National Cancer Institute

Interoperability and Semantics in Use-Application of UML , XMI and MDA to Precision Medicine and Cancer Research

Ian Fore, D.Phil.
Associate Director, Biorepository and Pathology Informatics
Senior Program Manager, Informatics Infrastructure for Biomedical Research

NCI Center for Biomedical Informatics and Information Technology

OMG TECHNICAL MEETING RESTON, VA
20 March 2013

Workshop & Information Day On Semantics
From Research To Reality: Implementing The Semantic Web

Two domain issues

- Precision medicine
- Reliability of molecular characterization of specimens

NATIONAL ACADEMY OF SCIENCES
NATIONAL ACADEMY OF ENGINEERING
INSTITUTE OF MEDICINE
NATIONAL RESEARCH COUNCIL

REPORT IN BRIEF

Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease

A new data network that integrates emerging research on the molecular makeup of diseases with clinical data on individual patients could drive the development of a more accurate classification of disease and ultimately enhance diagnosis and treatment. Recent advances in biomedical research have caused an explosion of data, offering the potential to develop a "New Taxonomy" that defines disease based on underlying molecular and environmental causes, rather than on physical signs and symptoms. This report outlines how research and clinical data can be captured in a "Knowledge Network" that will be broadly accessible to researchers and clinicians. As well as improving health care, the new data network could also improve biomedical research by enabling scientists to access patient information through electronic health records, while still protecting patient rights.

Today, in a clinic somewhere in America, a patient is learning that he has diabetes. Based on the patient's symptoms and lab tests that show high levels of insulin, his doctor diagnoses Type II diabetes—but this imprecise category serves only to distinguish the disease from diabetes that typically

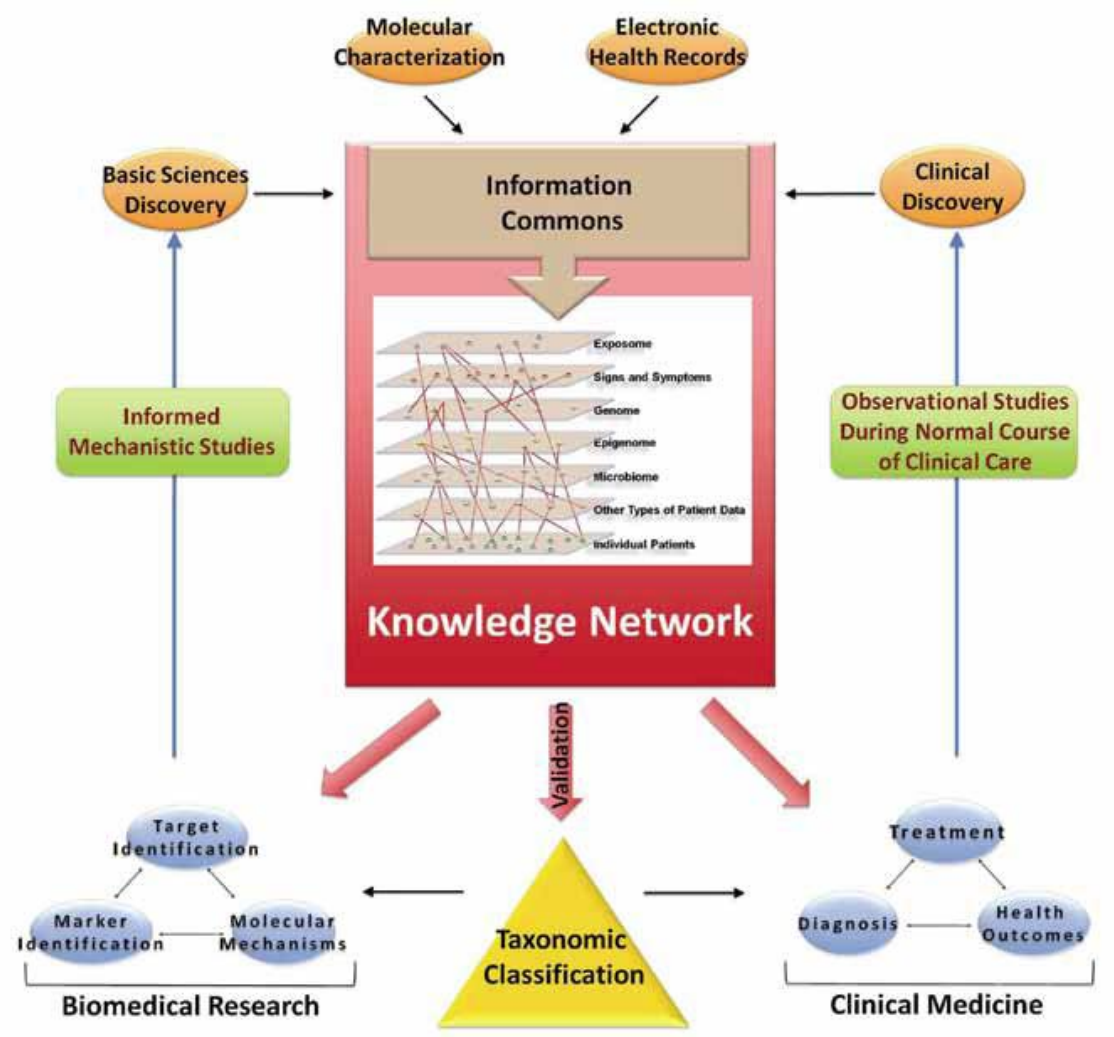
members will likely develop diabetes in the future, and if necessary could implement preventative care.

