DDI and GSIM – Impacts, Context, and Future Possibilities

Arofan Gregory
Metadata Technology

Overview

- The general situation for GSIM – DDI
- Implementing GSIM with DDI
- Detailed view of some GSIM areas and overlap/gaps with DDI
  - Describing data
  - Describing questionnaires
  - Describing codelists, categories, and concepts
  - Describing events and processing
- Looking forward

GSIM and DDI

- GSIM is a creation of the HLG-BAS group under UN/ECE
- DDI is a creation of the DDI Alliance
- There is no immediate formal relationship between these organizations
- However, both organizations have made statements that they will work together to make DDI a good implementation vehicle for GSIM

GSIM, DDI, and Official Statistics

- GSIM is a key standard for official statistics organizations
- Some official statistical organizations already use DDI or are planning to do so
  - IHSN Metadata tools (developing world)
  - DDI-Lifecycle (ABS, Stats NZ, INSEE, Eurostat)
- GSIM is a potential vehicle for the widespead adoption of DDI among official statistical organizations

Models at Different Levels

- GSIM is a *Conceptual Model*
  - It is technology and implementation-neutral
- DDI is an *Implementation Model*
  - It is cross-platform and application-neutral
  - It is implementated in XML (and soon, RDF), but isn't technology-neutral
- Specific applications have their own, internal models
  - These are bound to specific technologies and platforms

Implementing GSIM at a Technical Level

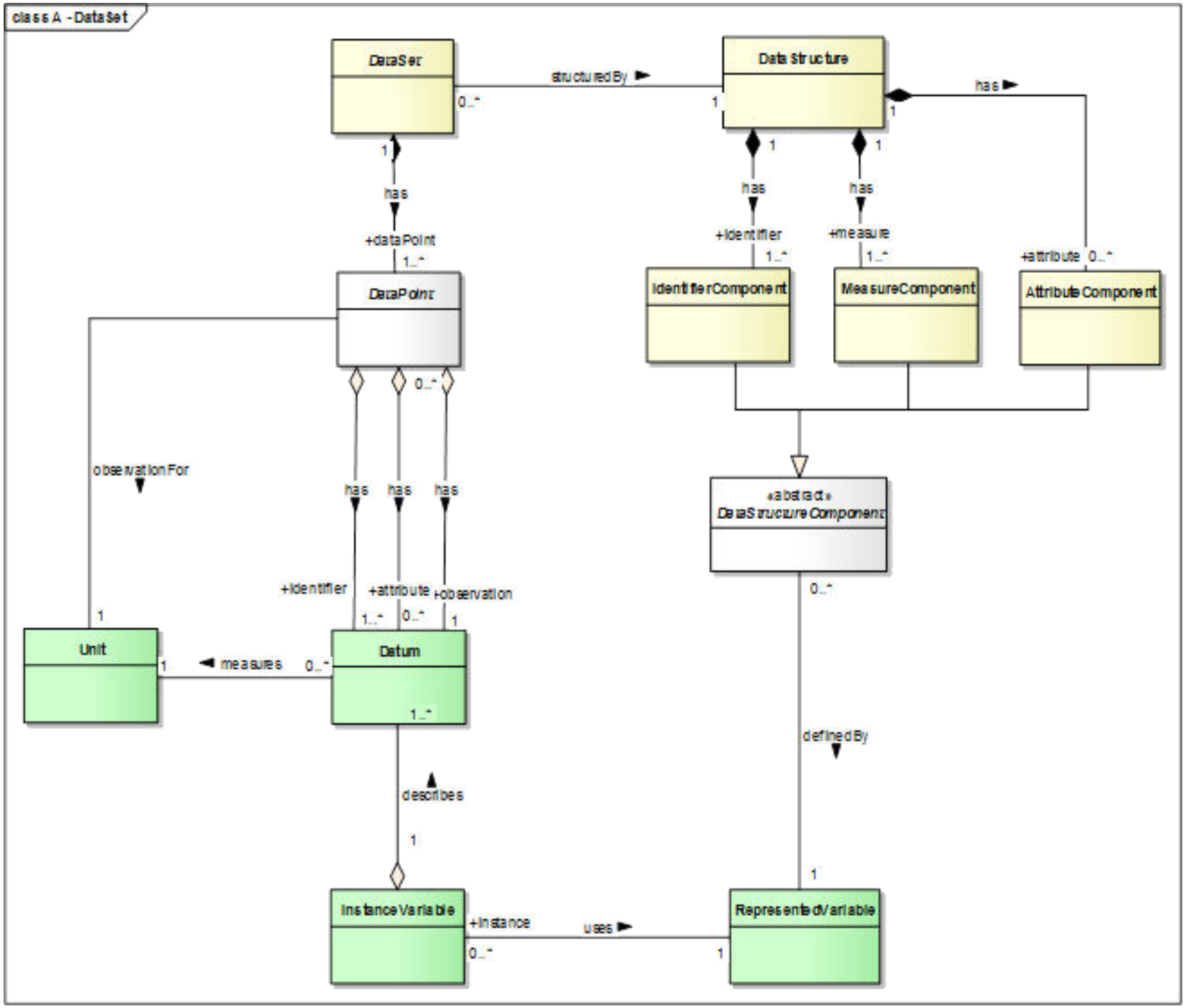
- To allow re-use of applications and services, agreements must exist on many levels
  - Conceptual models must match (GSIM)
  - Implementation models must match (DDI)
  - Application models must match (TBD – web services? Others?)
- There is still a lot of work around mapping DDI to GSIM, and then agreeing on how DDI XML will be used within applications before we have reusable, interoperable GSIM-based services and applications

What is the Usefulness of GSIM?

- To make applications work together on all levels, we will need to map existing application models to each other
  - On the basis of DDI
  - On the basis of GSIM
- From a technical perspective, this can be very difficult
  - Having an agreed base model at the conceptual and implementation level makes it easier/possible

