Toward an Unified Metadata Model for 'Research Objects'

Matthew Brush

Oregon Health and Science University

November 3, 2015

About Me

- **Bench Scientist Turned Data Scientist**

  Cancer Biology -> BioInformatics -> Biomedical Ontology

- **OHSU Ontology Development Group (ODG)**

  Development and application of ontologies to facilitate discovery and use of research data and resources

  - **eagle-i resource discovery platform**
    eagle-i.net
  - **Resource Identification Initiative**
    force11.org/node/4824
  - **Monarch Initiative**
    monarchinitiative.org

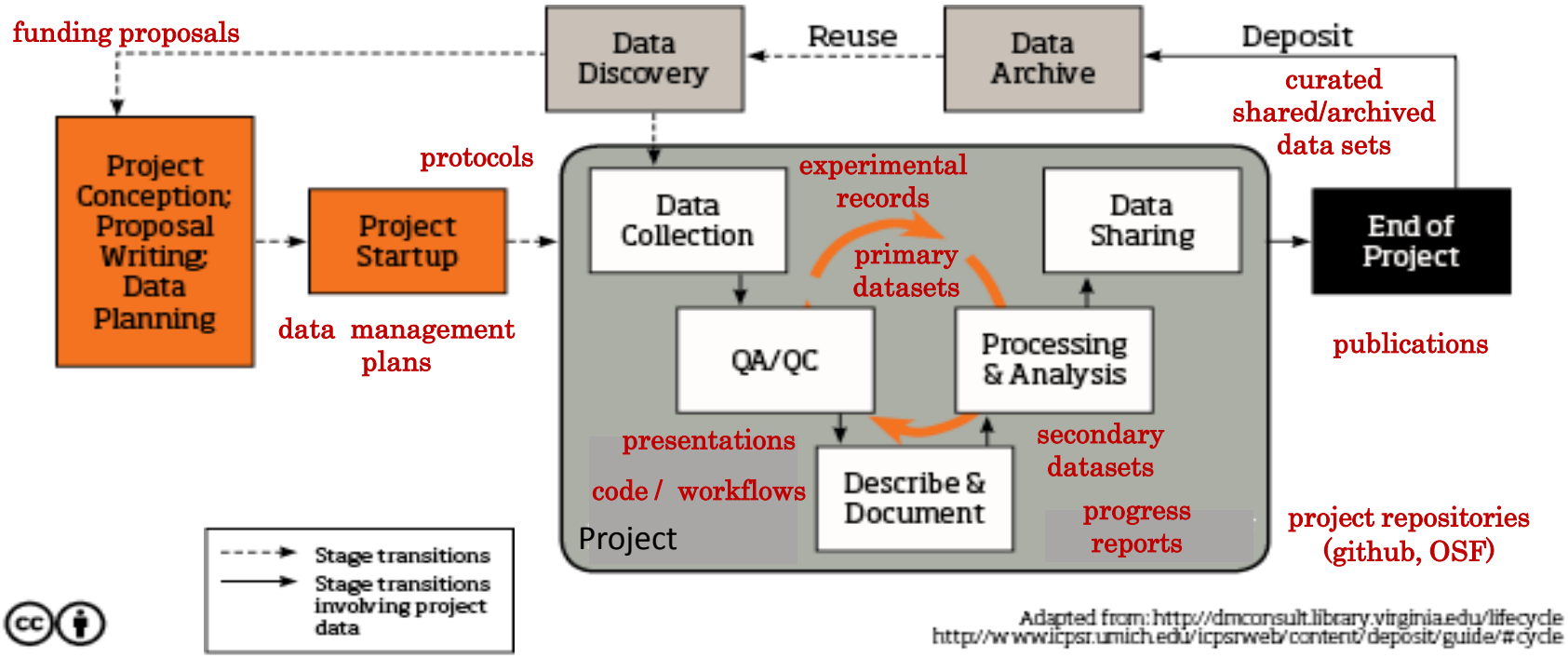
Project Overview

Three month contract between Elsevier and OHSU

- Full time pre-doctoral intern
  (Olga Giraldo)
- Part time faculty mentors
  (Matthew Brush, Melissa Haendel)

Goal: Develop a high-level model of experimental research concepts to serve as a framework for describing ‘research objects’

‘Research objects’ are information artifacts that are utilized or produced in the research lifecycle



Metadata about research objects is becoming increasingly important in today’s data intensive research paradigm

Applications for this Work

Facilitate Internal Efficiencies

1. Conceptual alignment and communication
2. Uniform collection of metadata across Elsevier systems

Support Community Integration

1. Metadata integration for search across external systems and research domains
2. Inform community metadata standards for research objects

Today's Topics

I. Research Object Landscape Analysis

II. Project Deliverables

A. Generic Metadata Recommendations

B. Reference Domain Model

C. Experimental Metadata Model (EMM)

III. Next Steps and Recommendations

I. Landscape Analysis

A. Initiatives

- How connected and aligned are community efforts?
- What partners might we engage in our future work?

B. Standards **(vocabularies, ontologies, data models and formats)**

- How diverse and aligned are existing models?
- Where are gaps in coverage?
- What might be re-used or inform our modeling efforts?

C. Systems **(data repositories, data journals)**

- What metadata is collected by schemas?
- What standards are used to structure this metadata?

Initiatives

1. Motivated community initiatives share goals and ideals but not resources
   - FORCE11, FAIRport, RDA, CASRAI
   - NIH BD2K Initiatives (Data Discovery Index)
2. Can learn from principles, successes, and failures
3. Engage key partners early to lay groundwork for future community integration efforts