In recent years, dramatic advances in biomedical research have created an explosion of data that could be used to move toward such a world of improved health outcomes. However, currently there is a discon-

The before…

- Health care has terminologies
  - It has many - ~70 (Ken Rubin)
- A particularly significant one
  - International Classification of Disease
    - Origin
      - 1893 Jacques Bertillon- *Bertillon Classification of Causes of Death*
    - Current
      - ICD-10
        - World Health Organization
- Original brief to the IOM report team
  - “A New Taxonomy of Disease”

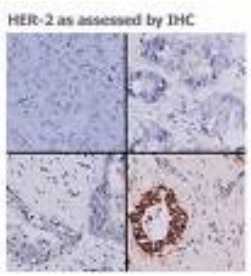
From “Toward precision medicine”

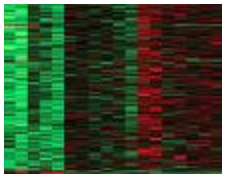


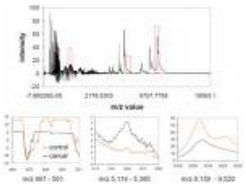
Collection and Processing of Biospecimens Impact Clinical and Research Outcomes

OBBR Office of Biorepositories and Biospecimen Research

- **Effects on Clinical Outcomes**
  - **Potential for incorrect diagnosis**
    - Morphological/immunostaining artifact
    - Skewed clinical chemistry results
  - **Potential for incorrect treatment**
    - Therapy linked to a diagnostic test on a biospecimen (e.g., HER2 in breast cancer)
- **Effects on Research Outcomes**
  - **Irreproducible results**
    - Variations in gene expression data
    - Variations in post-translational modification data
  - **Misinterpretation of artifacts as biomarkers**



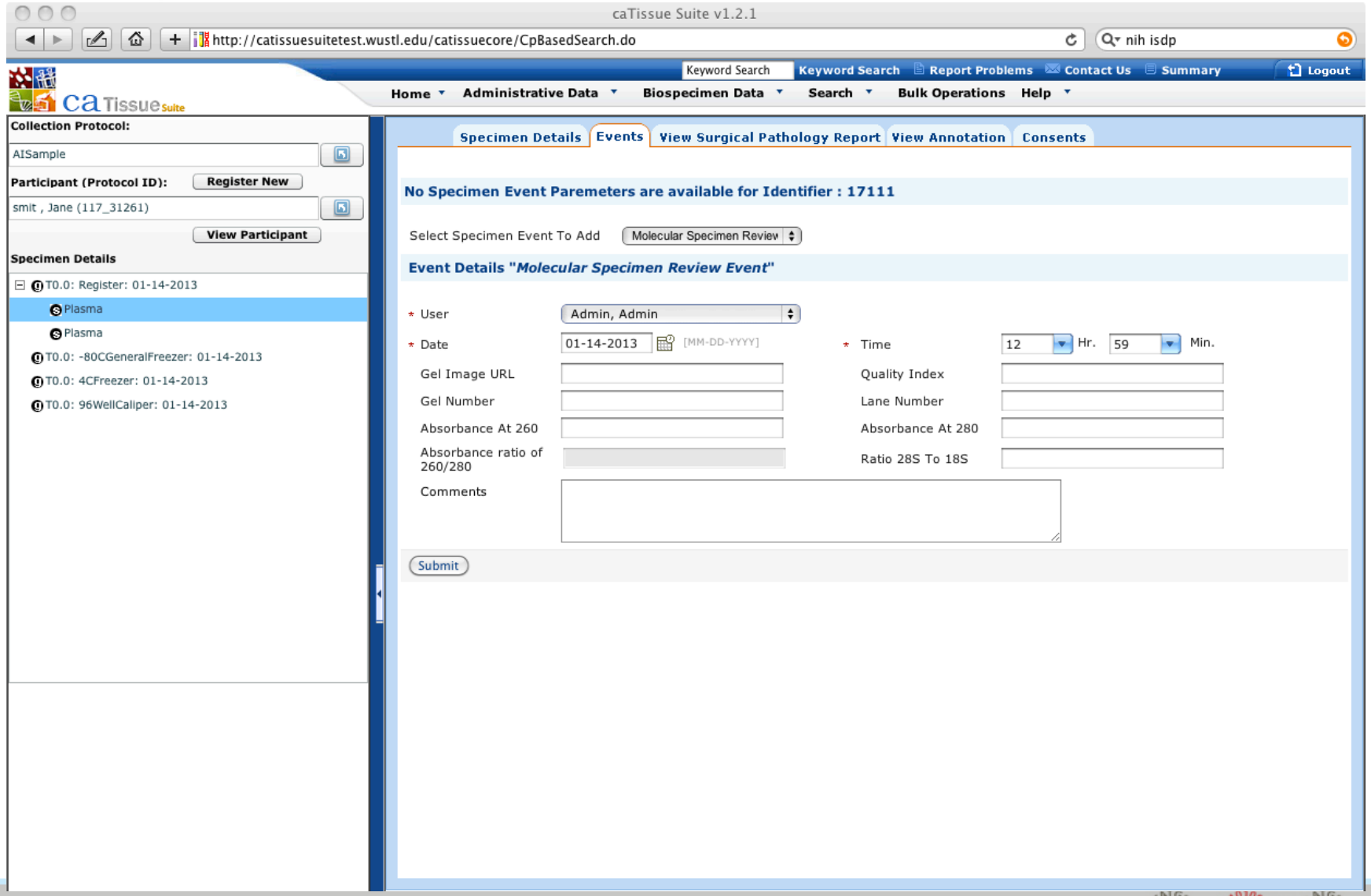




What we have used to address the issue

- Applications
  - caTissue – end user biobanking application
- NCI Semantic Infrastructure
  - EVS – Enterprise Vocabulary Services
  - caDSR – cancer Data Standards Repository
    - ISO11179 based common data elements
  - UML Domain models
- Interoperability Framework
  - caCORE SDK
    - Data service generation from XMI
  - caGrid
    - WSRF based grid
    - Data services
    - Analytical services

caTissue – Specimen Processing Parameters



Data Element Details

Public ID:	2513757
Version:	1.0
Long Name:	Molecular Specimen Review Parameter Object Gel Image Uniform Resource Locator java.lang.String
Short Name:	2513491v1.0:2178533v1.0
Preferred Question Text:	
Definition:	Relating to or produced by or consisting of molecules.:A sample or part of a thing, or of several things, taken to demonstrate or to determine the character of the whole, e.g. a substance, or portion of material obtained for use in testing, examination, or study; particularly, a preparation of tissue or bodily fluid taken for examination or diagnosis.:The act of appraisal, evaluation, or analysis.:Any factor that defines a system and determines (or limits) its performance._Jelly like material formed by the coagulation of a colloidal liquid. (from On-line Medical Dictionary):Any record of an imaging event whether physical or electronic.:An Internet address which tells a browser where to find an Internet resource. (from http://www.computeruser.com)_Generic value domain for a java datatype that is a class that represents character strings.
Value Domain:	java.lang.String
Data Element Concept:	Molecular Specimen Review Parameter Object Gel Image Uniform Resource Locator
Context:	caBIG
Workflow Status:	RELEASED
Origin:	
Registration Status:	Qualified
Direct Link:	https://cdebrowser.nci.nih.gov/CDEBrowser/search?elementDetails=9&FirstTimer=0&PageId=ElementDetailsGroup&publicId=2513757&version=1.0

Reference Documents

Document Name	Document Type	Document Text	Context	URL
There are no reference documents for the selected CDE.				