Describing Data

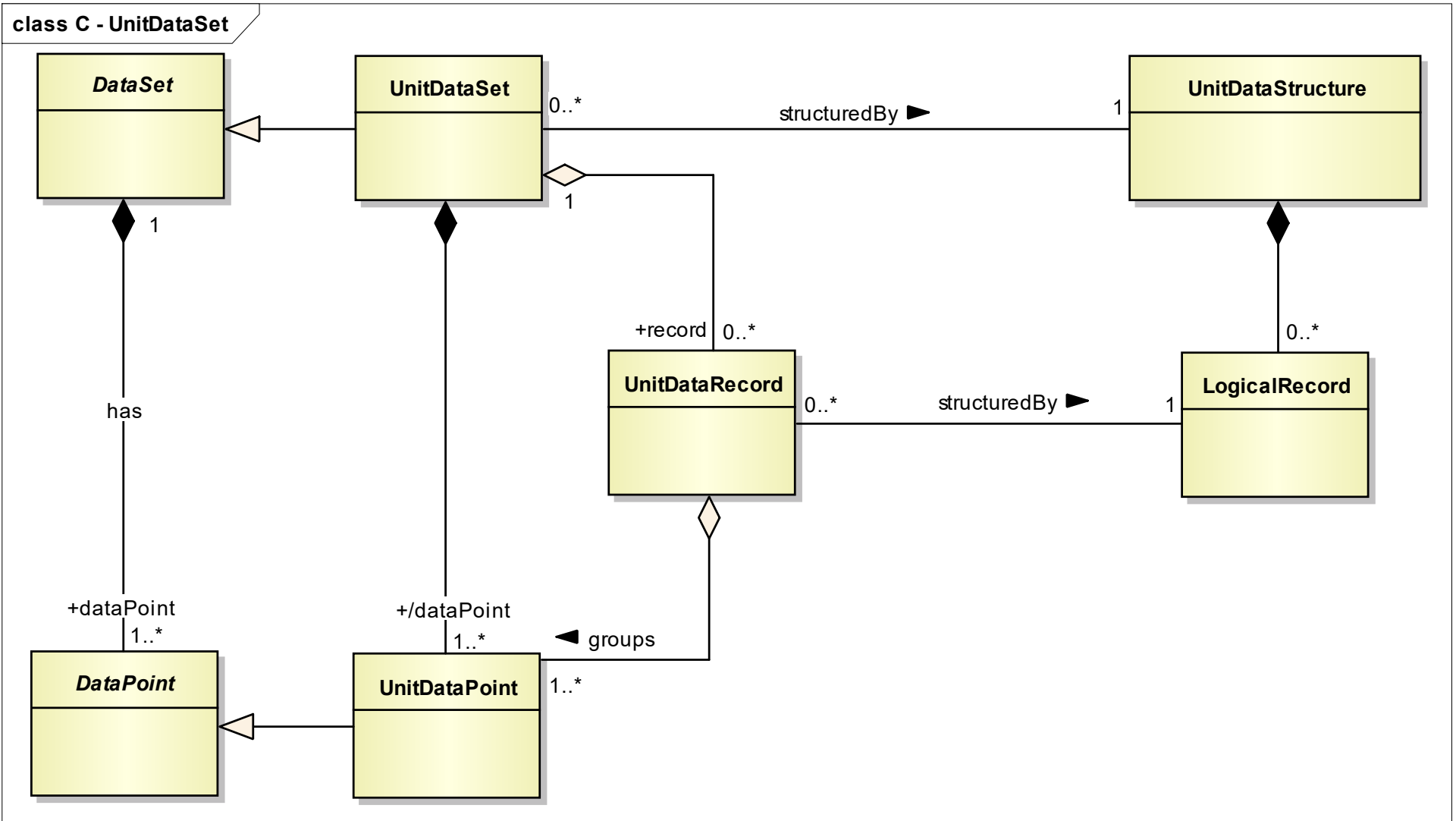
- DDI describes two kinds of data
  - Microdata sets
  - Aggregate (“dimensional”) data sets – Ncubes
- Both exist in GSIM
- In DDI microdata, each case/unit has a set of variables, at least one of which is the case identifier
  - Others hold observations or derived or supporting values (such as weights)
- Ncube structures use variables as dimensions, observations, and attributes to describe the matrix structure of tables


class A - DataSet
DataSet
structuredBy
DataStructure
0..*
1
1
1
has
1
1
has
has
+dataPoint
1..*
+identifier
1..*
+measure
1..*
+attribute
0..*
DataPoint
IdentifierComponent
MeasureComponent
AttributeComponent
0..*
observationFor
has
has
has
«abstract»
DataStructureComponent
+identifier
+attribute
+observation
1..*
0..*
1
0..*
Unit
1
measures
0..*
Datum
1..*
definedBy
describes
1
1
InstanceVariable
+instance
0..*
uses
1
RepresentedVariable

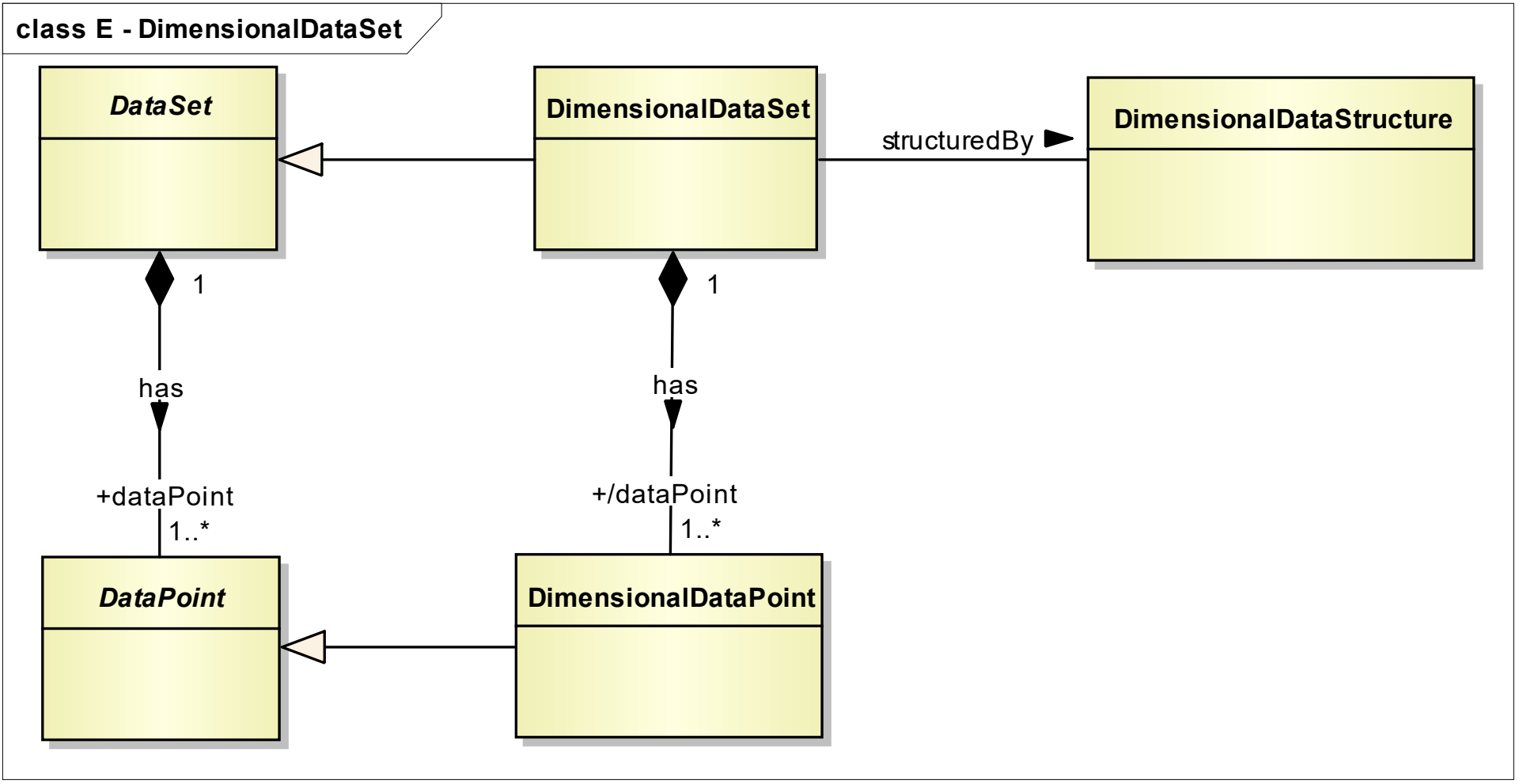
Other GSIM Constructs

- GSIM does make a distinction between “unit data structures” and “dimensional data structures”
  - GSIM supports hierarchical relationships in data sets
- Both are based on the core data model you have seem