Standards

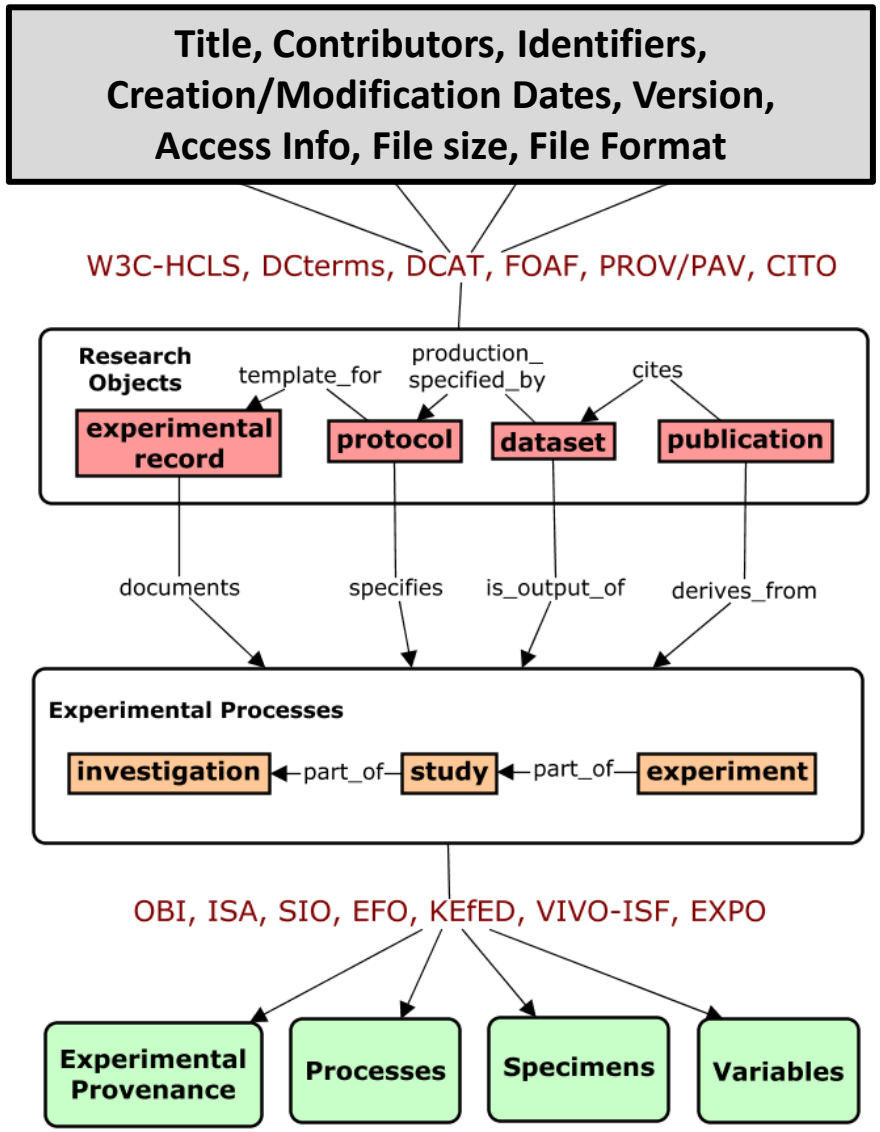
1. Generic Metadata

- Directly describes datasets
- Many mature and widely used standards exist, reasonably aligned

Two Levels of Metadata Standards for Research Objects

2. Experimental Metadata

- Indirectly describes datasets by capturing experimental provenance
- Existing standards not scoped or packaged for efficient metadata collection and search, and lack shared foundation for interoperability



Systems

1. Generic Data Repositories

- Dryad, FigShare, Dataverse
- Use internal schema, alignment with standards varies
- Structured metadata collected primarily for generic attributes of the dataset
- Minimal structured metadata about experiments that produced data (keywords)

Data from: The impact of gene expression variation on robustness and evolvability of a developmental gene regulatory network

dc.contributor.author	Garfield, David A.	
dc.contributor.author	Runcie, Daniel E.	
dc.coverage.spatial	Pacific Ocean	
dc.date.accessioned	2013-10-31T13:31:26Z	
dc.date.available	2013-10-31T13:31:26Z	
dc.date.issued	2013-10-29	
dc.identifier	doi:10.5061/dryad.7j9t5	
dc.identifier	doi:10.5061/dryad.7j9t5	
dc.identifier.citation	Garfield DA, Runcie DE, Babbitt CC, Haygood R, Nielsen WJ, Wray GA (2013) The impact of gene expression variation on robustness and evolvability of a developmental gene regulatory network. PLoS Biology 11(10): e1001696.	
dc.identifier.uri	http://hdl.handle.net/10255/dryad.54031	
dc.description	Regulatory interactions buffer development against genetic and environmental perturbations, but adaptation requires phenotypes to change. We investigated the relationship between robustness and evolvability within the gene regulatory network underlying development of the larval skeleton in the sea urchin	
dc.relation.haspart	doi:10.5061/dryad.7j9t5/3	
dc.relation.isreferencedby	doi:10.1371/journal.pbio.1001696	
dc.relation.isreferencedby	PMID:24204211	
dc.subject	gene regulatory network	
dc.subject	molecular evolution	
dc.subject	developmental evolution	
dc.subject	population genetics	
dc.subject	sea urchin	
dc.title	Data from: The impact of gene expression variation on robustness and evolvability of a developmental gene regulatory network	
dc.type	Article	*

http://datadryad.org/resource/doi:10.5061/dryad.7j9t5

Systems

2. Domain-Specific Repositories

- e.g. ArrayExpress, GEO

E-MTAB-3315 - RNA-seq of mouse splenocytes following infection with Listeria monocytogenes

Status	*Submitted on 6 October 2011, last updated on 6 May 2015, released on 30 June 2015*
Organism	Mus musculus domesticus
Experiment types	RNA-seq of coding RNA, pathogenicity design
Contact	Daniel Caffrey <daniel.caffrey@gmail.com>
MINSEQE	* Exp. design * Protocols * Variables * Processed * Seq. reads
Files	Investigation description: E-MTAB-3315.idf.txt Sample and data relationship: E-MTAB-3315.sdrf.txt Processed data (3): E-MTAB-3315.processed.1.zip, E-MTAB-3315.Sample24h.sort.bam, E-MTAB-3315.SamplePBS.sort.bam

Methods

Experimental Variables

Specimen Characteristics

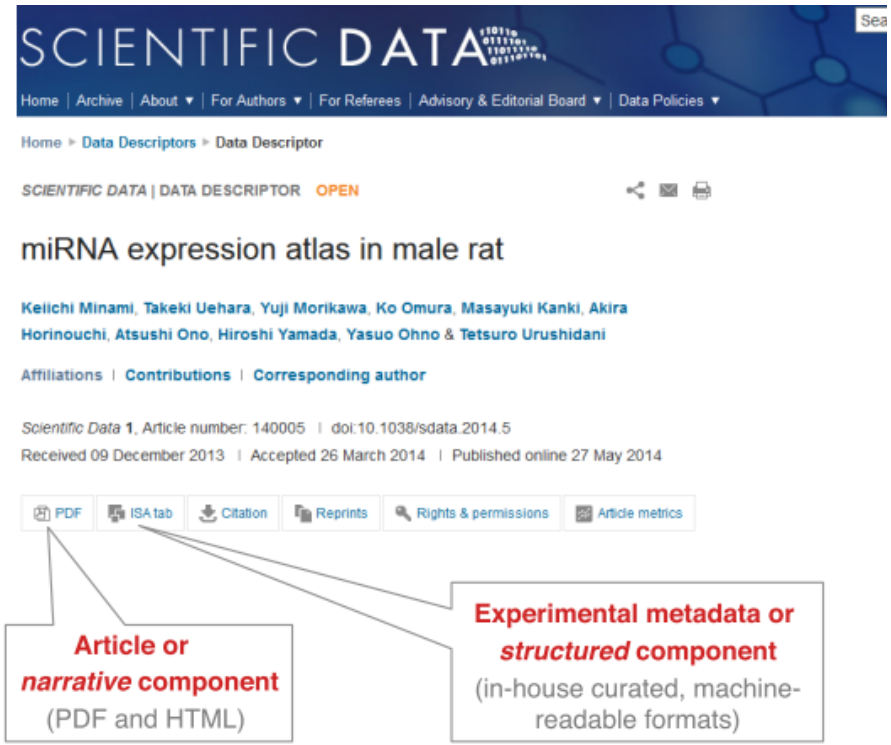
	Sample Attributes			Variables
Source Name	organism	cell type	infect	infect
Sample24h	Mus musculus domesticus	splenocyte	Listeria monocytogenes (clinical isolate 10403s)	Listeria monocytogenes (clinical isolate 10403s)
Sample24h	Mus musculus domesticus	splenocyte	Listeria monocytogenes (clinical isolate 10403s)	Listeria monocytogenes (clinical isolate 10403s)
SamplePBS	Mus musculus domesticus	splenocyte	none (PBS control)	none (PBS control)
SamplePBS	Mus musculus domesticus	splenocyte	none (PBS control)	none (PBS control)