Alternate Names and Definitions

CS* Long Name	CS* Definition	CSI* Name	CSI* Type
caTissue_Suite	caTissue Suite version 1.0	edu.wustl.catissuecore.domain	UML_PACKAGE_NAME

Alternate Names

Name	Type	Context	Language
edu.wustl.catissuecore.domain.MolecularSpecimenReviewParameters.gelImageURL	UML Qualified Attr	caBIG	ENGLISH
MolecularSpecimenReviewParameters:gelImageURL	UML Class:UML Attr	caBIG	ENGLISH

Alternate Definitions

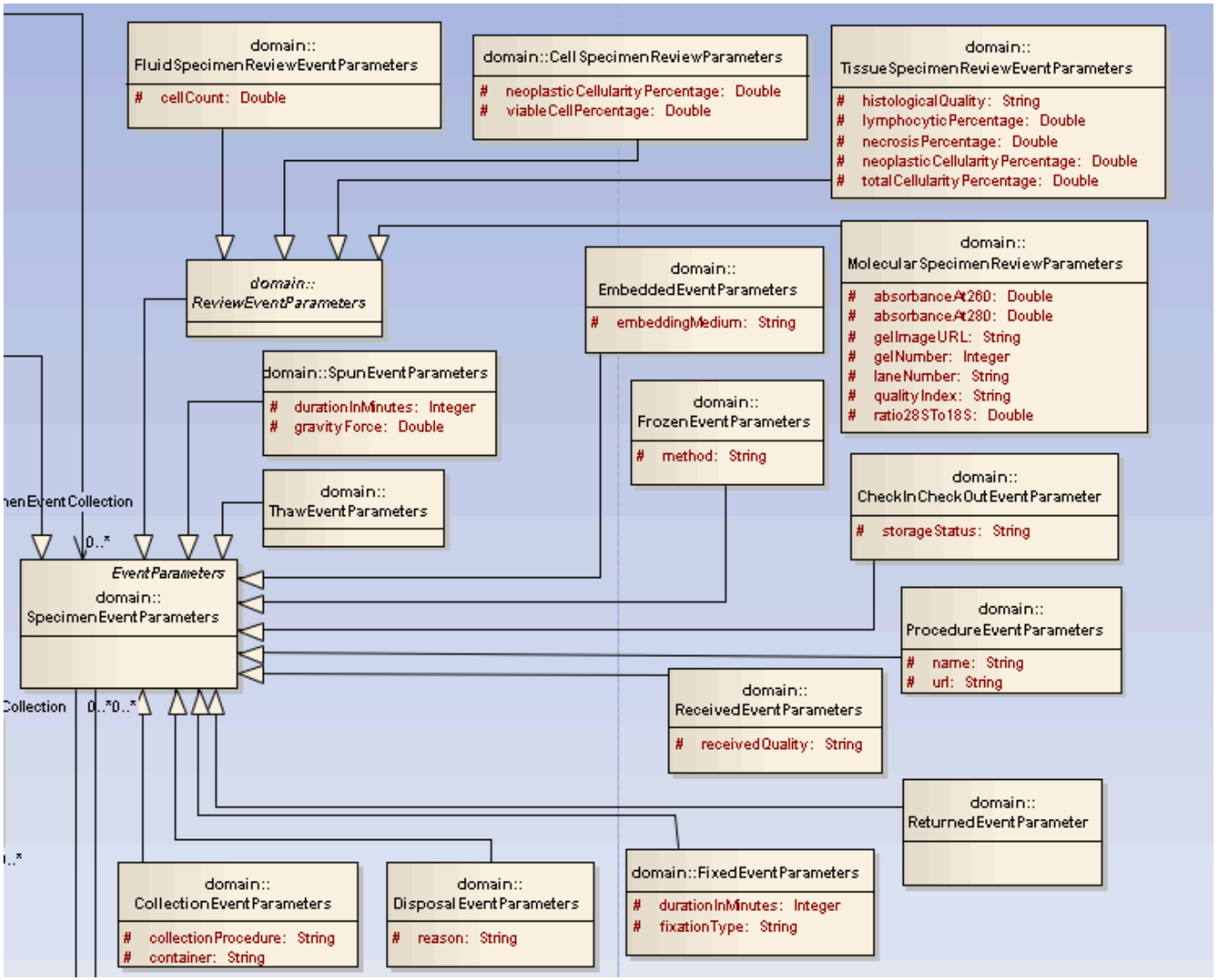
Name	Type	Context
A reference to the location of an electrophoretic gel image of the specimen.	UML Class:UML Attr	caBIG

CS* Long Name	CS* Definition	CSI* Name	CSI* Type
caTissue_Suite1_1	caTissue Suite 1.1 is a biospecimen banking and inventory tracking application based on the caTissue Suite 1.1 model.	edu.wustl.catissuecore.domain	UML_PACKAGE_NAME

