The Evolution*of DDI
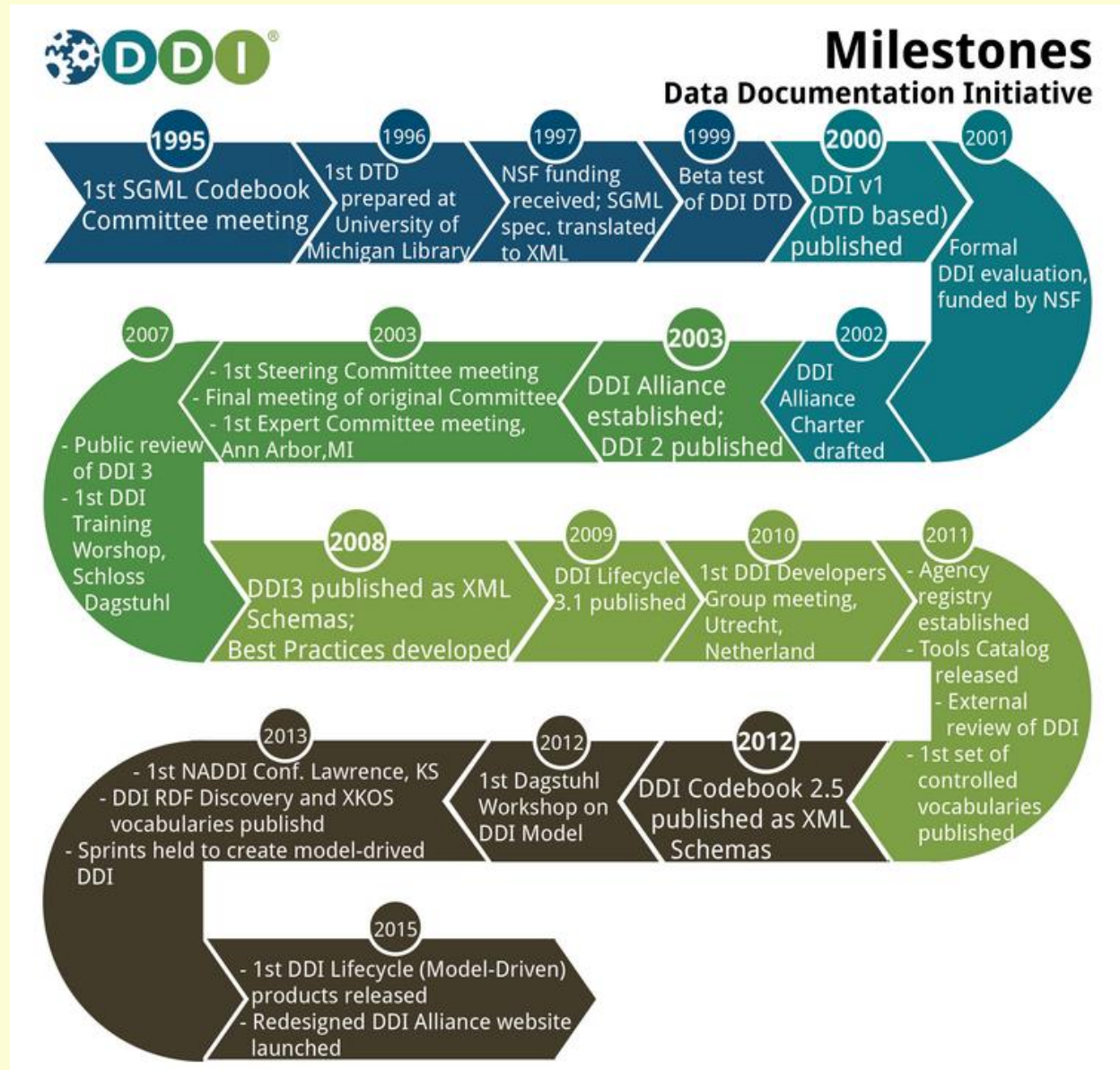
Concepts and Technology

* Yes some people from Kansas can talk about evolution

DDI Timeline (http://www.ddialliance.org/what/history.html)

Milestones
Data Documentation Initiative

- **1995** 1st SGML Codebook Committee meeting
- **1996** 1st DTD prepared at University of Michigan Library
- **1997** NSF funding received; SGML spec. translated to XML
- **1999** Beta test of DDI DTD
- **2000** DDI v1 (DTD based) published
- **2001** Formal DDI evaluation, funded by NSF
- **2002** DDI Alliance Charter drafted
- **2003** DDI Alliance established; DDI 2 published
- **2003** - 1st Steering Committee meeting - Final meeting of original Committee - 1st Expert Committee meeting, Ann Arbor, MI
- **2007** - Public review of DDI 3 - 1st DDI Training Worshop, Schloss Dagstuhl
- **2008** DDI3 published as XML Schemas; Best Practices developed
- **2009** DDI Lifecycle 3.1 published
- **2010** 1st DDI Developers Group meeting, Utrecht, Netherland
- **2011** - Agency registry established - Tools Catalog released - External review of DDI - 1st set of controlled vocabularies published
- **2012** DDI Codebook 2.5 published as XML Schemas
- **2012** 1st Dagstuhl Workshop on DDI Model
- **2013** - 1st NADDI Conf. Lawrence, KS - DDI RDF Discovery and XKOS vocabularies publishd - Sprints held to create model-driven DDI
- **2015** - 1st DDI Lifecycle (Model-Driven) products released - Redesigned DDI Alliance website launched

20+ year history
Includes:
- Dates of releases
- Administrative events
- Related events

See Also:
Vardigan, Mary *DDI Timeline*.
 IASSIST Quarterly
Volume 37 Number 1 – 4, 2013 pp. 51 – 56

In the Beginning There was Data
But much has changed in 20+ years

- Technology
  - Example: 1990 US Census STF3 on CDROM as DBASEIII+ files
  - Technical Documentation as text documents

  - On Tape: space conserving fixed column formats. Example: a hierarchy.

Example - 1990 U.S. PUMS Data Dictionary

1990 PUBLIC USE MICRODATA SAMPLES, U.S. DATA DICTIONARY
HOUSING UNIT RECORD

DATA		SIZE	BEGIN
D	RECTYPE	1	1
	Record Type		
V	H	.Housing Record	
D	SERIALNO	7	2
V	0000000..		
	9999999	.Housing unit/GQ person serial number unique	
		.identifier assigned within state or state group	
D	SAMPLE	1	9
	Sample Identifier		
V	1	.5% sample	
V	2	.1% sample	
V	3	.Elderly	
D	DIVISION	1	10
	Division code		
V	0	.Region/division not identifiable	
		.(Selected MSA/PMSAs on 1% sample)	
V	1	.New England (Northeast region)	
V	2	.Middle Atlantic (Northeast region)	
V	3	.East North Central (Midwest region)	
V	4	.West North Central (Midwest region)	
V	5	.South Atlantic (South region)	
V	6	.East South Central (South region)	
V	7	.West South Central (South Region)	
V	8	.Mountain (West region)	
V	9	.Pacific (West region)	

DATA		SIZE	BEGIN
D	MSAPMSA	4	20
	MSA/PMSA		
V	0040..		
V	9340	.FIPS/MSA/PMSA code, selected MSA/PMSA	
V		.(See appendix G)	
V	9997	.Mixed MSA/PMSA NONMSA/PMSA area	
V	9998	.2 or more MSAs	
V	9999	.Not in MA	
D	PSA	3	24
	Planning service area (elderly sample only - state dependent)		
V	000	.N/A (Elderly PUMS only)	
V	144..188	.Planning service area codes (See appendix G)	
D	SUBSAMPL	2	27
	Subsample number (Use to pull extracts - 1/1000/etc.)		
V	00..99	.See text. pp 4-45.	
D	HOUSWGT	4	29
	Housing Weight		
V	0000..		
V	1152	.Integer weight of housing unit	
D	PERSONS	2	33
	Number of person records following this housing record		
V	00	.Vacant unit	
V	01	.One person record (one person in household	

NADDI

Complex (Ostensibly) Human Readable Instructions

"Please note that the electronic data dictionary has the same information as the data dictionary in the basic paper documentation."

from http://www.archives.gov/research/census/1990-statistics.html

Digitized Human Readable Table Layout Descriptions

TABLE (MATRIX) SECTION

Table (matrix) number	Table (matrix)	Data dictionary reference name	Field size	Data type	Table (matrix) coordinates
P1.	PERSONS(1) [1]				
	Universe: Persons				
	Total	P0010001	9	N	1
P2.	UNWEIGHTED SAMPLE COUNT OF PERSONS(1) [1]				
	Universe: Persons				
	Total	P0020001	9	N	1
P3.	100 PERCENT COUNT OF PERSONS(1) [1]				
	Universe: Persons				
	Total	P0030001	9	N	1
P3A.	PERCENT OF PERSONS IN SAMPLE(1) [1]				
	1 implied decimal(s)				
	Universe: Persons				
	Total	P003A001	9	9	1

Somewhat Structured Data Dictionaries

http://www.icpsr.umich.edu/cgi-bin/file?comp=none&study=9592&ds=2&file_id=827049&path=ICPSR

```
G1  FILEID      8 A/N       1       1 0
G2        1 File Identification
G1  STUSAB      2 A         9       9 0
G2        9 State/US Abbreviation
G3        9       AK Alaska
G3        9       AL Alabama
G3        9       AR Arkansas
G3        9       AZ Arizona
G3        9       CA California
G3        9       CO Colorado
G3        9       CT Connecticut
G3        9       DC District of Columbia
G3        9       DE Delaware
G3        9       FL Florida
G3        9       GA Georgia
G3        9       HI Hawaii
G3        9       IA Iowa
```

1990 STF3 tape data dictionary

A text file which could be parsed, but sometimes tricky.

"Human Readable" Table Layout Descriptions
Not exactly formless or empty

```
P82. RACE OF HOUSEHOLDER(5) BY HOUSEHOLD
     INCOME IN 1989(9) [45]
   Universe:  Households
   White:
     Less than $5,000                    P0820001    9      N       1,1
     $5,000 to $9,999                    P0820002    9      N       1,2
     $10,000 to $14,999                  P0820003    9      N       1,3
     $15,000 to $24,999                  P0820004    9      N       1,4
     $25,000 to $34,999                  P0820005    9      N       1,5
     $35,000 to $49,999                  P0820006    9      N       1,6
     $50,000 to $74,999                  P0820007    9      N       1,7
     $75,000 to $99,999                  P0820008    9      N       1,8
     $100,000 or more                    P0820009    9      N       1,9
   Black:
     (Repeat HOUSEHOLD INCOME IN 1989)   P0820010   811     N       2,1
   American Indian, Eskimo, or Aleut:
     (Repeat HOUSEHOLD INCOME IN 1989)   P0820019   811     N       3,1
   Asian or Pacific Islander:
     (Repeat HOUSEHOLD INCOME IN 1989)   P0820028   811     N       4,1
   Other race:
     (Repeat HOUSEHOLD INCOME IN 1989)   P0820037   811     N       5,1
```

But not easy to parse either

Sometimes We Shared Code For Proprietary Platforms

```
/*   Setup to invoke the sdcpro3 macro to generate one or more profile reports using the sdcprofile3
     by the national SDC.
   */
options mprint source2;

 %include '/pub/sasmacro/sdcpro3.sas';
%macro sdcpro3(setin=,      /* name of input data set. If blank program will merge 3 sets based on &st
  sumlev=,stab=,profiles=,pages=,
  logrecno=0,   /* specify the logical rec no and program will read the single obs. */
  filter=, /* name of filter macro.  Ignored if logrecno specified*/
  CumulPaging=1,geocomps=00,
  RevDate=26Aug02,
  debug=0);

/* Adapted from MainsSF3.sas by Roy Williams, Mass. SDC (MISER).
   by John Blodgett,  Missouri Census Data Center.
   ------*/
```

Origin Context 1990-1993

Concepts and Features

Technology

1993

IASSIST Action Group

"**Codebook** Documentation

of Social Science Data."

