The Paradata Information Model

NADDI

April 2013

Lawrence, KS

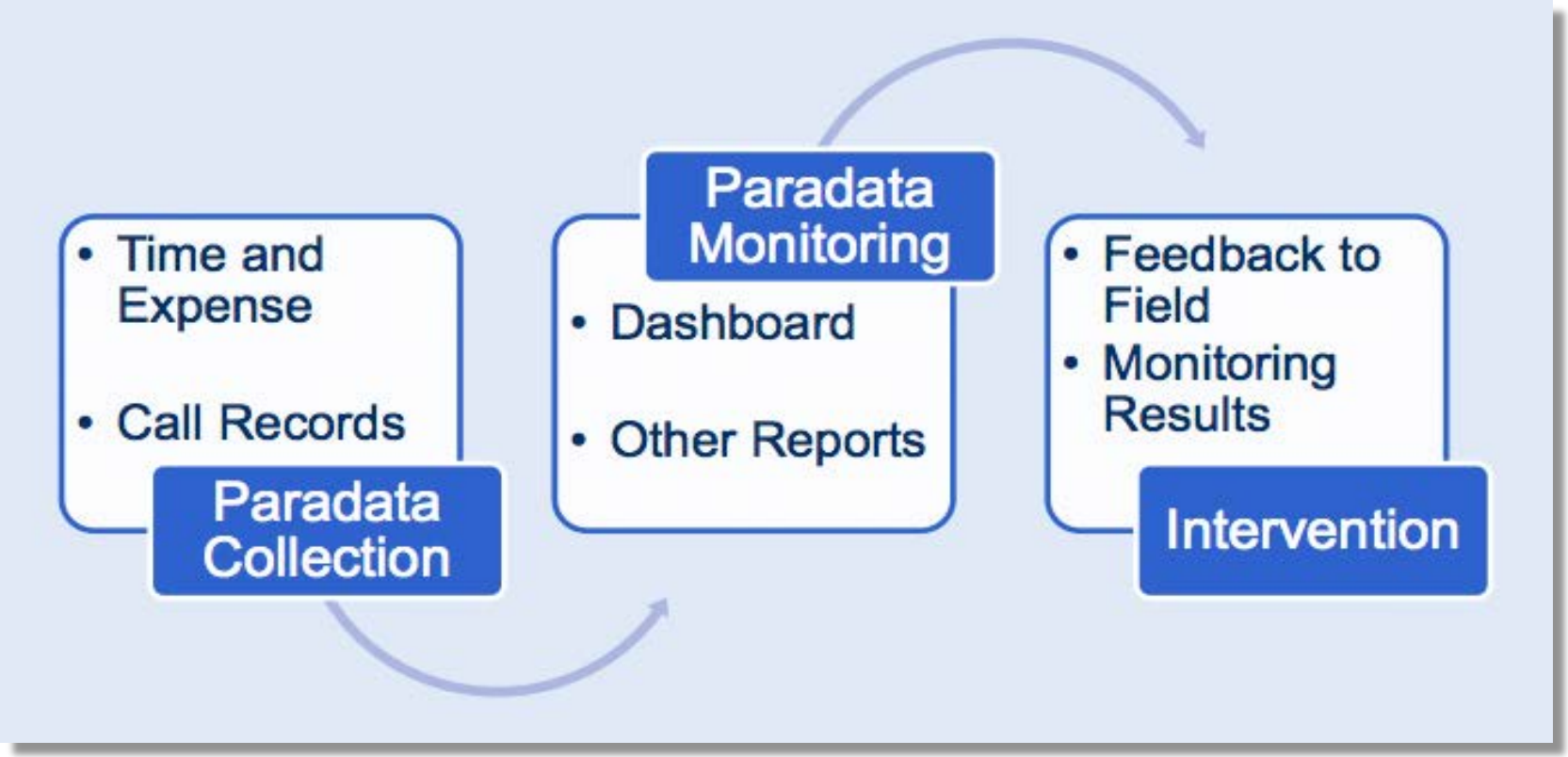
Agenda

- Motivation for the Paradata Information Model (PIM)
  - The National Children's Study (NCS)
  - The National Science Digital Library (NSDL)
- Model Building
  - Bricoleur and Bricolage
  - Meet the Strawman: Generic Longitudinal Business Process Model (GLBPM)
  - Specializing GLBPM
  - The PIM Formalism
- Sequencing Data Collection, Data Processing and Data Understanding Activities
  - The Microarray Experiment Use Case
  - Understanding Sequences
  - The Gamification of GSBPM
- Next Steps

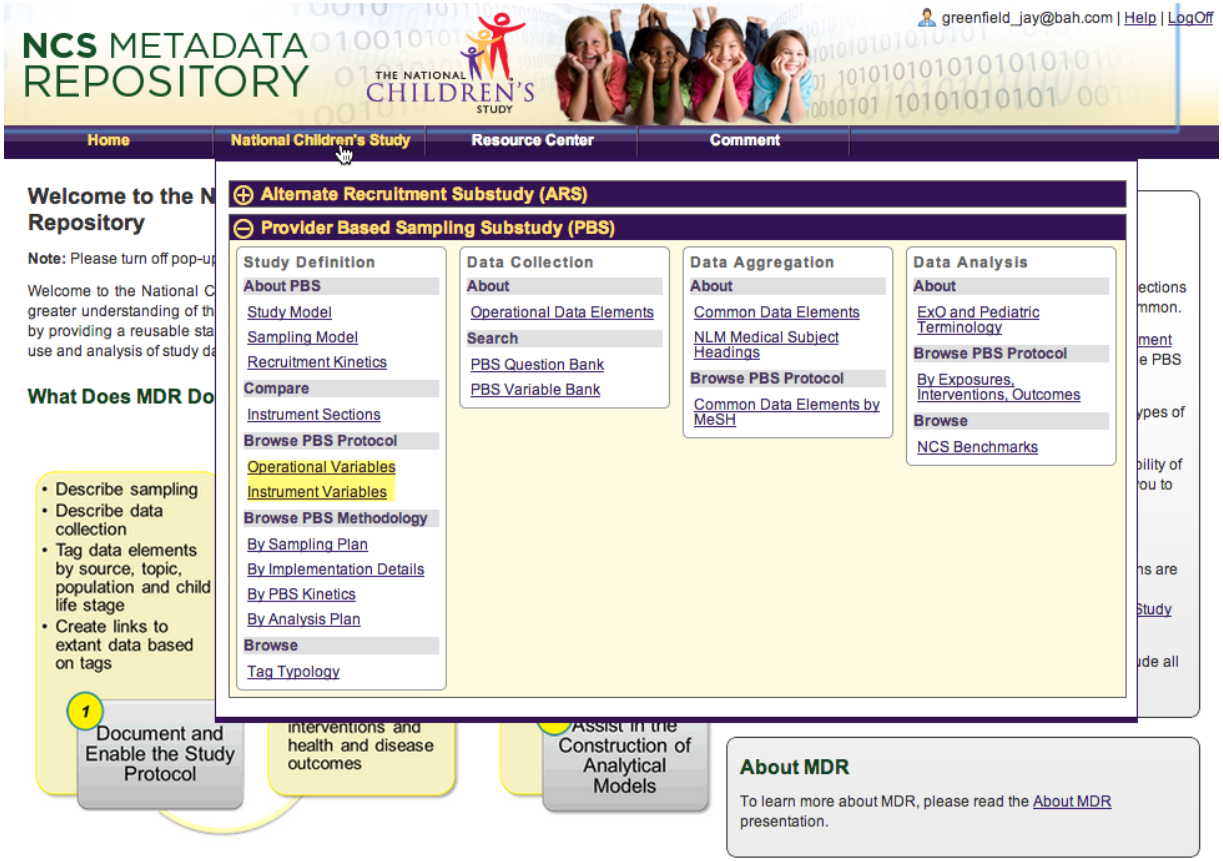
Use Case

PARADATA IN THE NATIONAL CHILDREN'S STUDY

In the NCS so-called operational data elements were defined and designed to assist in the assessment of feasibility, acceptability and cost



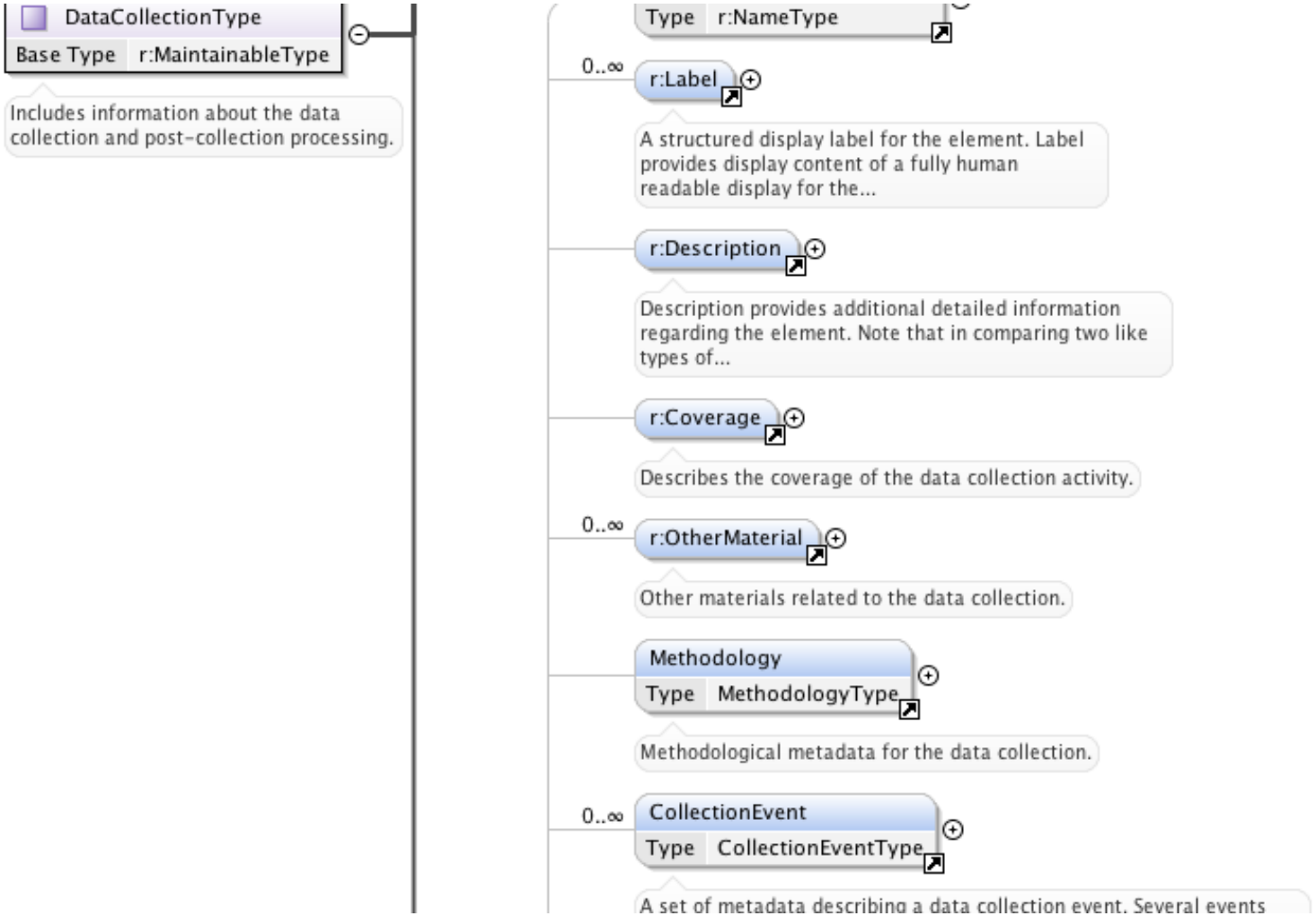
From Lepkowski and Kirgis based on work by Mosher and Groves