Alternate Names

Name	Type	Context	Language

caTissue – Specimen event model



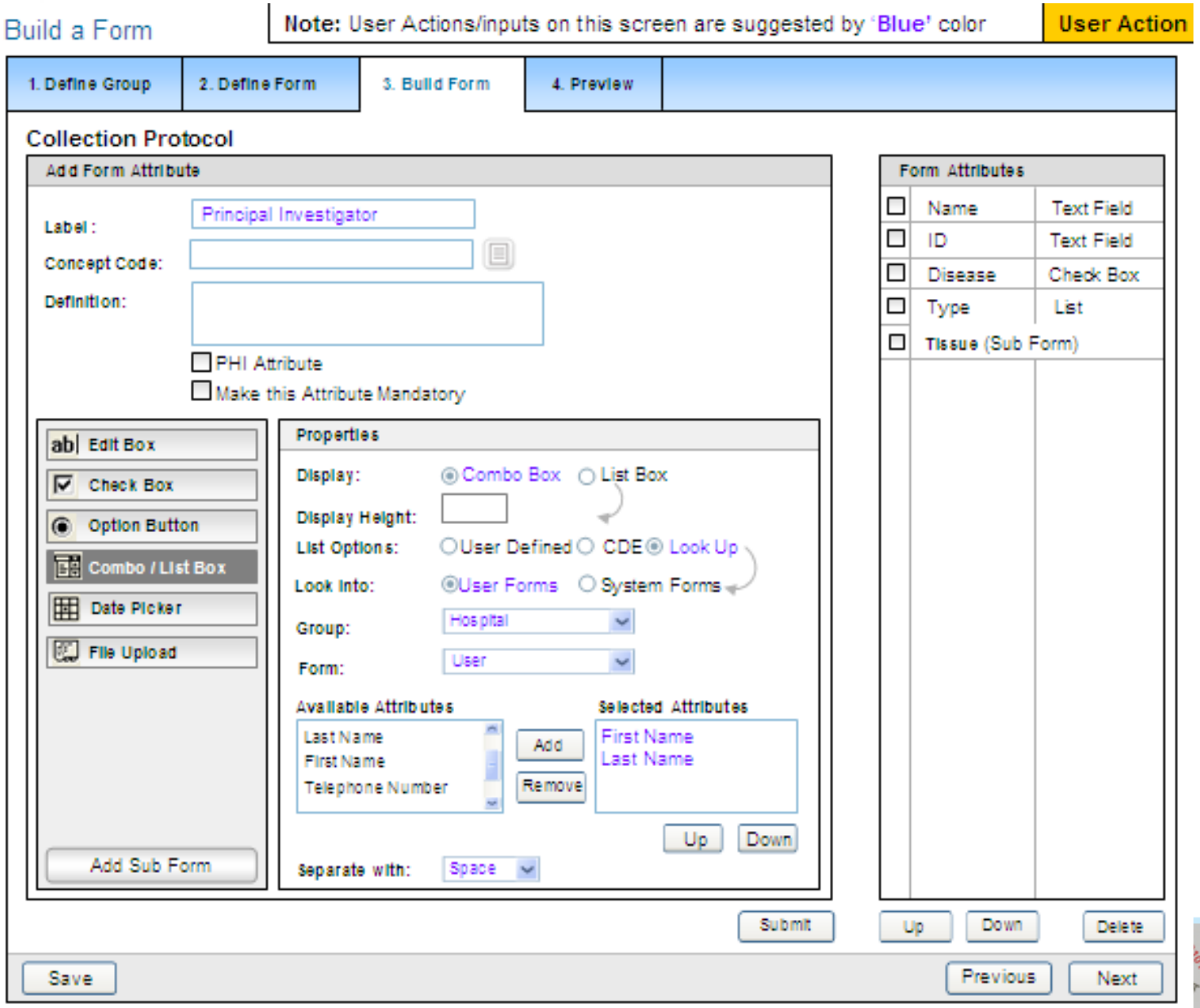
caTissue Dynamic Extension Principles

- Extend model in particular areas
  - E.g. specimen events
  - Hook entities
    - Participant, Specimen Collection Group, Specimen
- Model Driven Extension
- Save the extended model standard format
  - UML – XMI
- Extend within the application
  - Nice UI
  - Users aren't conscious of
    - Metadata
    - The model
    - Generating services

caTissue extensions: Combo/Listbox

caTissue extensions: Combo/Listbox

caTissue extensions: Edit Box (Association)