CESSDA

seminar on "**Variable** Level

Documentation" in Gothenburg

Some context:

Web Browser (1990-12-20, Berners-Lee)

Lynx Browser (1992)
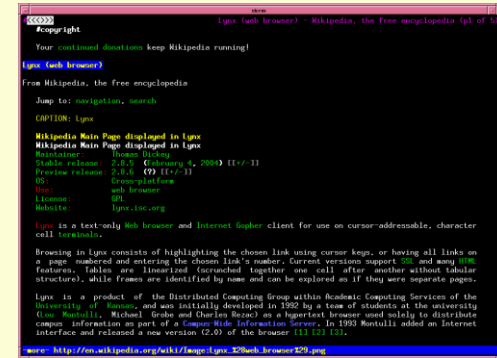
WWW made free (1993-04-20)

Mosaic Browser (1993)

https://en.wikipedia.org/wiki/File:Lynx-wikipedia.png

https://upload.wikimedia.org/wikipedia/en/a/a5/NCSAMosaic1.0Mac.png

https://en.wikipedia.org/wiki/World_Wide_Web

1994

1994

Technology

Dublin Core Metadata Initiative
(2nd International World Wide Web Conference, October 1994)

Metadata design and specification

http://dublincore.org/about/history/
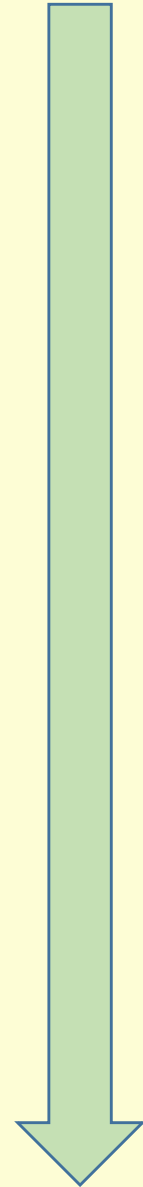
1995

1995

Technology

First Codebook Committee meeting

Standard Generalized Markup Language
SGML (then 9 years old)

A structured codebook
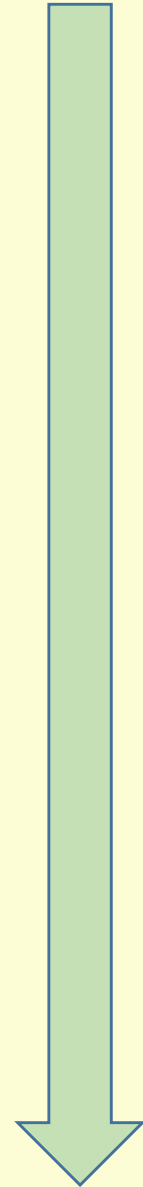
1996

1996

First codebook SGML **DTD**

XML working group subsets XML from SGML
http://www.w3.org/TR/REC-xml/

1997

1997

First codebook expressed as XML

XML

NADDI
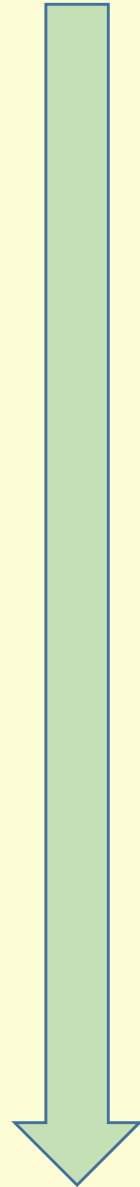
1998 - 2000

1998

Beta testing and revision

XML

(21 versions shown in final 1.0 DTD)

Mapping to Dublin Core documented

1999

Resource Description Framework (RDF)
as a W3C recommendation

2000

2000

2000

Version 1.0 published March 17, 2000

XML DTD

DTD - "Document Type Definition"

<!ELEMENT codeBook
   (**docDscr***
   , stdyDscr+
   , fileDscr*
   , dataDscr*
   , otherMat*)

A Codebook as a *document* describing dataset(s)

Machine Actionable

Coverage, universe, dataKind, methodology(collection), sampling, weighting, access conditions, summary statistics, Variable groups

5 Years From first Committee Meeting

DDI-C - Structure

<!ELEMENT codeBook

  (**docDscr***   About the Metadata (document)– citation, status, title, date,

  , stdyDscr+   About the Study – citation, abstract, method, other, etc.

  , fileDscr*   About the File – name, citation, structure, format, etc.

  , dataDscr*   About the Data – variables, cube description,

  , otherMat*)   Additional Material

```
PUMS5_20.TXT    ×
        1,0        2,0        3,0        4,0        5,0
1 H00019575202400100201009999999999999999090        44367497000
2 P00019570100001101000120230010110000010147031400110011
3 P00019570200001103011010050010110000010147050000212102
4 P00019570300001203011020020010110000010147050000000000
5 H00021975202400100201009999999999999999090        44367497000
6 P00021970100000801000010610010110000010147010010001300
7 P00021970200000702000420560011110000010147010010001000
8 H00103355202400100201009999999999999999090        44367497000
9 P00103350100002601000010690010110000010147010010000910
```

<docDscr>
  <citation>
    <titlStmt>
      <titl>A Sample Codebook</titl>
      <subTitl>About this Codebook</subTitl>
    </titlStmt>
  </citation>
</docDscr>

<stdyDscr>
  <citation>
    <titlStmt>
      <titl>A DDI-C example Study</titl>
      <subTitl>About this Study</subTitl>
    </titlStmt>
  </citation>
</stdyDscr>
<fileDscr>
  <fileTxt>
    <fileName>PUMS5_20.txt</fileName>
  </fileTxt>
</fileDscr>

dataDscr: DDI-C₍₂.₅₎ Var (the variable)

element <var> (global)

Namespace:	ddi:codebook:2_5
Type:	varType
Content:	complex, 36 attributes, 26 elements
Defined:	globally in codebook.xsd; see XML source
Used:	at 1 location

XML Representation Summary

```
<var
    ID                      = xs:ID
    xml-lang                = xs:NMTOKEN
    xml:lang                = xs:language
    source                  = ("archive" | "producer") : "producer"
    elementVersion          = xs:string
    elementVersionDate      = (xs:dateTime | xs:date | xs:gYearMonth | xs:gYear)
    ddiLifecycleUrn         = xs:anyURI
    ddiCodebookUrn          = xs:anyURI
    name                    = xs:string
    wgt                     = ("wgt" | "not-wgt") : "not-wgt"
    wgt-var                 = xs:IDREFS
    weight                  = xs:IDREFS
    qstn                    = xs:IDREFS
    files                   = xs:IDREFS
    vendor                  = xs:string
    dcml                    = xs:string
    intrvl                  = ("contin" | "discrete") : "discrete"
    rectype                 = xs:string
    sdatrefs                = xs:IDREFS
    methrefs                = xs:IDREFS
    pubrefs                 = xs:IDREFS
    access                  = xs:IDREFS
    aggrMeth                = ("sum" | "average" | "count" | "mode" | "median" | "maximum" | "minimum" | "percent" | "other")
    otherAggrMeth           = xs:NMTOKEN
    measUnit                = xs:string
    scale                   = xs:string
    origin                  = xs:string
    nature                  = ("nominal" | "ordinal" | "interval" | "ratio" | "percent" | "other")
    additivity              = ("stock" | "flow" | "non-additive" | "other")
    otherAdditivity         = xs:NMTOKEN
    temporal                = ("Y" | "N") : "N"
    geog                    = ("Y" | "N") : "N"
    geoVocab                = xs:string
    catQnty                 = xs:string
    representationType      = ("text" | "numeric" | "code" | "datetime" | "other")
    otherRepresentationType = xs:NMTOKEN
    >
    Content: location*, labl*, imputation?, security?, embargo?, respUnit?, anlysUnit?, qstn*, valrng*, invalrng*, undocCod*, universe*, TotlResp?, sumStat*, txt*, stdCatgry*,
             catgryGrp*, catgry*, codInstr*, verStmt*, concept*, derivation*, varFormat?, geoMap*, catLevel*, notes*
</var>
```

Many properties of a variable, tightly tied to a dataset

http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/field_level_documentation_files/schemas/codebook_xsd/elements/var.html

DDI-C$_{(1.0)}$ var

Variable elements:

location*,   labl* ,   imputation?,
security?,   embargo?,   respUnit?,
anlysUnit?,   qstn*,   valrng*,
invalrng*,   undocCod*,   universe*,
TotlResp?,   sumStat*,   txt*,
stdCatgry*,   catgryGrp*,   catgry*,
codInstr*,   verStmt*,   concept*,
derivation?, varFormat?,   notes*)

Variable Attributes:

name,   wgt,   wgt-var,
qstn,   files,   vendor,
dcml,   intrvl,   rectype,
sdatrefs, methrefs, pubrefs,
access

*ID references

DDI-C References

```
wgt-var          = xs:IDREFS
weight           = xs:IDREFS
qstn             = xs:IDREFS
files            = xs:IDREFS

sdatrefs         = xs:IDREFS
methrefs         = xs:IDREFS
pubrefs          = xs:IDREFS
access           = xs:IDREFS
```

Some use by reference.

ID unique only within **document**. (xs:ID)

In Hindsight DDI-C var as an *Instance Variable*

Version 1.0 (continued)

Variable somewhat tied to physical representation

Example - location:
StartPos, EndPos, width, RecSegNo,
fileid

GSIM Variable Cascade

http://www1.unece.org/stat/platform/display/GSIMclick/Represented+Variable
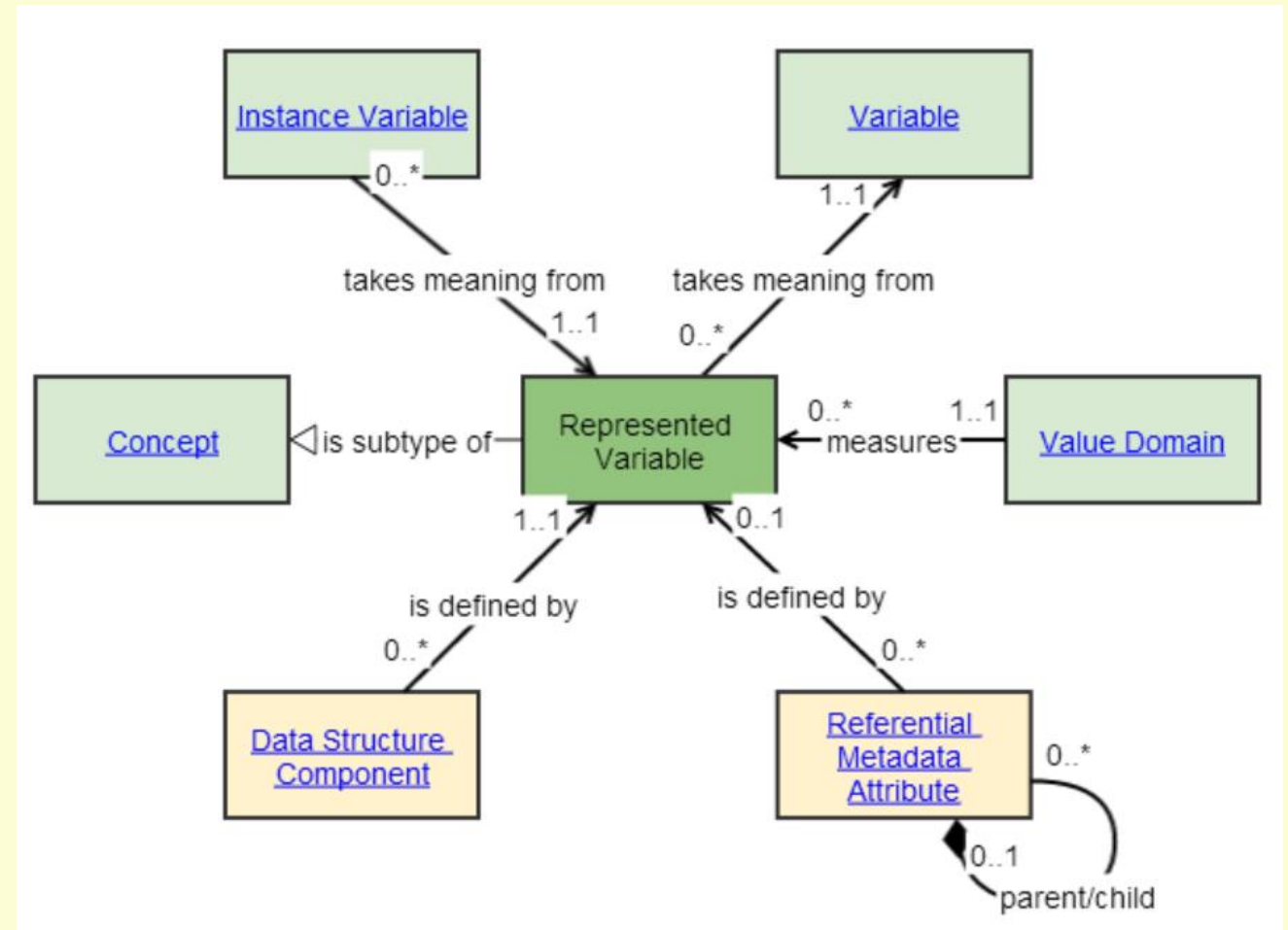
DDI-C Variables Sharing Questions

Version 1.0 (continued)

Variables can share questions

Question (qstn)

 elements: preQTxt, **qstnLit**, postQTxt, forward, backward, ivuInstr

 attributes: qstn, var, seqNo, sdatrefs

*ID references

Note that some

questionnaire

attributes are embedded in qstn

DDI-C Origins: Survey Centric

Version 1.0 (continued)

Variable Attributes:

Initial focus on Surveys

name, wgt, wgt-var, **qstn**, files,

vendor, dcml, intrvl, rectype, sdatrefs, methrefs, pubrefs, access
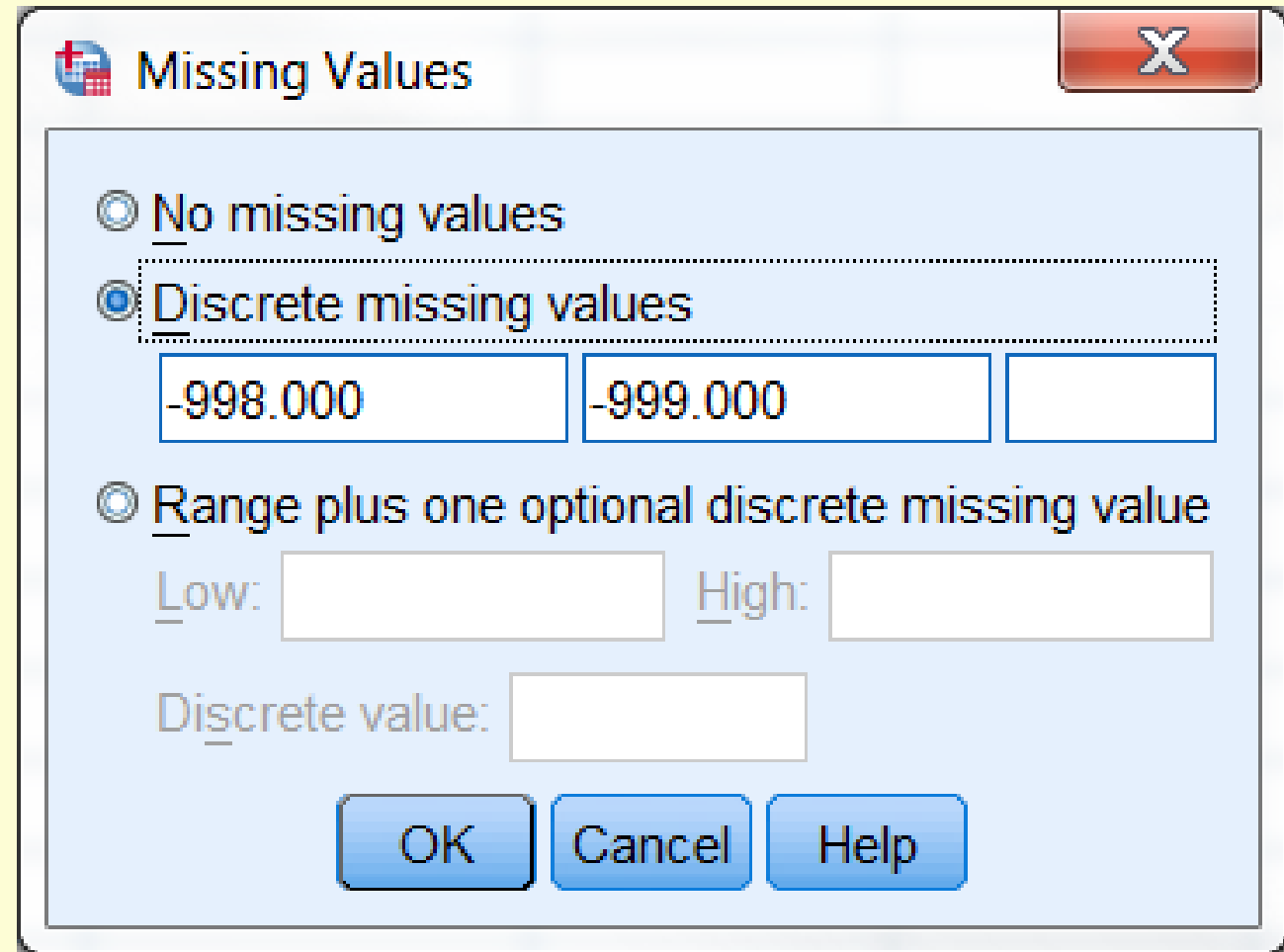
*ID references

DDI-C A Little SPSS Centric

Version 1.0 (continued)

<!ELEMENT invalrng ((range | item)+,
key?, notes?)

