GSK Experience Designing a SHARE-like Metadata Repository

PhUSE, Philadelphia

August 11, 2011

Diane Wold, PhD

Outline

- GSK's current standards and the need for change
- Current GSK work and the relationship to CDISC SHARE
- Information Model Foundational Standards
- Making the Information Model Real
- SHARE content versus GSK content
- Advice on gaining maximal benefits from SHARE
- Additional information on BRIDG and ISO21090 [not covered in presentation]

GSK'S CURRENT STANDARDS AND THE NEED FOR CHANGE

GSK's current standards

- Based on standard **datasets**
  - Include terminology, rules and structure
  - Originally based on SDS V2, more recently influenced by SDTM
- Standard datasets have corresponding standard CRFs, eCRFs, data extraction programs, statistical displays, algorithms etc.
- Standards managed in Lotus Notes databases. bespoke Dataset Manager tool and study specification tools, InForm libraries, etc.
  - Documentation in multiple systems
- Standards at global, therapeutic and indication levels
  - Exemption/ change request processes for all levels
  - Global standards managed by central group
  - Therapeutic area and indication standards managed by committees

Limitations of Current Standards

- Ambiguity (what is this?; how am I meant to use this?)
  - Documentation scattered and/or insufficient
- Distributed management of standards has led to variable and dataset libraries including duplicate and near-duplicates
- Datasets structures inconsistent
  - Original basis was SDS V2.1, which didn't address non-safety domains
  - Distributed responsibility for therapeutic area and indication standards
- Variable dataset structures limit opportunity for automation
- Multiple approaches to therapeutic area and indication standards make aggregation of data difficult
- Standard datasets do not always meet the different needs for data management and analysis
- Standards tightly coupled to tools, so tool replacement is painful

Difficulties Related to SDTM

- Mapping GSK standards to SDTM
  - GSK variables which don't map to SDTM variables
  - Different uses of GSK variables (subtle and not-so-subtle) require study-specific mappings
- Need to deliver SDTM, but don't want to build our standards on SDTM
  - SDTM designed for regulatory submission, not ideal for other purposes
  - SDTM has known limitations, is certain to change
  - Regulators may require some other format in future

Drivers for change

- Need to transform data between standards
  - FDA requirements
    - CDISC now, perhaps HL7 in future
  - More work with development partners
- More work with fewer resources
- Study-level flexibility vs. standardisation
  - Not a new problem; need better solutions
- GSK program to simplify clinical computing environment
  - Opportunity to change standards **now**

Anticipated Benefits

Regulatory/legal/public mandate:

- Ability to provide required data in required format to regulators and others
- Ability to respond quickly to regulatory queries

Operational efficiency:

- Systems share metadata to increase operational efficiency
- More “atomic” standards provide greater flexibility to study teams
- Clearer definitions and more consistent structure reduce ambiguity

Data Reuse:

- Ability to aggregate and analyze data more easily

Traceability:

- Ability to trace from a result back to CRF values, including computational methods at each step

CURRENT GSK WORK AND THE RELATIONSHIP TO CDISC SHARE

GSK Approach to Data Standards

- Want to use CDISC SHARE content in future
- Replacing systems now
- Building a Metadata Repository using a GSK-developed Information Model
  - Aligning with SHARE as much as possible
    - Same industry standards as foundation
    - Active in SHARE teams
  - Plan to eventually replace clinical data definitions with SHARE content

Metadata Repository Context

- Metadata Repository Content
  - Clinical data definitions
  - GSK global, therapeutic area, and indication usage decisions
  - Operational metadata (information needed to create eCRFs, datasets, SDTM datasets etc)
- Metadata driven approach to developing, executing and reporting clinical trials
  - eProtocol tool
  - Metadata Repository
  - Many systems consuming the metadata -- eCRF tool, reporting tools …

Information Model Foundational Standards

Industry Standards

- The BRIDG model (a collaboration between CDISC, HL7, FDA and the US National Cancer Institute (NCI))
- ISO21090 datatype standard (applicable across healthcare, not just regulated clinical research) … very similar to the HL7 abstract datatypes
- ISO11179 metadata registry standard

The BRIDG Model

- A standard way of representing the world of clinical research
- A subset of the model is particularly relevant for modeling the assessments and interventions on which the great majority of clinical data are based
- The model gives patterns that allow us to represent these things in a consistent way
- BRIDG modelling is independent of representation of data in any particular database structure

ISO 21090 Datatype Standard

- Many more datatypes than used in familiar database systems
- Each dataytpe consists of multiple related “components” each with a datatype of its own
  - For example, there is a datatype for addresses, which can express all the components of a multi-line postal address
  - Allows handling of related data elements as a collection
- Many datatype components are solutions for dealing with “awkward” data
  - Varieties of missing-ness, such as Other, Unknown, and Not Applicable
  - Entries that don’t match the desired format, such as text or ranges when a number is desired
  - Code, decode and score for the responses to a questionnaire response
  - Whether a range (such as a normal range) does or does not include the endpoints
- BRIDG uses ISO 21090 datatypes

ISO11179 Metadata Registry standard

- A sort of “Good Metadata Registry Practice” guide
- For most users, this will be background which doesn’t directly affect them

Sources of Information

- BRIDG site: http://www.bridgmodel.org/ (we are using 3.0.3)
- ISO21090 standard: http://gforge.hl7.org/svn/hl7v3/trunk/dt/iso/index.htm (logon with username= anonymous and blank password) … the 2011 published version is on the ISO website
- Enterprise Architect is the modelling software used by BRIDG. Here is a link to a free viewer: http://www.sparxsystems.com/bin/EALite.exe
- Additional information on BRIDG and ISO21090 is provided in 15 easy slides appended to this presentation.

