Reference and Metadata Models for Experimental Research

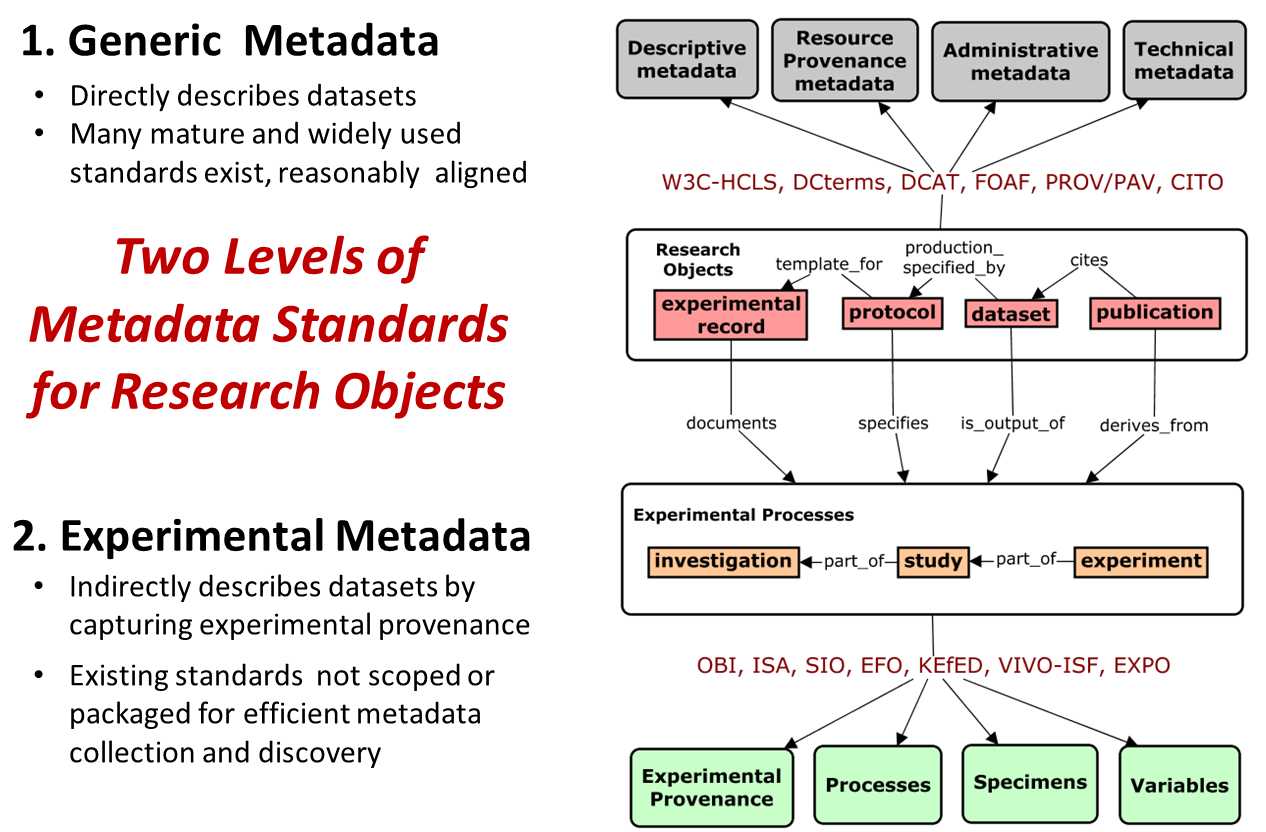
## CONTENTS

1. [Overview](#id.g9xe9dl785x1)
2. [Reference Model](#id.3n2kdnxt5851)
   1. [Concept Glossary](#id.3n2kdnxt5851)
   2. [Ontological Hierarchy](#id.peoqavdegdlu)
   3. [Concept Graph](#id.o3jkd9rxpop2)
3. [Metadata Model](#id.343vz3bsgq1s)
   1. [Hierarchical Metadata Schema with Exemplar Record](#kix.e01ko83d1de7)
   2. [Concept Graph with Exemplar Record](#id.343vz3bsgq1s)

# 

# I Overview

The goal of this document is to summarize work done to date to define a simple, domain-independent, and extensible experimental metadata model for describing research objects (ROs) such as datasets, protocols, experimental records, and publications. In these efforts we are clear to distinguish the notion of ***general research object metadata***, which describes attributes of a research object itself, from that of ***experimental research object metadata***, which describes an experiment related to the research object (Figure 1). Examples of the ***general metadata*** include things like the creator of a dataset, its title, identifiers, creation date, version number, license info, format, provenance information. There are many well established standards in place to describe such information (e.g. DC, DCAT, PROV, PAV), and recommendations about how to apply these (e.g. the [HCLS Dataset Description Standard](http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/)). The greater and more pressing challenge that remains is to similarly define standards around capturing ***experimental metadata*** - which includes things like the techniques, specimens, and variables applied in an experiment that generated a dataset, and the key relationships that contextualize the participation of these entities in the experiment.