Invalid values by range or list
(a little **SPSS centric?**)

2001 - 2002

2001

Technology

XML DTD

Working group on **aggregate data**

March 2002 - SDMX initiative begins
https://sdmx.org/?page_id=2703

2002

2003

2003

XML DTD

DDI 2.0

nCubes

Geographic Bounding Box, Polygon, Point, G-Ring
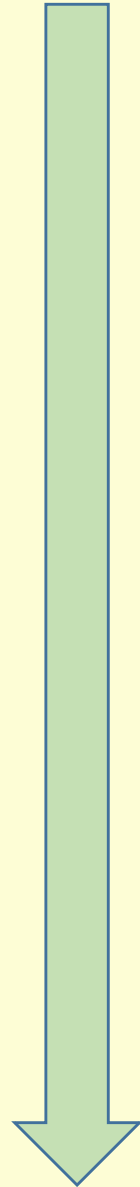
2004

2004

Technology

DDI 3 Conceptual Model Lifecycle

DDI underpinning an infrastructure

Figure 3: Combined Life Cycle Model

From http://www.ddialliance.org/system/files/Concept-Model-WD.pdf

RDF 1.0 published

Lifecycle View – New Approach to Metadata

DDI-C origins
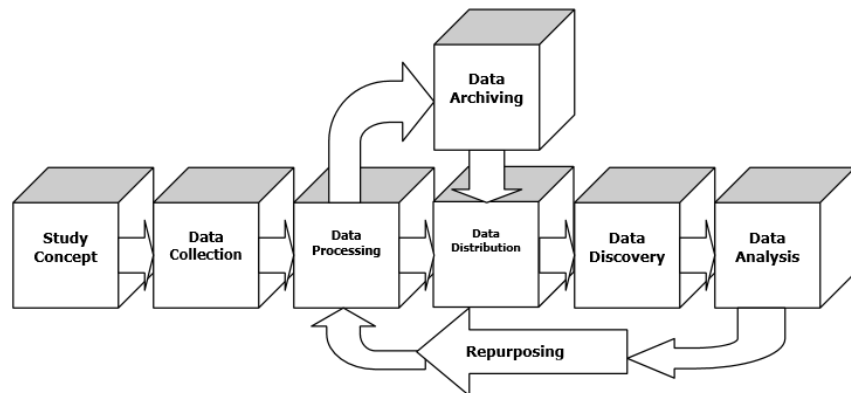
Data first

Metadata

Metadata

My Survey

What is your Name

Do you like NADDI?

○ Very Much

○ Meh

○ Not Very Much

Metadata

https://en.wikipedia.org/wiki/File:Crystal_Clear_app_ktip.svg

Metadata Driven Survey Design

Metadata-Driven Survey Design

Introduction

In current survey practice, the creation of a data collection instrument involves two distinct steps. The first is survey design, in which a researcher defines the questions and flow of a survey. The second is survey implementation, in which a researcher or programmer turns the design into an electronic or paper survey instrument.

Jeremy Iverson[1]

Once questions are defined, the survey designer can create the flow of a survey. A survey's flow could be as simple as a series of questions in a particular order.

Many surveys have a more complicated structure that includes loops, sampling, and conditional branching. For example, a certain section of a survey may only be administered to a random subsample of half the respondents in the full sample, or additional

Summer 2009

http://www.iassistdata.org/sites/default/files/iq/iqvol3312iverson.pdf

2004

2004

Figure 3: Combined Life Cycle Model
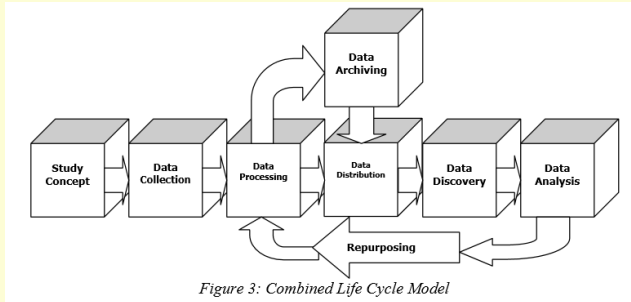
Modules
follow the conceptual design

Figure 4: Modules in the Conceptual Model

From http://www.ddialliance.org/system/files/Concept-Model-WD.pdf
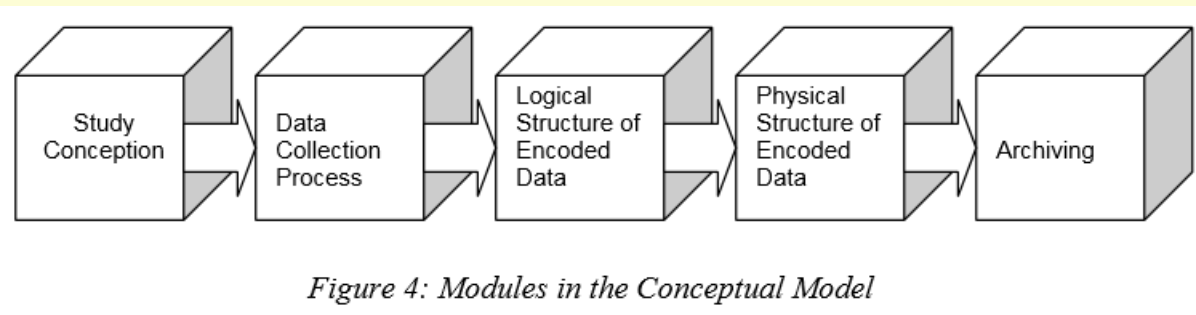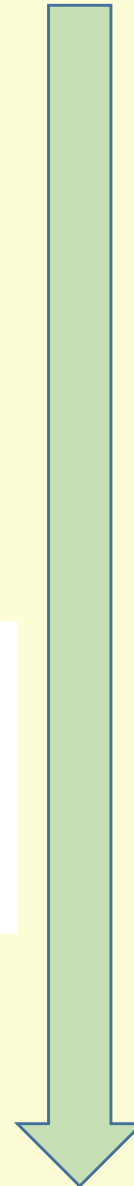
2005

2005

UML Model

XML Schema

Lifecycle model continues development

Conceptual Model **and Schemas**

ISO11179 metadata registry standard

http://www.ddialliance.org/DDI/committee-info/minutes/2005-05-22.html
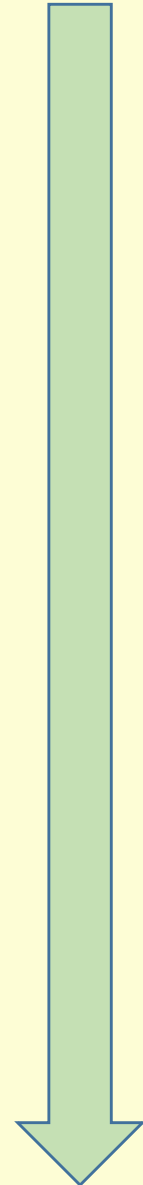
2006

2006

(XML DTD)

DDI 2.1 with Nested Categories

Additional var attributes

measunit,   scale,   origin,
nature,   additivity,   temporal,
geog,   geoVocab,   catQnty

NADDI

2007

2007

Technology
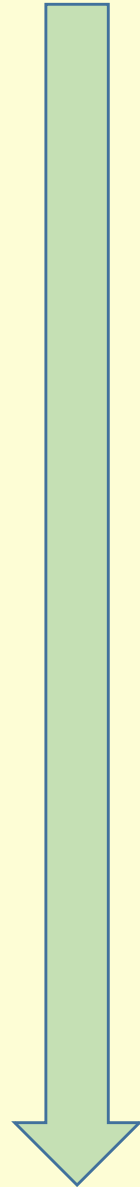
DDI 3.0 Public Review

- **Globally unique identifiers**
- **Maintainable, versionable, and identifiable classes.**

- # Reuse by reference

- **Modules**
  - **E.g. Logical and Physical**
- **Generalized "archive" notion**
- **Comparison**
- **Instrument / Control Construct**
- **More generalized grouping**
- **Fewer abbreviations in elmt nms**
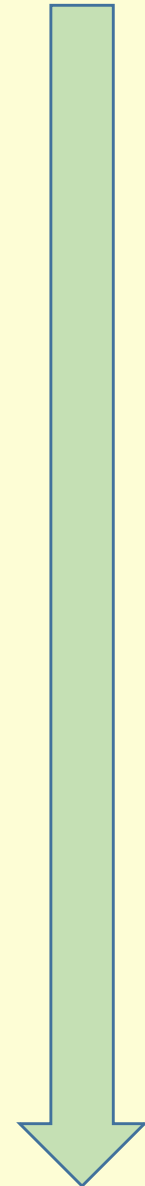
XML Schema

URN

Goal: UML model

"in Version 3.0 all aspects of the data life cycle will now be supported "

DDI_3.0_Part_I_Overview.pdf

2008

2008

XML Schema

DDI 3.0 Published

5 Years From DDI 2
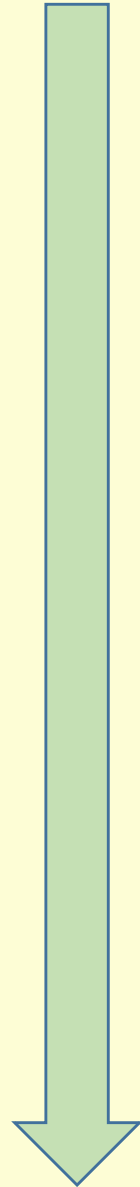
Identifiers

The structure of Identifiers has changed

across the major versions of DDI

NADDI

Larry Hoyle, IPSR, Univ. of Kansas

DDI-C IDs using xs:ID

- DDI-C Ids Unique within **Document**
- References within document

```
<dataDscr>
    <var name="MyVar"  ID="myVar1">
      <qstn ID="MyQ1">How tall are you?</qstn>
    </var>
    <var name="MyVar2" qstn="MyQ1">
    </var>
```

DDI-L Identifiers
(r:URN | (r:Agency, r:ID, r:Version))[1..2]

- DDI-L Identifiers **Internationally** Unique

- Managed within Agency

- **Agency, ID, Version**

- Using UUID possible (128 bit value) unlikely to be duplicated

- ISO 11179-6 compliant

See https://ddi-alliance.atlassian.net/wiki/display/DDI4/Identification

```
<c:Universe>
   <r:URN>URN:DDI:US.IPSR:c7b8959b-9aa3-4dc0-a2b8-b640347d3506:1.0</r:URN>
   <r:Agency>US.IPSR</r:Agency>
   <r:ID>c7b8959b-9aa3-4dc0-a2b8-b640347d3506</r:ID>
   <r:Version>1.0</r:Version>
```

DDI-L Identification and Maintainable Object

- **Identifiable** – Name and id
- **Versionable** – version, versionDate, versionResponsibility, versionRationale
- **Maintainable** – agency

- Context (Uniqueness) provided by Maintainables
  - Parent maintainable identifier part of identifier
  - Required 3.0 and 3.1, optional 3.2

- Schemas and Schemes
  - Schemes – Organizations, Concept, Universe, Geographic Structure, Geographic Location, Control Construct, Question, Category, Code, Ncube, Variable, Physical Structure, Record Layout

DDI-C(2.5) ddiLifeCycleURN

- Unique within Document

- Combines Agency, ID, Version

- **Can carry DDI-L URN for transition**

```
<dataDscr>
    <var name="MyVar" ID="myVar1">
      <qstn ID="MyQ1"
        ddiLifecycleUrn="URN:DDI:US.IPSR:17b8959b-9aa3-4dc0-a2b8-b640347d3506:1.0">
        How tall are you?
      </qstn>
    </var>
    <var name="MyVar2" qstn="MyQ1">
    </var>
```

DDI-L(3.2) Variable – Mostly References

element <Variable> (global)

Namespace:	ddi:logicalproduct:3_2
Type:	VariableType
Content:	complex, 9 attributes, 26 elements
Defined:	globally in logicalproduct.xsd; see XML source
Used:	at 2 locations

XML Representation Summary

```
<Variable
    inheritanceAction   = ("Add" | "Update" | "Delete")
    objectSource        = (xs:string | xs:string)
    scopeOfUniqueness   = ("Agency" | "Maintainable") : "Agency"
    isUniversallyUnique = xs:boolean
    versionDate         = (xs:dateTime | xs:date | xs:gYearMonth | xs:gYear |
                          xs:duration)
    isVersionable       = "true"
    isTemporal          = xs:boolean : "false"
    isGeographic        = xs:boolean : "false"
    isWeight            = xs:boolean : "false"
    >
    Content: (r:URN | (r:Agency, r:ID, r:Version))[1..2], r:UserID*,
            r:UserAttributePair*, (r:VersionResponsibility |
            r:VersionResponsibilityReference)?, r:VersionRationale?,
            r:BasedOnObject?, r:MaintainableObject?, VariableName*, r:Label*,
            r:Description?, r:OutParameter?, r:SourceParameterReference?,
            r:SourceVariableReference*, RepresentedVariableReference?,
            r:ConceptualVariableReference?, r:UniverseReference*,
            r:ConceptReference?, r:QuestionReference*, EmbargoReference?,
            SourceUnit?, r:AnalysisUnit?, VariableRepresentation?
</Variable>
```

All of these objects can be reused!