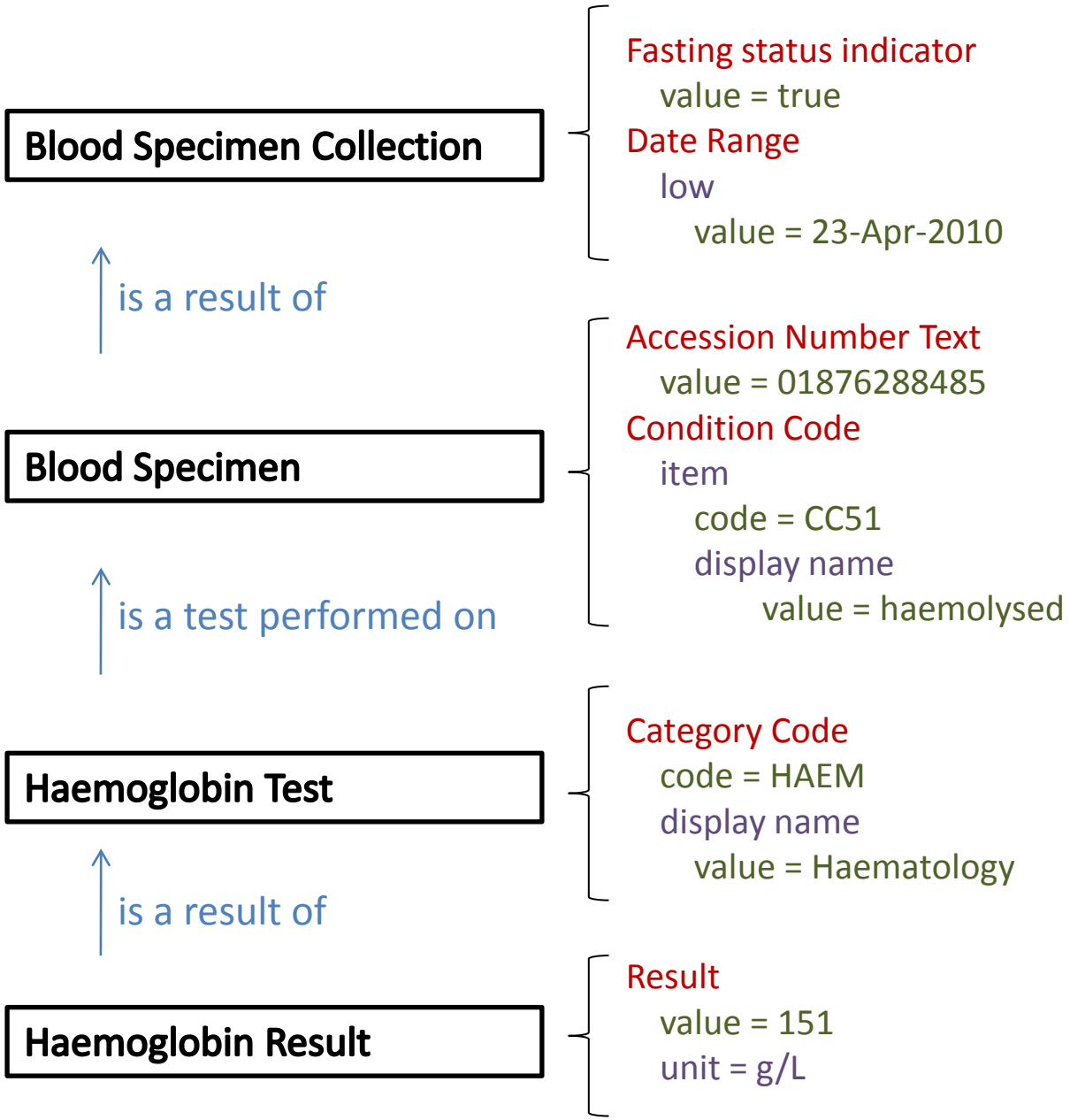
Making the Information Model Real

Information Model Benefits

- A well developed model of clinical research … there shouldn't be anything missing
- Provides a set of "templates" to use in the development of our standards
  - Data items (variables) are selected from a short list of possibilities

Laboratory Example

BRIDG based modelling of the clinical data includes clinical concepts (boxes) and associations between them (blue arrows)



Each of the concepts has attributes (red text) and each attribute has a datatype with components and sub-components

Most granular level of datatype components corresponds to variables

Layers of Metadata

- Concepts – clear definitions of clinical information (e.g. height, systolic blood pressure, weight result)
- Associations – how the concepts connect together
- BRIDG attributes – meaningful attributes for a piece of clinical data (e.g. method, date, anatomic site, result); some may have codelists
- ISO21090 decomposition: components of the complex datatype; provide grouping of related variables; some may have codelists
- Variables – clear, model based, unambiguous variables

Process for creating a model-based scientific concept

- Choose relevant clinical scenario template
  - in this case, one containing specimen, lab test & lab result
- Enter information about each new concept needed
  - name, description, definition …
- Choose which BRIDG attributes are needed
- Choose which bits of the ISO21090 decomposition are needed
- Enter the names of codelists needed
  - select from the codelist the codes needed for this concept
- Choose which associations are needed