***Figure 1: Metadata Framework Distinguishing General from Experimental Metadata***

In this document we begin to define the key concepts relevant for this goal, and initial proposals for assembling them into a metadata model that can be applied as a schema to collect metadata across disciplines and systems. This work was funded by Elsevier and requirements and use cases fed from developing Elsevier platforms and initiatives.

# II. Reference Model: Concept Glossary

## OVERVIEW

The purpose of this glossary is to identify and and define the core concepts that can inform an experimental metadata for research objects, and provide a preferred terminology for referring to these concepts. These concepts were chosen after lengthy process that included a landscape analysis of existing experimental metadata models and repositories, and consideration of metadata requirements that would support the discovery of research objects based on attributes of the experiments to which they are related. These 20 concepts were deemed most relevant for informing such an experimental metadata model that is intuitive, generic, extensible, parsimonious, aligned with existing frameworks, and scoped for supporting data entry and discovery.

Note that each entry in the glossary is a concept, not a term. Assigning terms as labels for these concepts is secondary, but obviously important for communication in our project and beyond. The concepts below are described with the following metadata elements:

1. **Label:** the preferred term for referring to the concept in discourse
2. **Alternative Term:** other labels by which the concept can be referred, including exact, broad, or narrow synonyms.
3. **Definition:** a definition of the concept that is crafted to make sense in the context our project and systems, often referring to or contextualized by other terms in our glossary. These definitions aim to provide criteria to identify and demarcate instances of the concept, and distinguish the concept from related concepts. Some are adapted from existing ontologies, but most are crafted de novo to make clear and specific distinctions relevant for our metadata model.
4. **Significance**: Highlights the relevance of the concept with respect to the proposed metadata model and our discovery and entry use cases, and other considerations for why it was chosen as significant for experimental descriptions.

-------------------------------------------------------------------------------------------------------------------

**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 | Specimen | Source Specimen | Precursor Specimen | Evaluated Specimen | Environment**

-------------------------------------------------------------------------------------------------------------------