- Direct Content
  - Agency, ID, Version (URN)
  - AnalysisUnit
  - Description
  - Label
  - SourceUnit
  - UserID
  - VariableName
  - VersionRationale
  - VersionResponsibility
- Contain References
  - BasedOnObject
  - OutParameter
  - VariableRepresentation
- References
  - ConceptReference
  - ConceptualVariableReference
  - EmbargoReference
  - MaintainableObject
  - QuestionReference
  - RepresentedVariableReference
  - SourceParameterReference
  - SourceVariableReference
  - UniverseReference
  - UserAttributePair
  - VersionResponsibilityReference

2008 DDI3.0 Modules

http://www.ddialliance.org/Specification/DDI-Lifecycle/3.0/XMLSchema/Documentation/

- Archive module
- Comparative module
- Conceptual components module
- Data collection module
- Dataset module
- Dublin Core Elements module
- DDI profile module
- Grouping module
- Instance module
- Logical product module
- Physical data product module
- Physical data product - inline n-cube module
- Physical data product - normal n-cube module
- Physical data product - tabular n-cube module
- Physical data product - proprietary module (beta)
- Physical instance module
- Reusable module
- Study unit module

Namespaces

2009

2009

Technology

DDI 3.1 Published
Tweaks and Bug fixes
    (but not backward compatible)
Resolution of URN structure
More things optional

ResearcherID,   cardinalities,
@isPublished attribute

http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/

Changes from 3.0:
  http://www.ddi-
  alliance.net/sites/default/files/ChangesVersion3.pdf

XML Schema

URN

43

NADDI

43
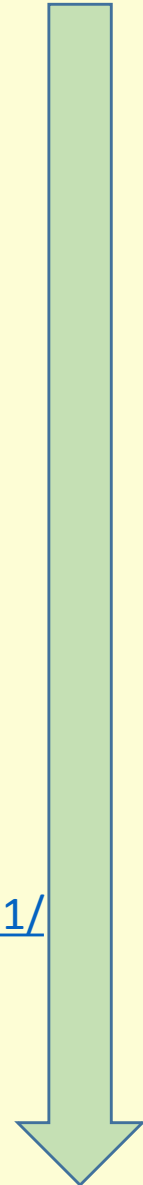
2011

2011

Genericode, HTML, XLS

DDI Controlled Vocabularies
- Currently 20 vocabularies

http://www.ddialliance.org/controlled-vocabularies

DDI Registry
- Register agency ID
- DNS based global resolution for DDI-related services

http://registry.ddialliance.org/
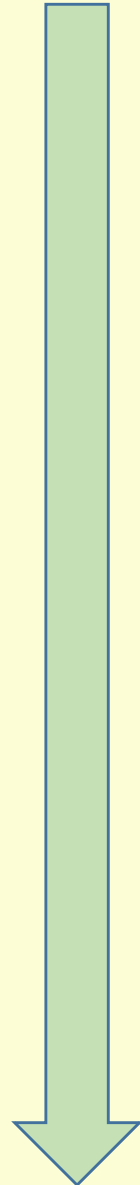
2012

2012

DDI 2.5 improved compatibility with Lifecycle

XML Schema

Citation – dc and dcterms
Controlled Vocabularies
DDI-L URN
Version everywhere
XHTML formatted text
QualityStatement
sampleFrame
Question: responseDomainType
Variable: RepresentationType

Changes:
http://www.ddialliance.org/Specification/DDI-Codebook
/2.5/detailed_changes_to_ddi-2.pdf

2012

2012

Technology
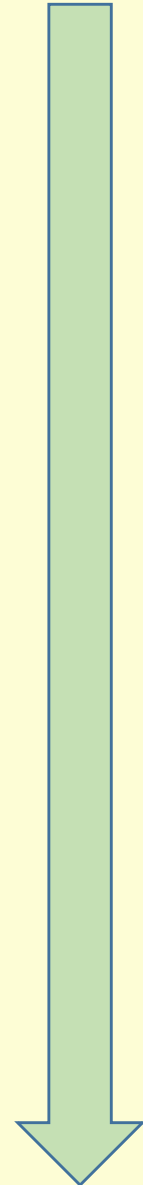
DDI 4 effort begins
A.k.a. DDI Model Driven, DDI Views

GSIM

Developed as a Model

UML (Unified Modeling Language)

(compose) – the existence of the parts is dependent on the aggregate
(aggregate) – parts belong to the aggregation but exist independently
(generalize) - i.e., inheritance, the specific inherits from the general
(reference)

Initial Bindings to XML and RDF
        to be automated

Informed by Information Models
        (e.g. GSIM)
                Variable Cascade
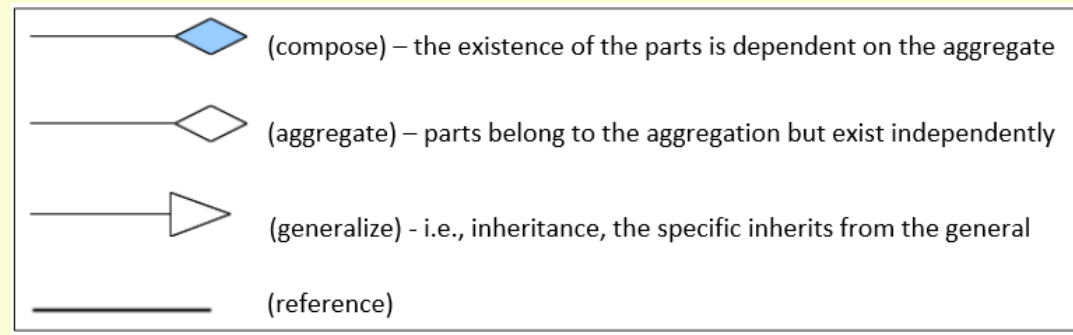
Automated binding to:
XML Schema
RDF
Other?

http://www.ddialliance.org/system/files/DevelopingaModel-DrivenDDISpecification2013_05_15.pdf

2013

2013

DDI RDF Vocabularies

RDF

- **DDI-RDF Discovery Vocabulary (DISCO)**
  - **Discovery In Linked Open Data**

- **XKOS - Extended Knowledge Organization System**
  - **Extends Simple Knowledge Organization System (SKOS)**
  - **Classifications, concept systems**

- **PHDD - Physical Data Description**
  - **Physical (table) layouts**
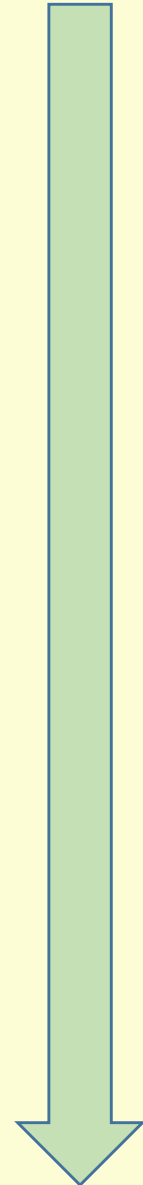
2013

2013

Moving Forward Sprints

NADDI

North American Data Documentation Initiative Conference

NADDI

Larry Hoyle, IPSR, Univ. of Kansas
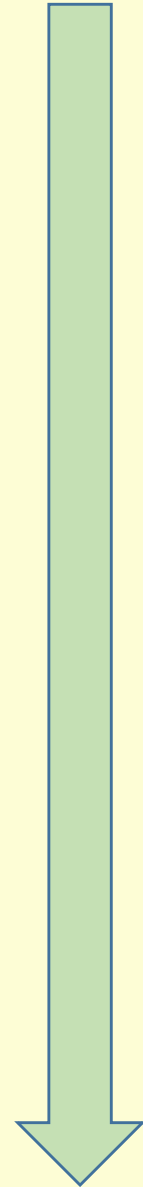
2014

Concepts and Features

Technology

DDI 3.2

Fragments
UserAttributePair
Variable Cascade
     Conceptual Variable
     Represented Variable
InParameter, OutParameter
Quality Descriptions
Managed Representations
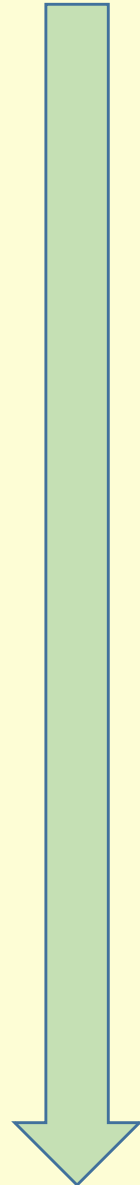
DDI4 Development Process – Drupal Site

Datum

View	What links here	Revision operations

Submitted by steve on Tue, 10/21/2014 - 15:19

Package: LogicalDataDescription
Extends: ValueDomain
Version: 0
Is Abstract: no
Status: Content review in progress

Definition:
A Datum is the designation of a value

Explanatory Notes:
A Datum is the actual instance of data that was collected or derived. It is the value which populates a Data Point. A Datum is the value found in a cell of a table. [GSIM 1.1]

NOTE: This is NOT datum from DDI3.2 (which is quite specific).

Property:

Inherited from: ValueDomain

Name	Cardinality	Datatype	Description
unitOfMeasurement	0..1	xs:string	The unit in which the data values are measured (kg, pound, euro).
label	0..n	DisplayLabel	A display label for the object. May be expressed in multiple languages. Repeat for labels with different content, for example, labels with differing length limitations.
definition	0..1	StructuredString	A definition of the object. May be expressed in multiple

Inherited from: Identifiable

Name	Cardinality	Datatype	Description
agency	1..1	xs:string	This is the registered agency code with optional sub-agencies separated by dots. For example, diw.soep, ucl.qss, abs.essg.
id	1..1	xs:string	The ID of the object. This must conform to the allowed structure of the DDI Identifier and must be unique within the Agency.
version	1..1	xs:string	The version number of the object. The version number is incremented whenever the non-administrative metadata contained by the object changes.

Object properties
Relationship:
Inherited from: AnnotatedIdentifiable

| Name	Target Object	Description	Source cardinality	Target cardinality	Relationship type
hasAnnotation	Annotation	Provides annotation information on the object to support citation and crediting of the creator(s) of the object.	0..1	1..n	Aggregation

Object relationships

Name	Target Object	Description	Source cardinality	Target cardinality	Relationship type
isDescribedBy	InstanceVariable	A Datum is described by an InstanceVariable.	0..n	1..1	Neither