SHARE CONTENT VERSUS GSK CONTENT

What we expect from SHARE

- Model based well-defined concepts
  - Including all the layers of metadata and controlled terminology (codelists)
- Information needed to represent these concepts in the CDASH and SDTM
- Facilities to download SHARE metadata
- Facilities to suggest or request additional concepts

What we don’t expect from SHARE

- The usage rules that GSK will want to apply
- Operational metadata needed to create study objects (GSK datasets, GSK eCRFs)

- In other words, GSK expect to **add** additional metadata to the GSK repository … we want to **augment** the SHARE content, not change it

GSK Usage Rules
Added to SHARE content

- Which variables are mandatory or optional to **include** in a study, or are conditional on the inclusion of other variables
  - the nature of conditionality, e.g. include variables for total daily dose and dose units or for dose, dose unit and frequency
- Which variables are mandatory or optional to **populate** in a study, or whose population is conditional on the population of other variables
  - e.g., if sex = F, must populate child bearing potential
- Based on associations
  - e.g., a WBC test requires a blood sample

Subject Disposition Example

Was the subject withdrawn from the study?

[Y] ☐ Yes [N] ☐ No

*If Yes, ✓ one **primary** reason for withdrawal:*

Tick this …

[1] ☐ Adverse event → *Record details on the Non-Serious Adverse Events or Serious Adverse Event pages as appropriate.*

(When the reason above is ✓, select any sub-reasons which may apply from the list below).

… and you MAY tick none, one or many of these

[?] ☐ Predefined sub-reason

Tick this …

[7] ☐ Investigator discretion *(Only ✓ if none of the other primary reasons are appropriate)*

… and you MUST enter text here

Specify: _

If the study includes pre-specified subreasons, an "other specify" subreason MUST be included and, if ticked, MUST be populated

[8] ☐ Withdrew consent *(Only ✓ if none of the other primary reasons are appropriate)* The investigator is required to provide a reason for withdrew consent either through a specify or a pre-hardcoded option.

(When the reason above is ✓, select AT LEAST ONE sub-reason from the list below).

[?] ☐ Predefined sub-reason

[OT] ☐ Other, specify: _ must be used if there **are** predefined sub-reasons used

If the study does not include subreasons, the "specify" MUST be included and populated

Specify: _ must be used if there are **no** predefined sub-reasons used

We should not expect SHARE to deliver these company specific rules

Operational metadata

- Mappings from other standards to concepts & concept variables
  - legacy data
  - development partners
- Mappings between SHARE terminology to GSK terminology
  - will use SHARE terminology as much as possible
  - there are certain to be occasions when we need to deviate
- Mappings from concepts and concept variables to GSK operational objects (e.g., eCRFs, datasets for data cleaning)