Unit Data Set



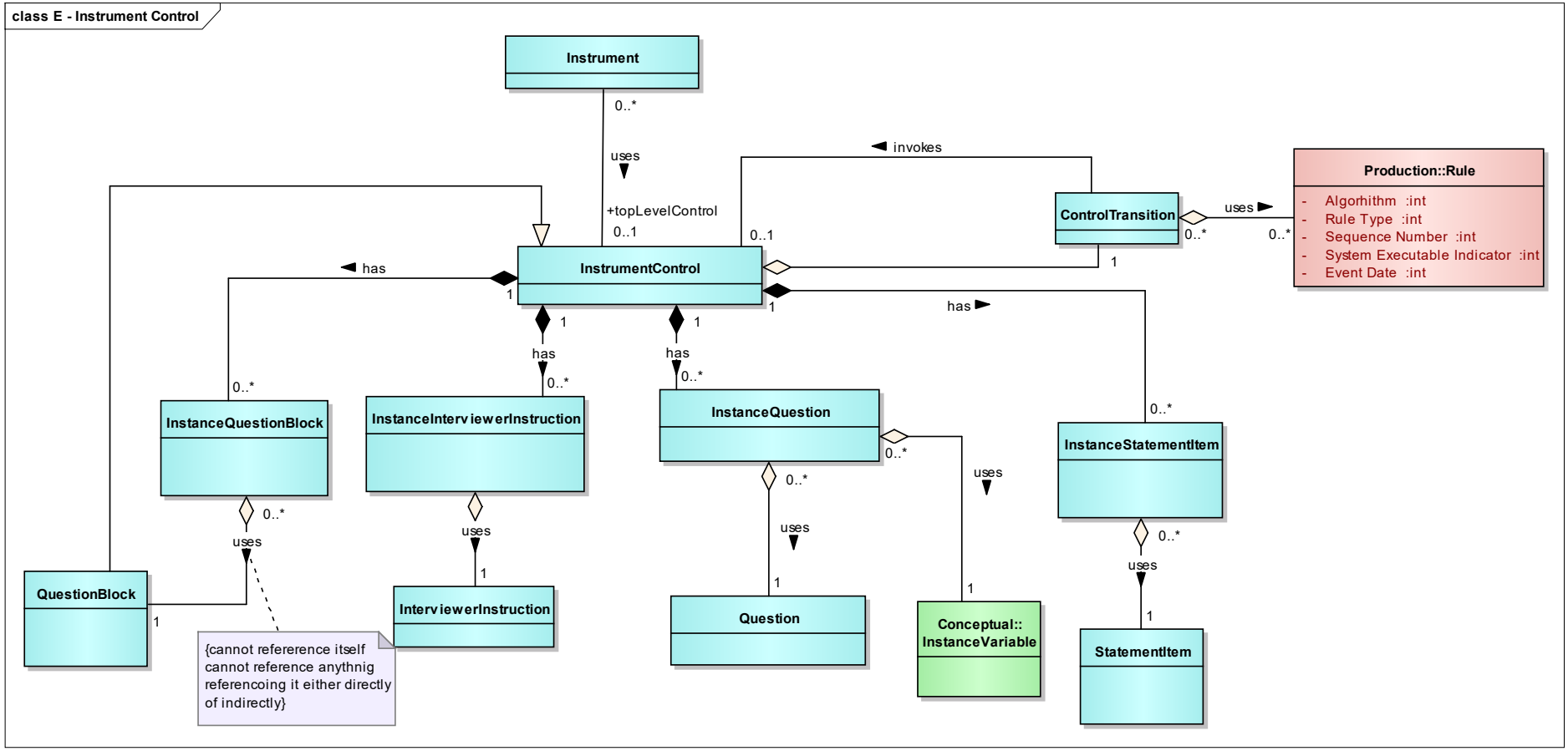
Dimensional Data Set



Describing Questionnaires

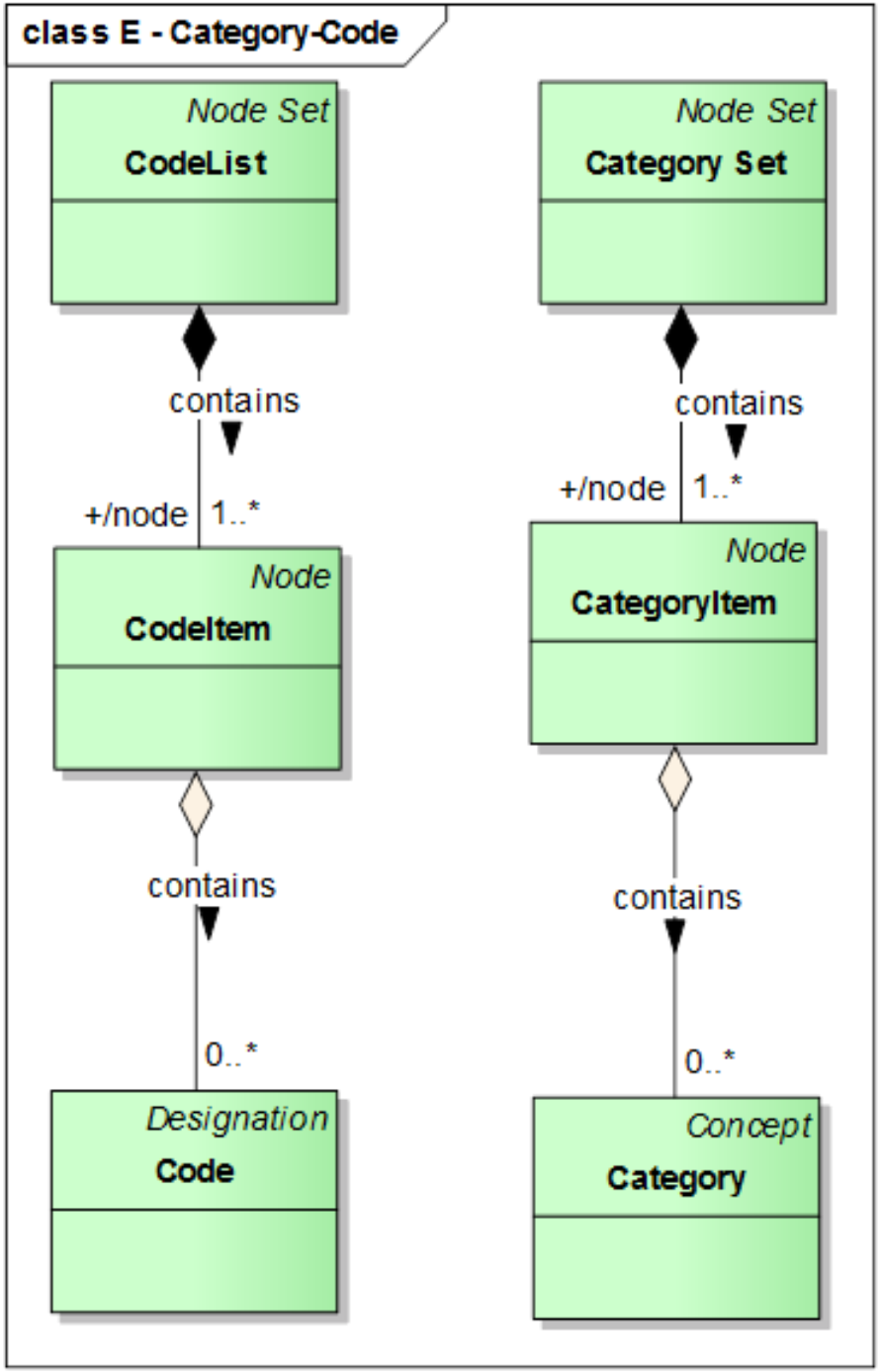
- DDI – Lifecycle has a very complete description of a questionnaire/instrument
  - Includes the mode and specifics of the instrument
  - Includes the questions, statements, and instructions used
  - Includes the flow logic of the questionnaire
  - Can have multiple-question “blocks”
- GSIM does the same
  - With less detail
  - Largely based on DDI