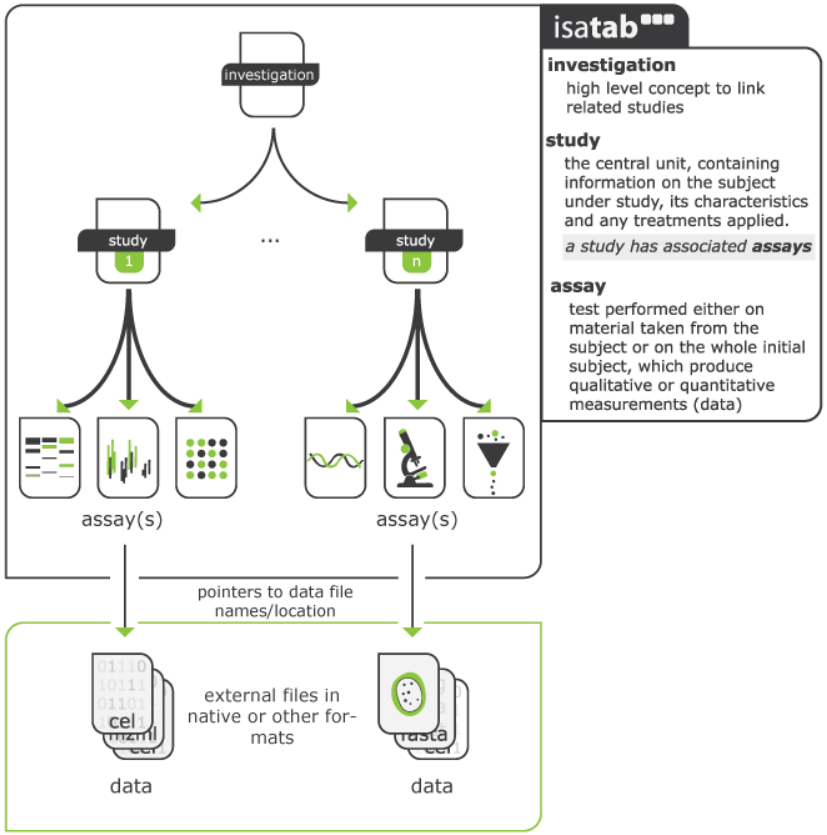
http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3315/

Systems

3. Data Journals

- Nature Scientific Data
- DataOne
- Biodiversity Data Journal





II. Lessons Inform Project Deliverables

1. **A diverse and informative body of work already exists in this area that can inform our efforts.**

   D1. Landscape Analysis : structured evaluation of initiatives, standards, and systems (link)

2. **A strong foundation of standards and practices generic metadata collection exists**

   D2. Generic Metadata Recommendations: identify core elements and propose standards to adopt (link)

3. **Key challenge lies in defining standards for experimental metadata that can support its collection and integration**

   D3. Reference Model of Experimental Concepts (link)

   D4. Experimental Metadata Model Straw Man (link)

D1. Landscape Analysis

Structured evaluation of initiatives, standards, and systems (link)

- High level summary of all relevant efforts
- Deeper evaluation of models and standards
- Overview of all groups working on "Research Objects"

D2. Generic Metadata Recommendations

Proposes set of core metadata elements and standards (link)

- Informed largely by W3C-HCLS Data Description Standard
- Can guide early iterations of metadata schema for developing systems
- And inform subsequent community integration efforts

D3. Reference Model of Core Experimental Concepts

Concept Glossary (link)

Experiment | Data Set | Research Study | Technique | Investigator | Research Organization | Research Facility | Study Design | Study Group | Experimental Variable | **Control Variable** | **Independent Variable** | **Dependent Variable** | Reagent | Instrument | Research Specimen | **Source Specimen** | **Precursor Specimen** | **Evaluated Specimen** | Environment

Conceptual Models

1. **Taxonomic hierarchy** organizes these concepts according to their ontological type (material, process, attribute, information)
2. **Ontology graph** illustrates relationships between concepts

D4. Experimental Metadata Model

Deliverables (link)

1. **Metadata Model Proposals:** Organize core domain concepts as a schema to support efficient metadata creation and search
2. **Exemplar Metadata Records -** Illustrates features of the metadata model using real data

Requirements Analysis

1. Scoped by identifying **priority use cases**
2. Driven be **competency questions** (link)
3. Informed by evaluation of **existing models**

Priority Use Cases

1. Search/Discovery: meet minimal criteria for discovery based on experimental attributes and relationships
2. Data Entry: intuitive and efficient collection of structured metadata with minimal curation burden
3. Interoperability: enable alignment with existing metadata schema

Competency Questions

1. *"Find datasets linked to publications from the Lin Lab"*
2. *"Find gene expression datasets produced using RNAseq technologies"*
3. *“Find imaging datasets from pathogen-exposed mice "*
4. *"Find datasets produced using Fugene reagent in cells derived from human colon"*
5. *"Find datasets investigating stroke in elderly smokers”*
6. *"Find datasets exploring effect of kinase knockdown on antigen expression in transformed vs normal primary cell lines”*

Evaluation of Existing Standards

- **OBI (Ontology for Biomedical Investigations)**
  Broad and rich, but focus in on biomedicine, and not suited or being used as a metadata model/schema
- **ISA (Investigation-Study-Assay) Format**
  A schema for metadata collection that is widely used, but model is too complex, under-constrained, and not scoped for our use case
- **KEfED (Knowledge Engineering from Experimental Design)**
  A high-level schema with appealing level of simplicity/abstraction, but lacks ontological underpinning, and structured for claims analysis use cases.
- **EBI (European Bioinformatics Institute) RDF Platform**
  Supports data collection and scoped for discovery use case, narrowly defined for specific datatypes and domains

Existing models do not serve our needs ...

. . . But They Still Have Much to Offer

Converge on central relationship between *techniques*, *specimens*, and *variables*

1. **Techniques**: material preparation, assays, analysis
2. **Specimens**: material input(s) processed and ultimately examined in an experiment
3. **Experimental Variables**: entities measured, varied, or held constant in an experiment
   - **Dependent variable**: specifies entity and attribute measured in an assay
   - **Independent variables**: specifies parameters intentionally varied to test affect on dependent variable
   - **Controlled variables:** parameters not varied across study groups

Driving Principles and Requirements

1. **Generic but Extensible:** accommodate diverse research disciplines while supporting extensions for specific domains and use cases

2. **Pragmatic for Data Entry and Discovery Use case:** balances efficiency of data entry with complexity needed for query resolution
   a. Captures experimental roles of specimens, techniques, and variables and the relationships between them

3. **Supports Interoperability and Re-Use**
   a. Support integration of data across existing repositories
   b. Use community CVs/ontologies to standardize data entry
   c. Facilitate data that is amenable for third party use (LOD)

Modeling an Exemplar Experiment

Hypothesis: *Differential stress-response pathways are activated in brains of normal vs memory-impaired mice.*