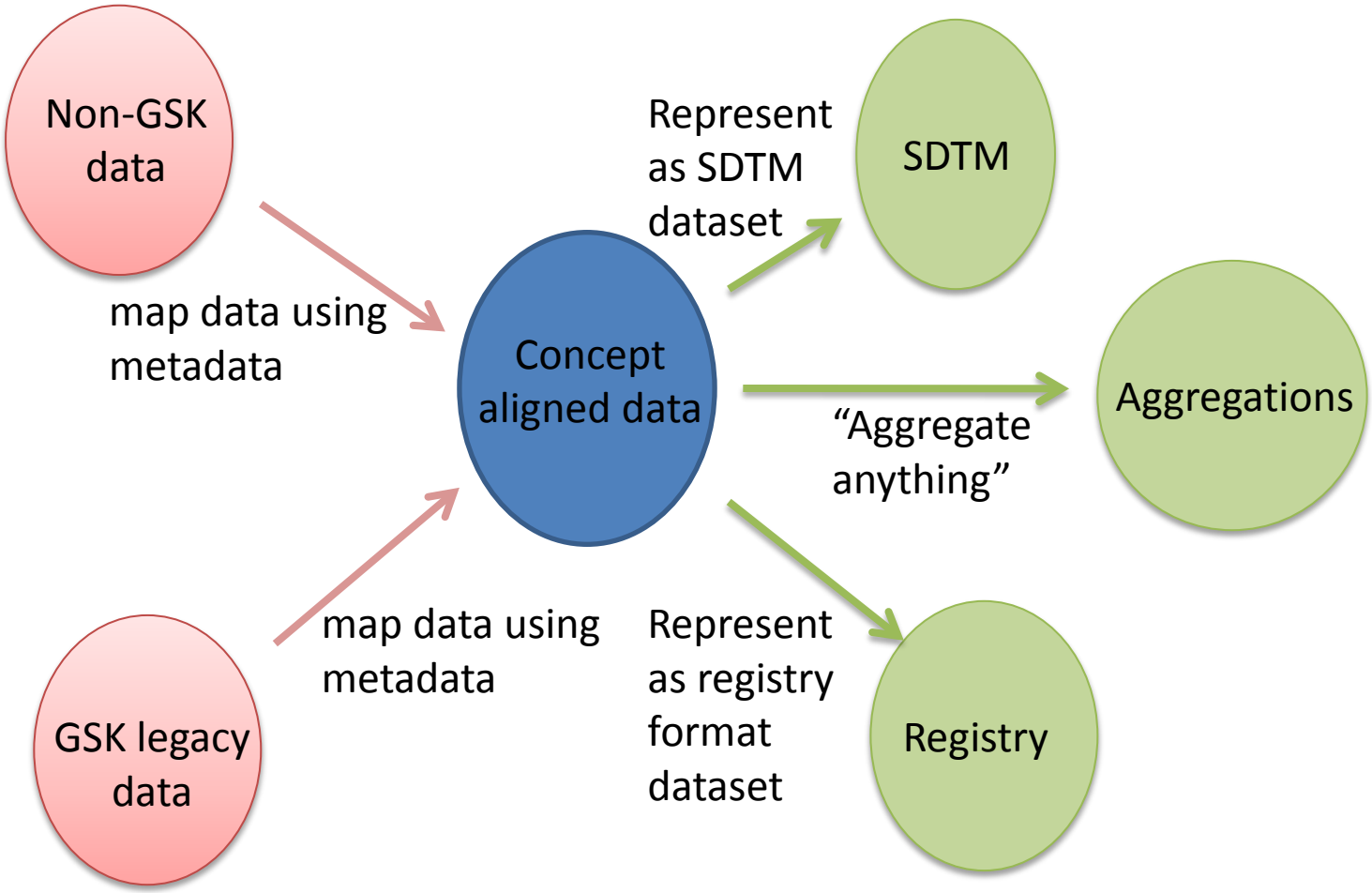
GSK Operational Metadata Added to SHARE

- Metadata needed to render the definitions in particular forms
  - e.g. an eCRF
    - length and precision for variables
    - whether a coded field should be represented as a drop down box or a radio button
    - dependencies between data items
  - e.g, a GSK dataset
- Detailed study specifications

SDTM

- Should get SDTM “for free”
  - concepts are associated with SDTM domains
  - concept variables are generated from BRIDG and ISO21090
  - BRIDG attributes/ISO decomposition mapped to SDTM variables
- Should get consistent SDTM, eliminating “wiggle-room”
- Will eventually get mapping to SDTM from SHARE

Metadata-driven Data Transformations



Setting Up Studies: Detailed Study Specifications

- Use BRIDG-based concepts and relationships
  - key to taking full advantage of the concept metadata
- Build using GSK and therapeutic/indication usage rules
- Include details for each visit/timepoint
  - which variables will be collected
  - for codelisted variables, which set of codes are available

Setting Up Studies: Benefits of Detailed Study Spec

- Integrates study-specific study design concepts (e.g, arms, epochs, visits, timepoints) and study-independent data collection concepts
- BRIDG modelling provides metadata/data driven navigation capability, guiding study investigators through even very complex study procedures
- The richness of the metadata in the study specification can be used to help creation of operational objects

ADVICE ON GAINING MAXIMAL BENEFITS FROM SHARE

Choices for SHARE users

- Minimal use: Just use the SHARE variables and forget about the rest of the metadata
  - you get consistent industry standard variables
  - you can keep your own processes
  - variables used out of context may not be used as intended, so your data may not be consistent with others' data
  - metadata layers and relationships not available for use for automation and other purposes
- Maximal use: Use all SHARE metadata and augment with additional company metadata
  - all SHARE metadata available for use in automation
  - Company usage decisions integrated with metadata to ensure consistent data collection across studies

Recommendations

- Always maintain a link from operational objects back to the SHARE definitions
- Use the SHARE objects beginning at study design
- Augment the SHARE metadata with company specific metadata, such as
  - usage rules (for consistent data collection)
  - additional operational metadata (may be somewhat tool dependent)
- Use all this metadata to create detailed study specifications

Anticipated Benefits

Regulatory/legal/public mandate:

- Ability to provide required data in required format to regulators and others
- Ability to respond quickly to regulatory queries

Operational efficiency:

- Systems share metadata to increase operational efficiency
- More “atomic” standards provide greater flexibility to study teams
- Clearer definitions and more consistent structure reduce ambiguity

Data Reuse:

- Ability to aggregate and analyze data more easily

Traceability:

- Ability to trace from a result back to CRF values, including computational methods at each step

Additional information on BRIDG and ISO21090

[not covered in presentation]

Two industry standards: BRIDG and ISO21090

A simple explanation of what these are and what they provide

Information Model?

- An information model is a combination of structure and nomenclature
  - modelling the structure of data
  - employing a set of terms to describe the objects
- A good information model will ensure that nothing is glossed over and that similar things will be described in a similar manner

GSK's rationale for using BRIDG and ISO21090

- Current GSK standards developed without underlying information model
  - right content (the info we need in GSK's clinical trials)
  - consistency of approach, avoidance of duplication and ambiguity are not as good as we would like
- In 2009, we started to develop an information model based approach to representing GSK's clinical trial standards
  - We originally intended to use BRIDG as a guide, rather than to implement it rigorously
  - We ran into various issues, and worked out solutions
  - By the end of the year, we realized that BRIDG could provide solutions to all our issues
  - In January 2010, we decided to implement BRIDG with the ISO datatype standard
    - addresses all our issues
    - better alignment with SHARE
    - better alignment with standards employed in the healthcare world

CDISC SHARE Project

- In the early days of the SHARE project, it was agreed that SHARE would use the BRIDG model, the ISO21090 datatype standard and the ISO11179 metadata registry standard as its information model
- Although SHARE could decide to implement these differently from GSK, currently the GSK and SHARE information models are very similar