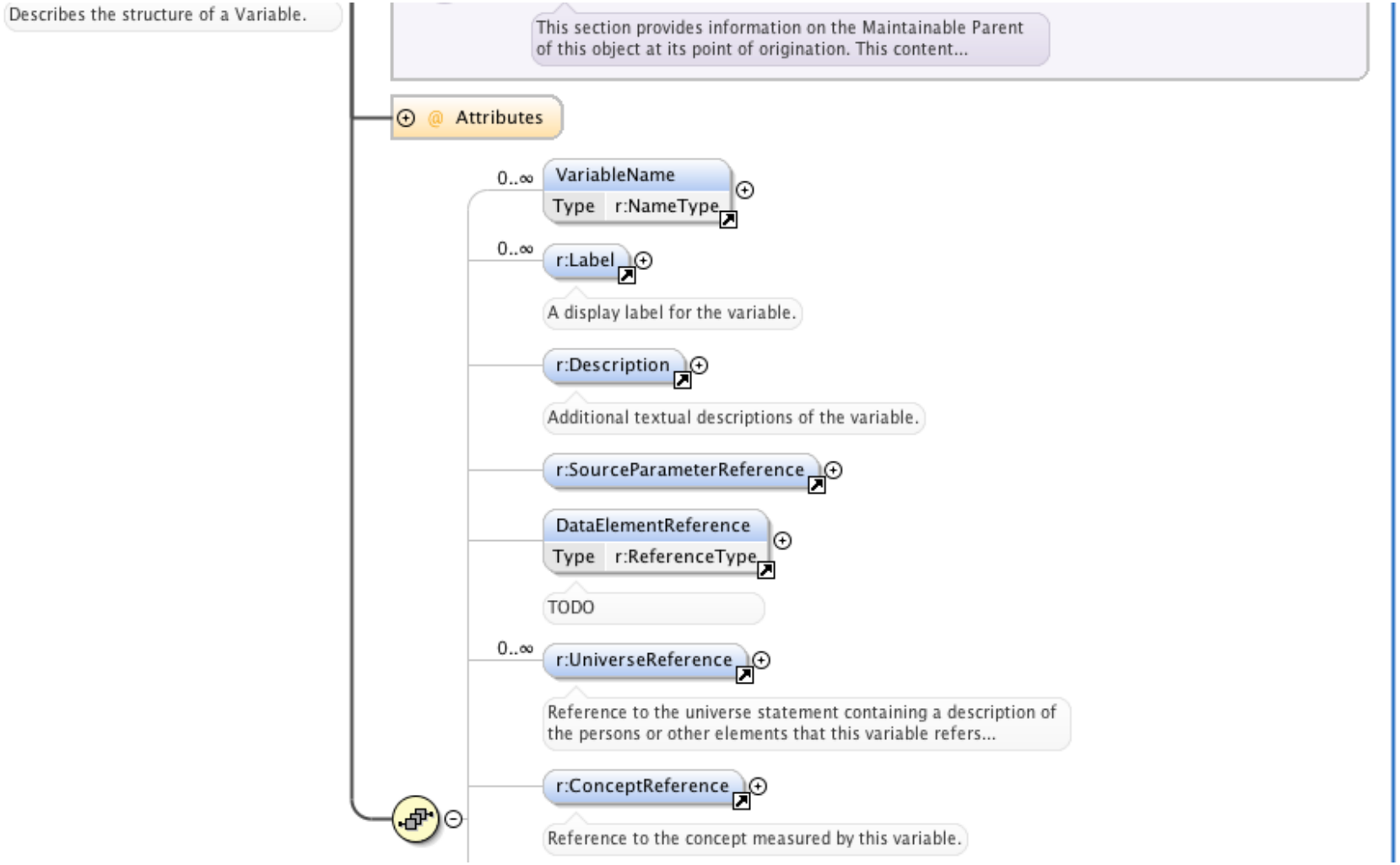
Demonstration

Instrument and Operational Data Elements

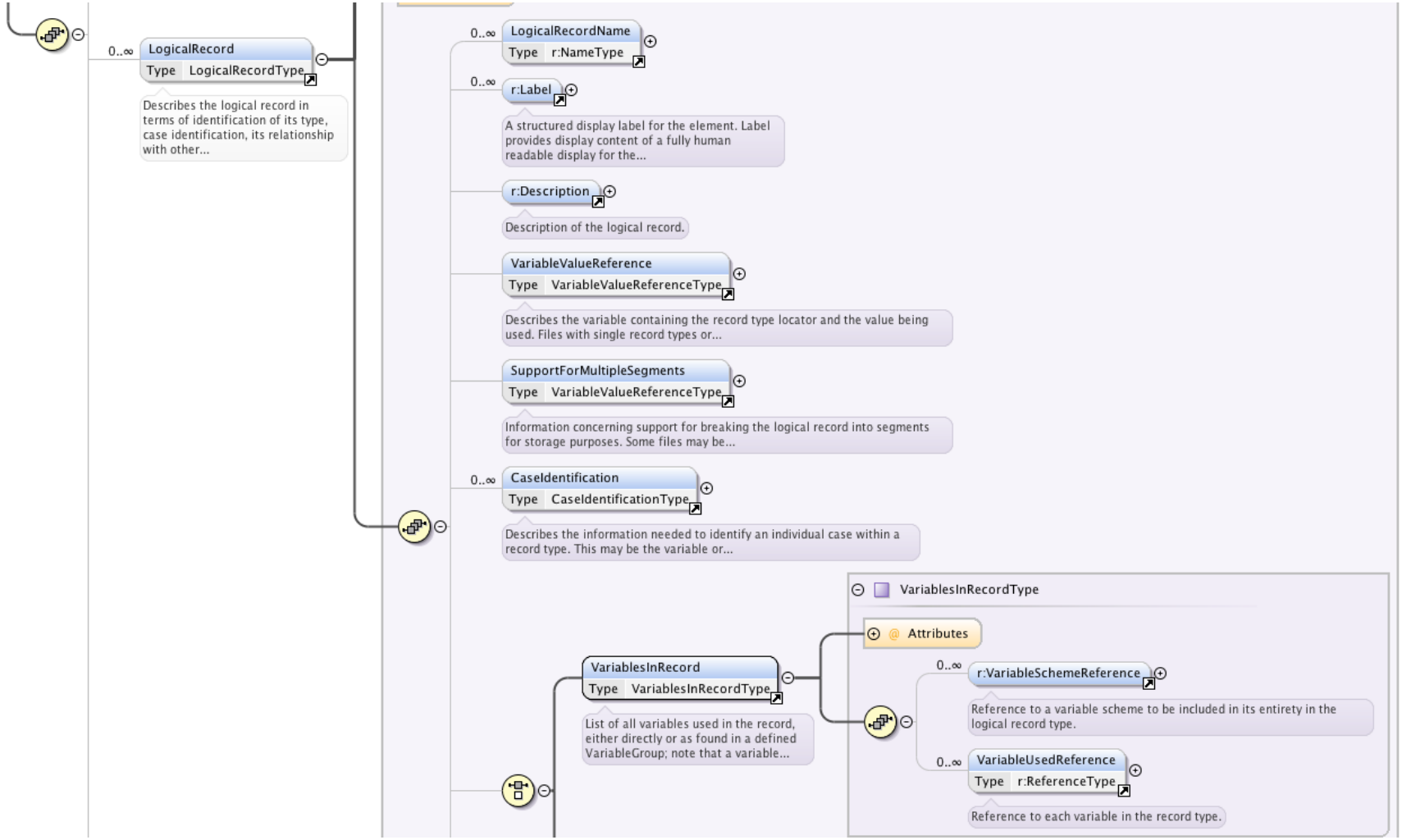
Paradata in the NCS is a DataCollectionType without any instruments…



In DDI our ODEs are variables…



…that we locate in logical records



There is in fact a better way…

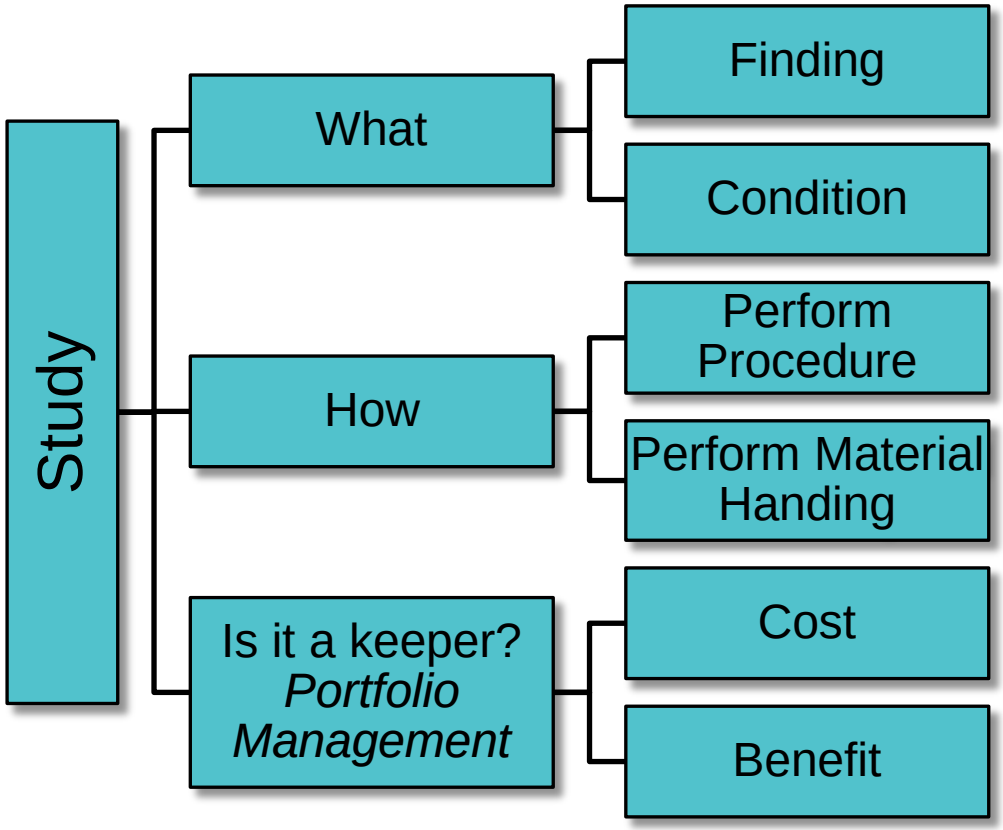
- What we haven't done is a best practice described in [Documenting a Wider Variety of Data Using the Data Documentation Initiative 3.1]
  - Here it is suggested that we first use ***<Instrument>*** to document a biospecimen data collection
  - Then it is suggested that we use a ***<ProcessingEvent>*** and its ***<GenerationInstruction>*** to tie blood draw with a ***<LogicalProduct>***
  - The ***<LogicalProduct>*** contains assay findings from the blood draw – certain cardiovascular variables
- This is useful and a great best practice
- But the cardiovascular variables are findings
  - ***Findings*** come at the end of a ***sequence*** in which first ***a procedure is performed*** at a visit and then ***the material collected during the procedure is handled and processed***

		B4BCHOL Blood Total Cholesterol (mg/dL)	B4BTRIGL Blood Triglycerides (mg/dL)	B4BHDL Blood HDL Cholesterol (mg/dL)
N	Valid	1244	1244	1242
	Missing	11	11	13
Mean		186.59	132.52	55.37
Std. Deviation		40.170	131.819	17.982
Range		348	3274	102
Minimum		91	25	19
Maximum		439	3299	121

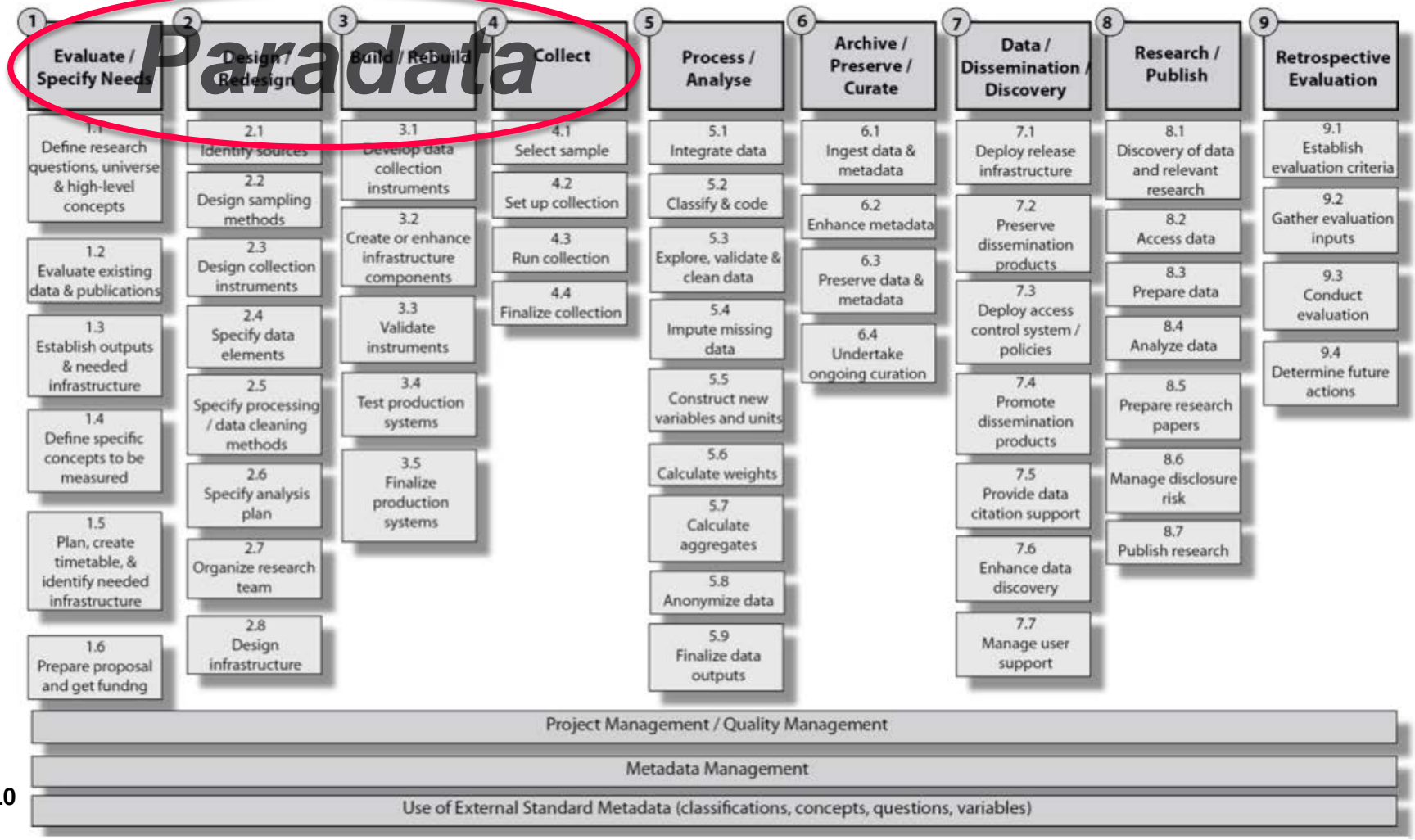
Table 1. Descriptive statistics of select MIDUS cardiovascular variables

Putting findings in context…

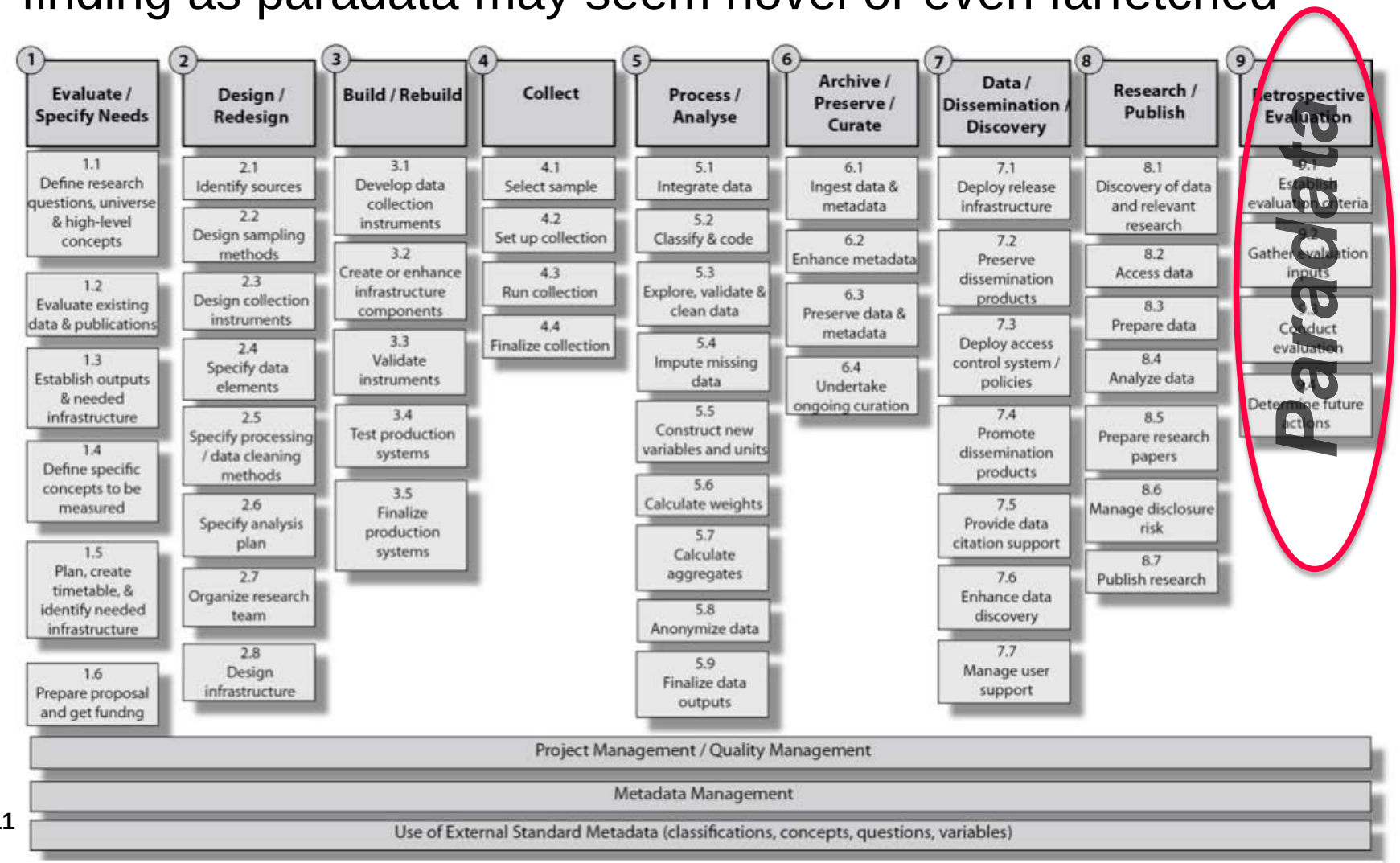
- What we need is for our ***<ProcessingEvent>*** with its ***<GenerationInstruction>*** to account not just for the finding, i.e. the ***data***
- Additionally, it needs to describe the process that makes the finding occur. i.e. the ***paradata***
- We specifically need the paradata to qualify the data
  - How does one go and ***reproduce*** a finding?
  - What did it ***cost***?
  - What ***value*** does it have?