GSIM Survey Instrument

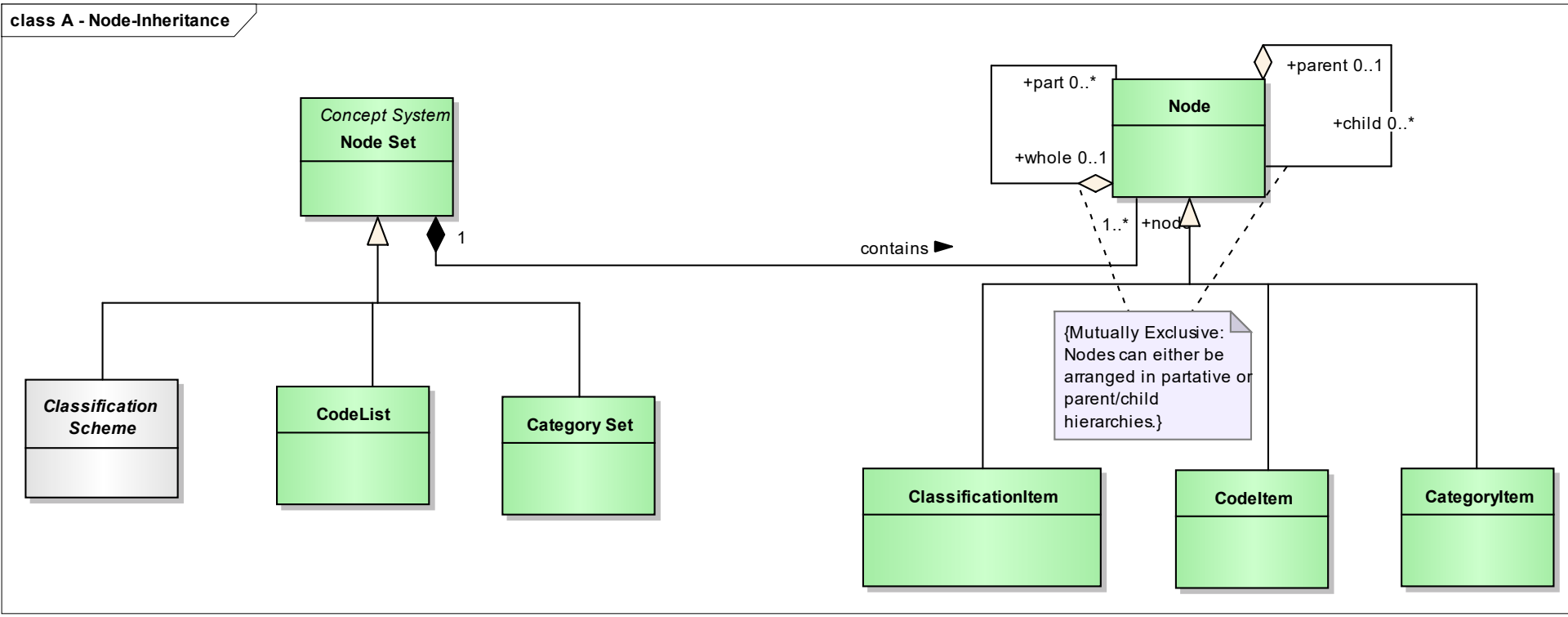


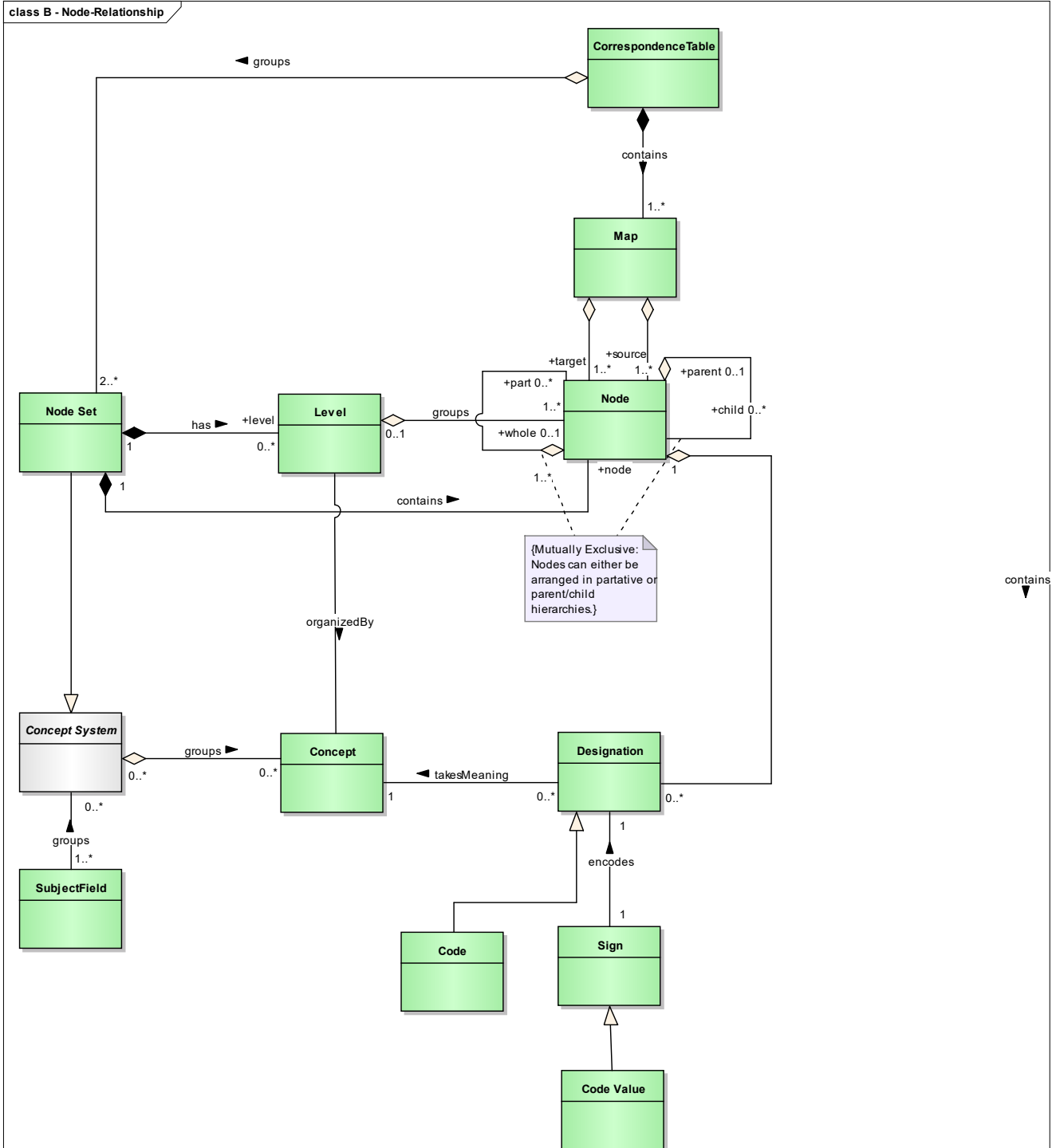
Classifications, Codelists, Categories and Concepts