Application

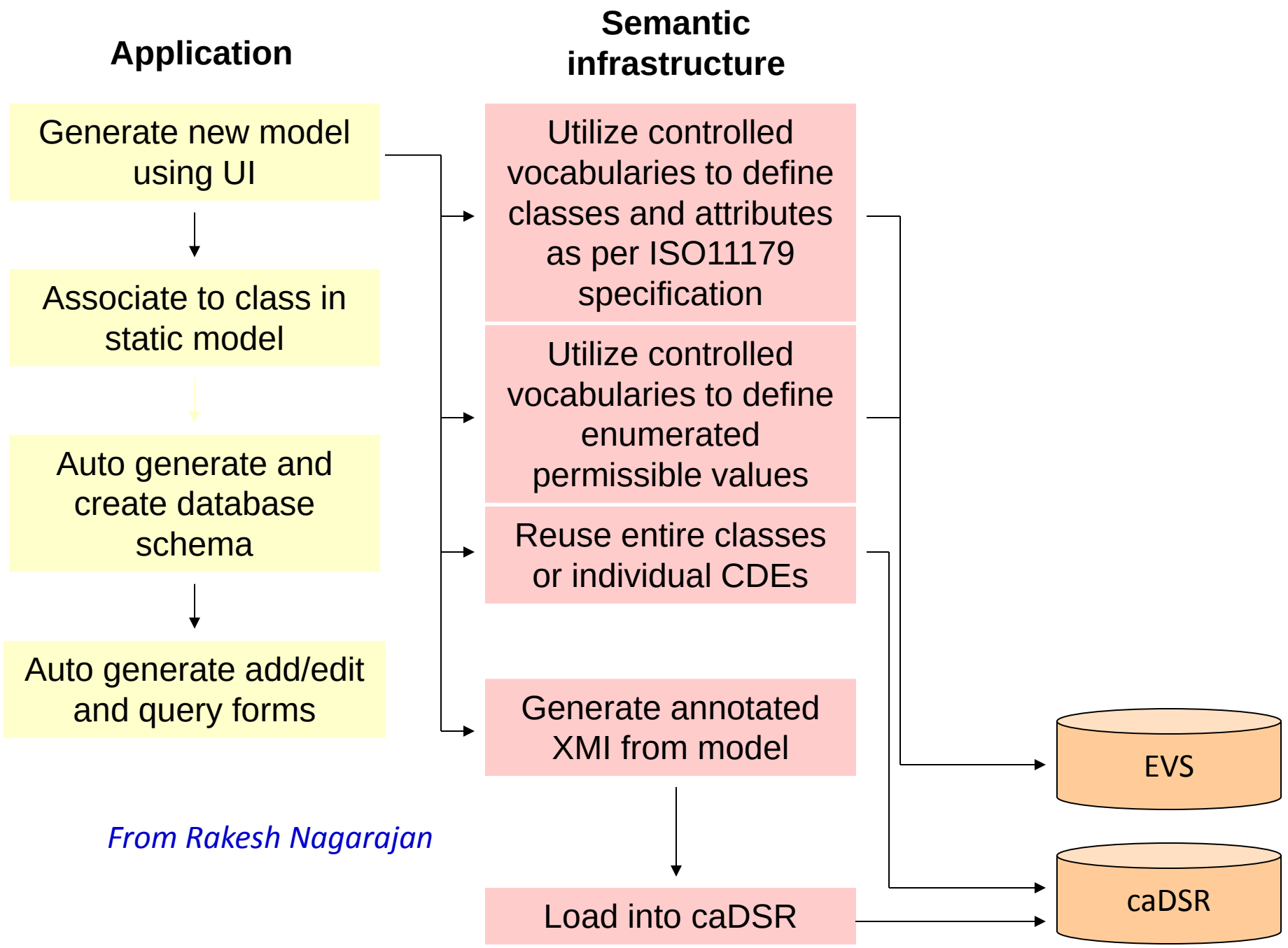
Generate new model using UI

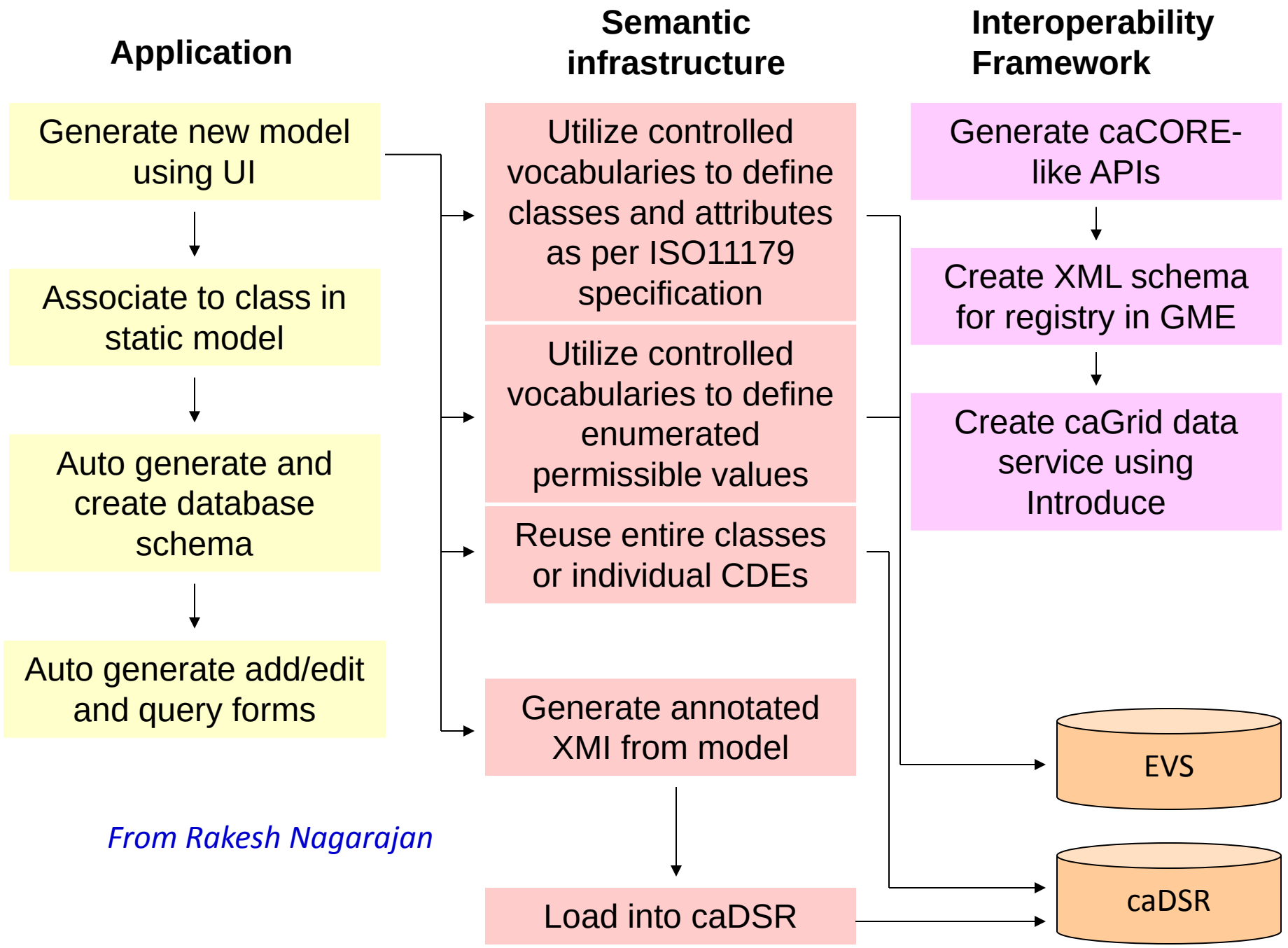