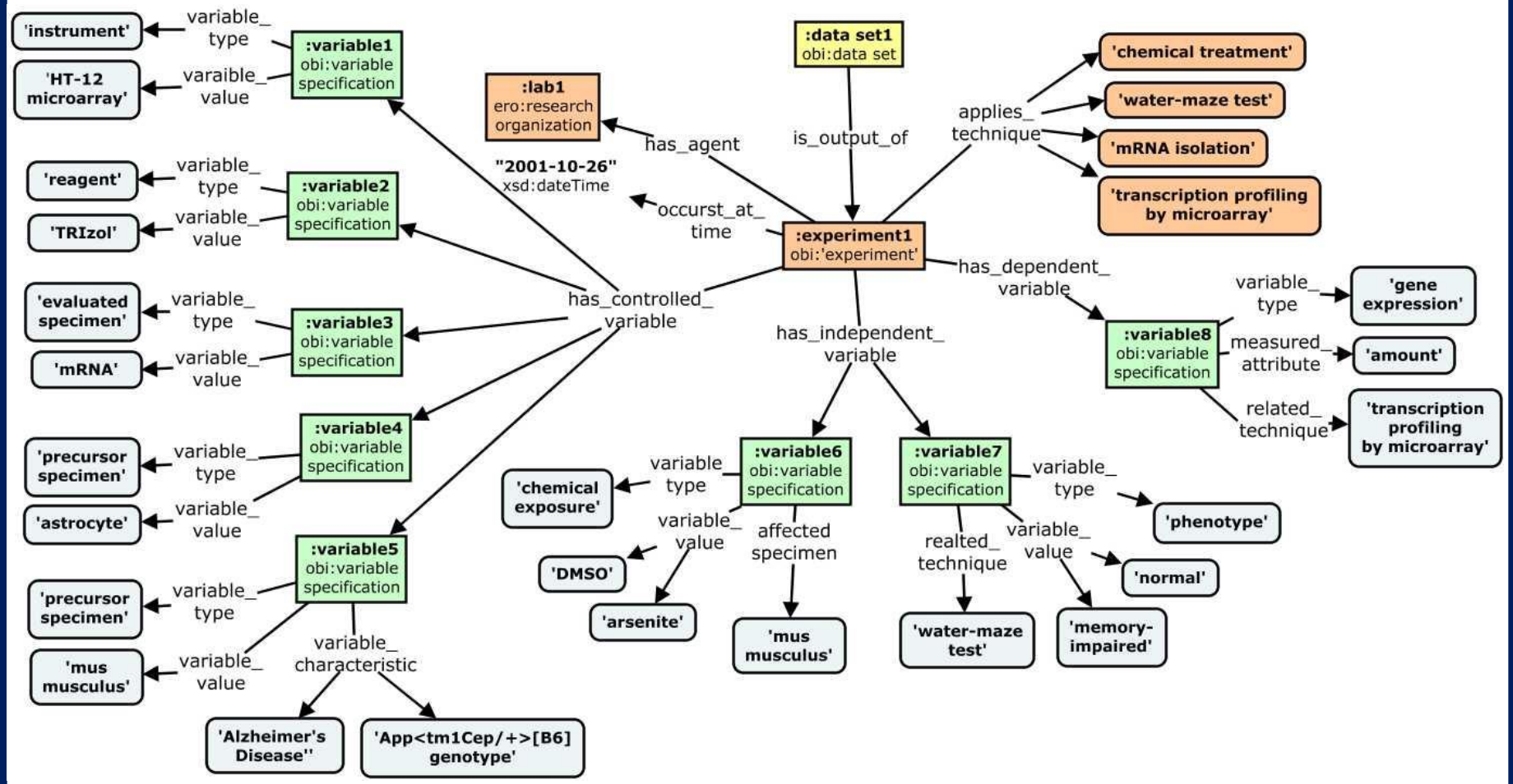
1. **Treatment:** exposure of a mouse model of Alzheimer's Disease to DMSO (control) or arsenite (oxidative stressor)
2. **Specimen Collection #1:** collection of astrocytes from hippocampi of mice displaying normal vs impaired memory (in water maze tests)
3. **Specimen Collection #2:** isolation of mRNA using TRIzol reagent to obtain mRNA samples
4. **Assay**: expression microarray analysis using an Illumina HT-12 array to identify patterns in level of gene expression

Controlled Variables **Independent Variables** **Dependent Variables**

Sidebar: The Challenges and Utility of a Variable-Centric Approach

- Most models of experimental metadata are organized to focus on experimental roles – what entities serve as specimens, reagents, instruments, and what are their key attributes.
- Our contention is that this may not be the most efficient approach for answering queries that are based on what an experiment tests and how.
- The notion of experimental variables (dependent, independent, control) is very useful here, because the relationships between these define what an experiment tests (e.g. the impact of treatment X on process Y in specimens of type Z).
- The key difference of the metadata model we propose is that it focuses on these experimental variables first instead of experimental roles – organizing the roles, techniques, and specimens around the variables
- It may be a minor conceptual hurdle for many to view an experiment through this lens – focusing on how the entities relate to variables rather than to roles. But given that the concepts of dependent, independent, and control variables are fundamental to experimental research and common across domains and communities, this hurdle should be a low one. (The highlighting of entities in the previous slide should help with this)
- And organizing metadata in this way may be more practical for answering queries/discovery in an efficient way that requires less complex models and la lower burden for data collection. This is the hypothesis we will set out to explore.

Modeling an Exemplar Experiment



Sidebar: A Tour Through the Model

1. **Different color for each 'level' in the model** – tried to limit the number of levels for ease of data collection/queries.

2. **Experiment in orange is focal point.** At this level we collect provenance metadata about the experiment rather than the dataset such as who performed it and when. We also link it to specific techniques applied in the experiment.

3. **The next level in green is where we clearly see the focus of the model around experimental variables.** Around these variable specification nodes hang terms from CVs and ontologies that represent key features of the experiment. We believe that this approach provides just the right context for supporting discovery use cases and competency questions.

4. **Dependent Variable Node (variable8)** - this represents the entity and attribute measured in an assay
    a. This specification describes entity and attribute measured in an assay (namely the level or amount of gene expression) as well as the optional indication of the specific technique used for measurement.

5. **Independent Variable Nodes (variable6, variable7)** – these represent parameters intentionally varied to test affect on dependent variable
    a. There are two independent variable specifications in this experiment, one is a chemical exposure that specifies dmso and arsenite to different study groups, and the other is a phenotype variable that specifies normal vs memory impaired.
    b. Additional properties can be used to record things like the type of specimen affected, and the specific technique used to apply or determine the variable.

6. **Controlled Variable Nodes (variables 1-5)** - these represent parameters not varied across study groups
    a. These specifications follow similar pattern. They allow description of reagents and devices that are not varied in the experiment, as well as specimen types and features that are held constant.
    b. Things get a bit more nuanced when it comes to describing specimens - here we define the notion of the evaluated specimen - that mRNA the material that was actually input into the assay, and precursor specimens that allow you to describe where it came from. Here this is astrocytes and mus musculus. We capture these because they represent potentially useful hooks into discovering this dataset.
    c. Note that the model supports additional descriptors and qualifiers to provide additional information about these variables, such as the mouse being a model of a disease, or its genotype.

Exemplar Metadata Record

```
{
dataset_uri:  "http://datadryad.org/resource/doi:xx.yyy/dryad.zzzz"
  {
  experiment:
        investigator: investigator1
        laboratory: lab1
        time: "2001-10-26"
        study  design: compound treatment design
        technique: chemical treatment
        technique: water maze test
        technique: mRNA isolation
        technique: transcriptional profiling by microarray

        controlled _variable: [
            {variable_type: instrument
             variable_value: Illumina HT-12v4 microarray
             resource_description: "lot #2314, 2001-02-15"}

            {variable_type: reagent
             variable_value: TRIzol
             resource_identifier: "Ambion 15596"}

            {variable_type: evaluated specimen
             variable_value: mRNA}

            {variable_type: precursor specimen
             variable_value: astrocyte}

            {variable_type: precursor specimen
             variable_value: mus musculus
             resource_description:  "RRID:MGI_5446893"}
            ]
```