We perform activities in the service of collecting, processing and evaluating data. We define paradata variables. There is paradata metadata. Should the design and collection of paradata be in this picture? This is GLBPM…

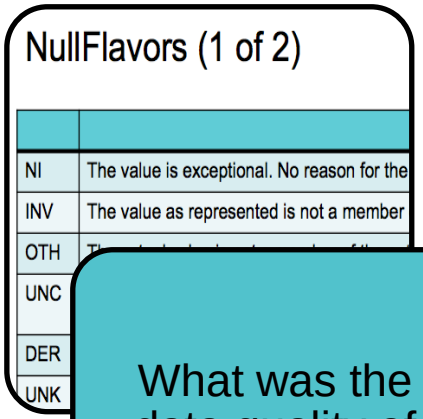


Let's also talk about how paradata figures into ***retrospective evaluation*** more specifically. This is perhaps a "reach" because the idea of considering the "***buzz***" surrounding a finding as paradata may seem novel or even farfetched



A retrospective evaluation of cow's milk…

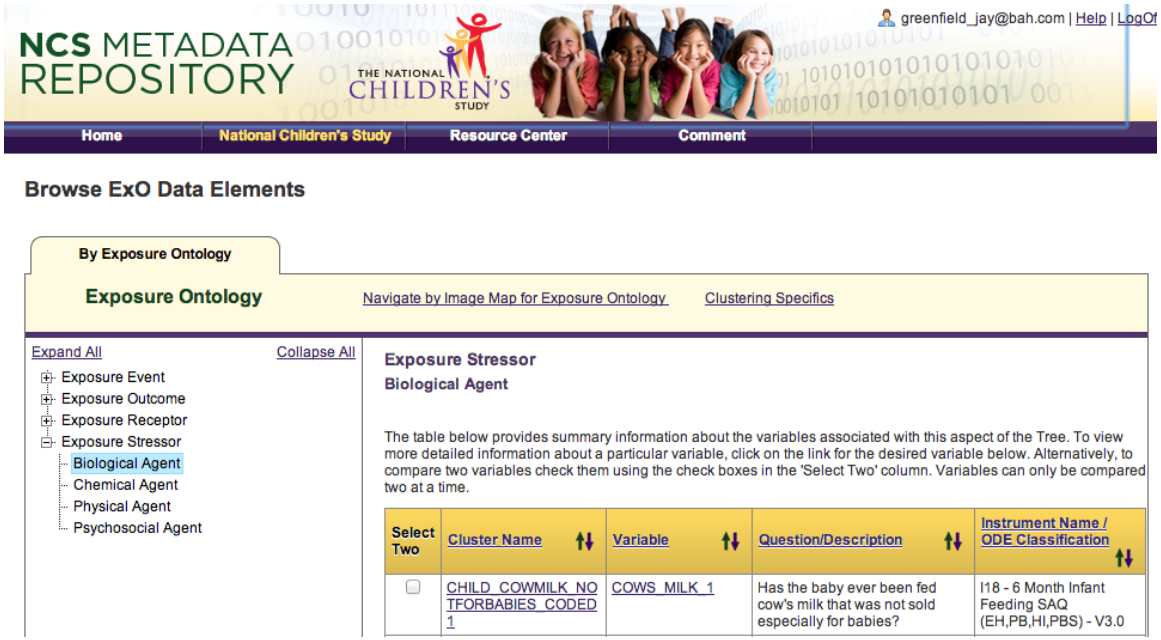
Variable	Question/Description	Instrument Name / ODE Classification
COWS_MILK_1	Has the baby ever been fed cow's milk that was not sold especially for babies?	I18 - 6 Month Infant Feeding SAQ (EH,PB,HI,PBS) - V3.0



What was the data quality of this question's variable?

Given the question's intent what chatter should we be listening for in the research community?

What is the chatter?

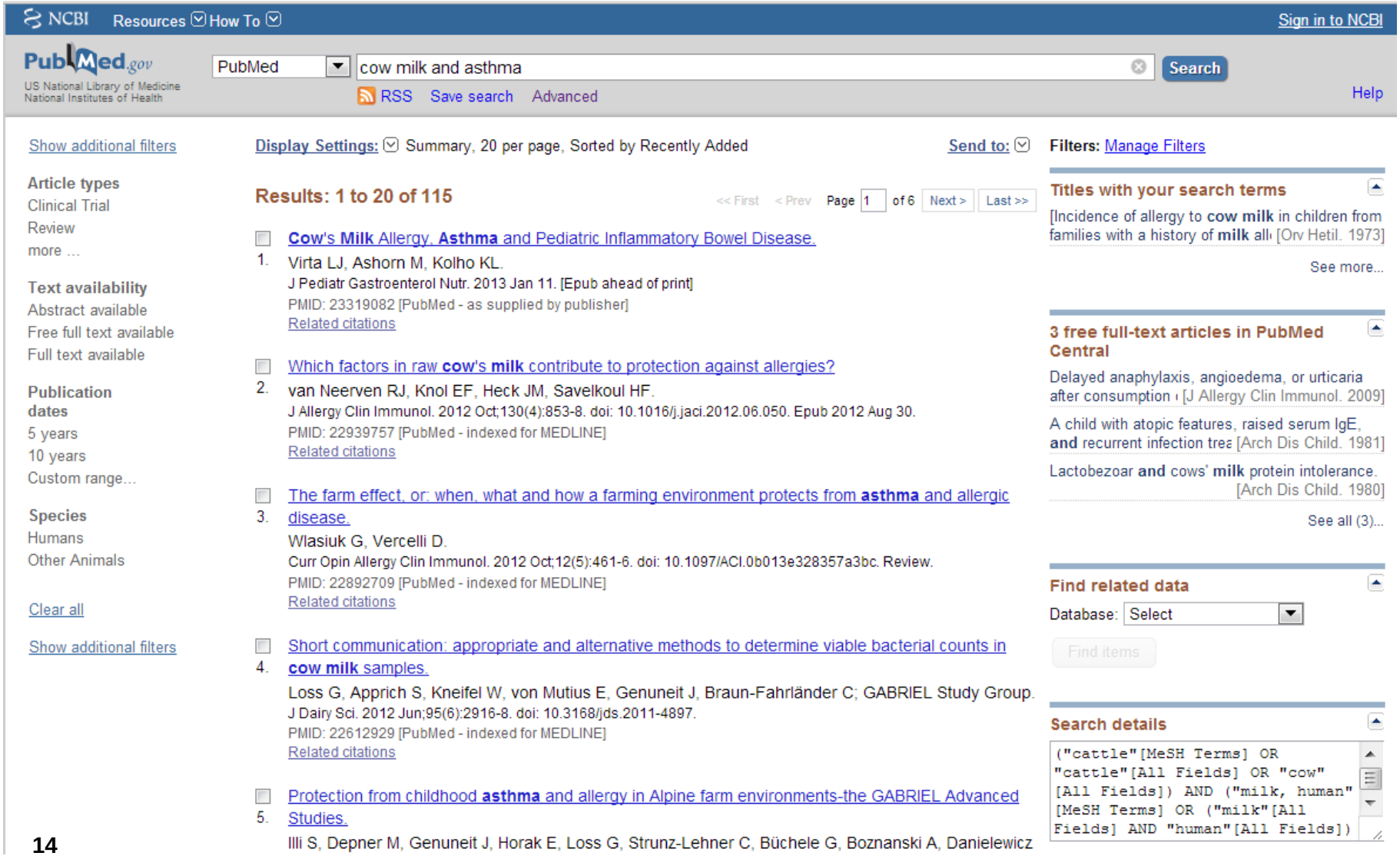


Demonstration

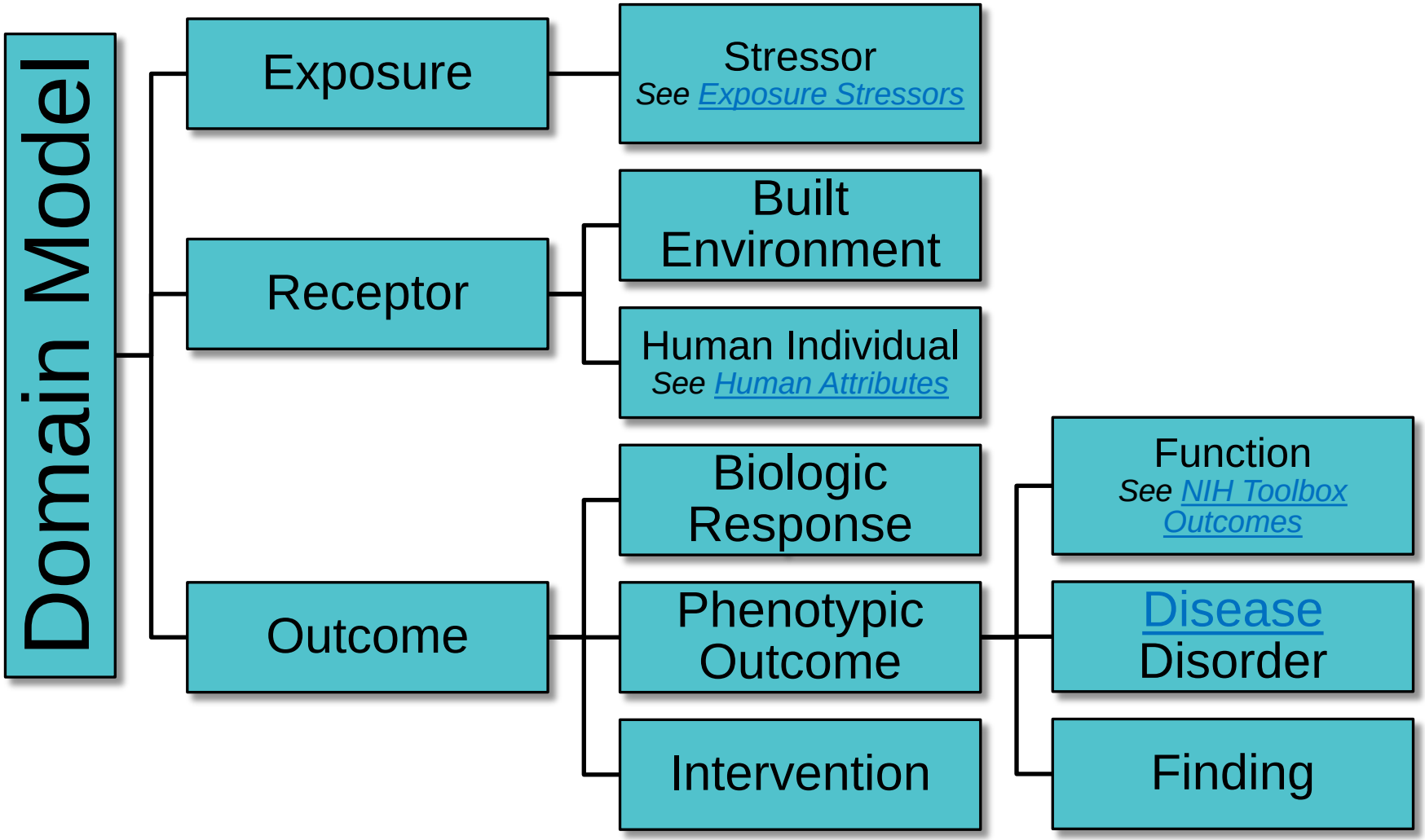
Cow's Milk

We can search PubMed using our MeSH terms

What is the chatter?