Associate to class in static model

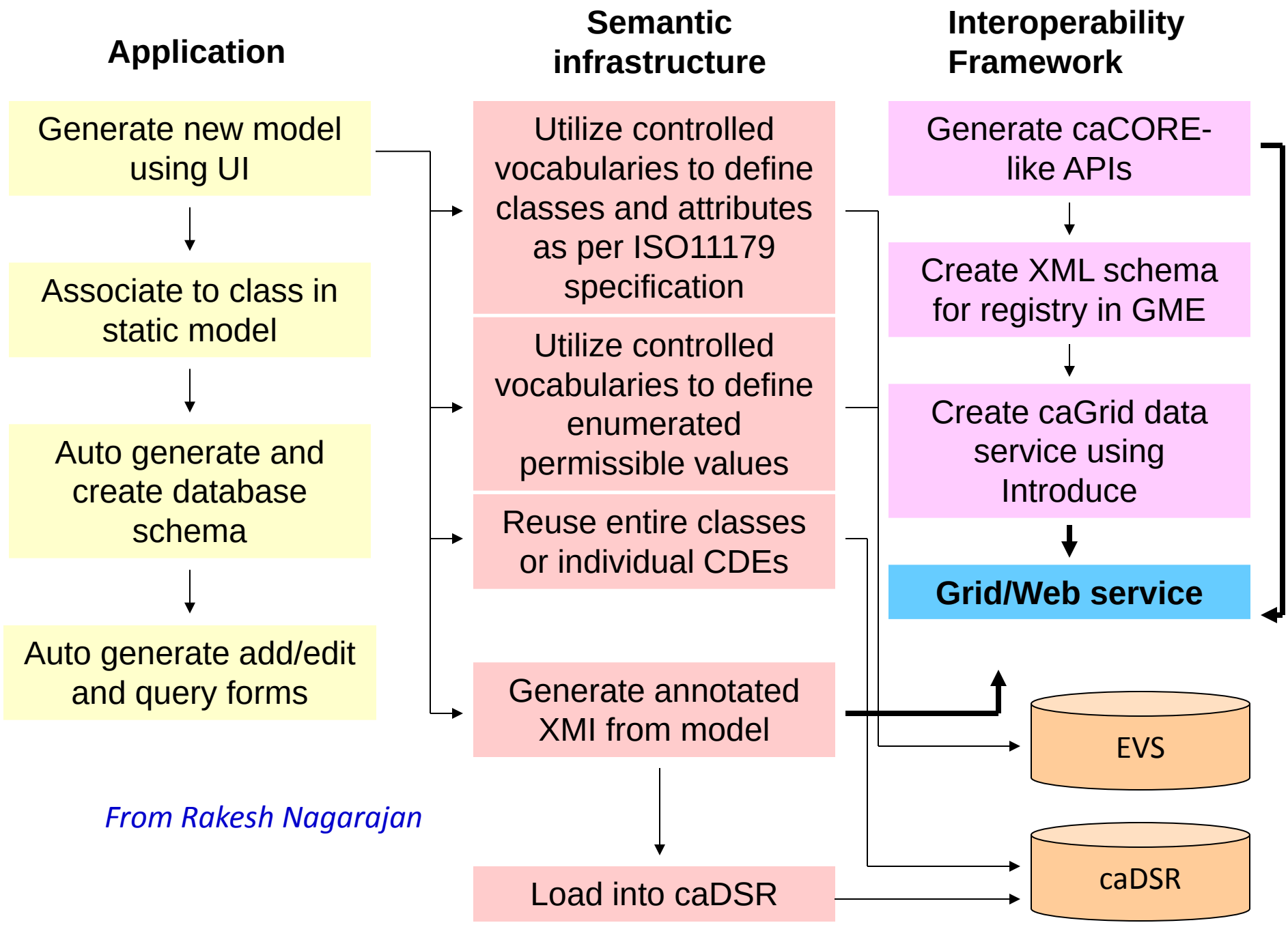
Auto generate and create database schema