|  |  |
| --- | --- |
| **CONCEPT:** | **EXPERIMENT** |
| Label: | experiment |
| Alt.Terms: | scientific experiment |
| Definition: | An orderly set of scientific procedures carried out with the goal of verifying, refuting, or generating a hypothesis. Experiments apply scientific techniques to identify or process specimens so as to test the effect of a defined set of independent variables on a measured dependent variable inhering in the experimental subject. |
| Significance: | Experiments produce data, are specified by protocols, and are documented by experimental records. Researchers need to discover these different research objects based on aspects of the experiment to which they are related (not just by attributes of the research object itself). Thus a clear definition of what constitutes an experiment, and its critical components, is essential for documenting and discovering research objects in ways most useful to scientists.  Experiments as research activities are inclusive of all processes leading to the production of a dataset about a given dependent variable, including: (1) specimen selection techniques that identify and/or isolate specimens of interest; (2) material processing techniques that modify or prepare specimens for measurement; (3) assays that produce raw data about specimens; and (4) data processing techniques that help to transform or analyze data to yield a conclusion (but not those that perform novel analyses or generate new data). Conceptually, we scope the extent of an experiment to include all processes that lead to the measurement of a single dependent variable (i.e. there is a 1:1 ratio between an experiment and a dependent variable). So a set of activities performed together can conceptually represent more than one experiment (if more than one dependent variable is being measured) |
| **CONCEPT:** | **DATA SET** |
| Label: | data set |
| Alt.Terms: | research data set |
| Definition: | A set of data records output from a single experiment, being the aggregate of data of the same type and generated in pursuit of validating an experimental hypothesis. |
| Significance: | Defining the scope of a single data set is important for identifying that which is output from an experiment or study, and distinguishing between different datasets when collecting metadata. Here an experimental dataset is tied to data output from a single experiment, implying the records are all about the same dependent continuant. But datasets defined at additional granularities may be needed or more appropriate in our model. This is a pressing question for many communities and use cases, and more work is needed in this area. |
| **CONCEPT:** | **RESEARCH STUDY** |
| Label: | research study |
| Alt. Terms: | study |
| Definition: | A set of one or more experiments that are aimed at answering a common research question (study hypothesis), through the application of apply different techniques and/or experimental variables. Studies can be more inclusive and broad in scope relative to experiments. While any experiment can qualify as a study in isolation, most studies include more than one experiment that aim to examine a common hypothesis. The hypothesis a study aims to explore is often broader than that probed by a single experiment, that thus can require many different experiments to validate. |
| Significance: | Documentation of a study according to the metadata model can be done by first separating it into its component experiments (based on identifying each dependent variables measured and the path of procedures that led to its measurement). But the model also supports linking a study directly to the experimental variables that define it (if this is the level at which the user wants to collect data) |
| **CONCEPT:** | **TECHNIQUE** |
| Label: | technique |
| Alt.Terms: | method, experimental method, scientific method, ,scientific technique, experimental process |
| Definition: | An established and reproducible process used in experimental research, to generate research materials (material processing technique), generate data (assay), or process data (data processing technique). |
| Significance: | Ontologically, techniques/methods are considered processes. A given experiment is comprised of the performance of many techniques to process, manipulate, and measure specimens. The types of techniques or methods applied in an experiment are critical metadata for discovery. But the order of their application, and the specific types and roles of participants in each are of secondary importance for this use case - and thus captured as optional features of variable specifications.  **Subtypes of scientific techniques include:**  (1) **Material processing techniques** have material inputs and outputs, and typically prepare or modify a specimen or other experimental resource for its participation in an experiment. Most interventions that apply specified experimental variables are material processing techniques (e.g. treatments with chemicals, genetic modifications);  (2) **Assays** measure a dependent variable to generate data about specimens (directly the evaluated specimen, but indirectly about its precursors) in an experiment. Assays take materials in and  generate data as output. they are more limited in scope than experiments in that they only include the processes that generate data, and not those preparing materials.  (3) **Data processing techniques** take data as input and transforms or analyzes it to yield data as output. |
| **CONCEPT:** | **INVESTIGATOR** |
| Label: | Investigator |
| Alt.Terms: | agent, research agent, experimenter, contributing researcher |
| Definition: | A role played by person(s) responsible for planning and/or performing an experiment (or parts thereof). |
| Significance: | The experimental investigators are those individuals who can be linked as direct contributors to an experimental research object and be attributed for its production. A individual must directly participate in planning, executing, or analysis to be considered a contributing agent for a given experiment (e.g. a PI who provided funding or lab space without direct intellectual input does not qualify). |
| **CONCEPT:** | **RESEARCH ORGANIZATION** |
| Label: | research organization |
| Alt.Term: | research laboratory, laboratory, lab |
| Definition: | An organization comprised of people with common research interests or goals who participate in a research program under common leadership, physical space, and/or funding.Model/metadata significance: A given experiment should have a primary association with only one lab, but can be associated with additional labs if contributing agents are affiliated with different laboratories. |