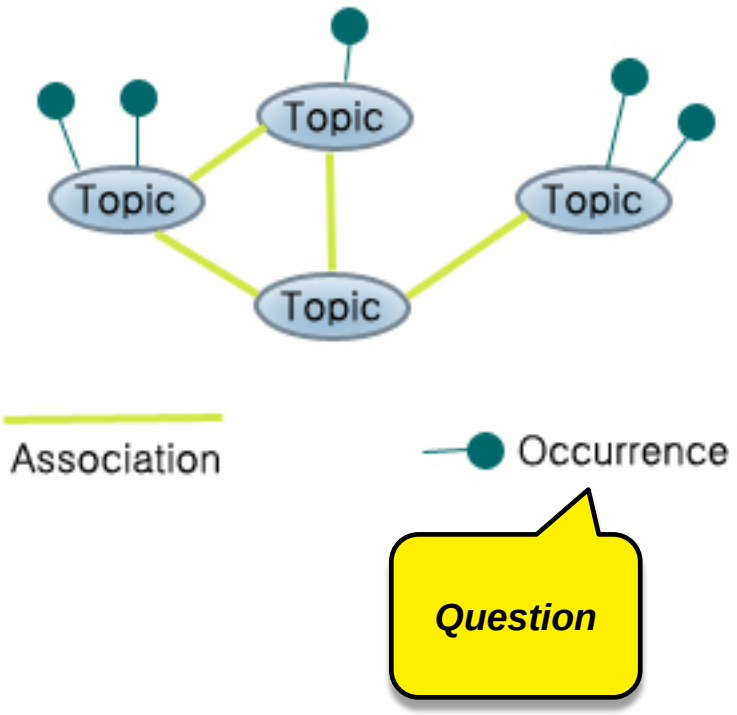


We return to our domain model and look for asthma under diseases and disorders



Documenting the chatter

- There is perhaps an association between cow's milk ingested by children and asthma under certain circumstances
- The type of association is a "pathway"
- Pathways make our data hypothesis aware



Data Discovery

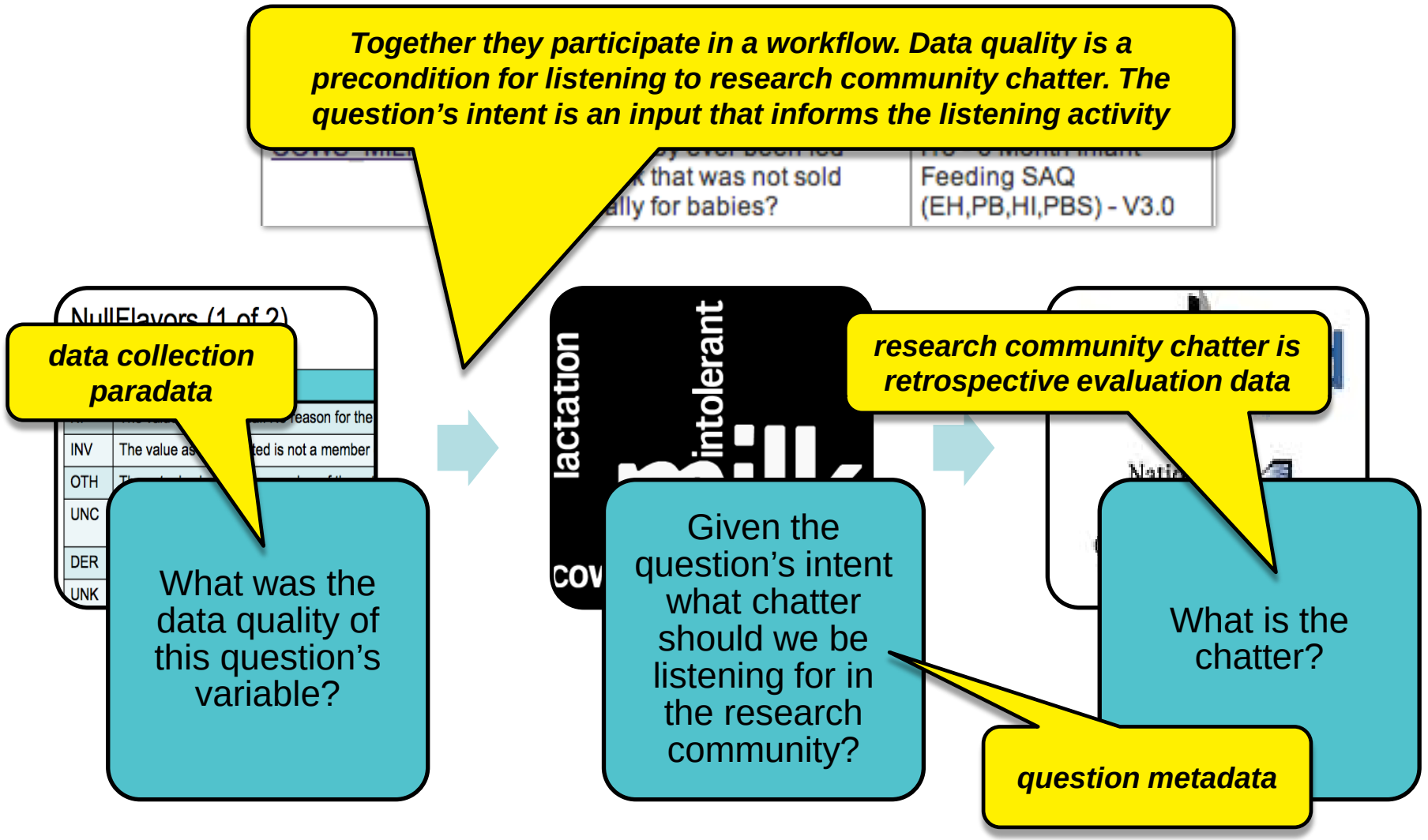
Revisit Data Elements

Modify PHI and PII tagging of data elements to facilitate access control?

Add dimensions to the knowledge space in which data elements are located?

If yes, modify data element tags to locate them in more dimensions

A retrospective evaluation of cow's milk…



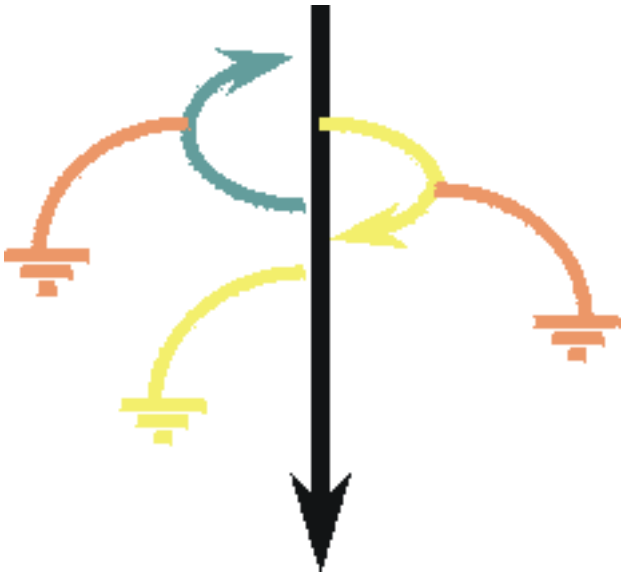
Lessons Learned: Paradata doesn't seem to know any boundaries ☺

From NSDL: In creating the concept of the STEM Exchange, we needed to distinguish between traditional, relatively static metadata that *describes* a digital learning object and the dynamic information about digital learning objects that is generated as they are used, reused, adapted, contextualized, favorited, tweeted, retweeted, shared, and all the other social media style ways in which educational users interact with resources. In this context, ***paradata captures the user activity related to the resource that helps to elucidate its potential educational utility***. We first presented paradata at a March 3, 2010 presentation as...

- a complement to metadata, not a replacement
- separate layer of information from metadata
- ***a means to automate information generation about resource use by using social networking tools***
- a means to create an open source and open access data space around resources
- emphasizes dissemination rather than description
- accommodates expert and user-generated knowledge

Lessons Learned: Research is a practice with steps. Paradata can be about one step at a time like call records, question navigation or question acceptability but in the end we all walk around with very long checklists and paradata is a ***multi-step process***

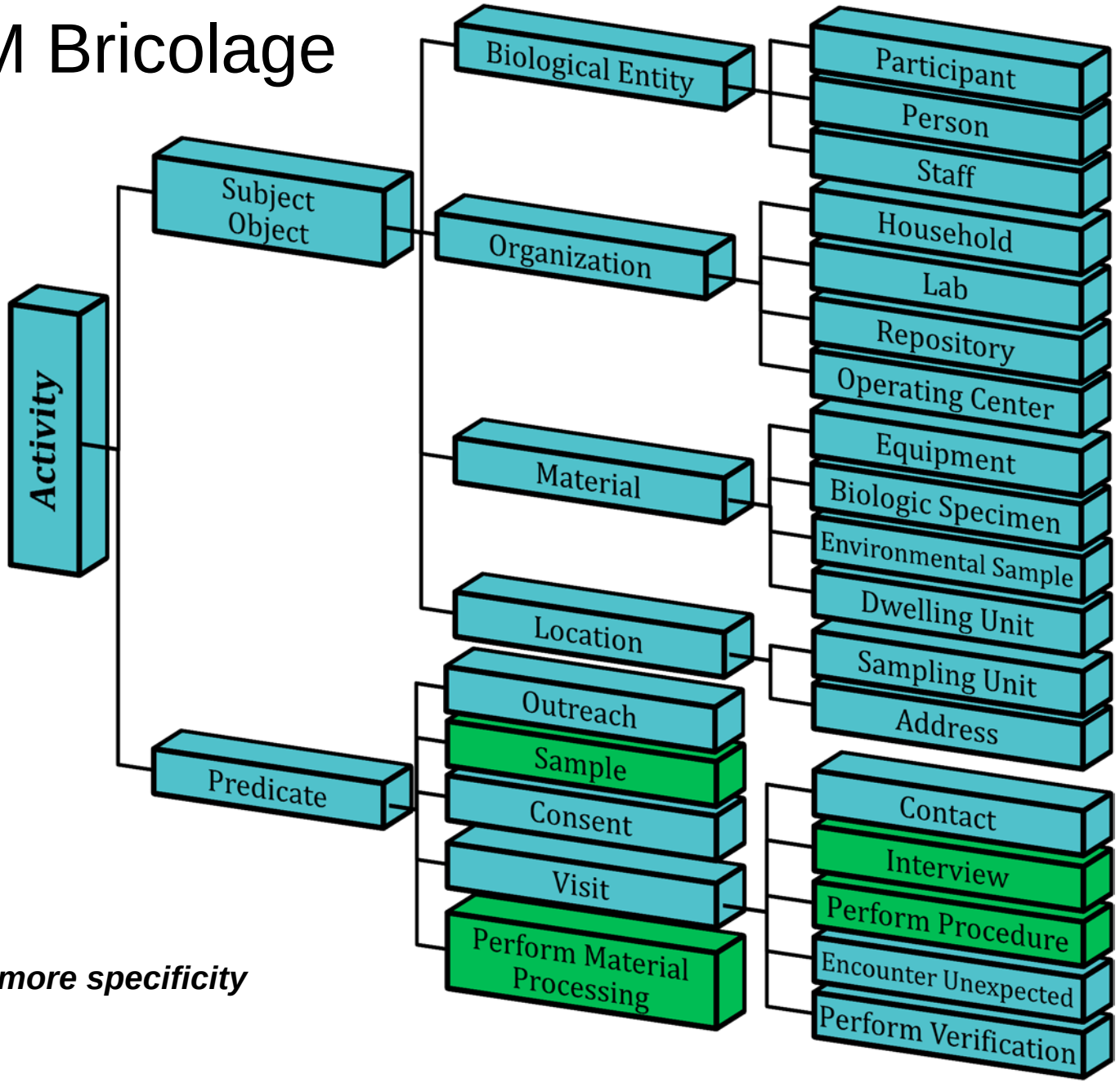
- Has an EHR record for this participant been received (***paradata***)?
- If not, what is the next visit in the protocol for this participant (***paradata + metadata***)?
  - Does it include performing a procedure (***metadata***)?
  - If so, what preparations have been undertaken (***paradata***)?
    - When was the equipment last calibrated (***paradata***)?
    - Is a kit required (***metadata***)?
    - If so, is this kit in store (***paradata***)?