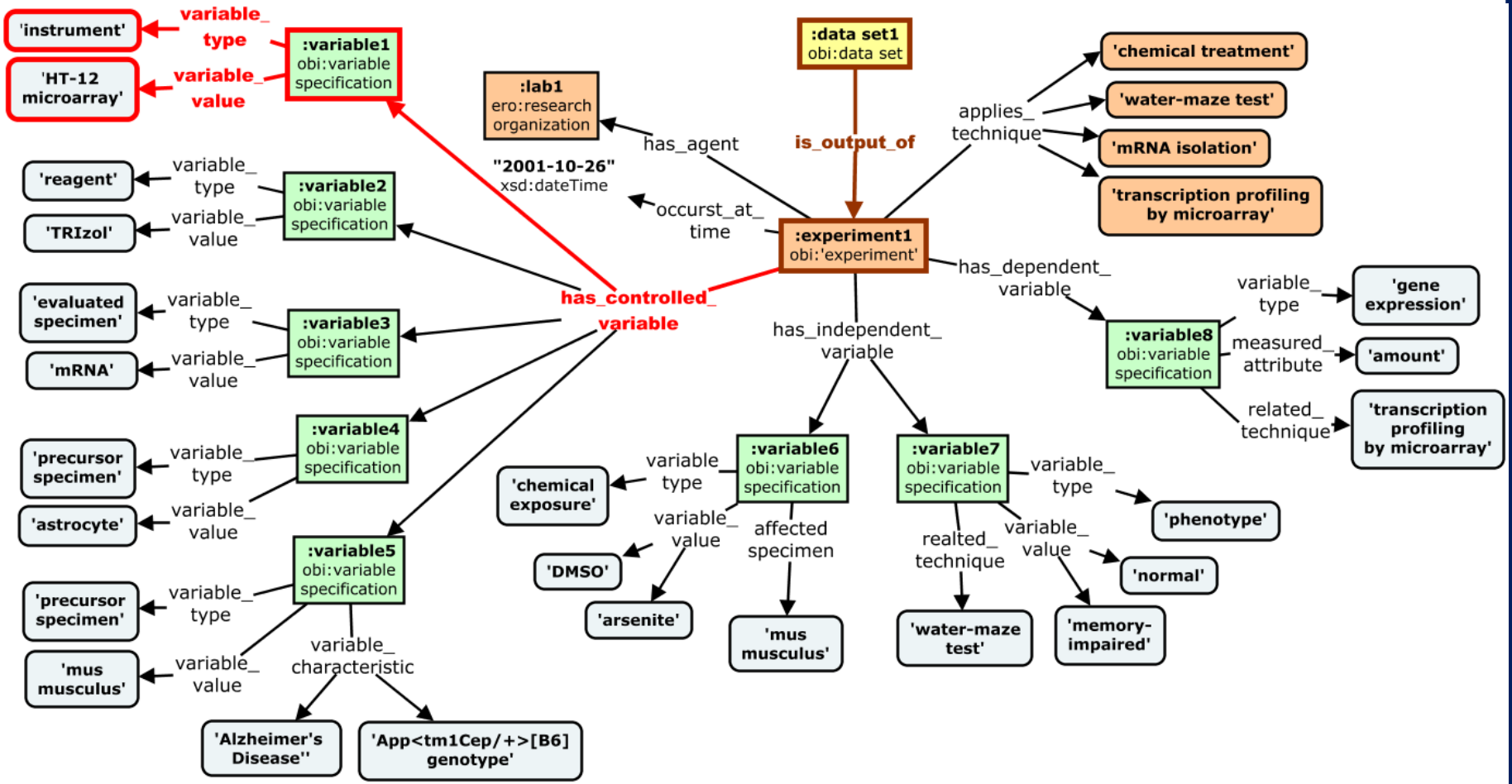
- ~JSON representation of full metadata record for exemplar experiment
- Relatively simple and flat, while capturing needed precision and variable context for core CQs and modeling requirements

```
        independent_variable: [
            {variable_type: phenotype
             variable_value: normal
             variable_value: memory-impaired
             affected_specimen: mus musculus
             related_technique: water maze test}

            {variable_type: chemical exposure
             variable_value: DMSO
             variable_value: arsenite
             affected_specimen: mus musculus
             related_conditiion: “37C, 5%CO2"}
            ]

        dependent variable: [
            {variable_type: gene expression
             measured_attribute: amount
             related_technique: transcriptional profiling by microarray
             has_scale: decimal}
        ]
    }
}
```

Key Features

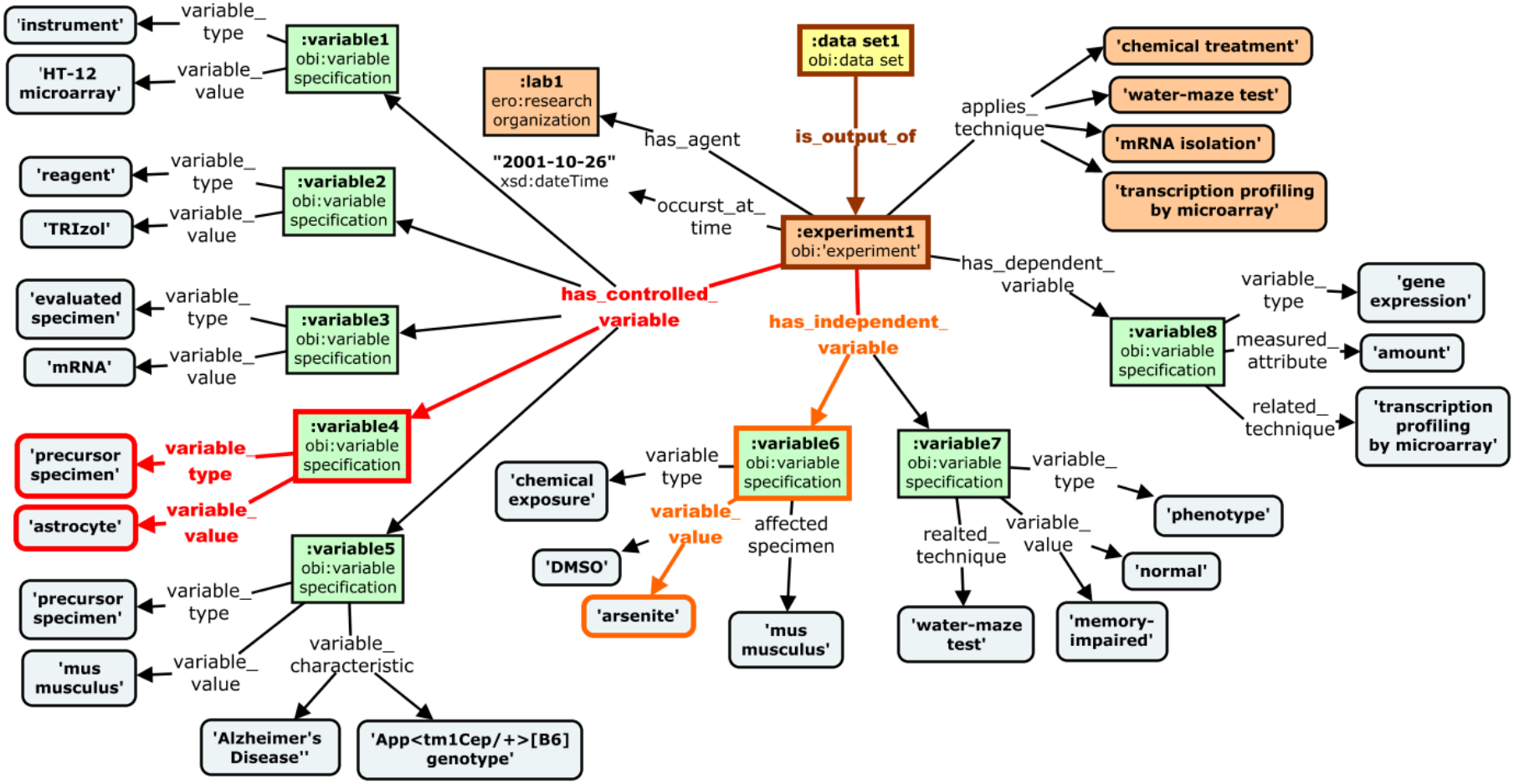
1. **Variable-Centric**
   - Experimental variables serve as central organizing nodes that describe variable type, values, and features.
2. **Central concepts are universal**
   - Experimental variables most proximal to the dataset are universal paradigm across disciplines that describes how entities participate in an experiment
   - Optional extensions record additional context/granularity
3. **Captures types over instances**
   - Abstracts away similarities between specimens to eliminate redundant representation of concepts.
4. **Direct re-use of use community vocabularies/ontologies**
   - Constrains data entry in variable descriptions, with terms coming from community CVs/ontologies
5. **Parsimonious model of manageable size and complexity**
   - Implements a relatively flat model using a relatively small set of classes and properties