Auto generate add/edit and query forms

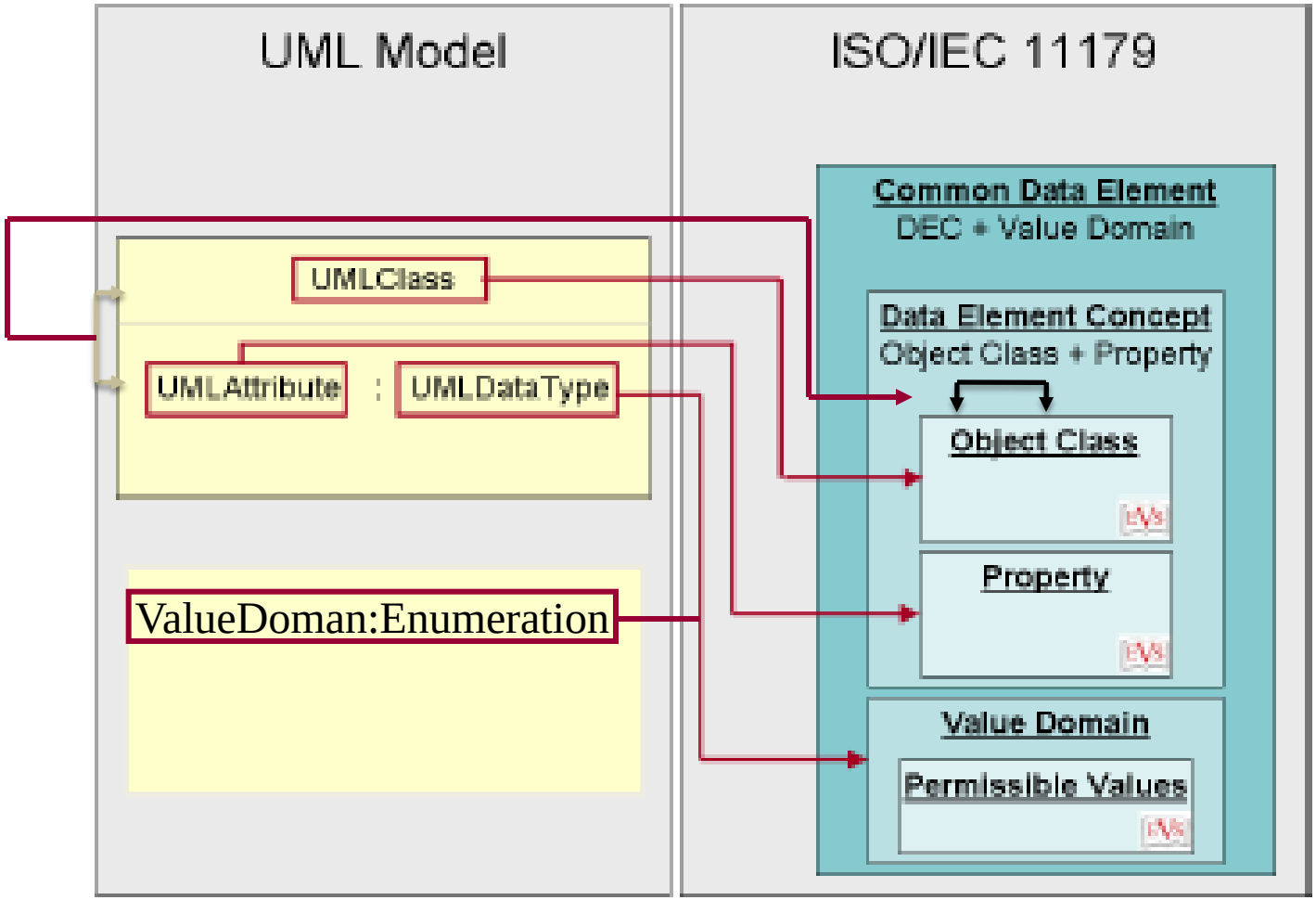
From Rakesh Nagarajan


Application
Semantic infrastructure
Generate new model using UI
Associate to class in static model
Auto generate and create database schema
Auto generate add/edit and query forms
Utilize controlled vocabularies to define classes and attributes as per ISO11179 specification
Utilize controlled vocabularies to define enumerated permissible values
Reuse entire classes or individual CDEs
Generate annotated XMI from model
Load into caDSR
EVS
caDSR
From Rakesh Nagarajan


Application
Semantic infrastructure
Interoperability Framework
Generate new model using UI
Associate to class in static model
Auto generate and create database schema
Auto generate add/edit and query forms
Utilize controlled vocabularies to define classes and attributes as per ISO11179 specification
Utilize controlled vocabularies to define enumerated permissible values
Reuse entire classes or individual CDEs
Generate annotated XMI from model
Load into caDSR
Generate caCORE-like APIs
Create XML schema for registry in GME
Create caGrid data service using Introduce
EVS
caDSR
From Rakesh Nagarajan


Application
Semantic infrastructure
Interoperability Framework
Generate new model using UI
Associate to class in static model
Auto generate and create database schema
Auto generate add/edit and query forms
Utilize controlled vocabularies to define classes and attributes as per ISO11179 specification
Utilize controlled vocabularies to define enumerated permissible values
Reuse entire classes or individual CDEs
Generate annotated XMI from model
Load into caDSR
Generate caCORE-like APIs
Create XML schema for registry in GME
Create caGrid data service using Introduce
Grid/Web service
EVS
caDSR
From Rakesh Nagarajan

UML → caDSR Mapping



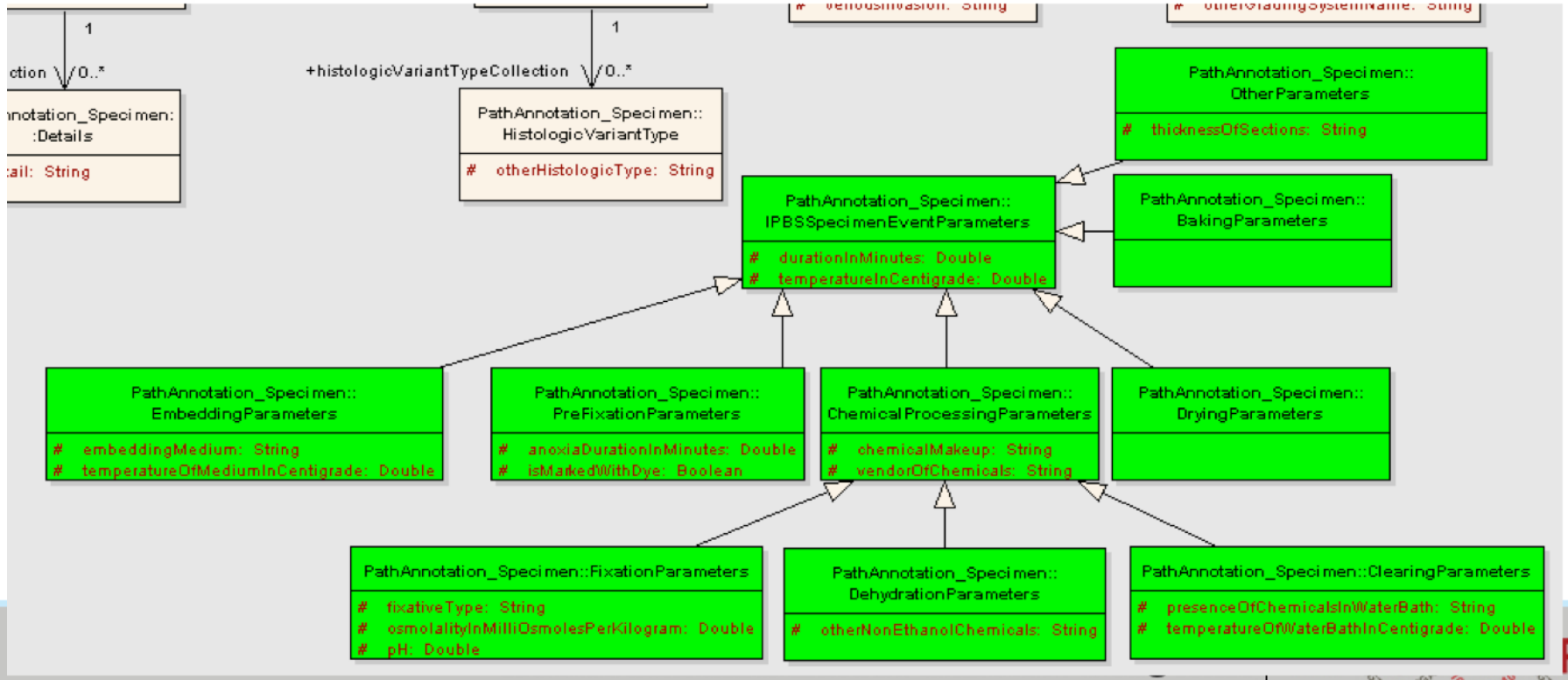
Dynamic Extensions - two routes to models

- Via the caTissue user interface as above
  - By domain experts
  - Capture model as part of doing business
  - Export to XMI
- Create model in modeling tool
  - Extend the hook classes
  - Save as XMI

Example of Dynamic Extensions

- Prostate SPORE Biospecimen Informatics Network specifically considered specimen handling variability and associated effects on biomarker fidelity

- Green classes/attributes were added off hook entity - specimen



Some lessons learned

- Validated the concept
  - Interoperability framework (caGrid) complex to deploy
- Allow rapid development of the service before labor intensive semantic curation
  - Iterate on the model
- Modeling is not part of end user workflow
  - Data elements – clinicians get it
  - One element at a time
- Data elements are too flat
- Resolve the dilemma by capturing models as part of doing domain business