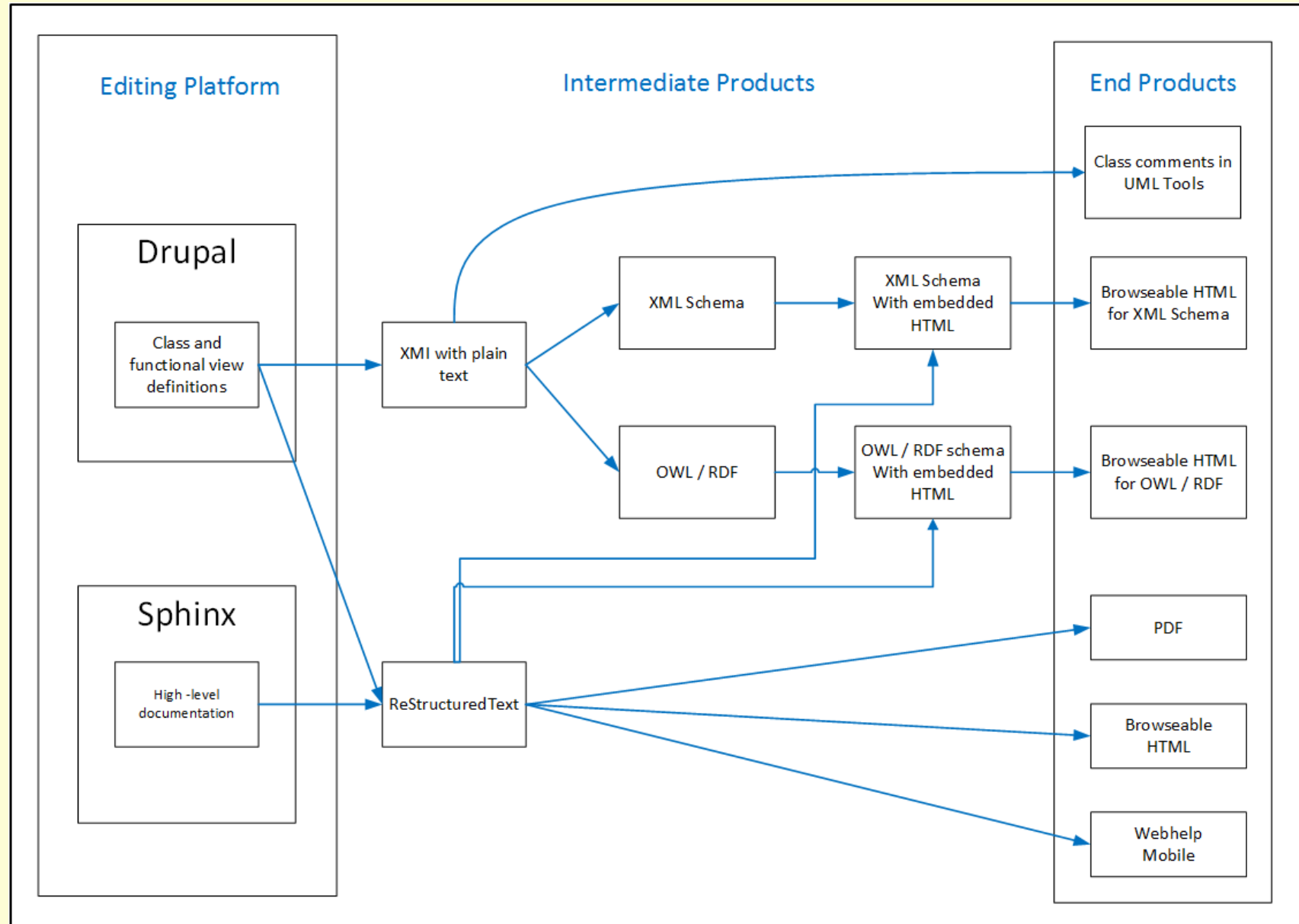# DDI4 Development Process – Drupal Site
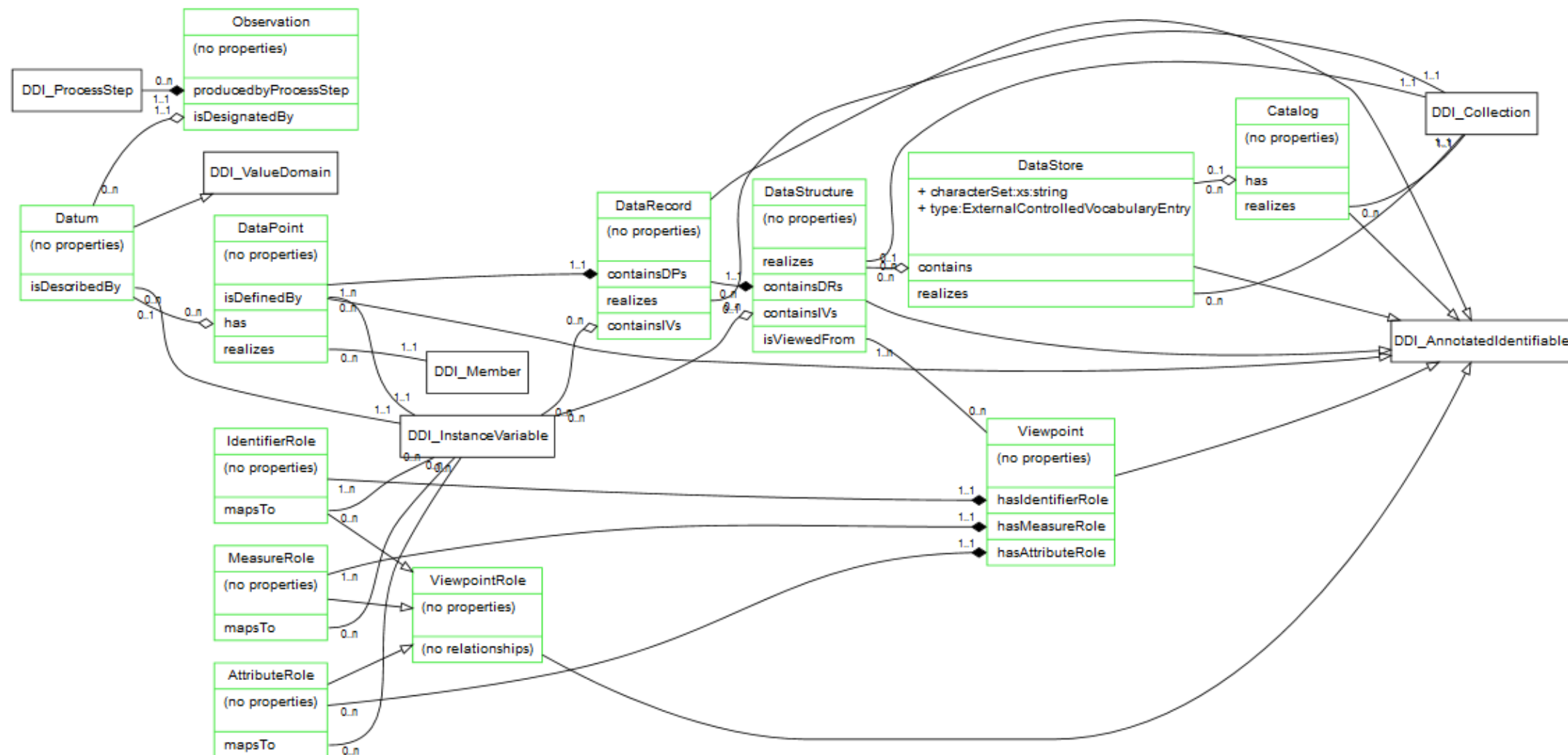
DDI4 Production Process - Bindings
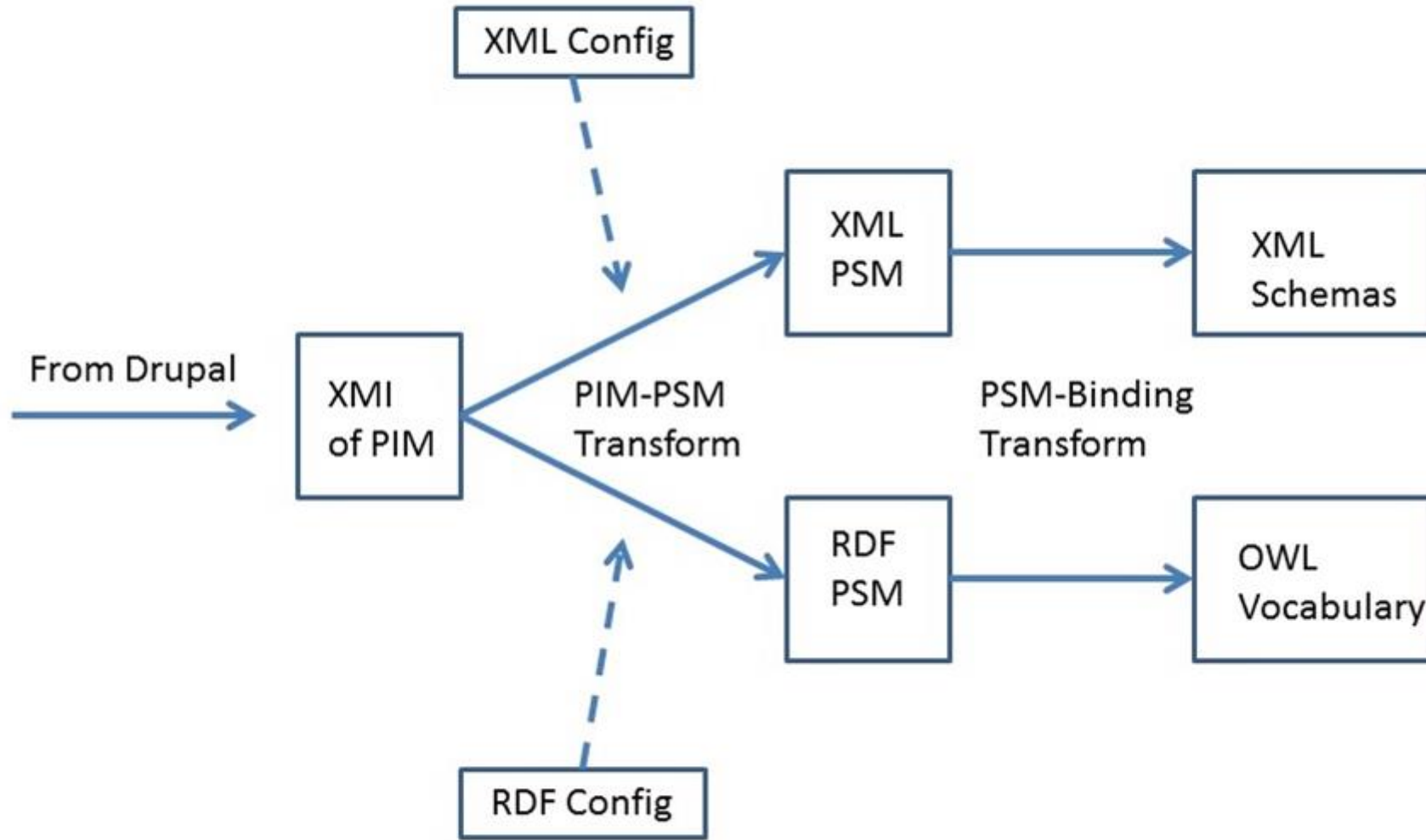
PIM – Platform Independent Model

PSM – Platform Specific Model

https://ddi4.readthedocs.org/en/latest/Introduction/modelproduction.html

DDI4 Production Process - Documentation

DDI4 Structure

- One Big Namespace

- **Views** with subsets of the overall model
  - Separate documentation
  - Separate XML and OWL/RDF schemas for each View

- Initially both XML and RDF bindings

     North American Data Documentation Initiative Conference   Larry Hoyle, IPSR, Univ. of Kansas

New in DDI4

- (Almost) Everything Identifiable (Agency, ID, Version)
- Variable Cascade (also in 3.2)
- Patterns
    - Collections and relations
    - Process
    - Citation information
- Data Capture (extending beyond surveys)
- Qualitative
- Custom Metadata & Controlled Vocabularies
- Datum, DataPoint, ViewPoint
- Substantive and Sentinel values (as in ISO11404)
- Multi-Stage Sampling

Variable Cascade

ConceptualVariable - Concept and UnitType

e.g. I want to measure height, of persons, numeric, record "refused"
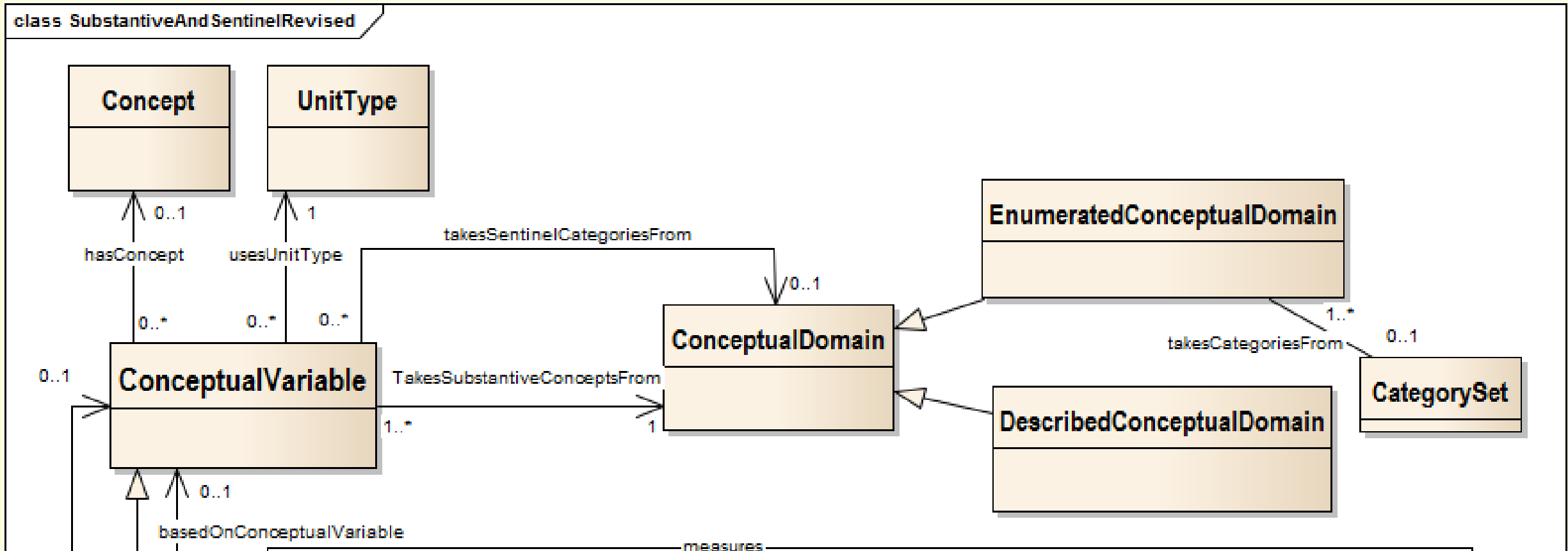
Substantive and Sentinel values as in ISO 11404
e.g. I want height in inches to 1 decimal place,
with -9 as refused.
Sample from schoolchildren in this city