Bricoleur and Bricolage

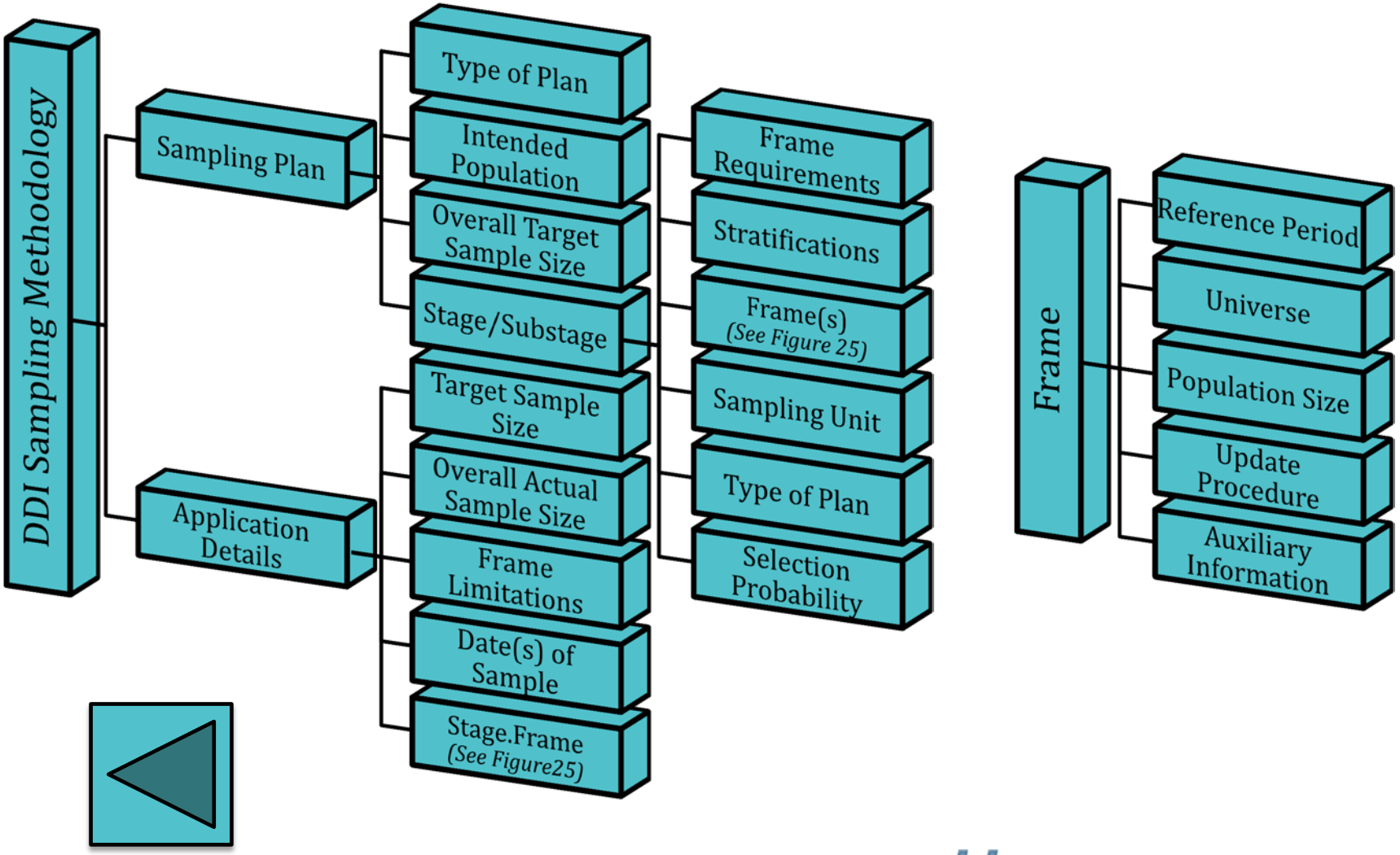
A PARTIAL PARADATA MODEL AND HOW WE BUILT IT

The PIM Bricolage

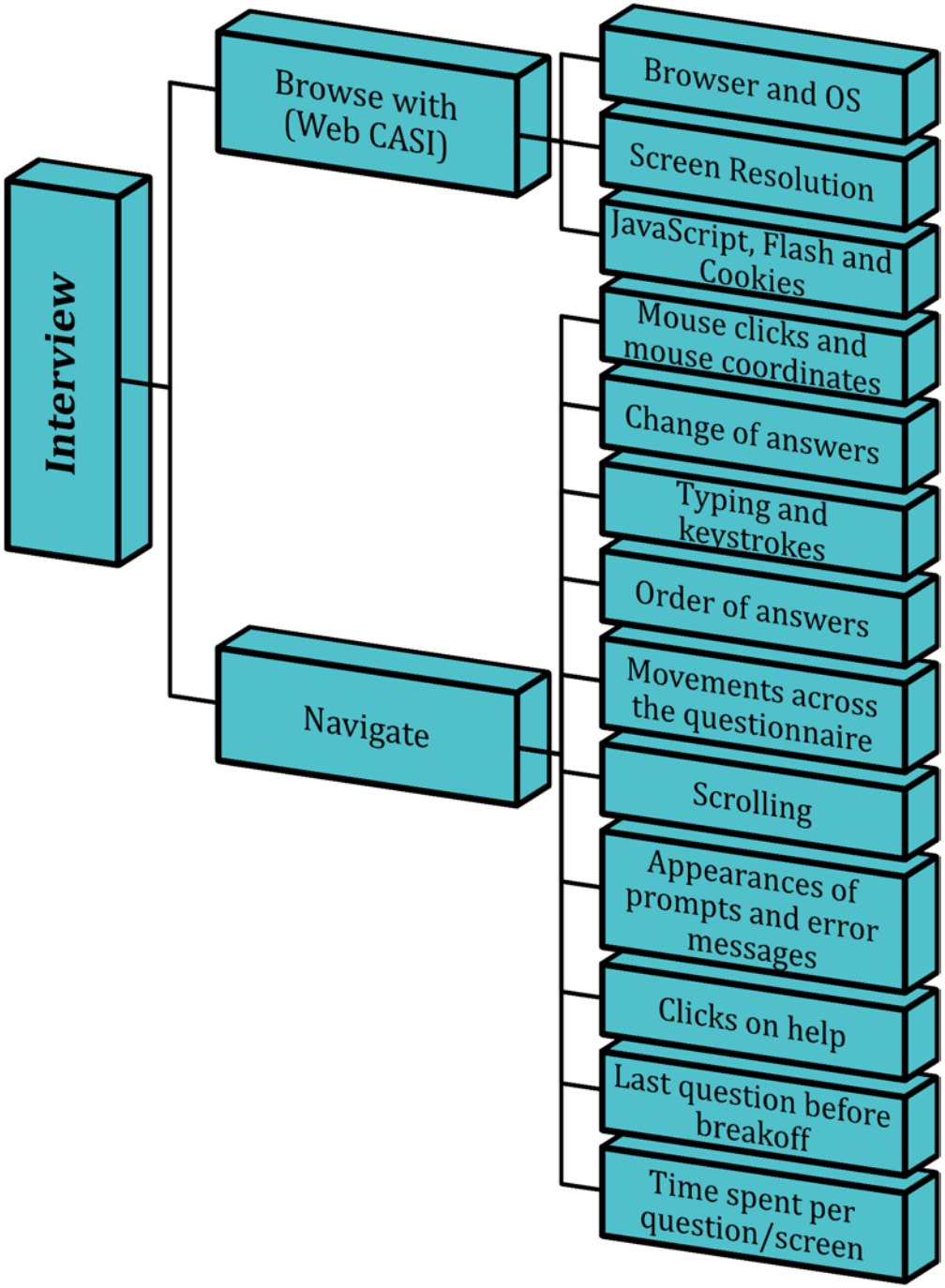


Click green for more specificity

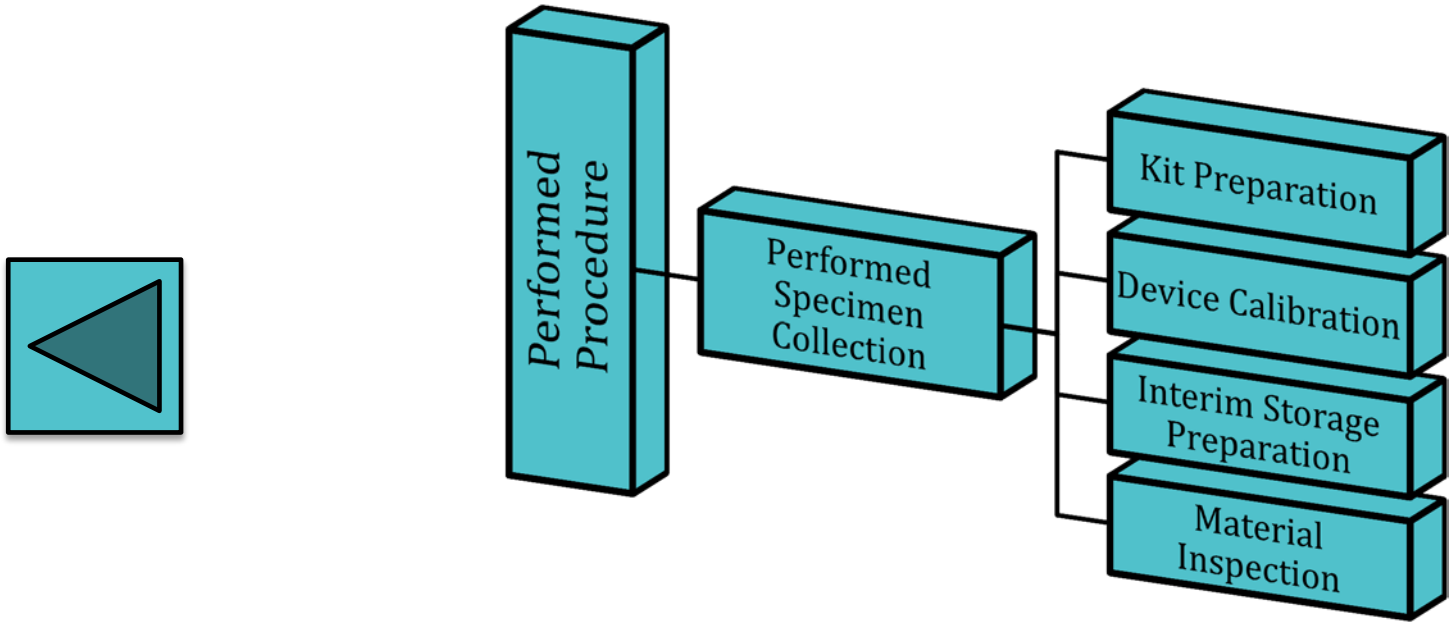
Sampling


DDI Sampling Methodology
Sampling Plan
Application Details
Type of Plan
Intended Population
Overall Target Sample Size
Stage/Substage
Target Sample Size
Overall Actual Sample Size
Frame Limitations
Date(s) of Sample
Stage.Frame (See Figure25)
Frame Requirements
Stratifications
Frame(s) (See Figure 25)
Sampling Unit
Type of Plan
Selection Probability
Frame
Reference Period
Universe
Population Size
Update Procedure
Auxiliary Information

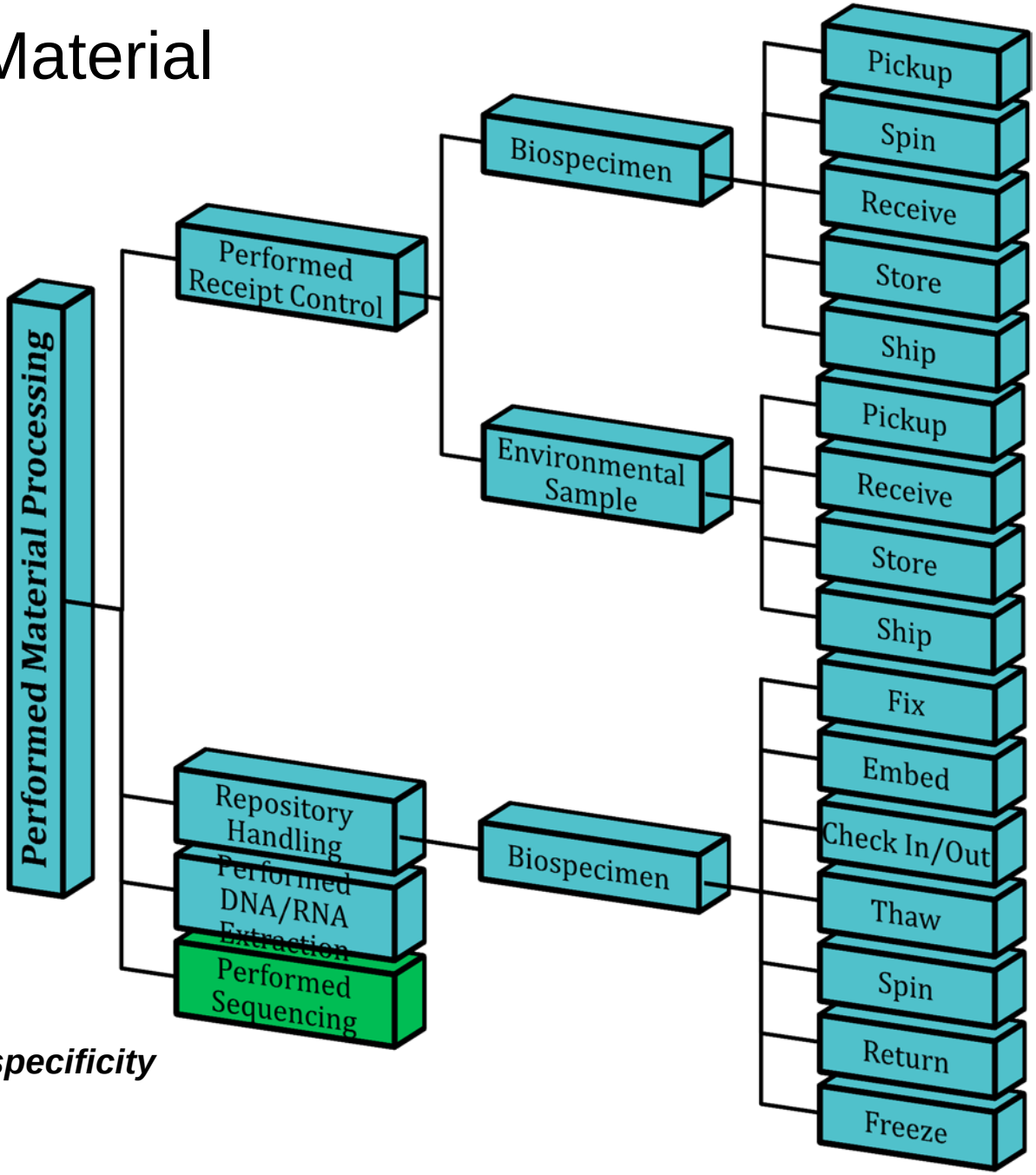
Interview


Interview
Browse with (Web CASI)
Navigate
Browser and OS
Screen Resolution
JavaScript, Flash and Cookies
Mouse clicks and mouse coordinates
Change of answers
Typing and keystrokes
Order of answers
Movements across the questionnaire
Scrolling
Appearances of prompts and error messages
Clicks on help
Last question before breakoff
Time spent per question/screen

Performed Procedure

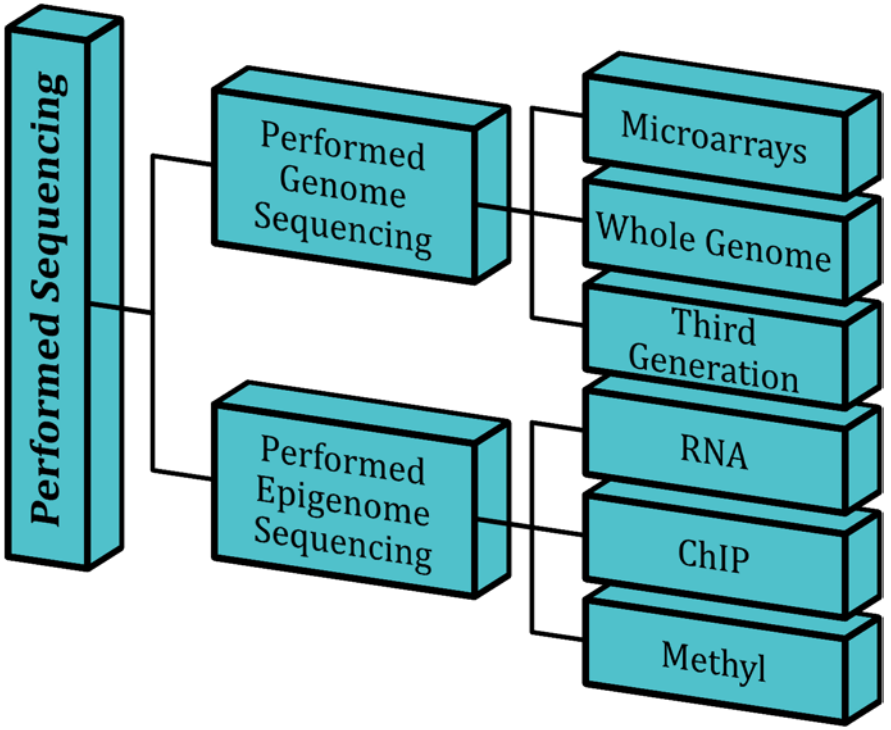


Performed Material Processing



Click green for more specificity

Performed Sequencing



Some paradata information model construction principles (1 of 2)

- We don’t need to be right. We just don’t want to be wrong
- We begin with the strawman
  - The strawman is an upper ontology or in the context of UML modeling the strawman consists of upper layer components in a conceptual model that is a work in progress
  - We don’t create our own strawman
  - We borrow a published one
- We specialize a strawman by once again borrowing from other published models
  - We are bricoleurs constructing bricolage
  - In his book *The Savage Mind* (1962, English translation 1966), French anthropologist Claude Lévi-Strauss used 'bricolage' to describe the characteristic patterns of mythological thought. In his description it is opposed to the engineers' creative thinking, which proceeds from goals to means. Mythical thought, according to Levi-Strauss, attempts to re-use available materials in order to solve new problems.[3]