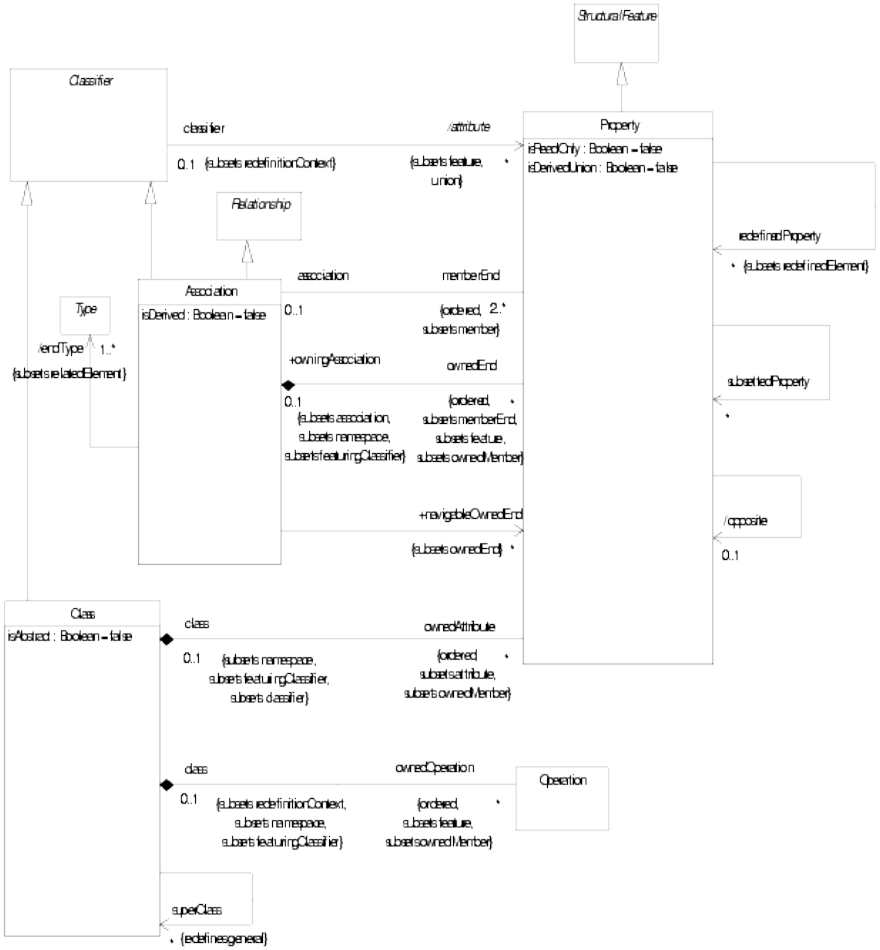
Opportunities

- Value of domain models
- Model driven architecture
  - To support dynamic domain models
- Domain driven models
  - Capture models as part of doing business
  - Not as a specific modeling exercise
- Process
  - Implementation of consistent modeling conventions
  - Or lock it down in the application
  - Need a flexible semantic process
- Interoperability framework
  - Looking at lighter weight technology
    - RESTful services
    - Semantic description of the above will be important

Relevant components of Semantic Infrastructure

- Metamodel
  - What do we need to know about data?
- Tools
  - Tools that generate interoperable, semantically annotated, data services
    - Multiple flavors – SOAP, REST, etc
    - Semantics driven
    - Model driven
- Process
  - The mechanisms we use to manage the metadata
  - How do we collect information about data from its source?

UML 2.0 metamodel – Classes, Properties, Associations



Metamodel

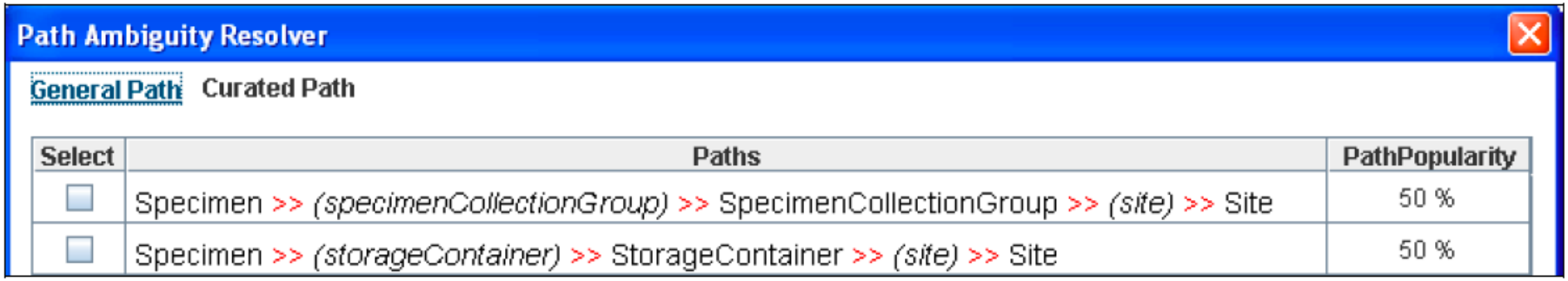
- Think about the metamodel you need
  - What metadata do your users need?
  - What do they need to know about data?

- How can we evolve the standards in this area?

UML metamodel

- Useful but likely to be a superset of the metamodel you need
- This use case here is focused on principally on data/information models
- UML still leaves a lot of choices open
  - Different ways of representing associations
- UML is used at different levels
  - Physical/Application/System
  - Domain models
- Need to pick what is needed for the purpose

How association metadata affects users



Specimen >> Specimen Collection Group >> Site

Specimen >> Storage Container >> Site

Specimen >> *collected as part of* >> Specimen Collection Group >> *collected at* >> Site

Specimen >> *stored in* >> Storage Container >> *located at* >> Site

Summary

- Model Driven Architecture
  - Data architecture –
  - Helps address dynamic implementation
- Domain Driven Models are key
  - Capture models in the course of business
- UML Class Models are a relevant technology
  - Establish a subset of UML
- Integrate metamodel of terminologies, data elements and data models