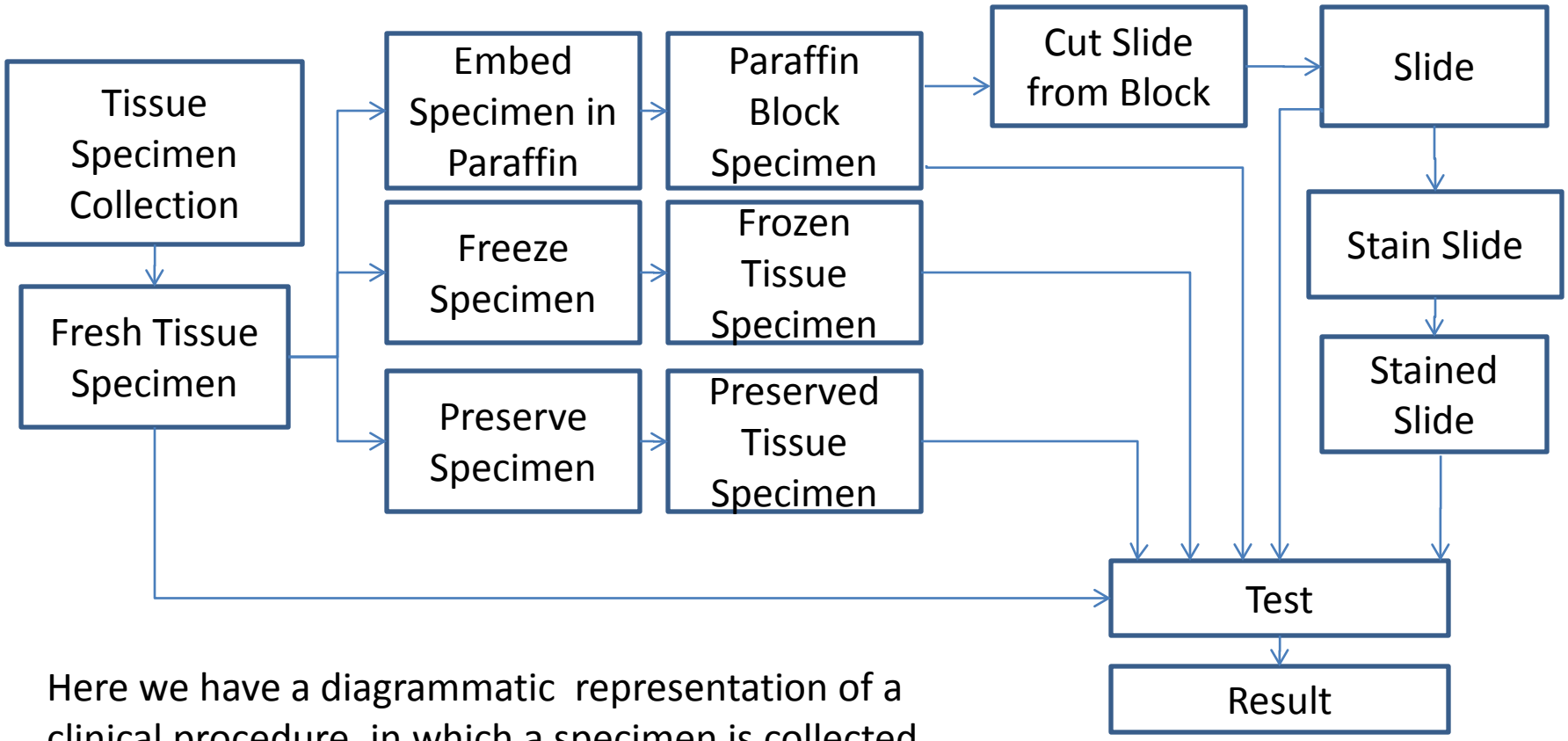
BRIDG

- An information model
- Targeted at protocol driven research
- Reasonably mature
- Key collaborators: CDISC, HL7, NCI, FDA

Key Features

- BRIDG is a model of protocol driven research
  - entities (animal, person, organisation, material)
  - activities (any action that can, in the context of a study, be planned, scheduled or performed e.g. a surgical procedure, a laboratory test, or the administration of a drug)
  - participation or functional role of an entity in an activity
  - relationships between activities (both simple and complex relationships