CQ1: *"Find **data sets** produced using **Illumina microarrays**"*



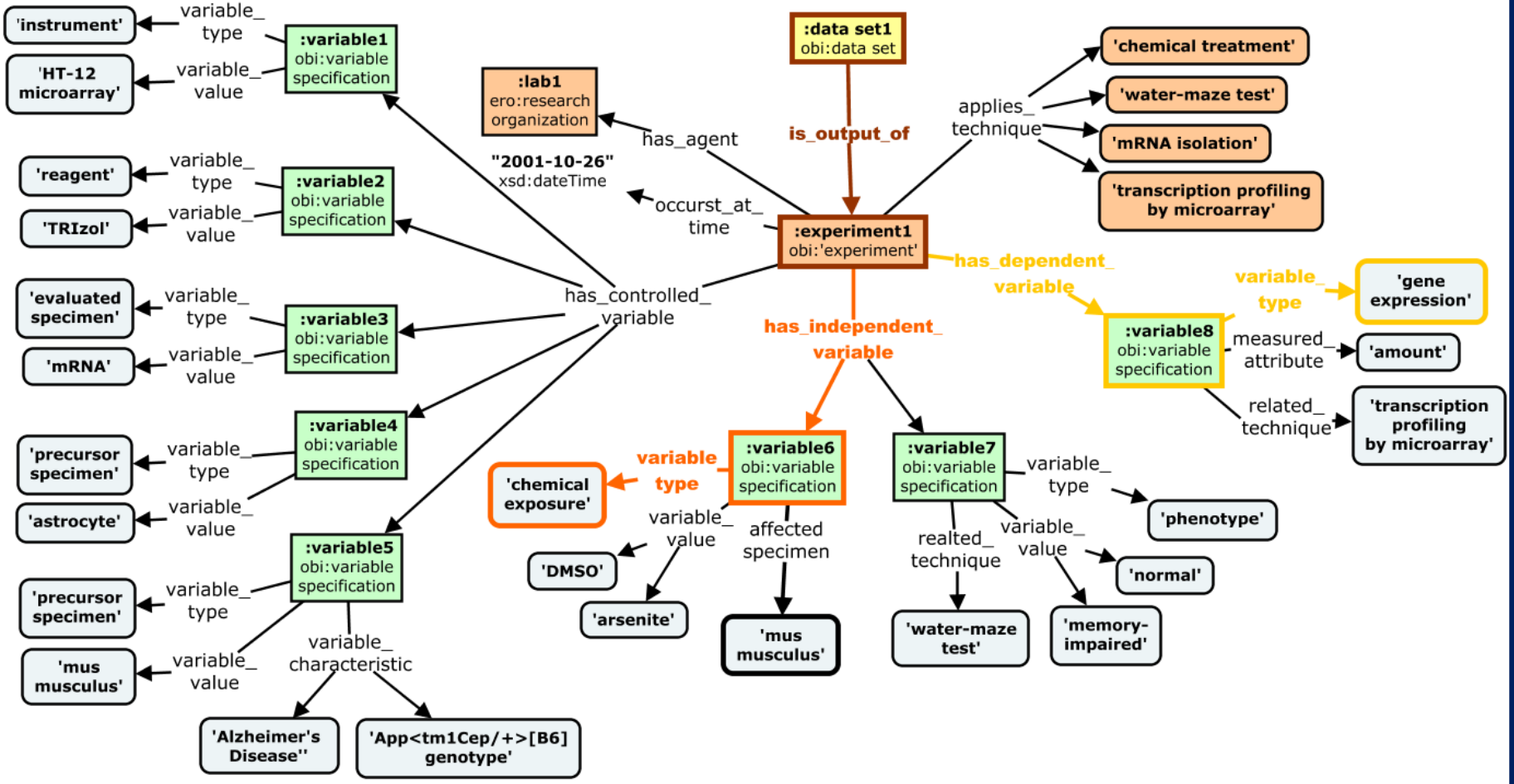
research object | *control variable* | *independent variable* | *dependent variable*

CQ2: *"Find **data sets** about the effects of **arsenite exposure** on the **brain**"*



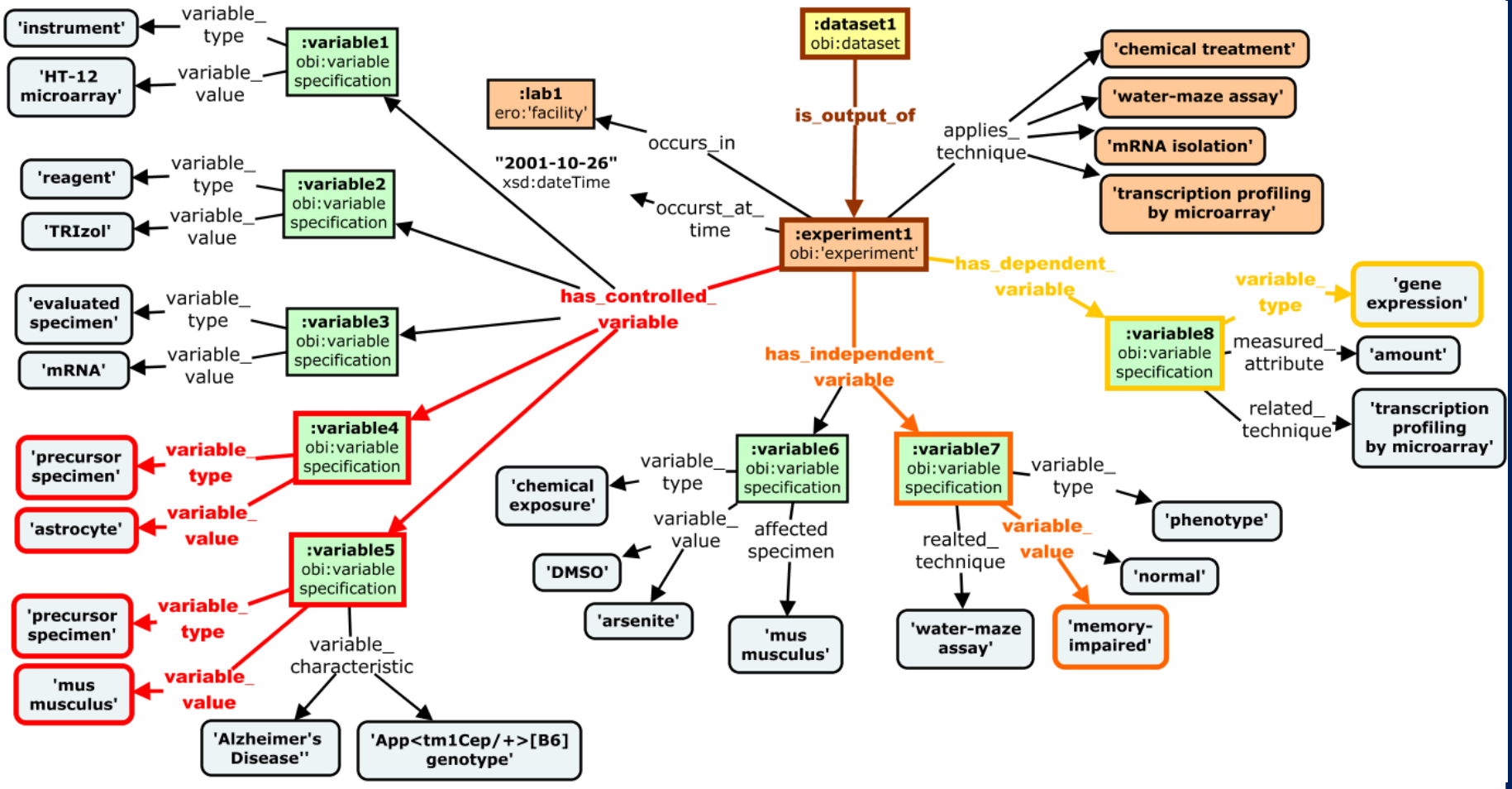
research object | *control variable* | *independent variable* | *dependent variable*