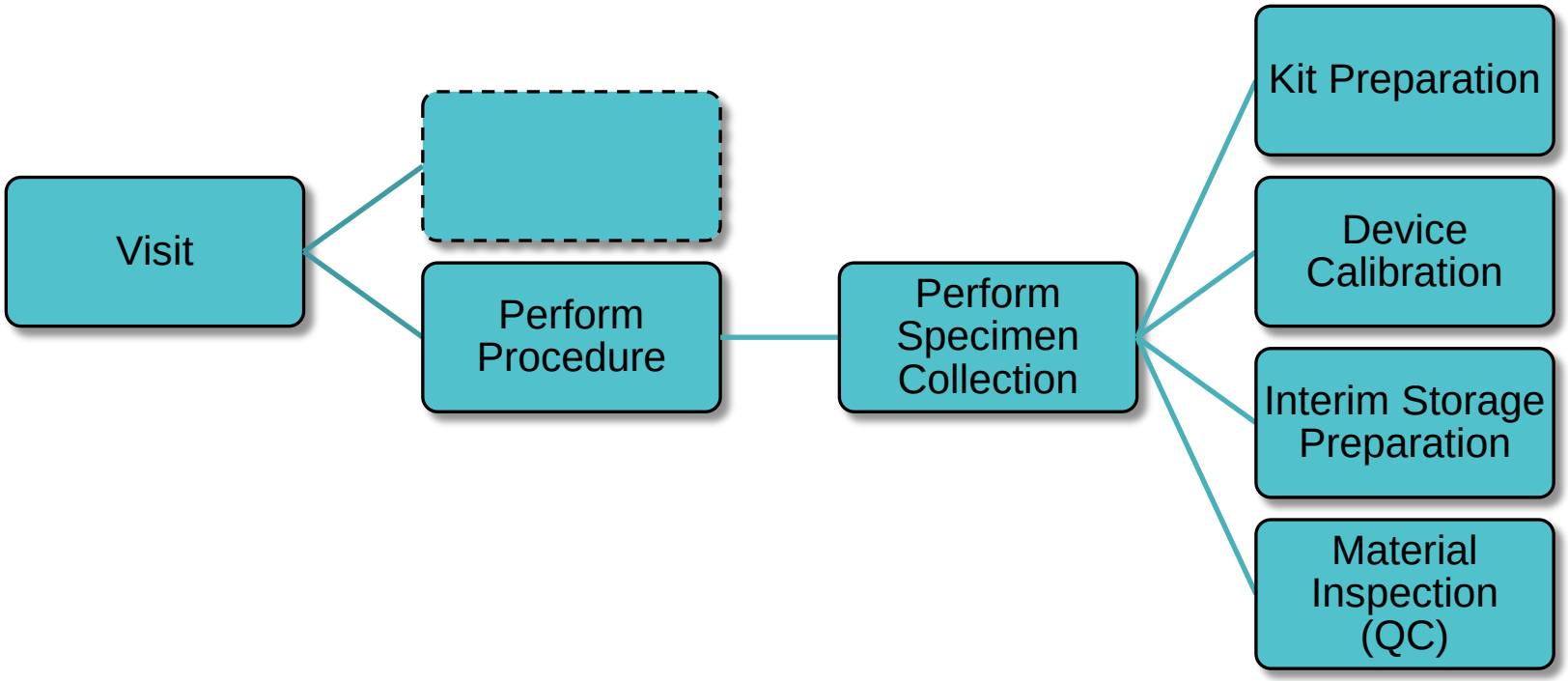
Some paradata information model construction principles (2 of 2)

- In information systems, bricolage is used by Claudio Ciborra to describe the way in which strategic information systems (SIS) can be built in order to maintain successful competitive advantage over a longer period of time than standard SIS. By valuing tinkering and allowing SIS to evolve from the bottom-up, rather than implementing it from the top-down, the firm will end up with something that is deeply rooted in the organisational culture that is specific to that firm and is much less easily imitated.[6]
- In her book *Life on the Screen* (1995), Sherry Turkle discusses the concept of bricolage as it applies to problem solving in code projects and workspace productivity. She advocates the "bricoleur style" of programming as a valid and underexamined alternative to what she describes as the conventional structured "planner" approach. In this style of coding, the programmer works without an exhaustive preliminary specification, opting instead for a step-by-step growth and re-evaluation process. In her essay "Epistemological Pluralism", Turkle writes: "The bricoleur resembles the painter who stands back between brushstrokes, looks at the canvas, and only after this contemplation, decides what to do next."[7]

- Our strawman is GLBPM…

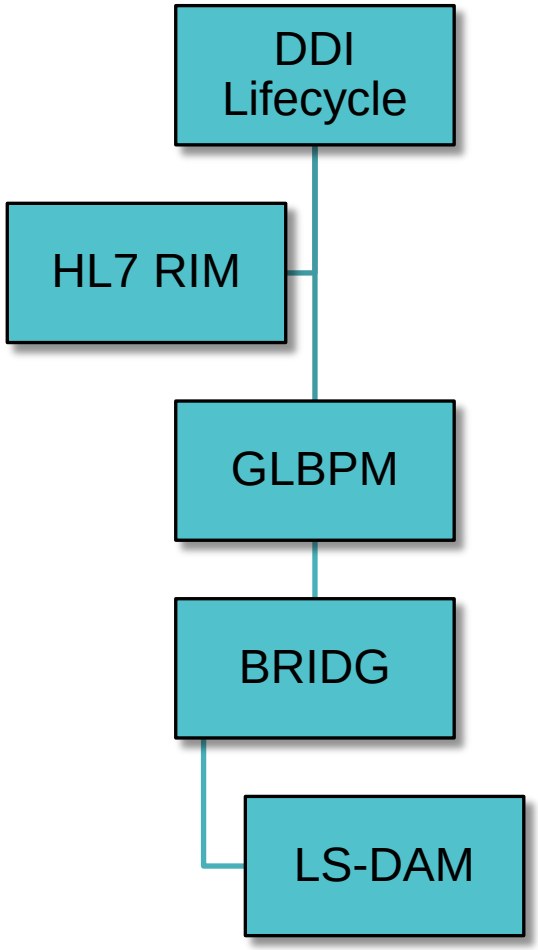
Specializing the GLBPM (1 of 3)

- PIM is a taxonomy of the activities performed in the service of research data collection, processing and understanding. These activities are described with increasing specificity as we traverse the PIM where activities go from broad to narrow:



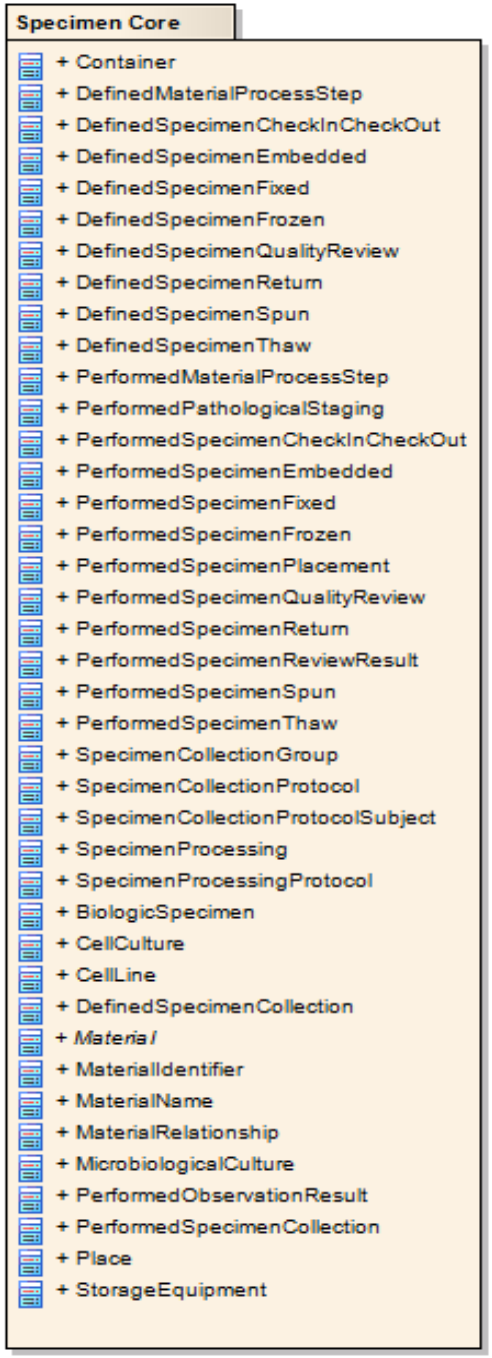
Specializing the GLBPM (2 of 3)

- We have taken the broad from first DDI Lifecycle and HL7 and then from the DDI GLBPM
- Next we specialized parts of the model further so it could describe biomedical research paradata
  - Here we have been guided by the caBIG Life Sciences Domain Analysis Model (LS-DAM) which is itself based on the CDISC Biomedical Research Integrated Domain Group (BRIDG) Domain Analysis Model
- LS-DAM provided specificity with its Specimen Core.
  - The Specimen Core identifies both the players and the acts these players engage in to handle and process biologic specimens once they are collected:



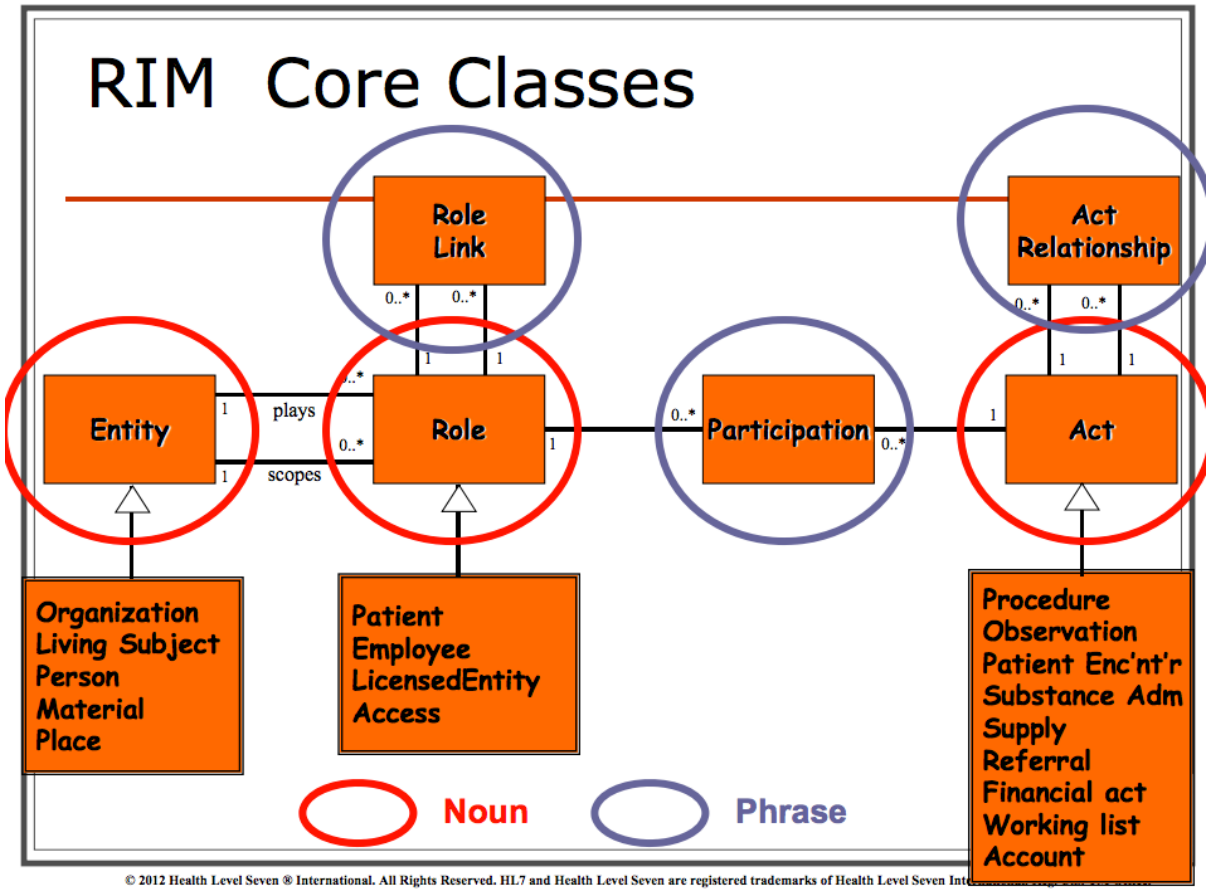
Specializing the GLBPM (3 of 3)

- We have in turn specialized the LS-DAM by observing practices in the National Children's Study (NCS)
  - By way of example LS-DAM includes a PerformedSpecimenCollectionStep which has remain largely unspecialized because the focus of the Specimen Core is more on the PerformedMaterialProcessStep.
  - However, the NCS Operational Data Elements include a number of activities that detail the PerformedSpecimenCollectionStep including:
    - Precision Thermometer Calibration and Verification
    - Preparation of refrigerators and freezers
    - Preparation of a breast milk collection kit
- The PIM, like all bricoleur, is nothing but opportunism with a dash of good judgment



PIM has a formalism (1 of 3)

- Actually it is a blend of two formalisms.
- The first formalism is HL7 RIM
- In the RIM entities play a role and participate in acts. In the predicate there can be multiple acts, and multiple actors so in the same act there can be a doctor and a patient, a researcher and a participant and so forth



PIM has a formalism (2 of 3)

- The second model of an activity is in widespread use across the semantic web.
  - In this model there is an ACTOR, a VERB and an OBJECT or, again, a SUBJECT, PREDICATE and an OBJECT.
  - In the second model an activity is a *triple*. Here is an example from JSON. JSON is JavaScript Object Notation:

There are three main parts to a basic paradata statement: ACTOR, VERB, and OBJECT. These words work pretty much like they do in regular English grammar. An "actor" *does* "verb" *to an* "object". For example, "A teacher taught the lesson located at some URL."