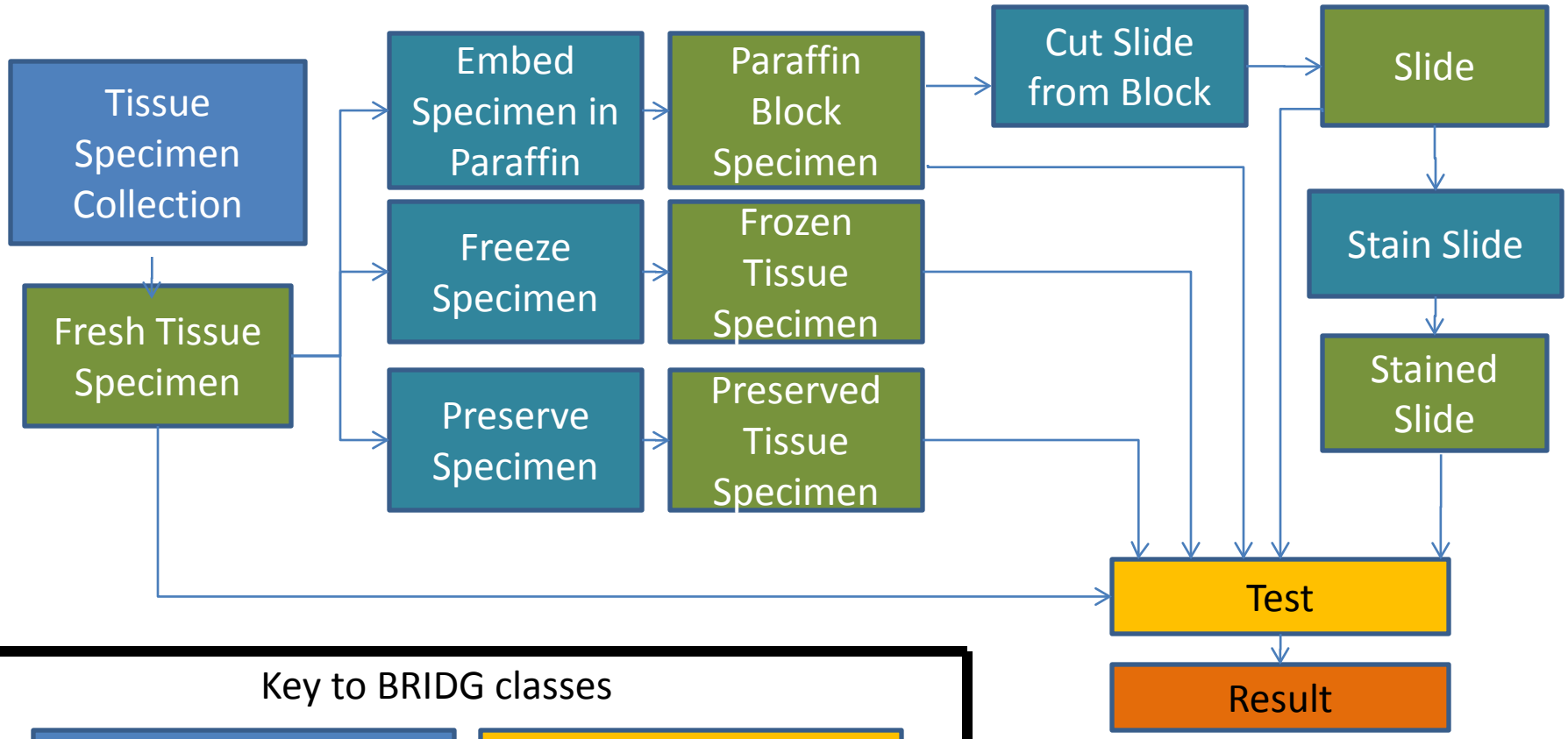
Example: Tissue Samples



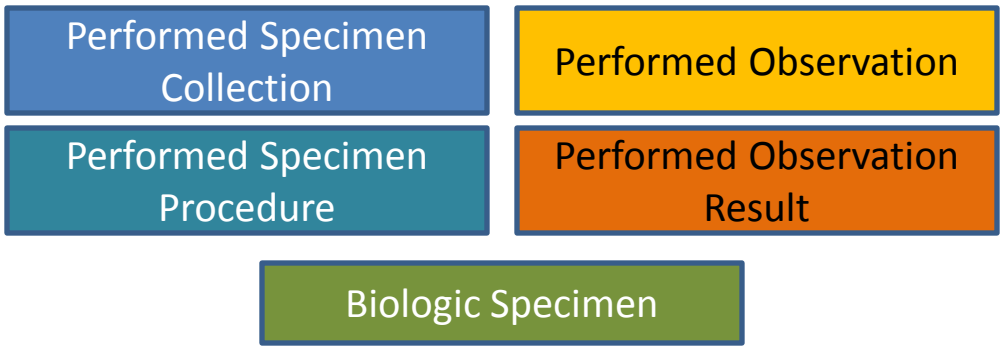
Here we have a diagrammatic representation of a clinical procedure, in which a specimen is collected from a subject, some processing of the specimen may occur, and then the specimen is tested and a result obtained

We may need data about some or all of the steps in this process, as well as about the test and its result