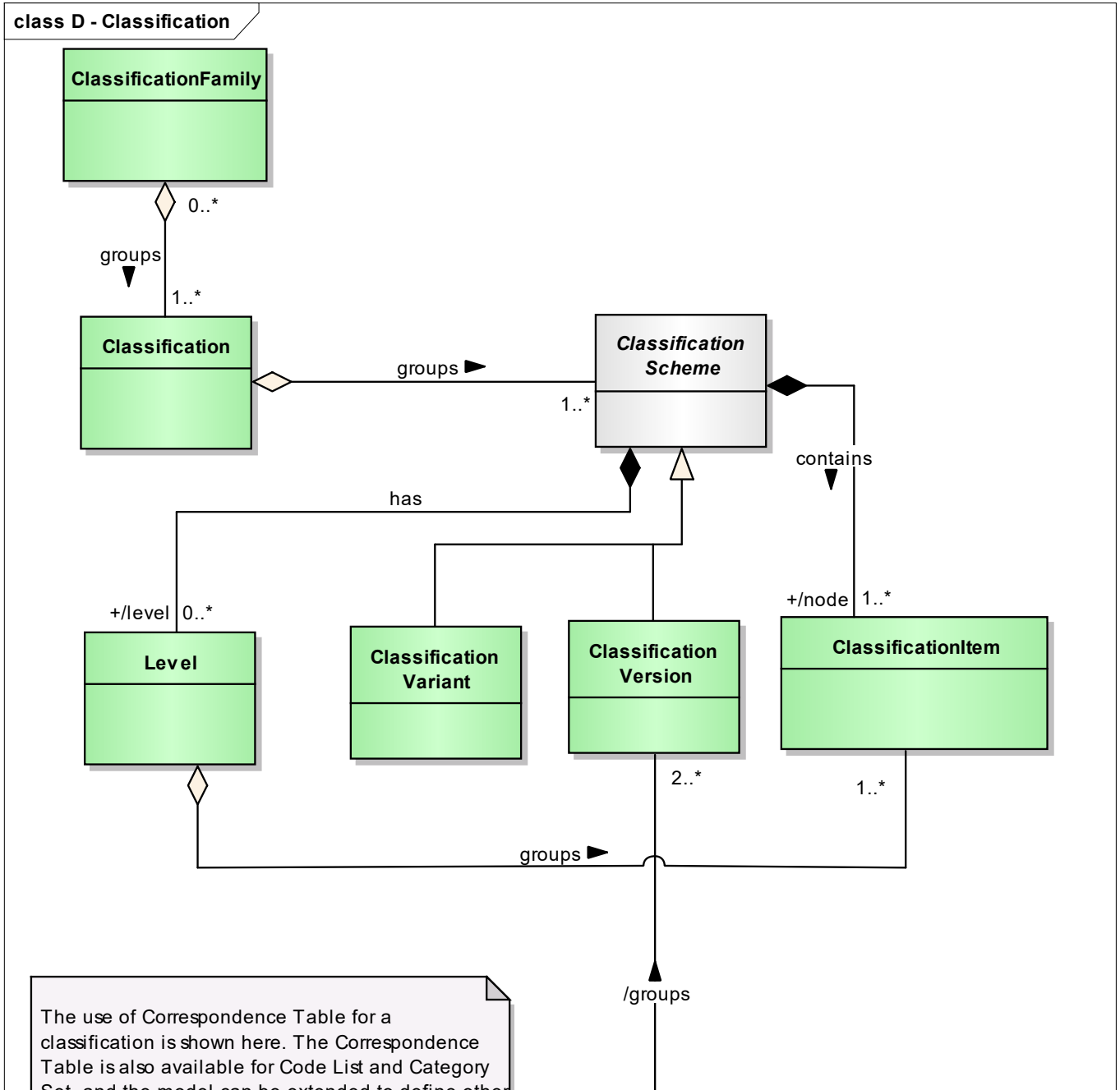
- DDI has codelists which take their meaning from categories.
- DDI has concepts associated with variables and questions.
- GSIM has all of this, and more!
  - GSIM is “concept-rich”
  - GSIM also has a pure classification model, which is not as complete in DDI-Lifecycle (a bit in 3.2)


class E - Category-Code
Node Set
CodeList
contains
+/node
1..*
Node
CodeItem
contains
0..*
Designation
Code
Node Set
Category Set
contains
+/node
1..*
Node
CategoryItem
contains
0..*
Concept
Category