CQ3: *"Find data sets about the impact of chemical exposure on gene expression"*



research object | *control variable* | *independent variable* | *dependent variable*

CQ4: *"Find **data sets** about **gene expression** in **astrocytes** of **mice** with **cognitive-defects**"*



research object | *control variable* | *independent variable* | *dependent variable*

IV. Applications and Future Directions

1. Partnering with the Community

A. Establish a FORCE11 Working Group

B. Inform Open Research Information Framework (OpenRIF) efforts
   - evolution of VIVO/eagle-i developer community

C. Work with BD2K Data Discovery Index Metadata WG (bioCADDIE)

BD2K Data Discovery Index (DDI)

- Grant funded the bioCADDIE group (UCSD) to build a system to index metadata about biomedical datasets across repositories using a common model, and provide search interface for discovery
- Metadata working group was tasked to define a model that could be used to aggregate metadata across existing repositories under unified framework
- Version 1 of their metadata specification is here.
  DDI white paper is here.

IV. Next Steps and Future Directions

2. Applications for Elsevier Projects

A. Landscape analysis
   - Who to partner with? What models to re-use or align with?

B. Glossary and conceptual model
   - Shared terminology for internal communication, documentation, models
   - Guide communication and alignment with broader community

C. General Metadata Requirements
   - Guide early iterations of internal metadata schema for developing systems
   - Inform selection of existing standards to adopt or align with

D. Experimental Metadata Model
   - Guide later iterations of internal metadata schema for systems
   - Inform development of community standards to unify metadata representation for research objects

IV. Next Steps and Future Directions

3. Broader Challenges for Community

A. Adopting a common conceptual framework to underlie model and terminology

B. Achieving consistent and robust use of community CVs/ontologies as value constraints

C. Documentation standards and practices

D. Mechanisms for translating between schema

E. Incenting researchers to contribute

Thank You!

