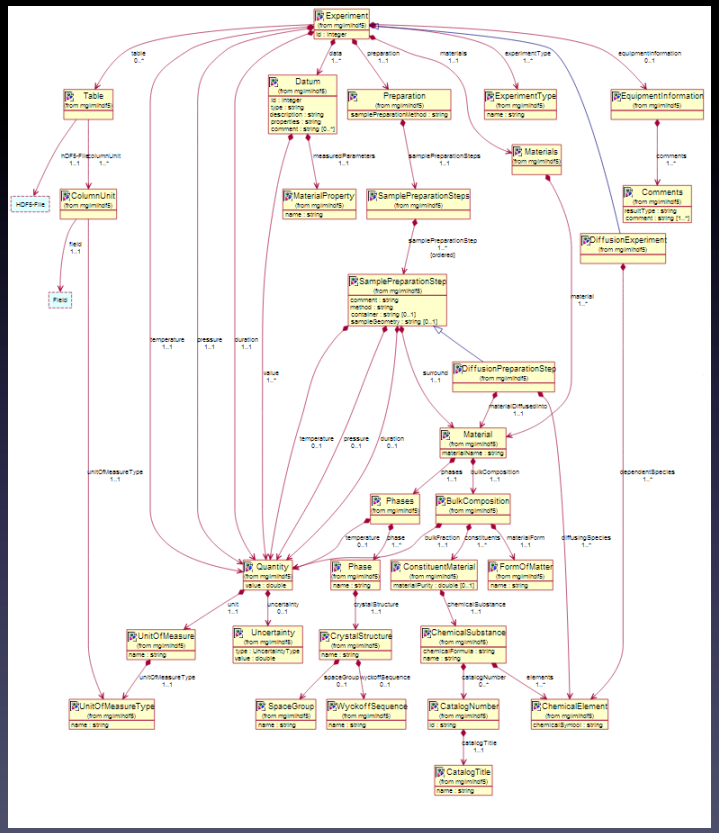
- Ability to store templates
- Schema management tools
- API interface

Data Curation: Tracer Diffusivity Test Schema

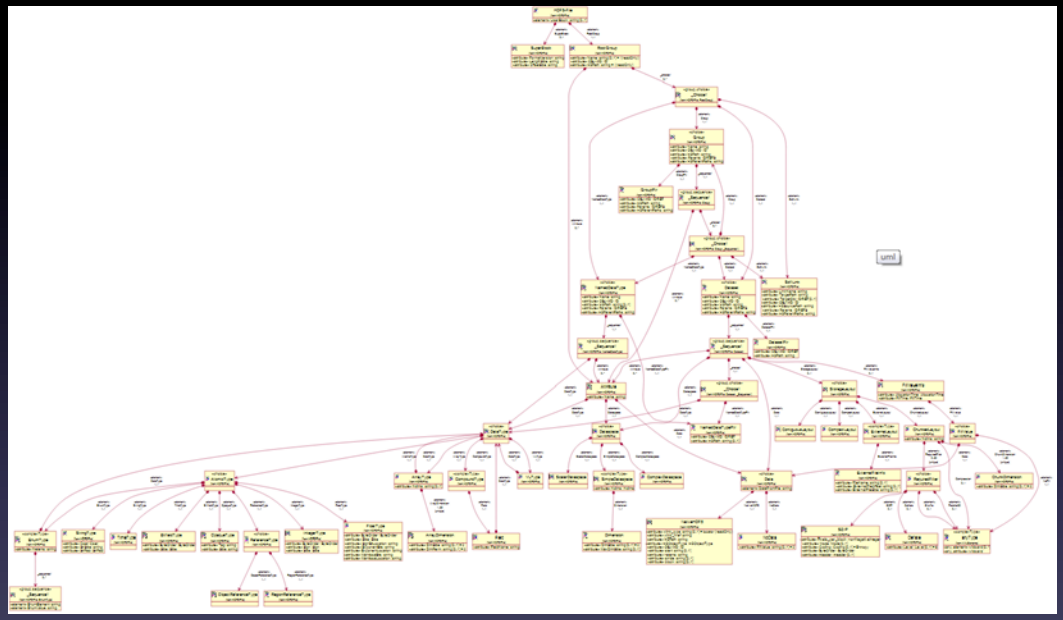


Encoding Phase-Based Data

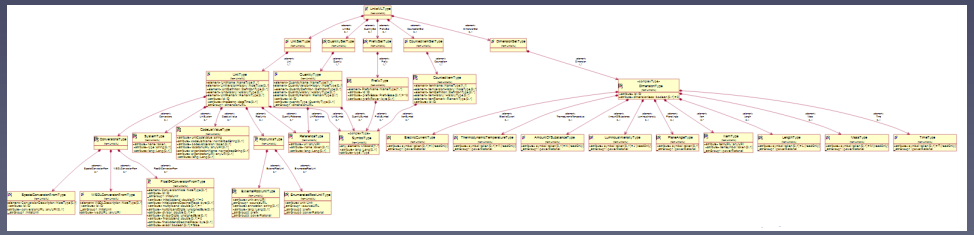
Core



Tabular Data (HDF5)



Measurement Units (Units ML)



Future of Materials Data Informatics

- Most materials data are only semi-structured
- Ontologies and the Semantic web offer opportunities to link unstructured data and enable knowledge creation.

Opportunities to Get Involved: (Lots of work to do)

- Share your data and tools
  - https://materialsdata.nist.gov/dspace/xmlui
  - http://nist.matdl.org
- Curate your data
- Participate in working groups to develop ontologies and XML schemas
  - (NIST Diffusion & CALPHAD Workshop, April 28-30, 2014)