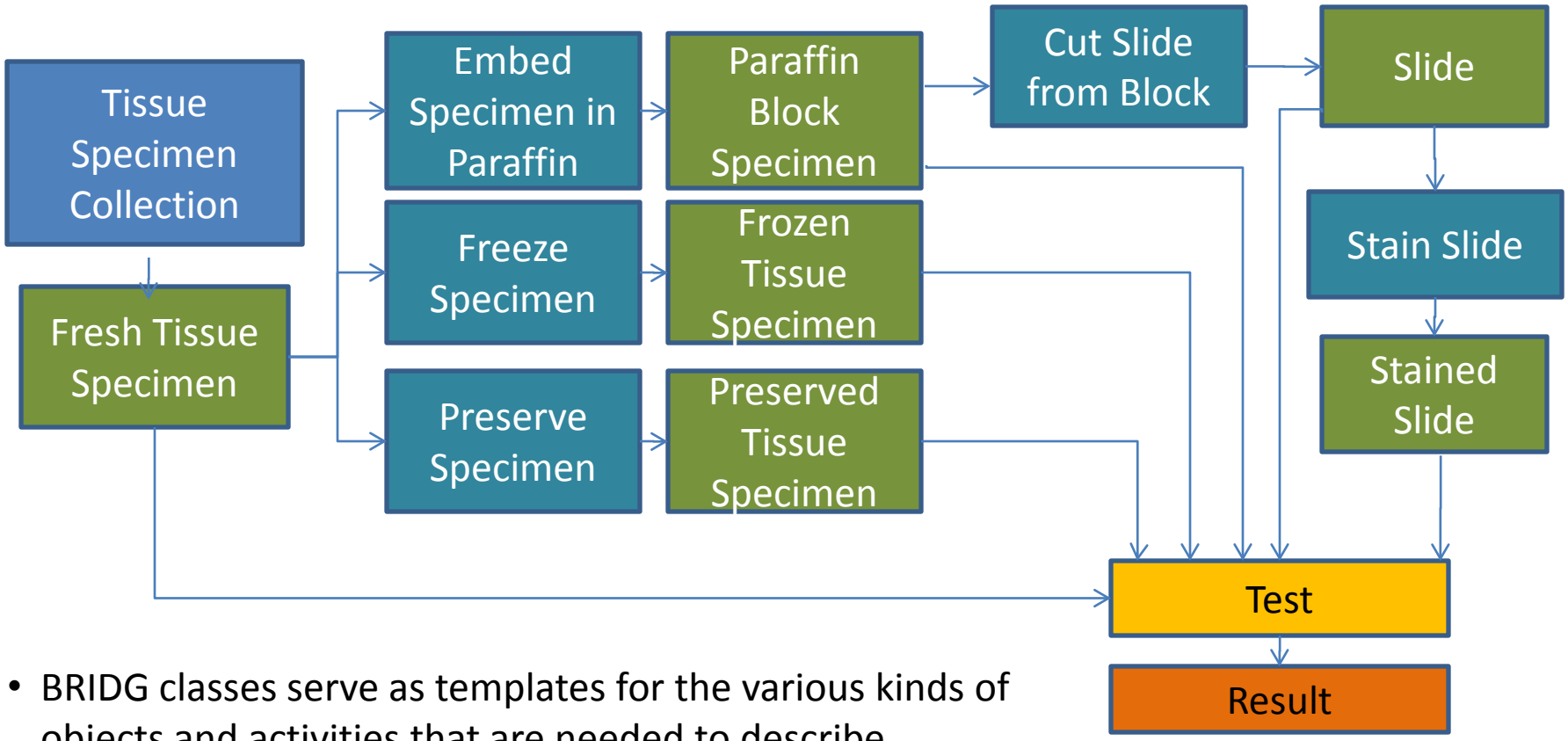
Same example, but indicating BRIDG classes



Key to BRIDG classes

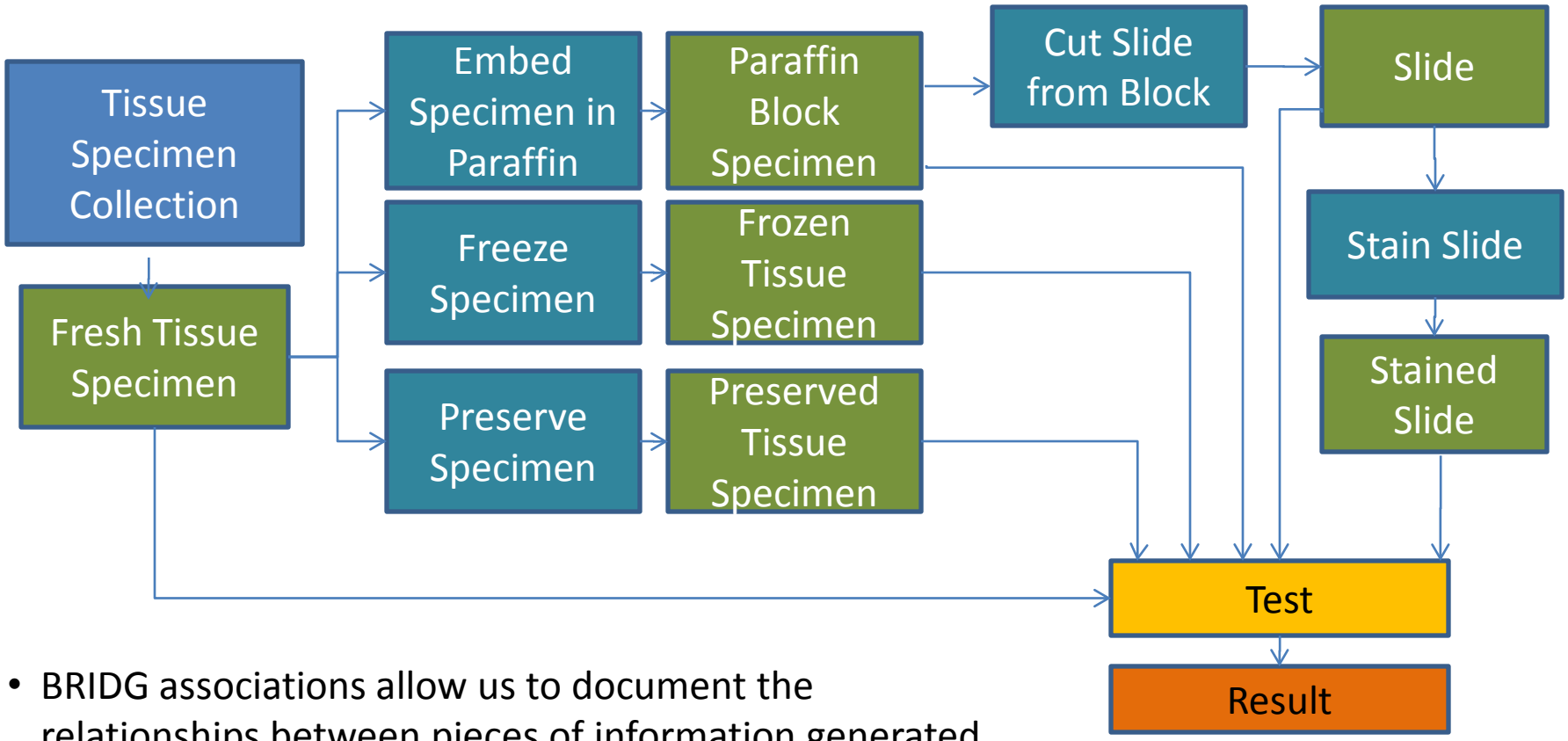


How BRIDG classes help



- BRIDG classes serve as templates for the various kinds of objects and activities that are needed to describe protocol driven research. We do not need to model each individual objects or activity from scratch.
- In effect, we have taken copies of BRIDG templates and make these specific to our clinical process