| Significance: | The agents performing an experiment or study that produced data is important provenance metadata that should be described for any dataset. |
| **CONCEPT:** | **RESEARCH FACILITY** |
| Label: | research facility |
| Alt.Terms: | facility, research location, location |
| Definition: | The physical facility or location where an experiment or research program is executed. Typically be specified by reference to a research organization that is linked to a physical location or address (e.g. a laboratory, department, university, institution). Not a geographical location - this would be a separate element. |
| Significance: | The facility and location where an experiment was performed is important provenance metadata that should be described for any experiment. This is a reference to the physical facility where an experiment was performed, not the organization that performed it - however often the latter is described in terms of the former (e.g. when a lab is tied to a specific physical location). |
| **CONCEPT:** | **STUDY DESIGN** |
| Label: | study design |
| Alt.Terms: | experimental design |
| Definition: | A plan that specifies in detail the goal and execution of an experiment or study, including the hypothesis explored, techniques used, specimens processed, variables applied, and measurements taken. |
| Significance: | Study designs apply to experiments or studies, as experiments can be considered to be studies comprised of a single experiment instead of many. Thus there is no distinction needed between a 'study design' and 'experimental design', and this concept covers both. A taxonomy of defined study design types is defined in OBI can be useful as a fast way to find or group an experiment or related research object, as a study designs are often indicative of the techniques and variables applied.  **Experimental protocol** is a more narrowly scoped subtype of a study design, that detail the steps to be executed in a given technique or experiment. A protocol is more limited in scope than a study design, specifying a specific combination of experimental designs, techniques, and variables necessary to do something, without defining goals or research hypothesis. Protocols are a core type of research object that has high value as sharable and re-usable artifacts in the scientific community. |
| **CONCEPT:** | **STUDY GROUP** |
| Label: | study group |
| Alt.Terms: |  |
| Definition: | A collection of subjects or specimens in a study that are defined by a common set of experimental variables (controlled, independent variables, and dependent variables) that apply to them according to the study design.. |
| Significance: | Another approach to modeling experimental metadata is study group centric. Given that study groups are defined by experimental variables, this approach is somewhat consistent with a variable-centric approach, but potentially more verbose in terms of the metadata collected for a given study. |
| **CONCEPT:** | **EXPERIMENTAL VARIABLE** |
| Label: | experimental variable |
| Alt.Terms: | variable, variable specification, parameter, experimental factor |
| Definition: | Experimental variables are specifications that describe the parameters according to which an experiment is executed. These parameters are captured as part of an experimental protocol, and include things that are to be experimental constants (controlled variables), things that will be varied to affect an outcome (independent variables), and the phenomena that will be observed or measured (dependent variables). |
| Significance: | Experimental variables are central to our model because (a) the paradigm of controlled, independent, and dependent variables as experimental parameters is shared across domains and disciplines, and (b) these concepts can be exploited to build a relatively lightweight and intuitive model that is organized around variable specification nodes, which summarize and contextualize the essential aspects of an experiment and capture important 'keywords' for discovery. Note that the phrase 'application of a variable' is used throughout our documentation, and refers to the process through which a variable specification is executed as part of an experiment. |
| **CONCEPT:** | **CONTROL VARIABLE** |
| Label: | control variable |
| Alt.Terms: | constant, constant variable, experimental constant, control variable specification, facilitatir |
| Definition: | An experimental variable that is specified to be held constant throughout and experiment (applied in uniform way to all specimens). A control variable might alter the dependent or independent variables, but is not the focus of the experiment and as such is kept constant or monitored to try to minimise its effect on the experiment. |
| Significance: | Controlled variables typically specify the participation of facilitating entities such as that are not varied between measurements, including the type or identity of certain reagents, instruments, or common attributes of specimens. For the purposes of capturing experimental metadata, we advocate representing controlled variables in terms of the type of entity held constant (instruments, reagents, specimen attributes), and the value that it takes in the experiment. |
| **CONCEPT:** | **INDEPENDENT VARIABLE** |
| Label: | independent variable |
| Alt.Terms: | experimental factor, factor, applied variable, predictor variable, explanatory variable, manipulated variable |
| Definition: | An experimental variable that is specified to be held constant throughout and experiment (applied in uniform way to all specimens) |
| Significance: | Controlled variables typically specify the participation of facilitating entities such as that are not varied between measurements, including the type or identity of certain reagents, instruments, or common attributes of specimens. For the purposes of capturing experimental metadata, we advocate representing controlled variables in terms of the type of entity held constant (instruments, reagents, specimen attributes), and the value that it takes in the experiment. |
| **CONCEPT:** | **DEPENDENT VARIABLE** |
| Label: | dependent variable |
| Alt.Terms: | measured variable |
| Definition: | An experimental variable that specifies the measurement made in a given experiment to determine an effect of varying one or more independent variables. |