Actor = A teacher

Verb = Taught

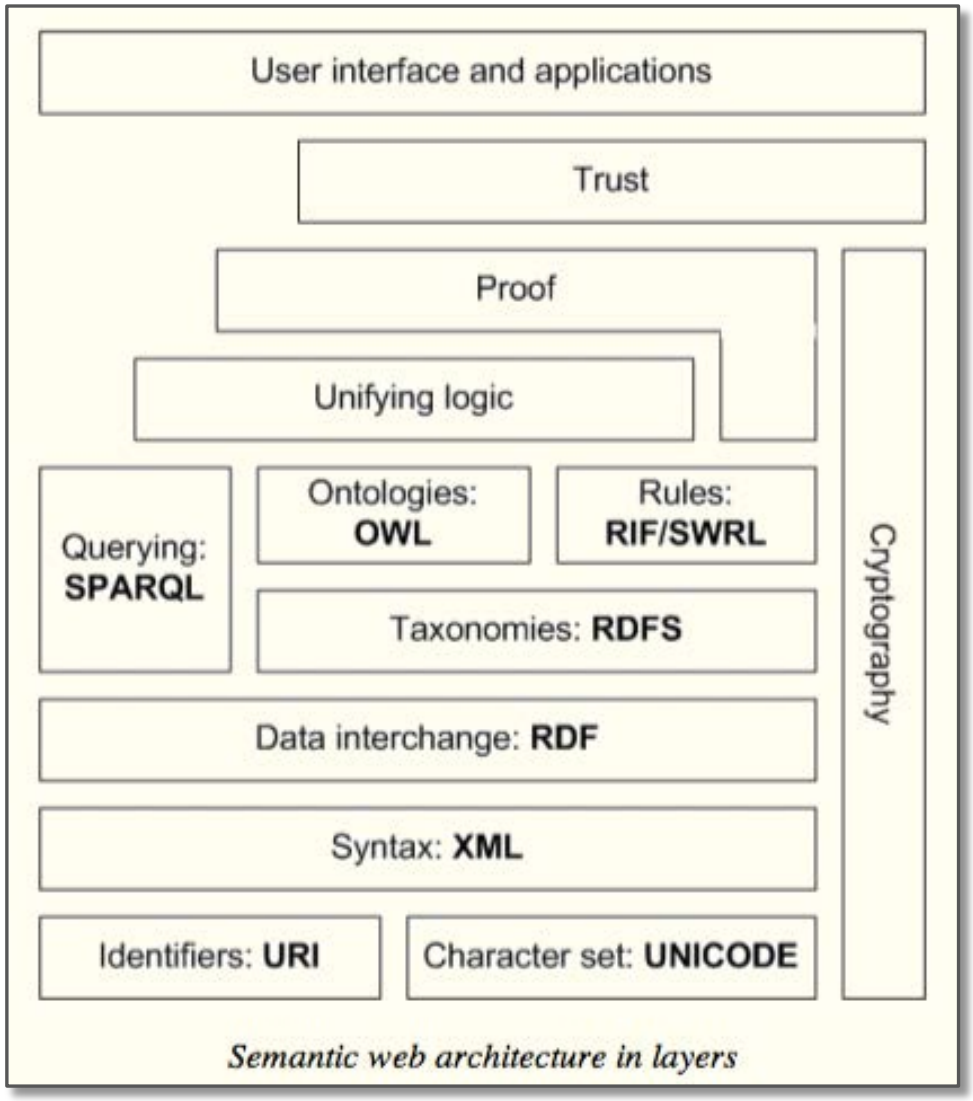
Object = The lesson located at some URL

Let's see what the JSON would look like:

```
{
    "activity": {
        "actor": "teacher",
        "verb": "taught",
        "object": "http://URL/to/lesson/"
    }
}
```

PIM has a formalism (3 of 3)

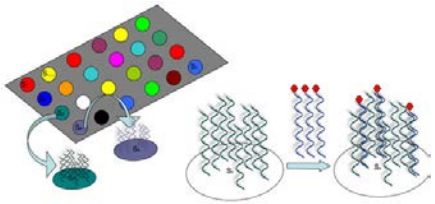
- We might represent the PIM as a UML model or as an ontology using OWL
  - In either case PIM is on a path that supports an RDF layer
- That being said, this formalism is lacking
- What it lacks is a layer that groups and sequences data collection, data processing and data understanding activities
- Our next section is a detailed use case that will provide the requirements for this process model



Semantic web architecture in layers

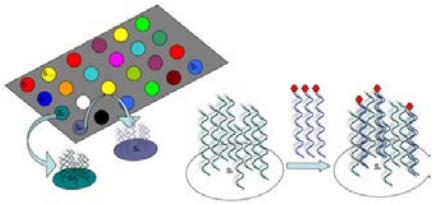
A Deep Dive into Sequencing Paradata

THE MICROARRAY EXPERIMENT USE CASE



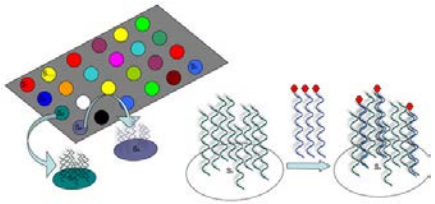
The Microarray Experiment (1 of 4)

a. *Investigator collects patient cells.*[3]
    i. *Investigator collects buccal cell sample.*[4]
        1. *Investigator thaws tubes containing DNA extraction solution.*
        2. *Subject rinses out mouth (twice, with water).*
        3. *Investigator collects tissue (with buccal brush, from cheek).*
        4. *Investigator stores brush (RT, up to one month or -20 for 6 months).*
b. Investigator isolates DNA
    i. Isolation method: proteinase K high-salt ethanol precipitation
    ii. *Investigator retrieves brush.*
    iii. *Investigator isolates DNA (from tissue, using high-salt ethanol precipitation*[5]*)*
    iv. *Investigator freezes DNA.*



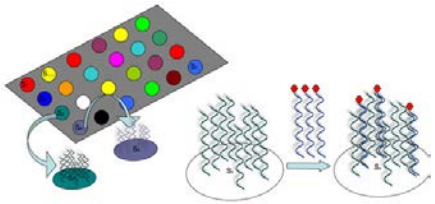
The Microarray Experiment (2 of 4)

c. Investigator purifies DNA
   i. Method: Montage PCR plates on PerkinElmer MultiPROBE II HT EX instrument
   ii. *Investigator thaws DNA.*
   iii. *Investigator purifies DNA (method, Montage PCR filter plate[6]).*
   iv. *Investigator measures DNA concentration (method, Nanodrop[7]).*
   v. *Investigator measures DNA quality (method, Nanodrop[8]).*
   vi. *Investigator dilutes DNA.*
   vii. *Investigator aliquots DNA.*
   viii. *Investigator freezes DNA aliquots.*



The Microarray Experiment (3 of 4)

d. Investigator does genotyping array.
   i. Array type: Affymetrix 100K Human Mapping array
   ii. *Investigator thaws DNA aliquot.*
   iii. *Investigator fluorescently labels DNA aliquot (protocol, reagents).*
   iv. *Investigator freezes DNA aliquot.*
   v. *Investigator freezes fluorescently labeled DNA.*
   vi. *Investigator thaws fluorescently labeled DNA. Investigator hybridizes fluorescently labeled DNA (to microarray, array number, protocol).*
   vii. *Investigator scans microarray (File name, scanner, settings).*
   viii. *Investigator extracts features (From array, software, version, settings, probe set).*



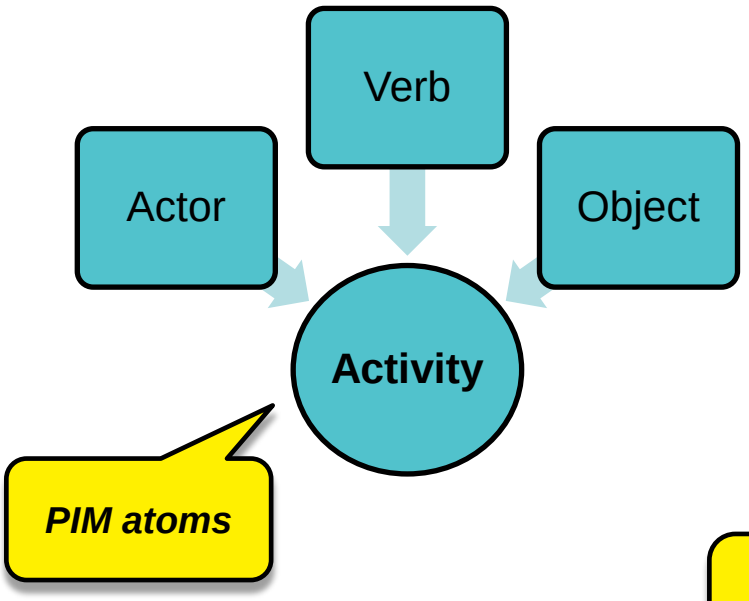
The Microarray Experiment (4 of 4)

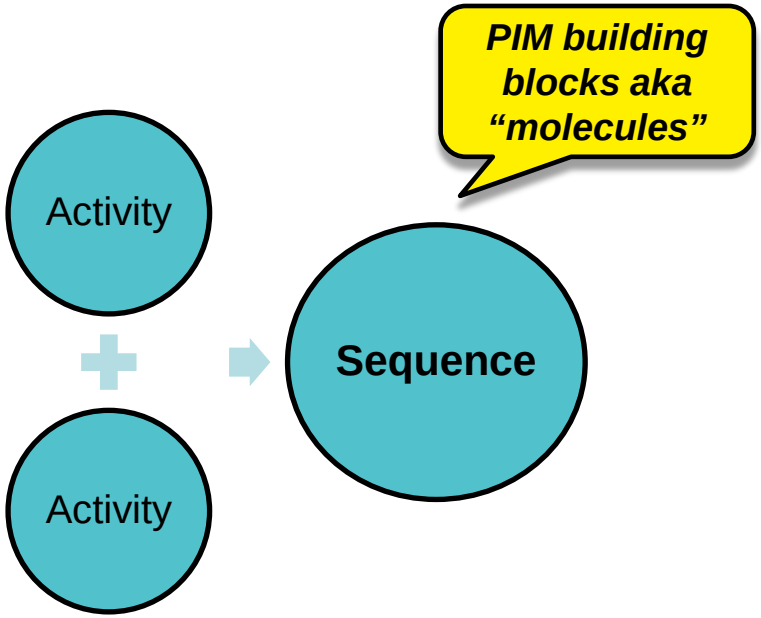
- At each step paradata is collected to document and assure data quality
  - In general, a researcher should record the date on which each step of an experiment occurred and any information necessary to allow another scientist to reproduce the experiment. This is the content that is typically recorded into a *lab notebook*.
  - This information is also important when checking the quality of data produced by the experiment.
    - For example, if a researcher has too many samples in a microarray experiment to complete all at the same time, he or she may break the set of samples up into smaller batches.
    - If the researcher then notices while doing later analysis that all of the microarrays with the highest fluorescent signal came from one batch, he or she would be concerned that the signal he or she observes is due to technical error, a "batch effect," rather than true biological signal.
- In the genotyping experiment above, the majority of steps shown in parts A, B, and C are shared with many other types of biological samples, and many of these actions are already encompassed by LS-DAM.
  - This suggests that much of the paradata of interest in a genomic workflow can be adopted by modifying an existing scientific domain model.
  - Only Part D contains steps that specifically relate to genomic analyses.

Lessons Learned from the Microarray Experiment Use Case

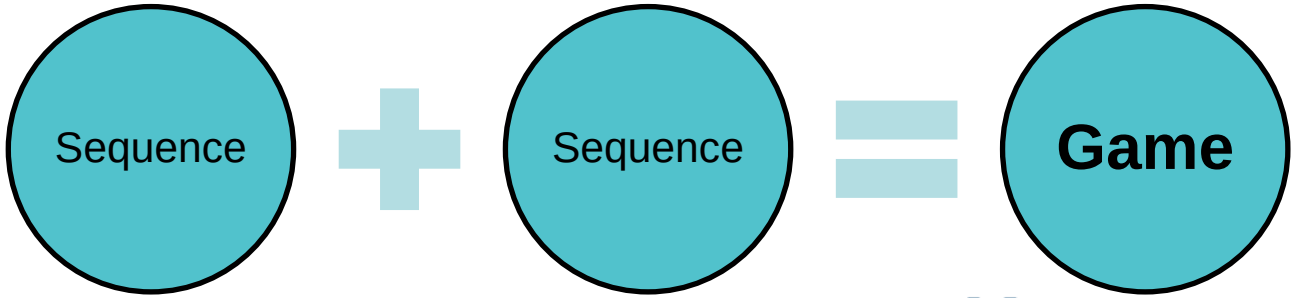
PIM REDUX

PIM Pong (1 of 2)





PIM is a composer who plugs-'n-plays sequences. PIM compositions include "the microarray experiment game", "the whole genome sequencing game", etc.



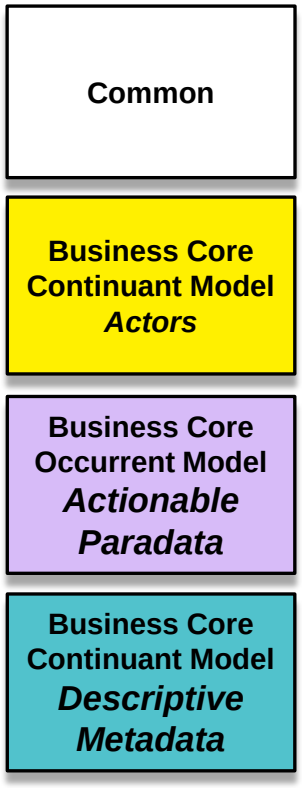
PIM Pong (2 of 2)

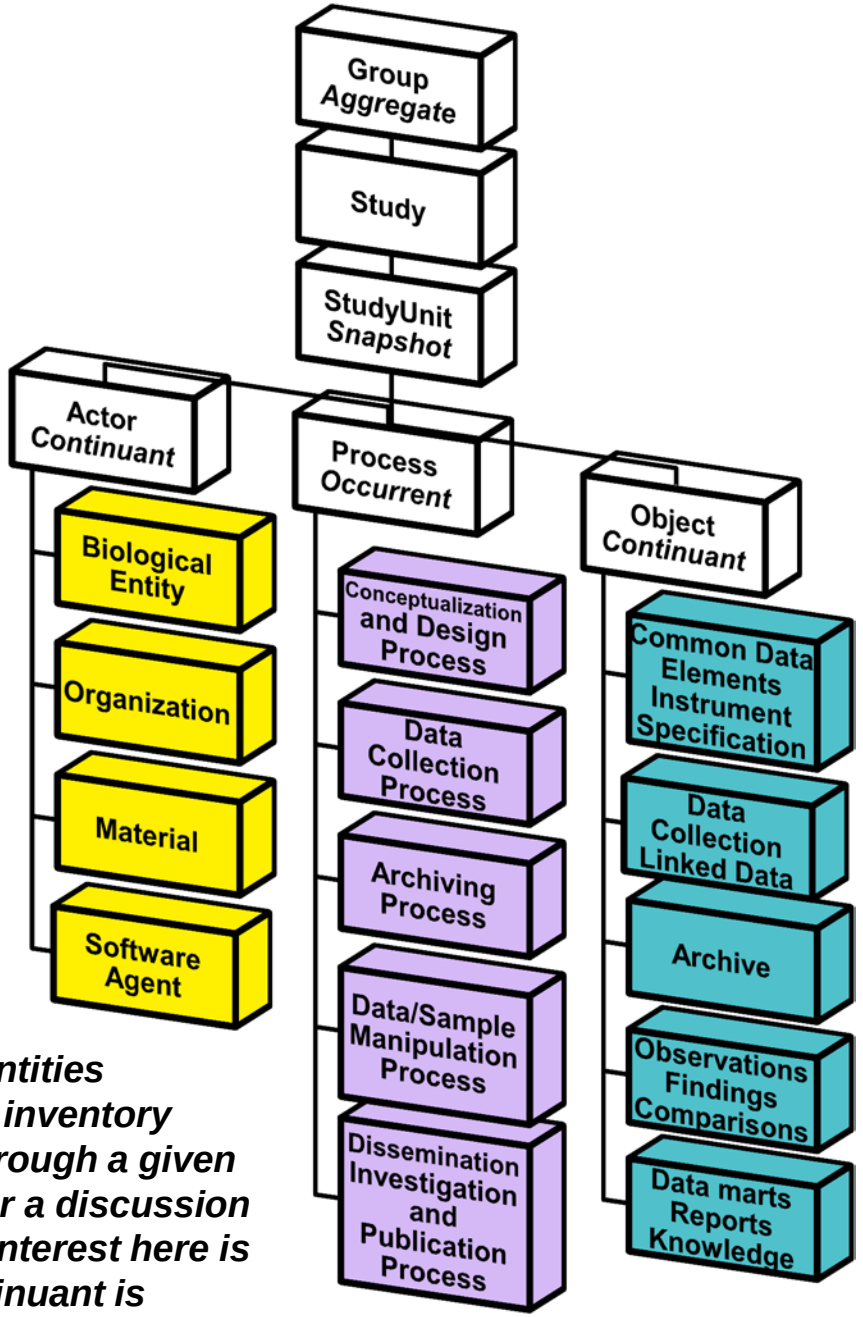
- To put this in words…
  - Imagine a log
    - It might be the log a case manager keeps by participant and visit
    - It might be the lab log a research assistant keeps who is conducting a genomic experiment with a collected biospecimen
  - The log is smart: it knows *sequences*
    - With case management it know the sequence of activities for scheduling a visit
      - Some of the activities may be conditional on the outcome(s) of other activities
    - In the genotyping experiment there are sequences like “investigator isolates DNA” and “investigator purifies DNA”
  - In order to further the conduct of research science workers assemble sequences into compositions they perform in real time
    - The log is a record of what happens in the course of these compositions