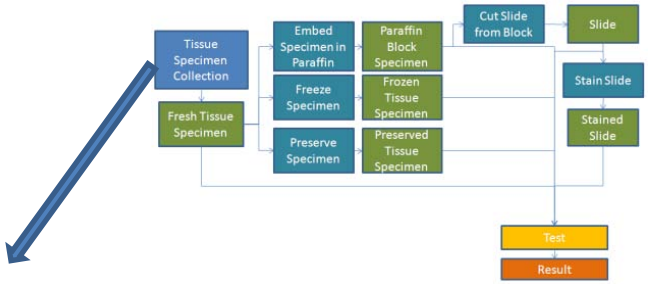
How BRIDG associations help



- BRIDG associations allow us to document the relationships between pieces of information generated through a clinical process. For example:
  - a specimen collection results in specimen(s)
  - some tests can have more than one result, but a result can only have one test

More on BRIDG classes

Here is one of those templates/classes:



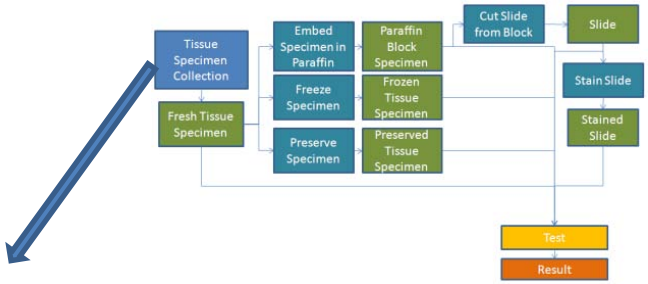
Tissue Specimen Collection:
- Collection Method
- Site from which the specimen was taken
- Date on which the specimen was taken
- .....

- Each BRIDG class has its own defined set of attributes (placeholders within the template)
- For example, the Specimen Collection class includes attributes for the collection method, the site from which the specimen was taken, the date on which the specimen was taken … plus another 14 possible attributes
- When we build a particular specimen collection concept, we choose the attributes we want to use
- Some attributes with be associated with a specific codelist
- Attributes should cover ALL the information we might want to record

What attributes give us

Here is one of those templates/classes:

Tissue Specimen Collection:
- Collection Method
- Site from which the specimen was taken
- Date on which the specimen was taken
- .....



- Every time you copy a particular template, and make it specific to your situation, you choose from the same set of attributes
- For some attributes, all copies of the template will use the same codelist, e.g. the anatomic location attribute always uses an anatomic location codelist, though different copies may use different subsets of the codelist
- This "copy, choose and make specific" process makes it easier for both computers and humans, and enforces a consistent approach

Attribute datatypes

Here is one of those templates/classes:

Tissue Specimen Collection:
Collection Method CD
Site from which the specimen was taken CD
Date on which the specimen was taken IVL<TS.DATETIME>
.....

- For each attribute, we need to know more:
  Is Collection Method text? Or is it coded? Is it in English?
- Datatypes answer these questions
- BRIDG uses an ISO standard (ISO21090) for the datatypes
- These datatypes are complex – not like SAS datatypes (character, numeric, date etc)
- Examples are shown in black in the diagram: CD and IVL<TS.DATETIME>
- Each datatype has a several components … it is through these components that we get down to the variable level

Datatype components

Here are components of one of these datatypes:

Selected attributes of the CD datatype:	
nullFlavor : NullFlavor,	<used if there is no code>
code : characterstring,	CODE
codeSystem : characterstring,	
codeSystemName : characterstring,	
codeSystemVersion : characterstring,	
valueSet : characterstring,	CODELIST
valueSetVersion : characterstring,	
displayName : ST,	DECODE
originalText : ED	<original text>

The CD datatype is for coded information (though CD stands for "Concept Descriptor")

You can see that it has all the attributes you need for coded data

You can also see that some of the datatype attributes are themselves datatypes (e.g. originalText is of datatype ED). So components of a datatype can have components too … we have to "decompose" all these levels to get down to what we know as variables.

Datatype components

Another example

And here is another of these datatypes:

Selected attributes of the PQ datatype:
originalText : ED.TEXT,
uncertainty : QTY,
uncertaintyType : UncertaintyType,
uncertainRange : IVL(QTY)
value : Decimal, VALUE
codingRationale : CodingRationale,
unit : characterstring, UNIT

The PQ datatype is for physical quantities (things like sodium concentration, systolic blood pressure, number of lesions)
You can see that it has the attributes you need

Why complex datatypes rather than nice simple ones?

- Because we get extra benefit!
- The datatype we use for a physical quantity (e.g. the sodium concentration in the blood) has several component parts including Value and Unit
- When we use a sodium concentration result e.g. in a SAS dataset, we always keep Value and Unit together as the value is meaningless without a unit
- Use of these ISO datatypes gives us the facility to keep these sort of things together