| Significance: | For the purposes of capturing experimental metadata, we advocate representing dependent variables in terms of the attribute measured and the entity in which it inheres. |
| **CONCEPT:** | **REAGENT** |
| Label: | reagent |
| Alt.Terms: | research reagent, facilitator |
| Definition: | A role played by a material that participates chemically in some experimental process to facilitate the processing of a specimen or the application of an experimental variable. |
| Significance: | Only critical or distinctive reagents should be captured in experimental metadata, as these will be most useful for queries for and comprehension of an experiment. In our model a 'reagent role' is represented as the variable type for controlled variable specifications, to describe chemical or biological entities applied as general facilitators rather than independent variables that are varied across specimens. |
| **CONCEPT:** | **INSTRUMENT** |
| Label: | instrument |
| Alt.Terms: | device, scientific instrument, research tool |
| Definition: | A role played by an entity that participates inertly (does not react chemically) in some experimental process with a specified function that facilitates the processing of a specimen, application of an experimental variable, measurement of a phenomenon of interest, or processing of data. |
| Significance: | Only critical or distinctive instruments should be captured in experimental metadata, as these will be most useful for queries for and comprehension of an experiment. In our model 'instrument' is used as a variable type for controlled variable specifications, to describe devices and tools applied as general facilitators rather than independent variables that are varied across specimens.  **Software** is a considered a subtype of instrument, in that it is an inert facilitator of an experiment. |
| **CONCEPT:** | **SPECIMEN** |
| Label: | specimen |
| Alt.Terms: | research specimen, experimental specimen |
| Definition: | A role played by a material entity that serves as an exemplar instance of its kind, and is ultimately processed into an evaluant that is measured in an assay. This concept covers inanimate biological or abiotic samples, as well as living subjects of behavioral or epidemiological experiments. |
| Significance: | The concept of a 'research specimen' here is used specifically to refer to entities that are themselves, contain as part, or are derived into the evaluant (measured material entity) in which the dependent variable inheres. Conceptually, a given experiment may start with a source specimen, which gets processed into derived specimens, culminating in the production of the evaluated specimen (evaluant) that is directly measured in an assay. The notion of the 'experimental lineage' of a specimen captures this idea by capturing the type of entity represented at each stage in the processing of a source specimen into the evaluant. Capturing the type of entity represented at each step in the processing of a specimen is a critical aspect of experimental metadata, as these represent important keywords for discovery. Minimal context around the derivation of specimens must be captured in the model to support precise queries. |
| **CONCEPT:** | **SOURCE SPECIMEN** |
| Label: | source specimen |
| Alt.Terms: | originating specimen |
| Definition: | A specimen role played by the original specimen from which all other specimens are derived - typically an organism, population, or environmental sample. |
| Significance: | From the perspective of metadata capture, the source specimen does not have to be part of the performed experiment, but may be important to reference because the experimental evaluant was derived from it at some point in its history outside the experiment. For example, an experiment may start with the processing of a biospecimen, but it is useful to indicate that it derived from some 'homo sapiens' in its past by listing 'homo sapiens' as the source specimen. So for biological samples, the source specimen will typically be an organism. Another example is recording the brain as a source specimen in an experiment that collects glial macrophages from the brain but does not actually extract and document the brain as a physical sample. |
| **CONCEPT:** | **PRECURSOR SPECIMEN** |
| Label: | precursor specimen |
| Alt.Terms: | derived specimen |
| Definition: | A specimen role played by any specimen derived from the source in order to produce the evaluated specimen. |
| Significance: | As for the source specimen, from the perspective of metadata capture the derived specimen does not have to be physically generated during an experiment, but may still be important to reference because the experimental evaluant was derived from it at some point in its history. For example, recording the brain as a source specimen in an experiment that collects glial macrophages from the brain but does not actually extract and document the brain as a physical sample. This is because it is typically relevant to capture the entire lineage/derivation chain through which an experimental evaluant was derived in order to provide keywords for discovery across granularities (organism, organ, tissue, cell subcellular component, biomolecule). |
| **CONCEPT:** | **EVALUATED SPECIMEN** |
| Label: | evaluated specimen |
| Alt.Terms: | evaluant, experimental evaluant, unit of measurement |
| Definition: | A specimen role played by the specimen that is ultimately input into and subject to measurement in an assay. |
| Significance: | The evaluant is a critical aspect of an experiment to capture in the metadata, as it is relevant to what the experiment is about and what is measured in an assay. |
| **CONCEPT:** | **ENVIRONMENT** |
| Label: | environmental system, environmental context |
| Alt.Terms: | evaluant, experimental evaluant, unit of measurement |
| Definition: | The external elements and conditions which surround, influence, and affect a material entity organism, population, inanimate object. An environment is considered a system that is comprised of all materials that have the disposition to affect the life, development, structure, or qualities of a material within it. |
| Significance: | Environments can themselves be the subject of a study or experiment, but more often are relevant as metadata that contextualizes the origin of a specimen under investigation. Subtypes of environments relevant to our model include natural environments (where the specimen is natively found), and experimental conditions (applied in a laboratory/experimental setting). |