Acknowledgements

Elsevier Research Data Services

Anita deWard

Oregon Health and Sciences University

Melissa Haendel

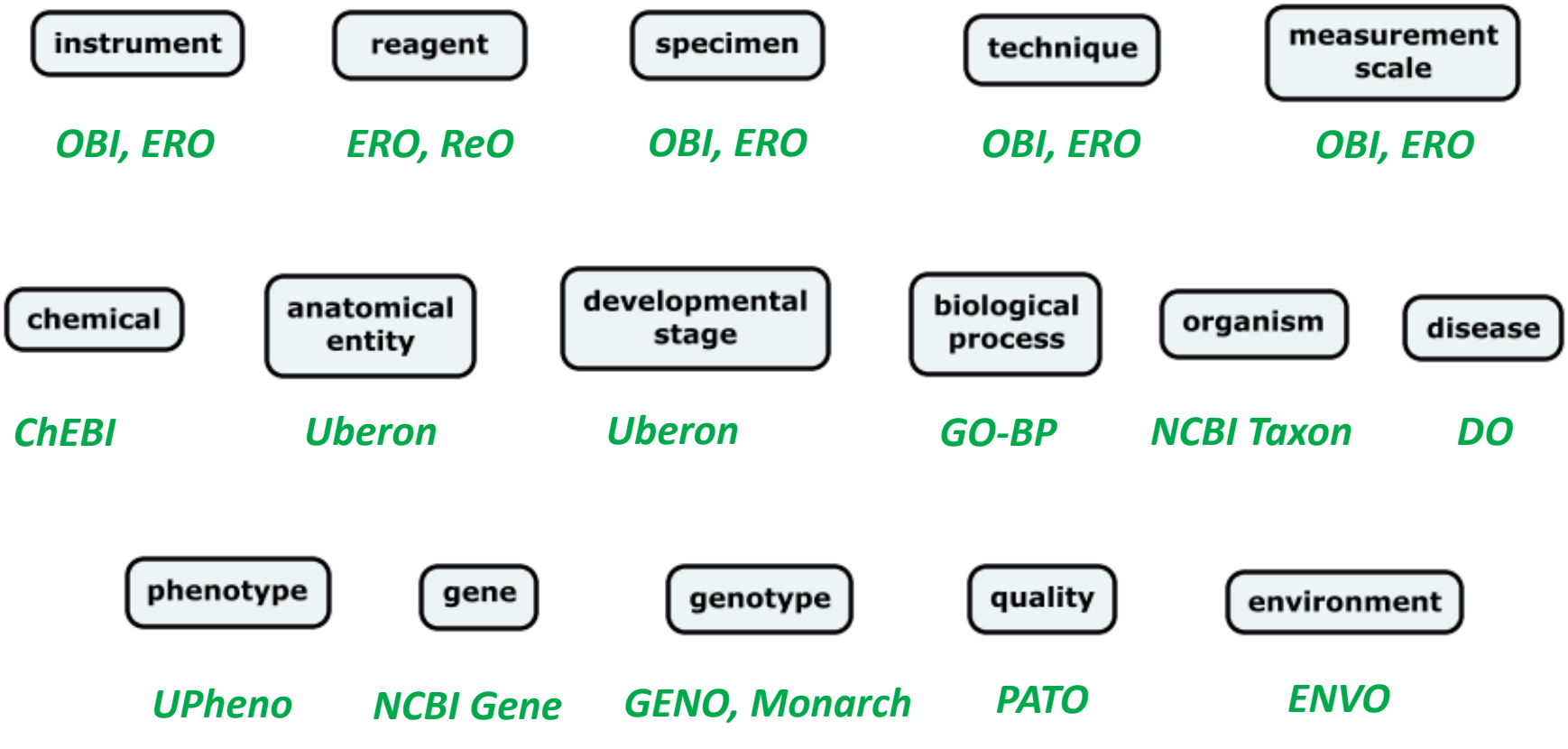
Universidad Politécnica de Madrid

Olga Ximena Giraldo

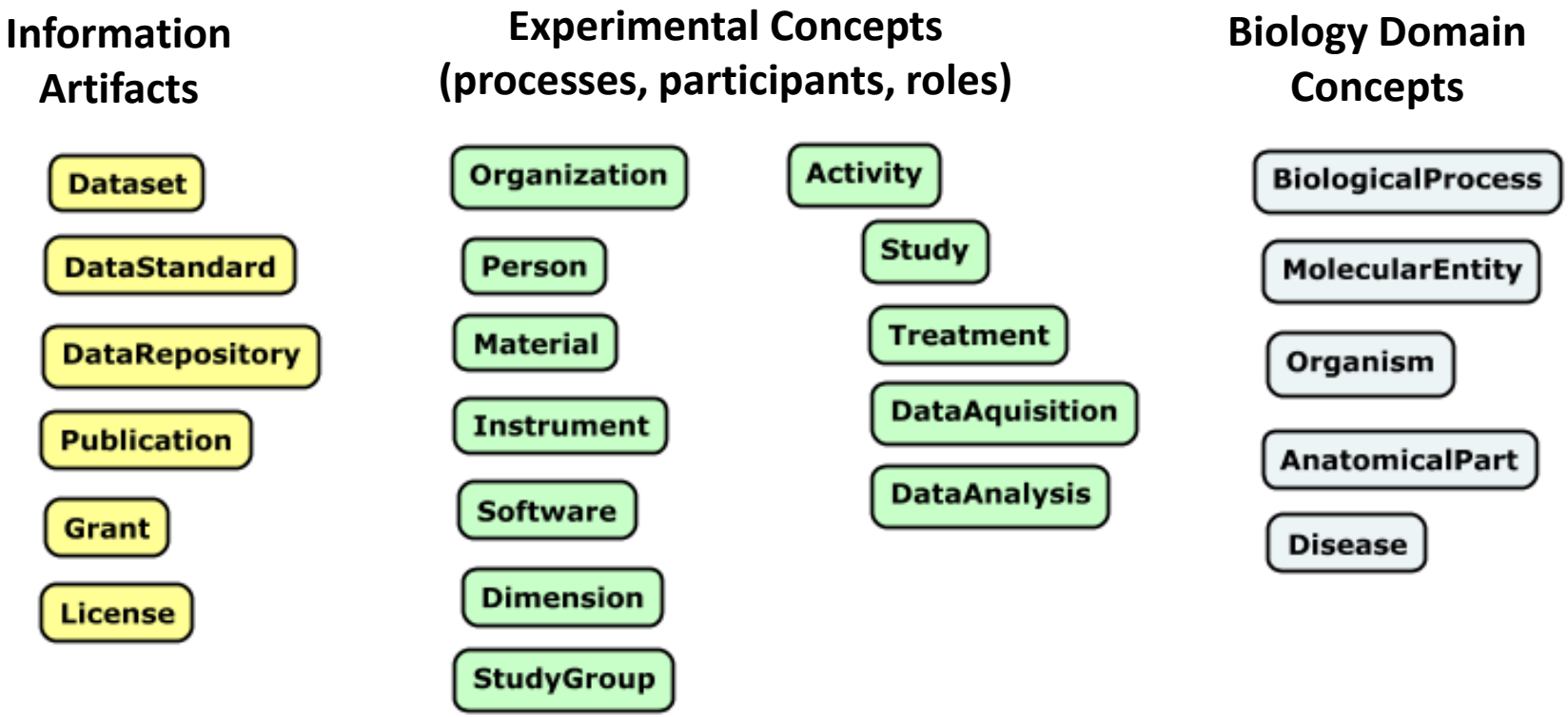
Ongoing work on this project hosted on GitHub:

https://github.com/OHSU-Ontology-Development-Group/experimental-metadata-model

Ontologies for Controlled Value Entry

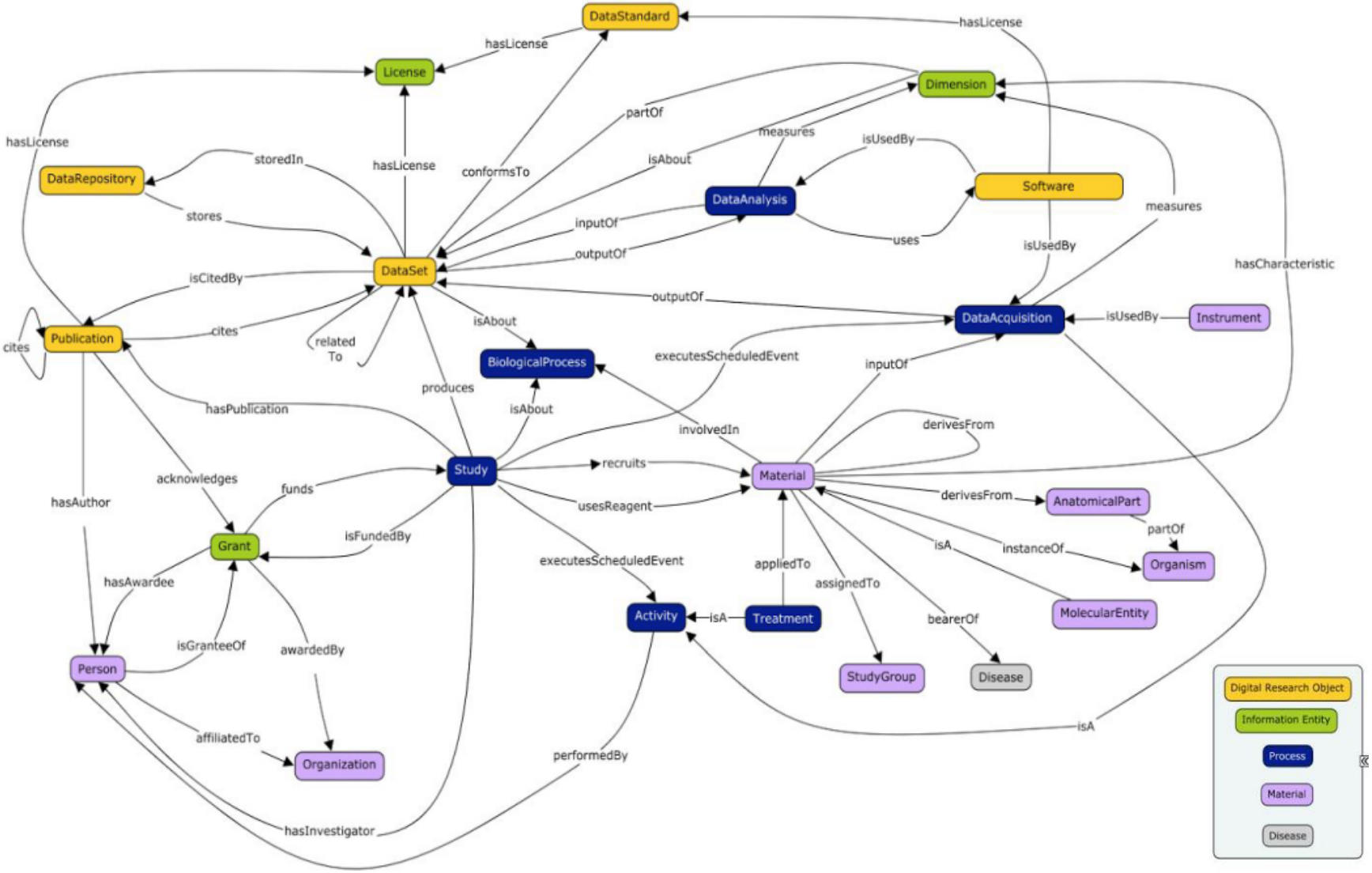


Core Concepts in bioCADDIE Model



- Similar set of core concepts represented when compared to our Experimental Metadata Model (EMM)
- But key areas where the model created from these concepts don't align with our model.

Overview of bioCADDIE Concepts and Relationships



https://github.com/biocaddie/WG3-MetadataSpecifications

Another Overview of bioCADDIE Concepts and Relationships