Nodes and Node-Sets

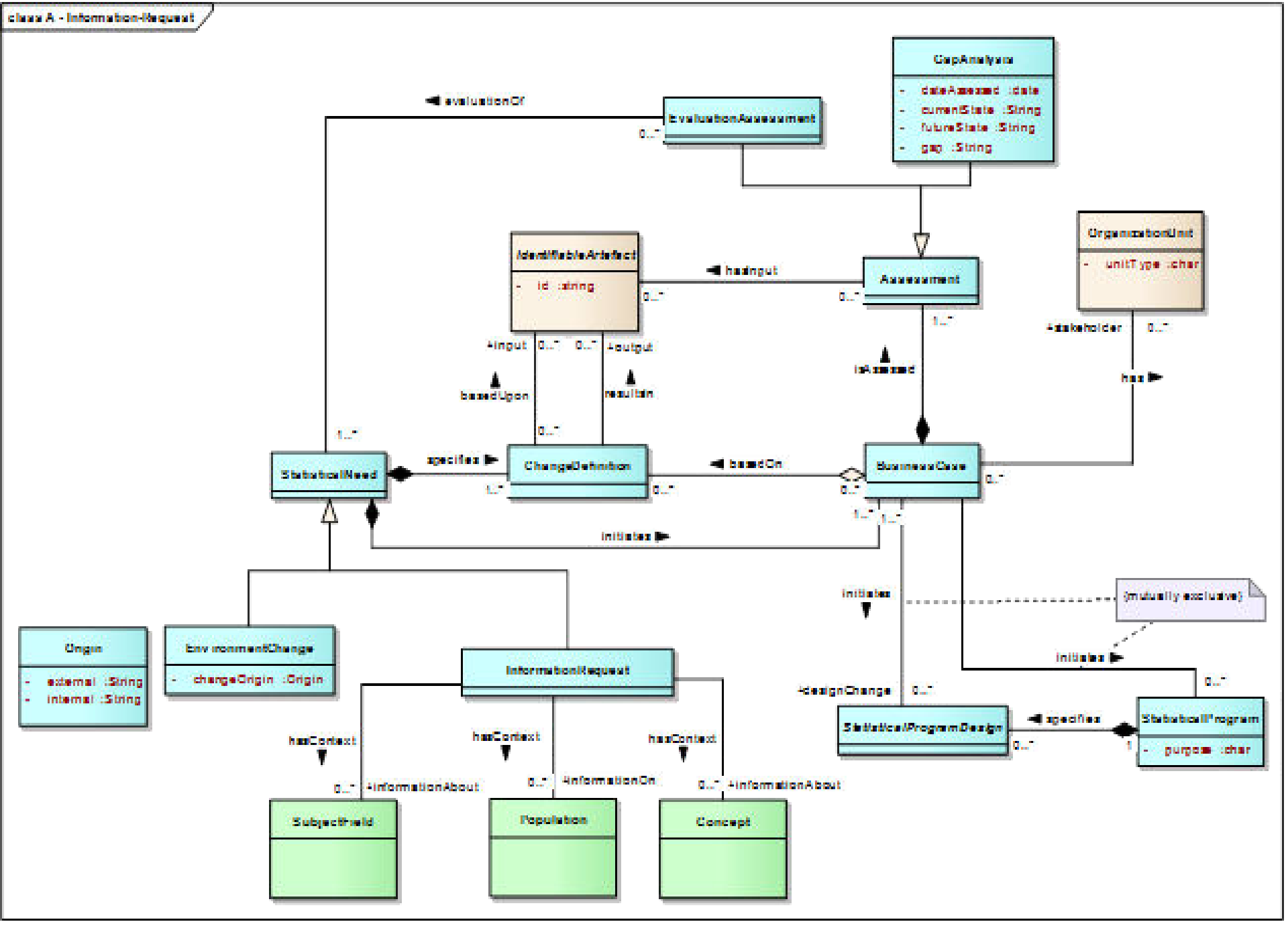


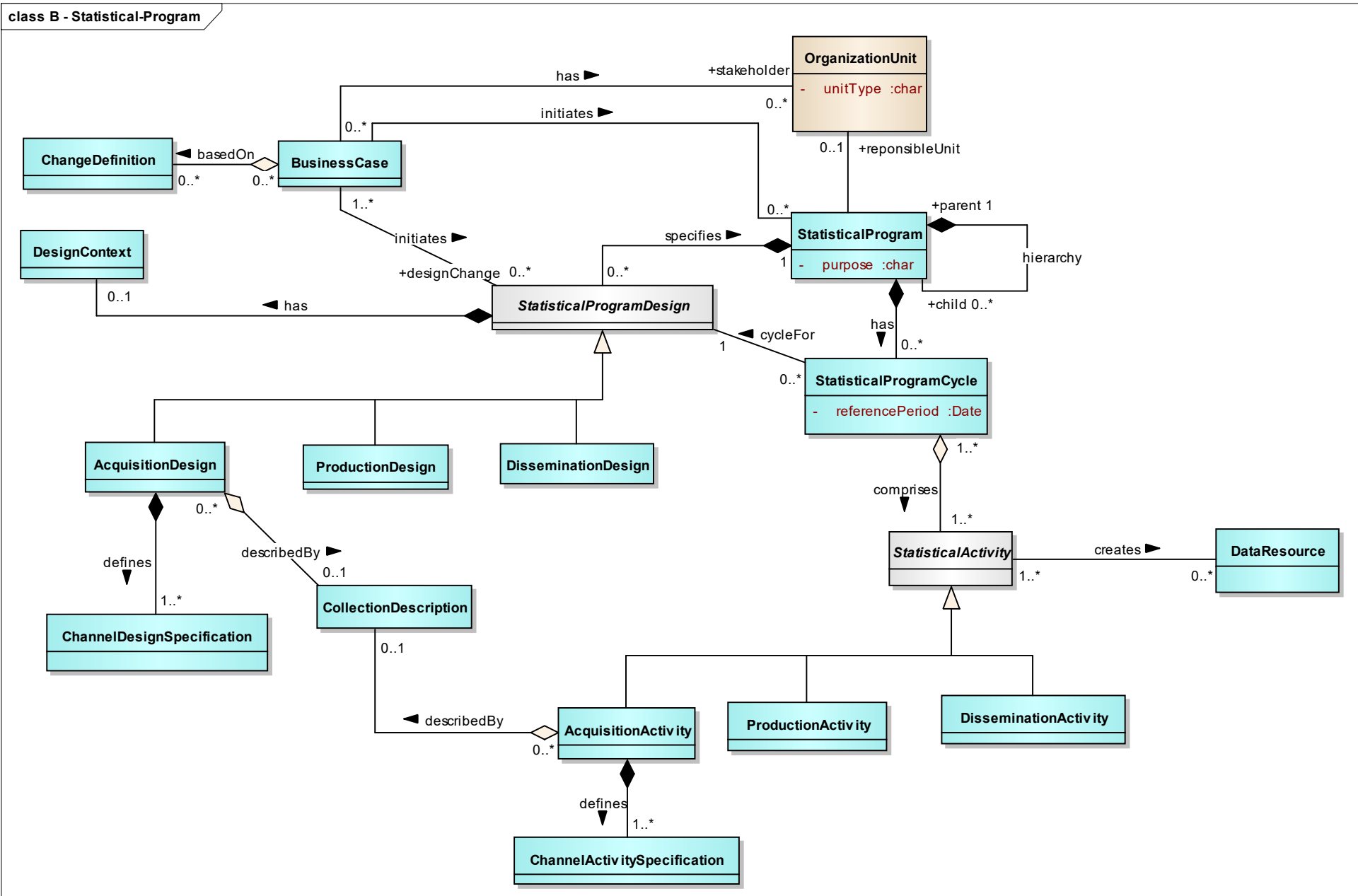

class B - Node-Relationship
CorrespondenceTable
groups
contains
1..*
Map
+target
+source
1..*
1..*
+parent 0..1
+child 0..*
+part 0..*
1..*
+whole 0..1
Node
2..*
Node Set
has
+level
0..*
1
Level
groups
0..1
1
contains
+node
1
1..*
{Mutually Exclusive: Nodes can either be arranged in partative or parent/child hierarchies.}
organizedBy
contains
Concept System
groups
0..*
0..*
Concept
0..*
takesMeaning
1
Designation
0..*
0..*
groups
1..*
SubjectField
1
encodes
1
Code
Sign
Code Value