Note that an earlier but incomplete version of this glossary with additional terms and content can be found [here](https://docs.google.com/document/d/1BFYCicmgJ0-9Cm7ZTbHqQofp2hvXCB_0cAANNYD6ros/edit#heading=h.qhzllqemfrmz).

III. Reference Model: Ontological Hierarchy

The hierarchy below **presents our core domain concepts in a subsumption hierarchy** according to their basic ontological types (abstract information entities, attributes, material, processes). The goal is to clarify the ontological nature of each concept, and make apparent the subsumptive relationships between them.

**Informational Entity**

data set

study design

experimental protocol

experimental variable

control variable

independent variable

dependent variable

**Attribute/Role**

experimental role

investigator

reagent

instrument

experimental specimen

source specimen

precursor specimen

evaluated specimen

**Material Entity**

research organization

research facility

**Process**

research study

experiment

technique

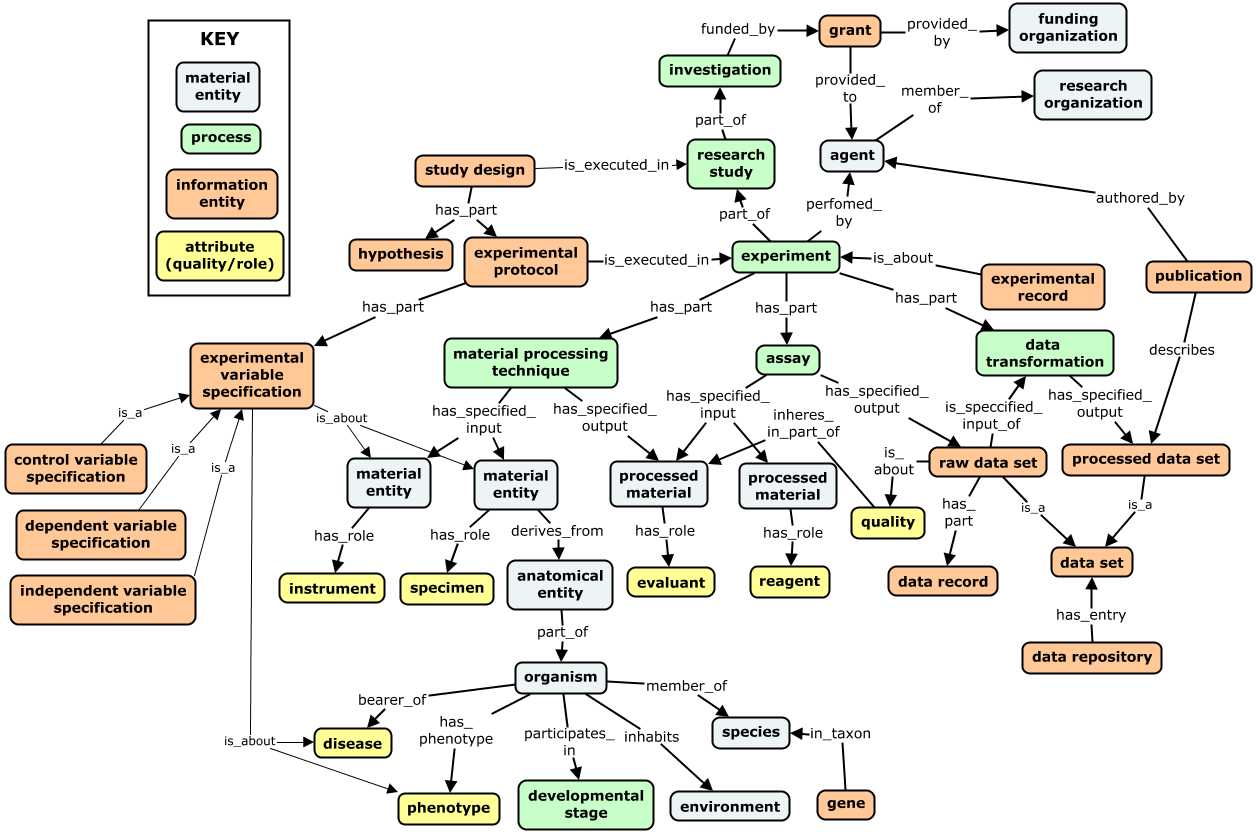
material preparation technique

assay

data processing technique

# IV. Reference Model: Concept Graph

The reference model **presents the relationships between core domain concepts** as they exist in the ‘real world’, to facilitate a shared understanding of their meaning, scope, and experimental roles. The model is ‘process-centric’ (focuses on experimental processes, and their participants and the roles inhering in them) - because this is how scientists think and work. This modeling perspective is useful for aligning our understanding of concepts because it presents them as we experience them in space and time. But due to its structure and complexity, this model is not well-suited for collecting experimental metadata and supporting our discovery use cases. Section VI presents these same concepts organized into a model more practically suited for implementation as a metadata schema.



# V. Metadata Model Specification: Hierarchical Schema

The informal metadata schema below presents an **entirely different perspective on our core domain concepts** - linking them in a way that is most efficient for the collection of metadata and returning query results. For this, we need to identify key component and abstract away extraneous details, to give a model that can be implemented to best meet our discovery and data entry requirements. Note that the design pattern of organizing data around variable specification nodes (minimally variable type and values) is utilized in many existing models, as it is an efficient way to conceptualize key concepts/keywords a meaningful way to scientists.

Core concept elements are distinguished in bold typeface. Some additional elements are included that specify how these core concepts are represented in the model (e.g. variable type and variable value elements are used to describe experimental variables). Cardinality specifications.constraints are indicated next to each element (e.g. ‘independent variable (1 . . . m)’ indicates that each experiment must have one, and can have multiple, independent variables). Note that some concepts will be captured as values in the schema rather than elements (e.g,. reagent, instrument, evaluant), so not all are depicted here.