Variable Cascade
RepresentedVariable – Universe and ValueDomain
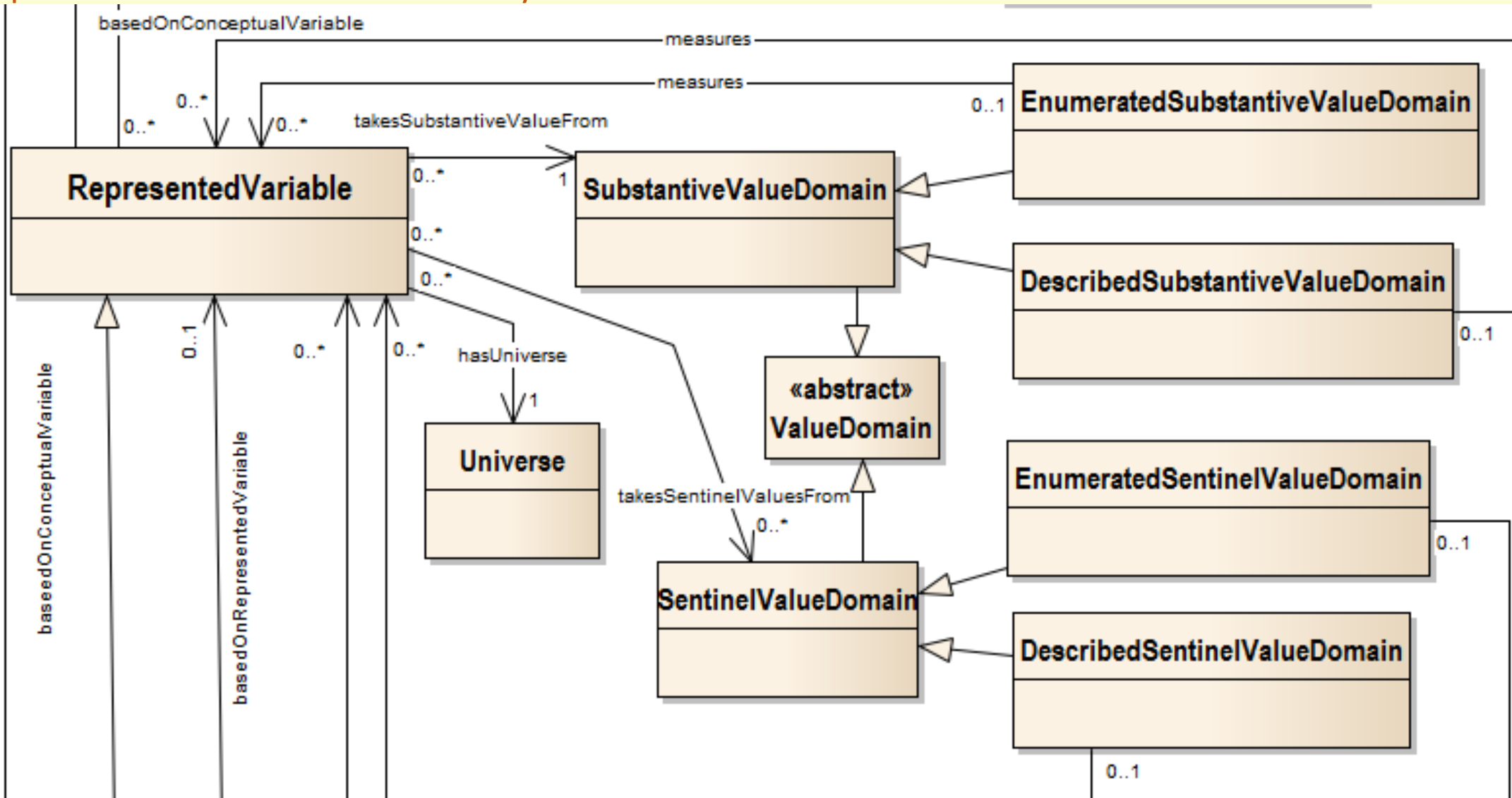
Variable Cascade

InstanceVariable – Population, Capture, DataType

e.g. We found children in these schools.... ,

Recorded as Stata double,

Measured with a Stanley tape measure....
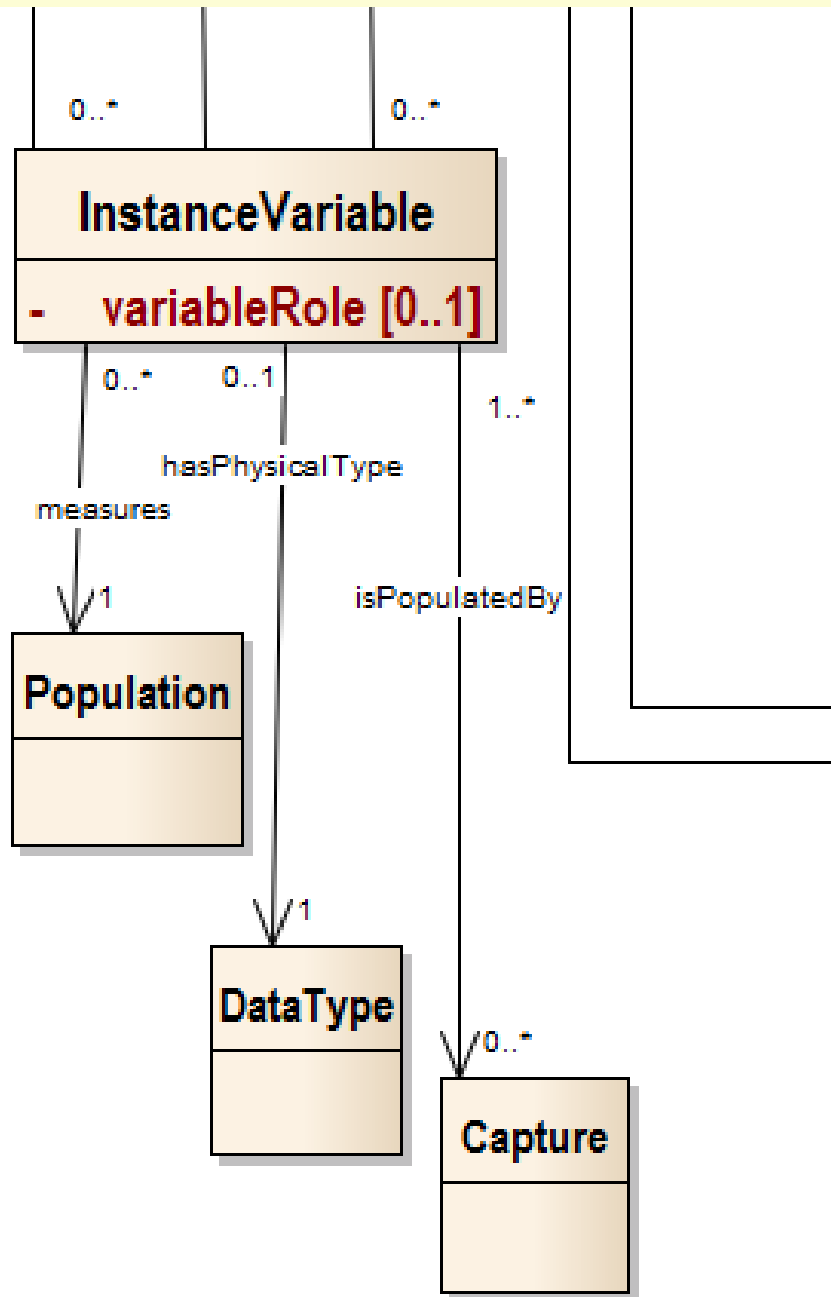
Variable Cascade

ConceptualVariable - Concept and UnitType
RepresentedVariable – Universe and ValueDomain
InstanceVariable – Population, Capture, DataType

Substantive and Sentinel values as in ISO 11404
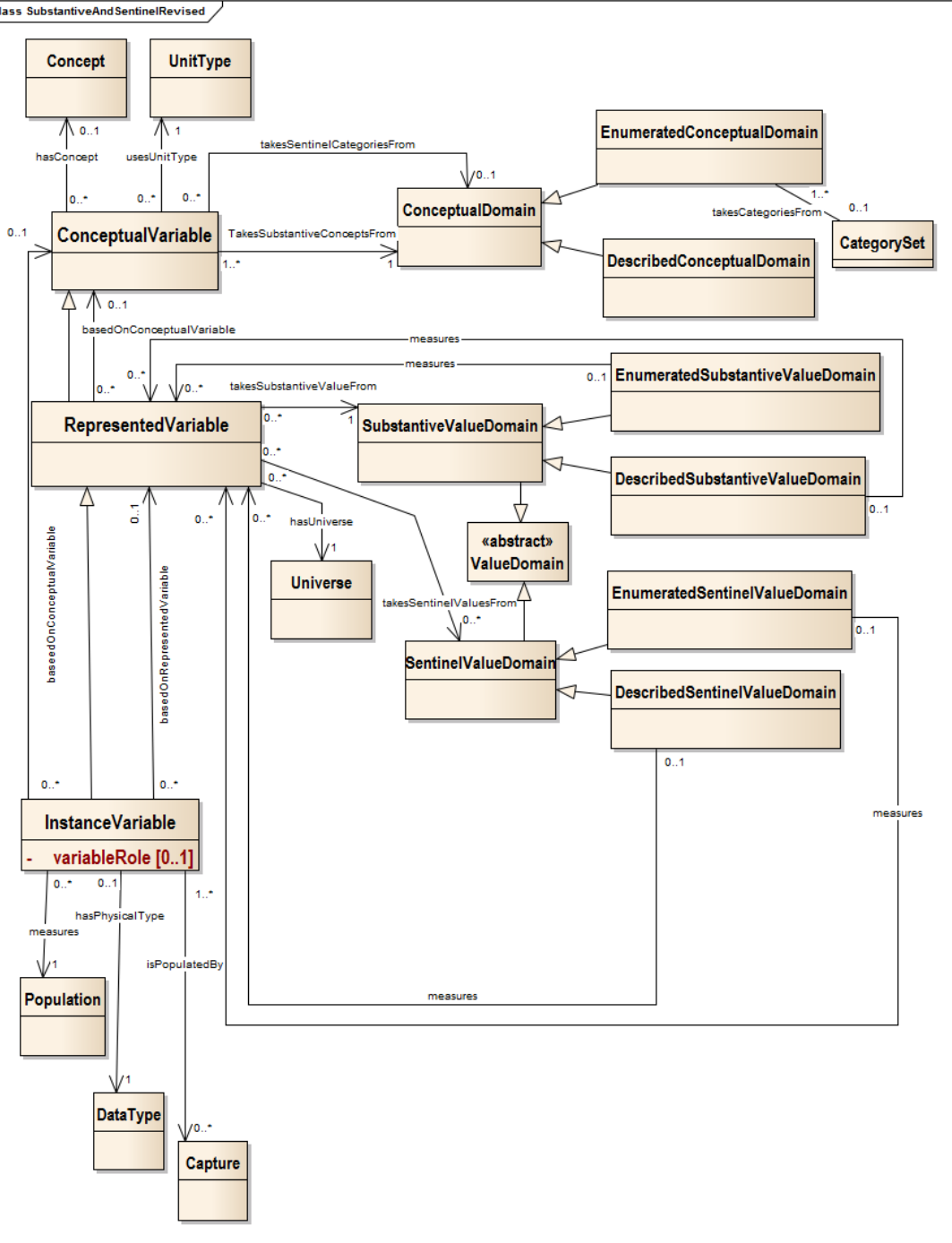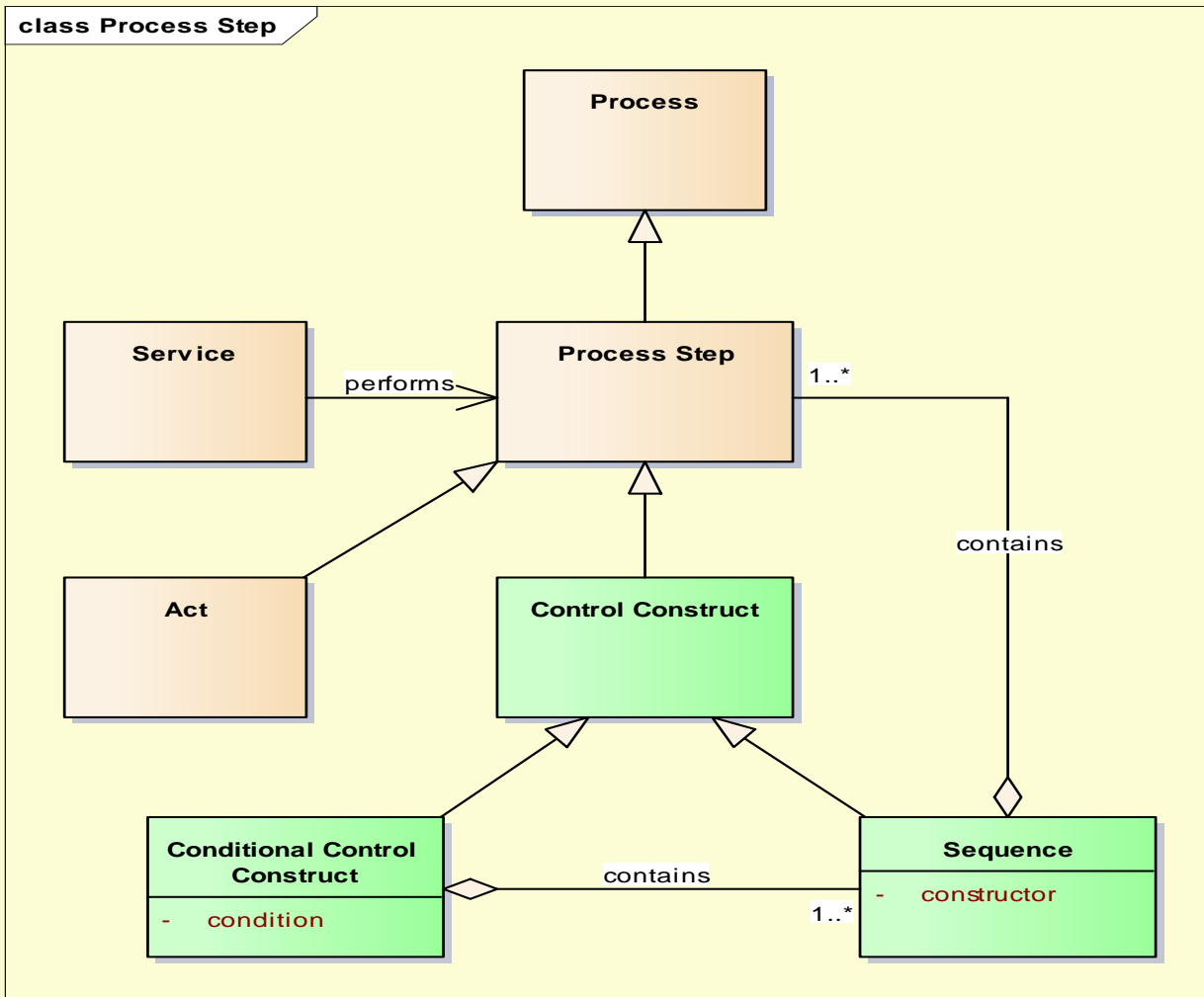
Collections and Relations Pattern
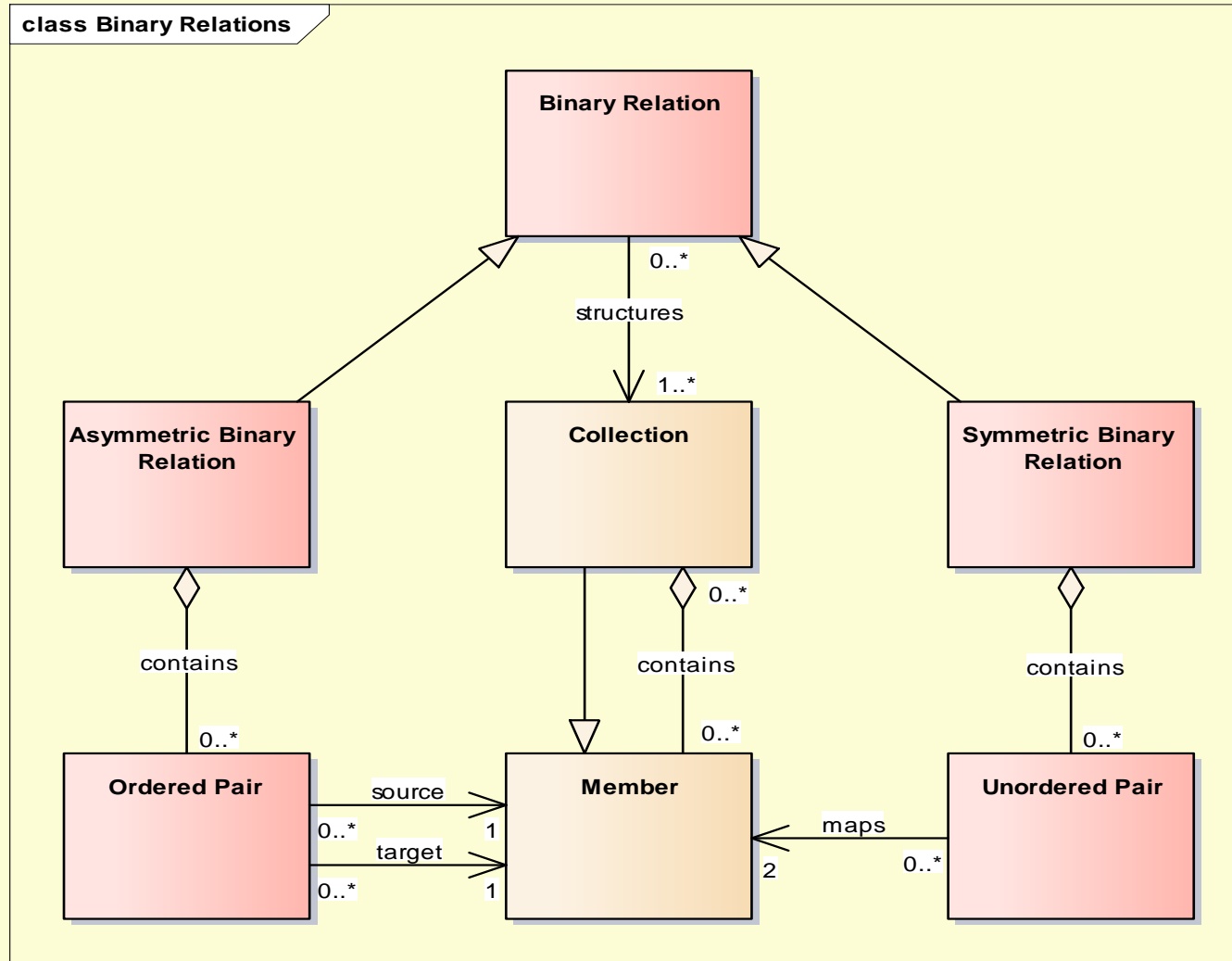
A Collection is a set or bag of things (Concepts, Categories, Variables, Process Steps, Qualitative Resources, Segments …)