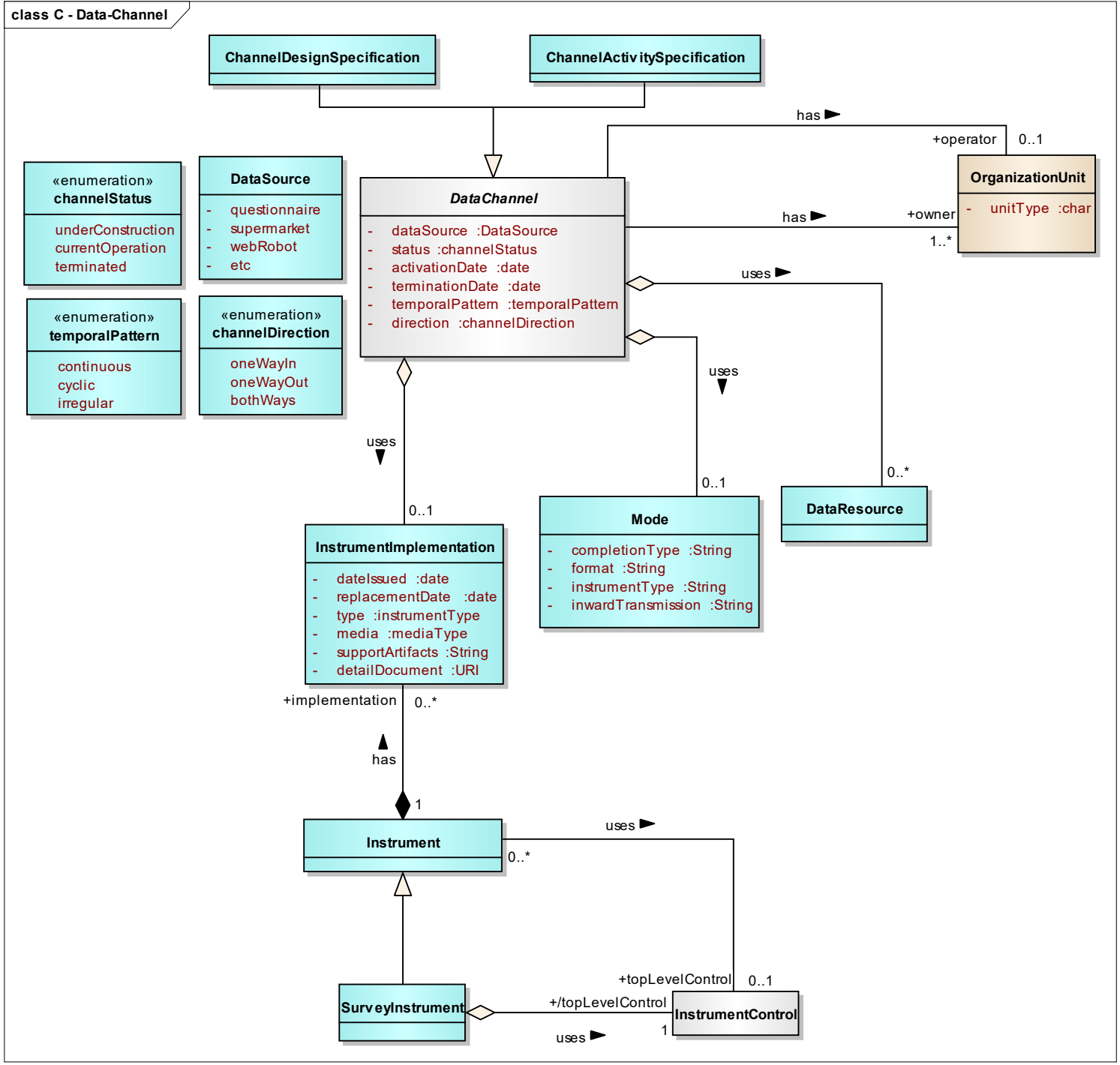

class D - Classification
ClassificationFamily
0..*
groups
1..*
Classification
groups
Classification Scheme
1..*
contains
has
+/level
0..*
Level
Classification Variant
Classification Version
+/node
1..*
ClassificationItem
2..*
1..*
groups
/groups
The use of Correspondence Table for a classification is shown here. The Correspondence Table is also available for Code List and Category