**dataset**

**experiment** (1 . . .m)

**investigator** (1 . . . m)

**research organization** (1 . . . m)

**research facility** (1 . . .m)

**occurs at time** (1 . . . 1)

**technique** (0 . . . m)

**part of study (0...m)**

**study design (0...m)**

**control variable** (**reagent**) (0 . . . m)

variable type (1. . .1)

variable value (1 . . . 1)

variable characteristic (0 . . .m)

resource description (0 . . . 1)

**controlled variable** (instrument) (0 . . . m)

variable type (1. . .1)

variable value (1 . . . 1)

variable characteristic (0 . . .m)

resource description (0 . . . 1)

**controlled variable** (evaluated specimen) (0 . . . m)

variable type (1. . .1)

variable value (1 . . . 1)

variable characteristic (0 . . .m)

resource description (0 . . . 1)

**independent variable** (1 . . .m)

variable type (1. . . 1)

variable value (1 . . . m)

affected specimen (0 . . .m)

related technique (0 . . .m)

related condition (0 . . . m)

**dependent variable** (1 . . .1)

variable type (1. . .1)

measured attribute (1 . . . 1)

related technique (0 . . .m)

has scale (0 . . . 1)

related condition (0 . . . m)

The representation of an **exemplar metadata record** for an exemplar experiment highlights fact that the metadata is relatively simple and flat, while capturing needed precision and variable context for core CQs and modeling requirements. The record is presented in a pseudo-json format. Note that all values not in “quotes” are terms from community CVs/ontologies.

{

**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"}

]

**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}

]