PUTTING IT ALL TOGETHER

DDI Study Model

From Wendy's Core





Each continuant model is an inventory of all entities existing at a time. Each occurrent model is an inventory (processory) of all the processes unfolding through a given interval of time. See Basic Formal Ontology for a discussion of dependencies. One type of dependency of interest here is realizable/realized by where a dependent continuant is realized by an occurrent

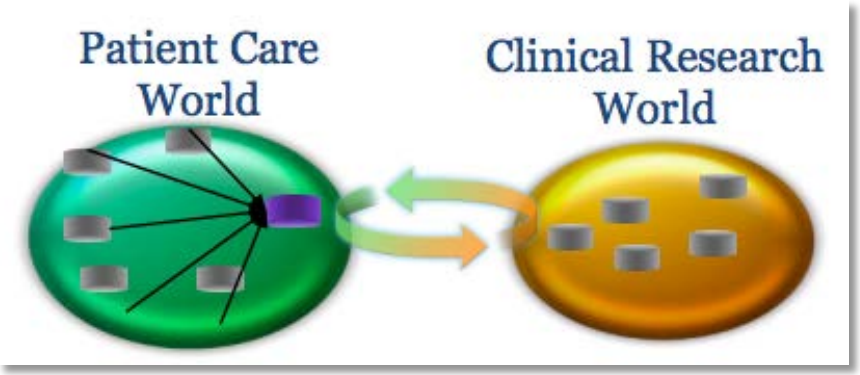
Continuants, Occurrents and their Dependencies

- Actor(s)
- DDI Study Object(s)

DDI Study Object(s)

Continuants, Occurrents and their Dependencies: Automatic Data Linking

- Today CDISC has launched its Healthcare Link that aims to leverage EHR information programmatically in clinical trial systems.
  - In the Healthcare Link there is a Retrieve Form for Data Capture (RFD).
  - An RFD Profile that negotiates with a host EHR system fuels the RFD.
  - The RFD works like a questionnaire only the EHR is the respondent and the interviewer is the RFD Profile. RFD is enabled to mark partial completes, breakoffs and comment on data quality in a way that conforms to the 21CFR11 quality tracking standard.



Data Linking Instance

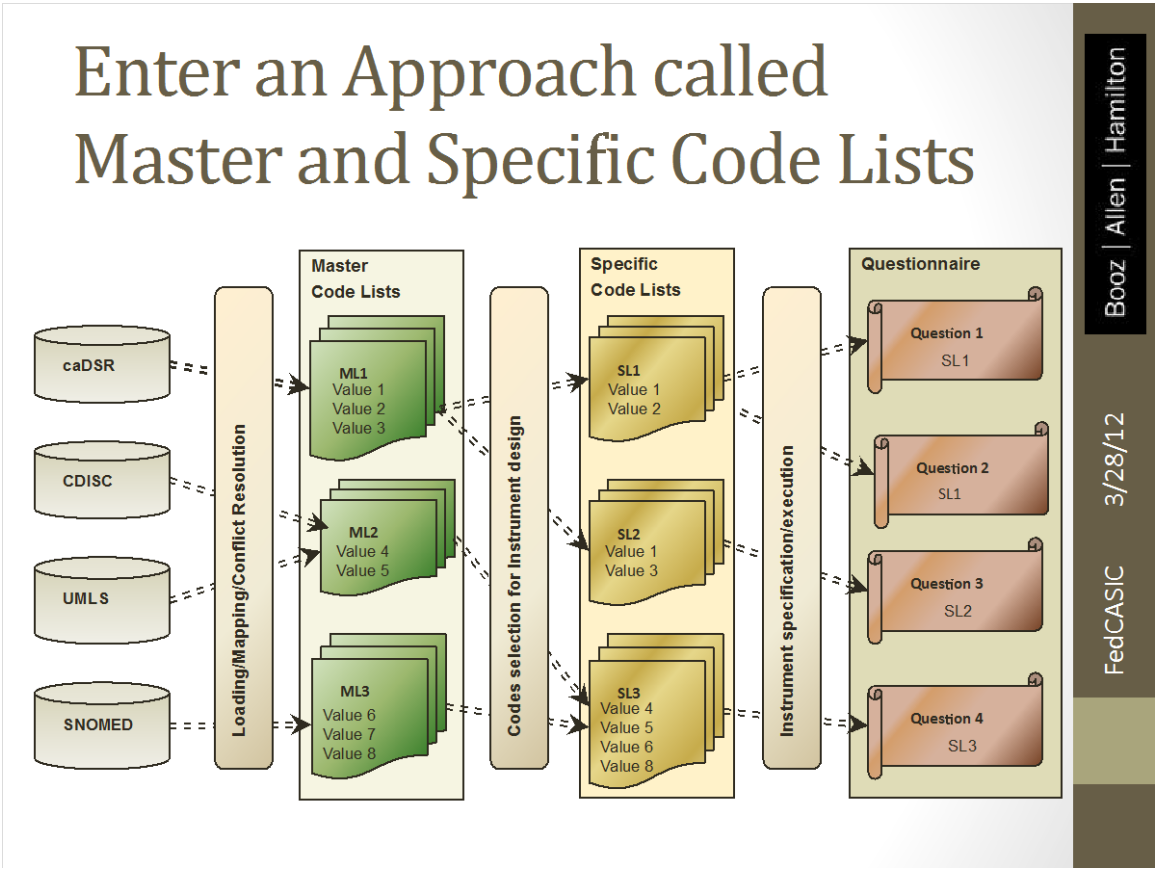
- EHR
- RFD Profile

Linked Data

Audit Trail

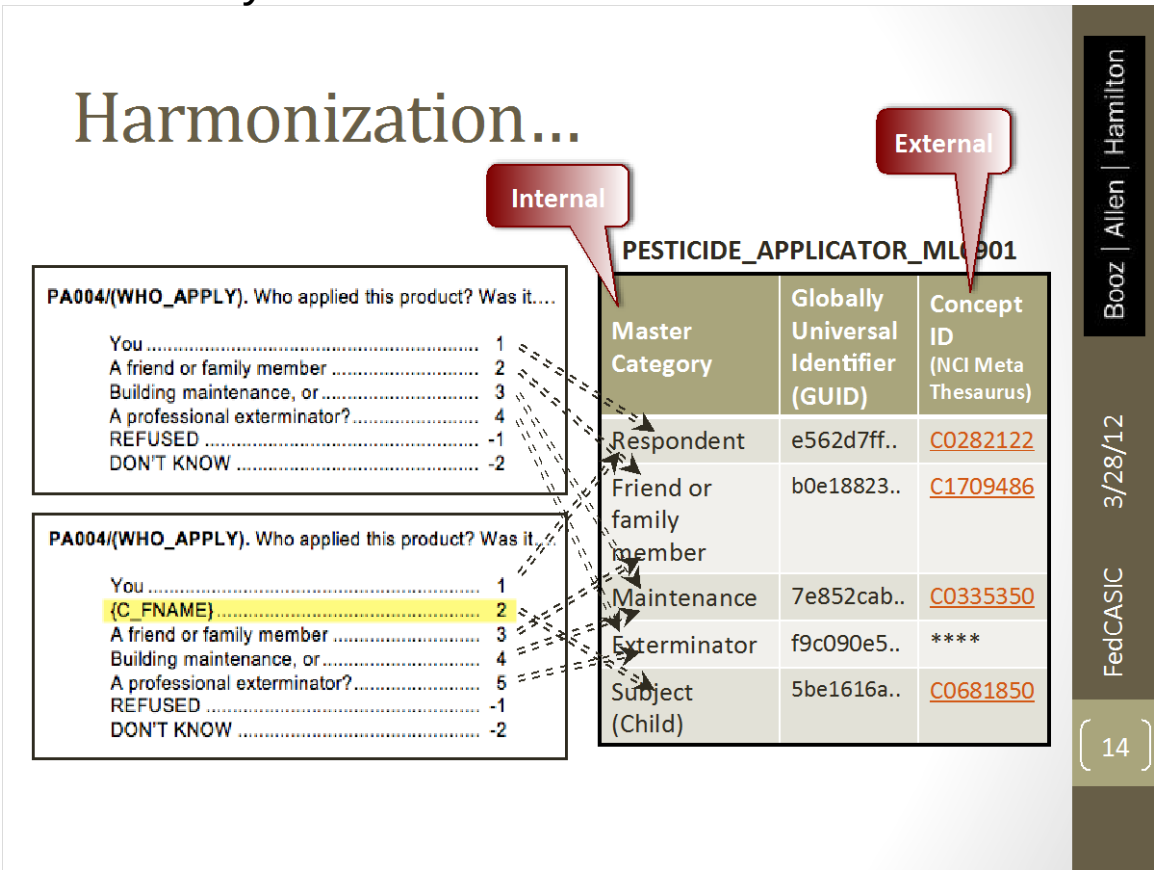
Continuants, Occurrents and their Dependencies: Automatic Comparisons (1 of 3)

- DDI provides a metadata framework for represents the code lists that are used in connection with single select and multi-select questions:



Continuants, Occurrents and their Dependencies: Automatic Comparisons (2 of 3)

- Using this framework it is possible to perform comparisons between questions with code lists. In this example, a code list has changed over time in a longitudinal study:



Continuants, Occurrents and their Dependencies: Automatic Comparisons (3 of 3)

- Using processing steps with preconditions and postconditions, comparisons like this may be performed by software agents informed by business rules and a decision tree that guides the workflow.

Comparison Instance

- Software Agent
- Code List #1
- Code List #2
- Mapping Algorithm

Map

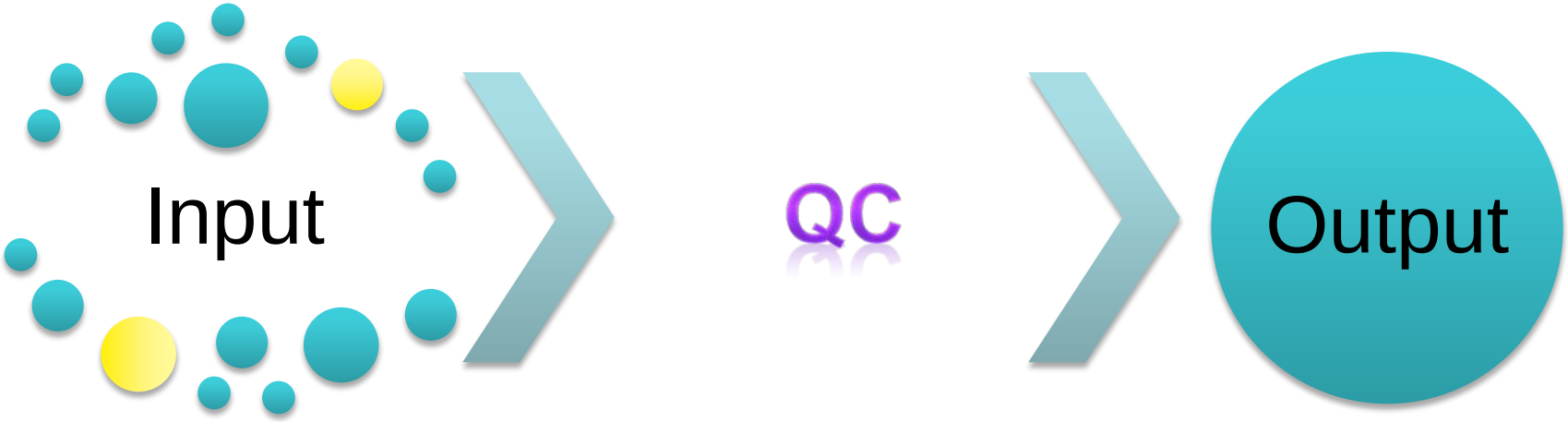
Harmonized Code List

Continuants, Occurrents and their Dependencies: Automatic Biospecimen Management

- Today in the area of PerformMaterialProcessing much biospecimen handling is performed by machines with complex control systems that humans can configure.

Automated Storage and Retrieval

Biospecimen QC Instance



- Brooks BioStore™
- Biospecimens
- BioStore™ Program Instructions

Quality Reports

Audit Trails

Continuants, Occurrents and their Dependencies: Automatic Experiments (1 of 2)

- Today high-throughput technologies are used to gather large quantities of biological data.
- Microarrays and next-generation sequencing allow investigators to observe an individual's genome and epigenome
- And mass spectrometry provides access to the proteome (the complete set of proteins of an organism) and metabolome (the complete set of metabolites of an organism).
- Robotic equipment is available to automate sample processing for these technologies, and relies on vendor- and user-provided workflows of steps having preconditions and postconditions.

Continuants, Occurrents and their Dependencies: Automatic Experiments (2 of 2)

- Users frequently turn to software which allows custom combination of building blocks for data analysis and visualization.
  - These pipelines allow scientific investigators with biological questions but minimal computational expertise to extract meaningful information from these large data sets.
  - Additionally, output information from these experiments can be electronically annotated with links to ontologies containing biological processes, cellular components and molecular functions to provide further biological insight.
  - Linking to standardized ontologies allows investigators to summarize large, complex quantities of information with simpler, well-documented terminology.

Automatic Genomic Experiment Instance

- Robot
- Annotator
- Biospecimens
- Standardized Ontologies
- Experiment Instructions

Annotated Findings

Audit Trails

Questions?