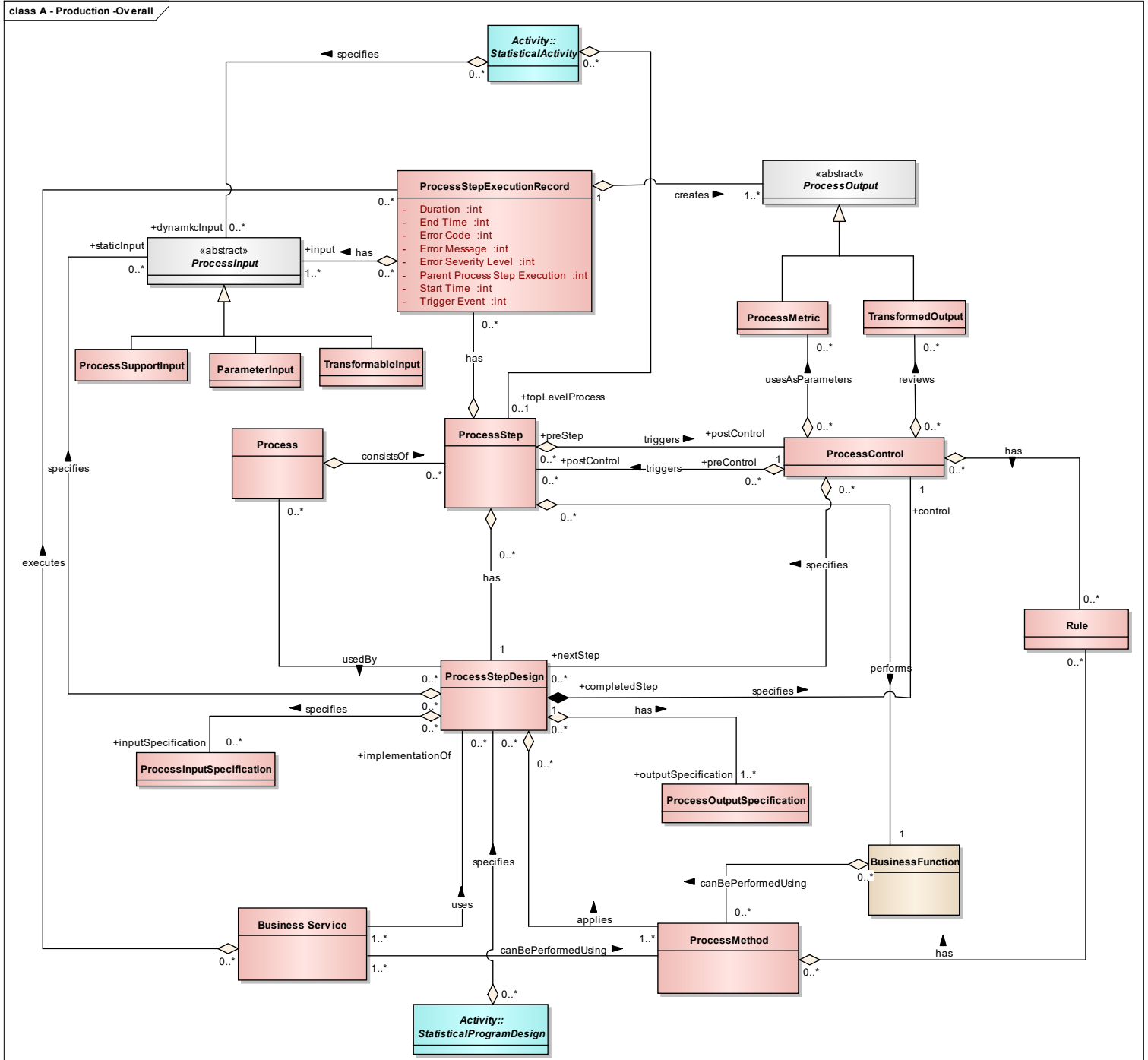
Events and Processing

- DDI provides us with several ways to describe events and processing
  - Lifecycle Events
  - Collection Events
  - “Coding” Elements
    - Generation Instructions
    - General instructions
- GSIM gives us much, much more!
  - Some of this is very specific to statistical agencies


class A - Information-Request
GapAnalysis
- dateAssessed :date
- currentState :String
- futureState :String
- gap :String
evaluationOf
EvaluationAssessment
0..*
OrganizationUnit
- unitType :char
IdentifiableArtifact
- id :string
hasInput
Assessment
0..*
0..*
1..*
+stakeholder
0..*
+input
0..*
0..*
+output
basedUpon
resultsIn
isAssessed
has
1..*
StatisticalNeed
specifies
ChangeDefinition
basedOn
BusinessCase
0..*
1..*
0..*
0..*
1..*
initiates
initiates
{mutually exclusive}
Origin
- external :String
- internal :String
EnvironmentChange
- changeOrigin :Origin
InformationRequest
+assignChange
0..*
StatisticalProgramDesign
specifies
StatisticalProgram
- purpose :char
1
0..*
hasContext
hasContext
hasContext
0..*
+informationAbout
0..*
+informationOn
0..*
+informationAbout
SubjectField
Population
Concept


class B - Statistical-Program
OrganizationUnit
- unitType :char
has
+stakeholder
0..*
initiates
0..*
0..1
+reponsibleUnit
ChangeDefinition
basedOn
0..*
0..*
BusinessCase
1..*
0..*
+parent 1
hierarchy
+child 0..*
DesignContext
initiates
+designChange
0..*
specifies
1
StatisticalProgram
- purpose :char
0..1
has
0..*
StatisticalProgramDesign
cycleFor
1
has
0..*
0..*
StatisticalProgramCycle
- referencePeriod :Date
AcquisitionDesign
ProductionDesign
DisseminationDesign
1..*
comprises
1..*
0..*
describedBy
0..1
defines
1..*
StatisticalActivity
creates
1..*
0..*
DataResource
ChannelDesignSpecification
CollectionDescription
0..1
describedBy
0..*
AcquisitionActivity
ProductionActivity
DisseminationActivity
defines
1..*
ChannelActivitySpecification


class C - Data-Channel
ChannelDesignSpecification
ChannelActivitySpecification
has
+operator
0..1
OrganizationUnit
- unitType :char
«enumeration»
channelStatus
underConstruction
currentOperation
terminated
DataSource
- questionnaire
- supermarket
- webRobot
- etc
DataChannel
- dataSource :DataSource
- status :channelStatus
- activationDate :date
- terminationDate :date
- temporalPattern :temporalPattern
- direction :channelDirection
has
+owner
1..*
uses
«enumeration»
temporalPattern
continuous
cyclic
irregular
«enumeration»
channelDirection
oneWayIn
oneWayOut
bothWays
uses
uses
0..1
0..*
0..1
InstrumentImplementation
- dateIssued :date
- replacementDate :date
- type :instrumentType
- media :mediaType
- supportArtifacts :String
- detailDocument :URI
Mode
- completionType :String
- format :String
- instrumentType :String
- inwardTransmission :String
DataResource
+implementation
0..*
has
1
Instrument
uses
0..*
+topLevelControl
0..1
SurveyInstrument
+/topLevelControl
InstrumentControl
1
uses


class A - Production -Overall
Activity::
StatisticalActivity
specifies
0..*
0..*
ProcessStepExecutionRecord
- Duration :int
- End Time :int
- Error Code :int
- Error Message :int
- Error Severity Level :int
- Parent Process Step Execution :int
- Start Time :int
- Trigger Event :int
creates
1..*
«abstract»
ProcessOutput
+dynamicInput
+staticInput
«abstract»
ProcessInput
+input
has
1..*
ProcessSupportInput
ParameterInput
TransformableInput
ProcessMetric
TransformedOutput
usesAsParameters
reviews
has
+topLevelProcess
0..1
Process
consistsOf
ProcessStep
+preStep
+postControl
triggers
+preControl
ProcessControl
+control
specifies
executes
has
usedBy
+nextStep
ProcessStepDesign
+completedStep
performs
Rule
+inputSpecification
ProcessInputSpecification
+implementationOf
+outputSpecification
ProcessOutputSpecification
BusinessFunction
canBePerformedUsing
uses
applies
Business Service
canBePerformedUsing
ProcessMethod
Activity::
StatisticalProgramDesign

Looking Forward

- DDI and GSIM have some very strong alignments
- There are also some gaps
- DDI may need to add support for some functionality
  - But maybe not everything – maybe SDMX can fill some gaps
- This is a two-way alignment
  - GSIM may need to adjust to better fit DDI implementation

Looking Forward (cont.)

- As we look to the next major re-design of DDI, we will be working proactively with GSIM
  - Representative from GSIM were invited to the first working session this year at Schloss Dagstuhl
- DDI will continue to attend events around GSIM sponsored by HLG-BAS
  - Like the Geneva meeting this past November
- Possibility for proactive engagement at a technical level
  - SDMX-DDI Dialogue
  - DDI Working Groups?
  - Others?
- GSIM may also provide a strong basis for other types of work within the DDI Community, less focused on official statistics
  - Like the “Generic Longitudinal Process Model”, which was based on GSBPM

Looking Forward

- Some external projects involve both archives and statistical agencies
  - Data without Boundaries (DwB) is a prime example
  - DwB is using a DDI-based metadata model
  - May lead to production implementations in future
- If archives and statistical agencies use the same metadata…
  - Archiving of official data becomes much easier
  - Both communities can leverage the same tools. Approaches, and resources (where appropriate)
    - Microdata access is an obvious point of stnergy

Questions?