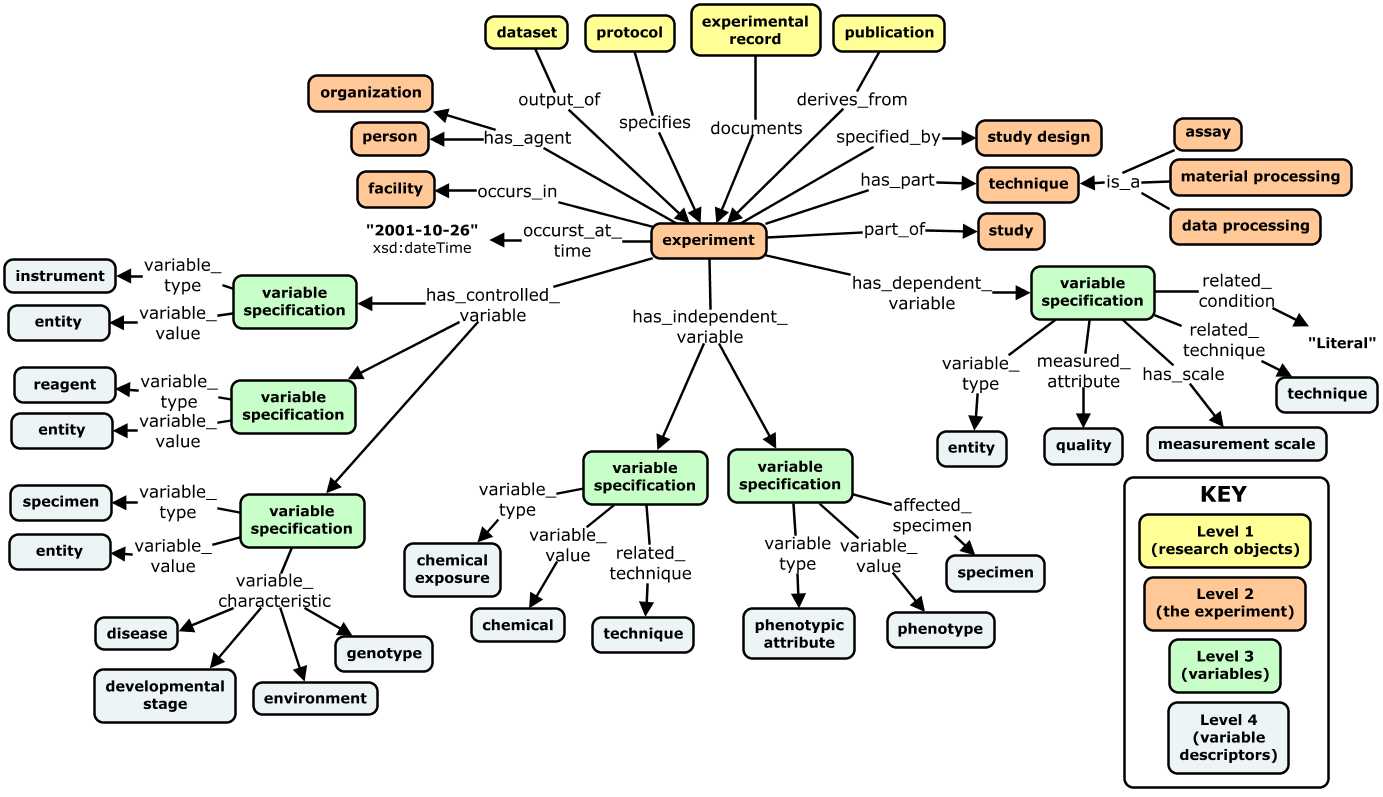
}

}

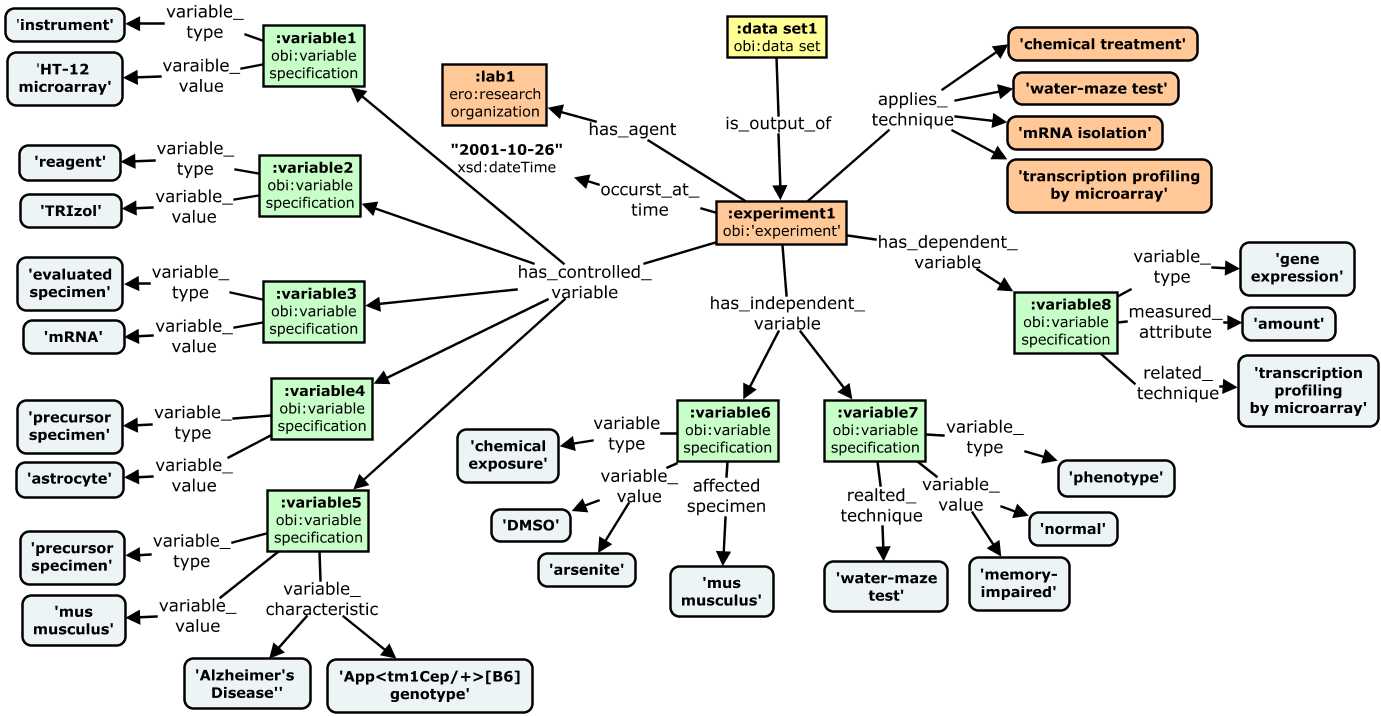
# VI. Metadata Model as a Concept Graph and Exemplar Record

As noted above, organizing core concepts as a metadata model links them in a way that makes sense for collecting and experimental structuring metadata. For this, we need to identify key components and abstract away extraneous details, to give a model that can be implemented to best meet our discovery and data entry requirements. Note that the design pattern of organizing data around variable specification nodes (minimally variable type and values) is utilized in many existing models, as it is an efficient way to conceptualize key concepts/keywords a meaningful way to scientists.

**Class-level model (with some exemplar variable types and variable values)**



**Metadata Record for an Exemplar Experiment (that is described in** [**Appendix 1**](#id.py6trkjlbqkr)**)**



-----------------------------------------------------------------------------------------------------------------------

Appendix 1: Description of Exemplar Experiment

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

**Procedure:**

1. **Treatment:** exposure of mouse model of Alzheimers to DMSO (control) or arsenite (oxidative stressor)
2. **Specimen Collection #1:** collection of astrocytes from hippocampi of normal vs memory-impaired mice
3. **Specimen Collection #2:** isolation of mRNA using TRIzol reagent to obtain mRNA samples
4. **Assay**: expression microarray analysis performed using an Illumina HT-12 v4 array to identify patterns in level of gene expression