Various types of relations can be described on the members of the Collection – equivalence, ordering (including hierarchy), networks
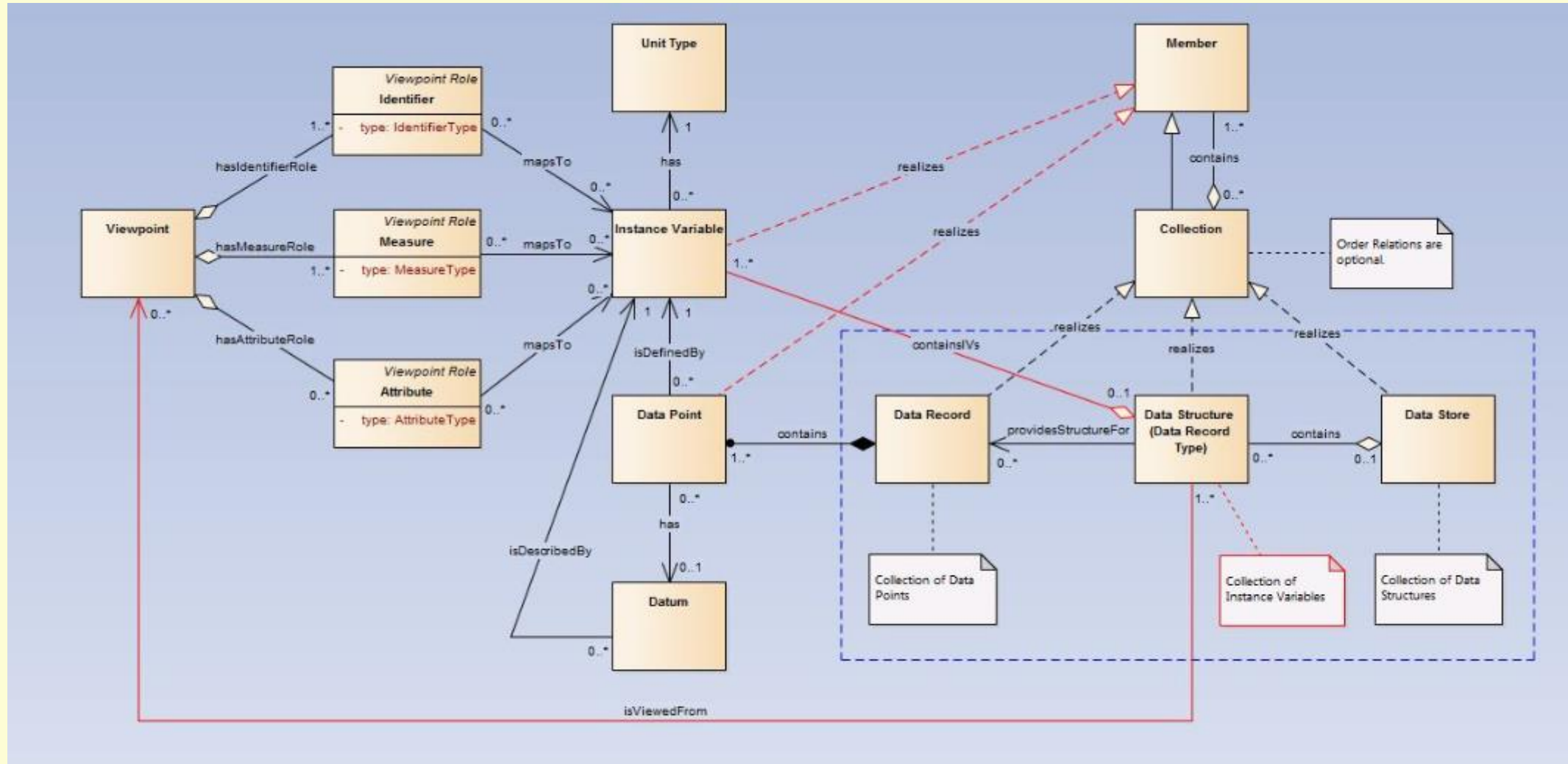
From: Rizzolo, Flavio. *Introduction to patterns in DDI-Views*

Process
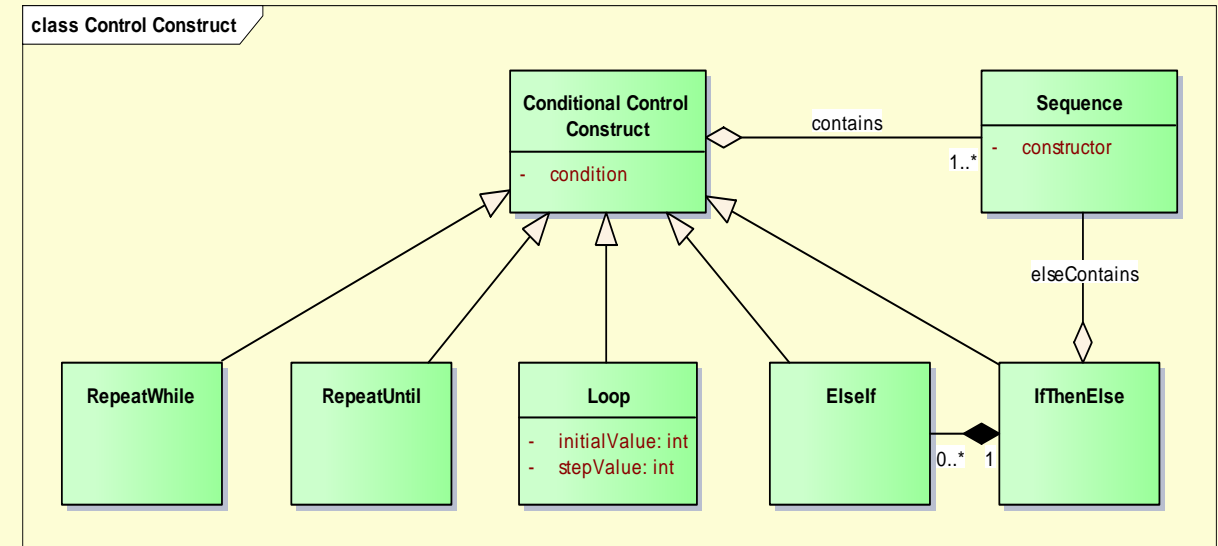
From: Rizzolo, Flavio. *Introduction to patterns in DDI-Views*
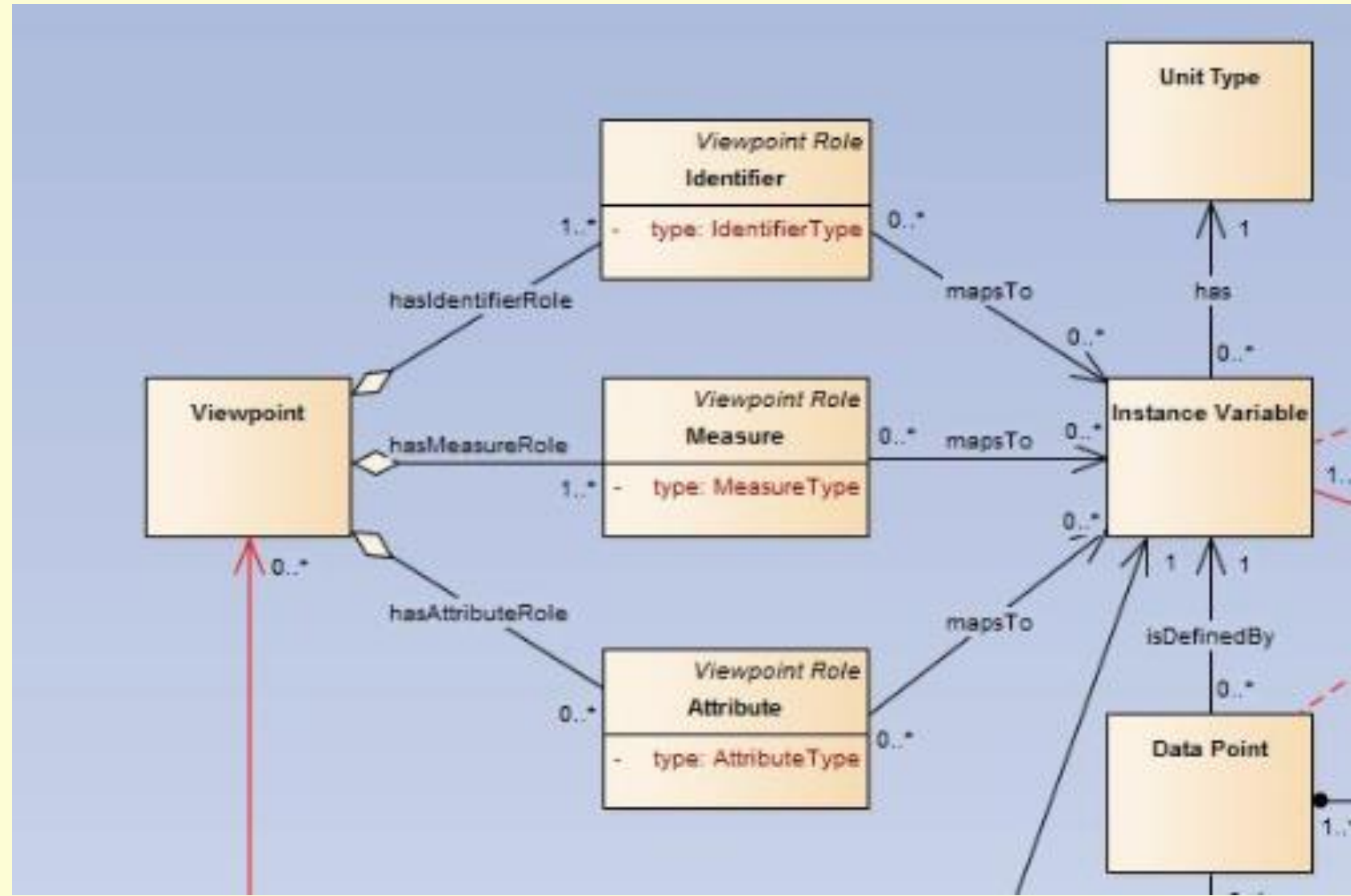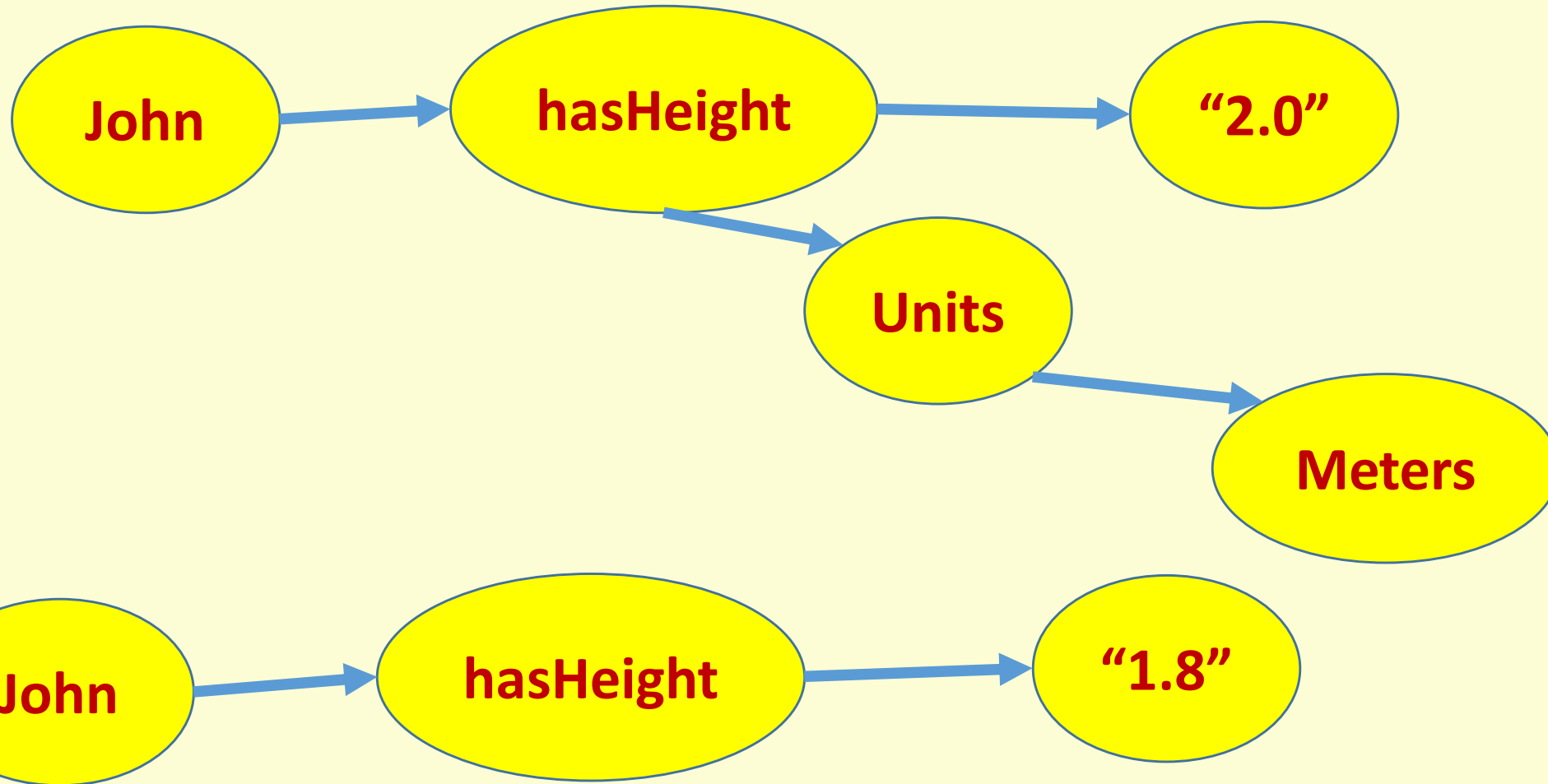
Datum

ViewPoint

Variables can take on different roles in different contexts (Viewpoint)
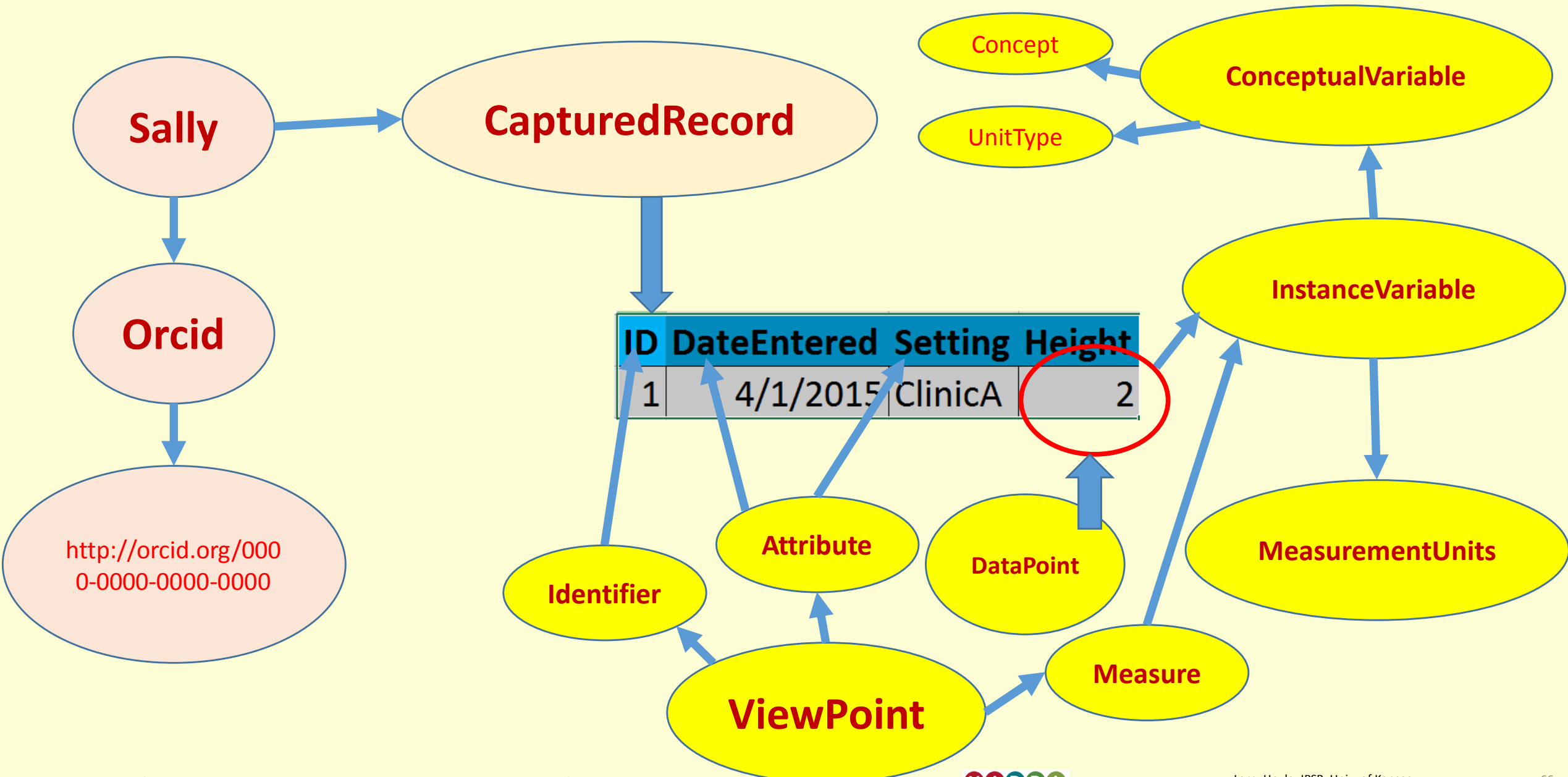
e.g. RAIRD project
http://www.raird.no/

Semantic Web – "Facts"

Links an entity to a value through a concept

Not "reproduced"

DDI4 View Describes Data in Context

From Notes to Custom Metadata
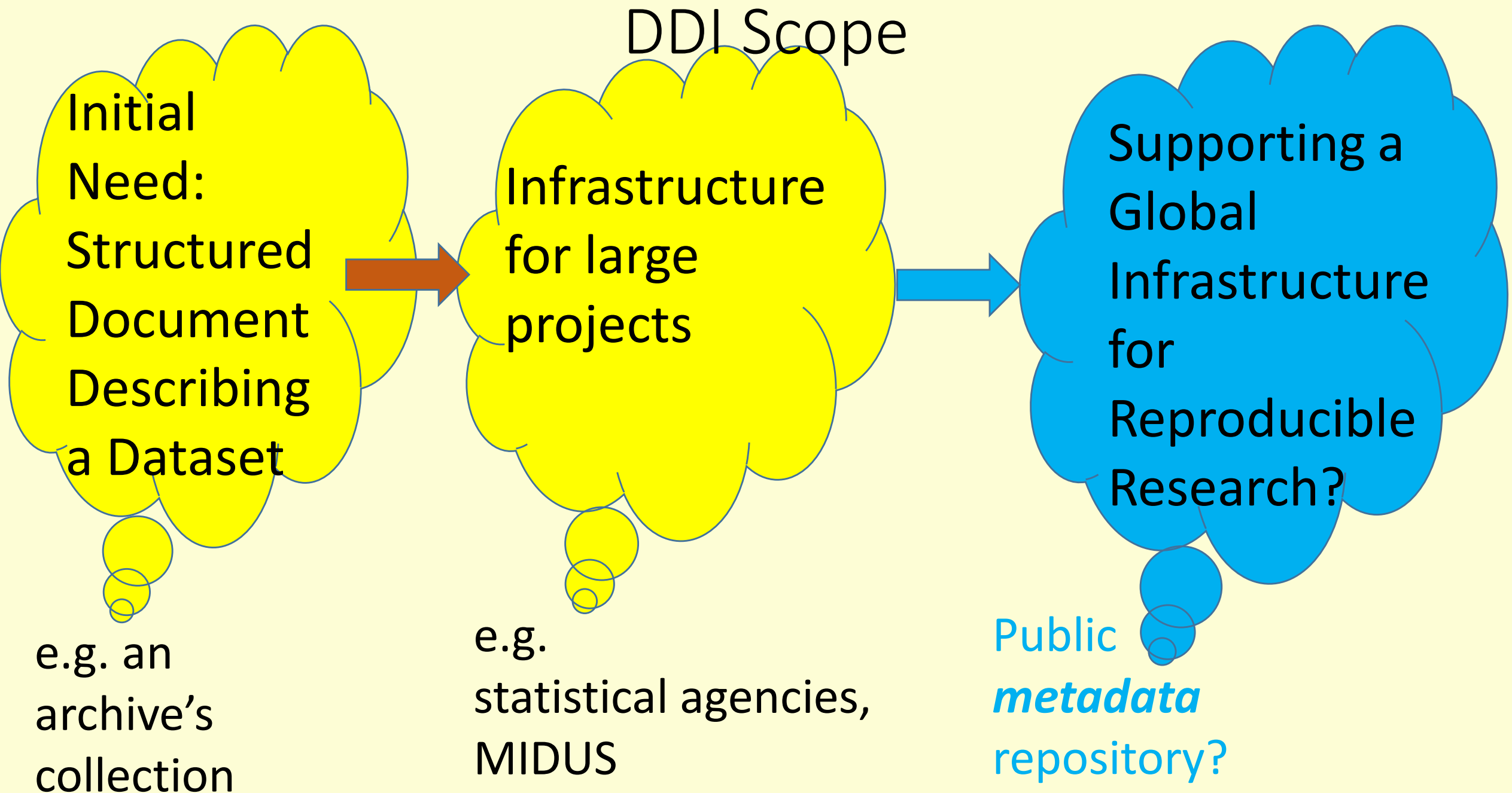
DDI 2 - Notes everywhere as attributes

DDI 3.0. 3.1 – A Note element with a Relationship

DDI 3.2 – UserAttributePair – key, value pair, external vocabulary

DDI 4 – CustomMetadata – key, value pair, structure/vocabulary in DDI

Final Thoughts

- A 20 + year effort, still going strong

- Founded in experience of archives

- Evolving conceptual model

- Moving to meet a changing environment
  - Multi-discipline studies
  - Increased data management emphasis
  - Reproducible research movement
  - Semantic web

- Metadata as a research product

DDI Scope

Initial Need: Structured Document Describing a Dataset

e.g. an archive's collection

Infrastructure for large projects

e.g. statistical agencies, MIDUS

Supporting a Global Infrastructure for Reproducible Research?

Public *metadata* repository?

Some References

- DDI Alliance. *History of the Standard* http://www.ddialliance.org/what/history.html accessed 2016-01-15

- Green, Ann and Chuck Humphrey. *Building the DDI* IASSIST Quarterly, Volume 37 – Number 1- 4 – 2013. pp. 36-44

- Rasmussen, Karsten Boye. *Social Science Metadata and the Foundations of the DDI* IASSIST Quarterly, Volume 37 – Number 1- 4 – 2013. pp. 28-35

- Vardigan, Mary. *DDI Timeline* IASSIST Quarterly, Volume 37 – Number 1- 4 – 2013. pp. 51-55

Contact

Larry Hoyle

Senior Scientist

Institute for Policy & Social Research, University of Kansas

http://orcid.org/0000-0002-8262-2393

LarryHoyle@ku.edu

1541 Lilac Lane Suite 607 Blake

Lawrence, KS 66045-3129

38.9562, -95.24333