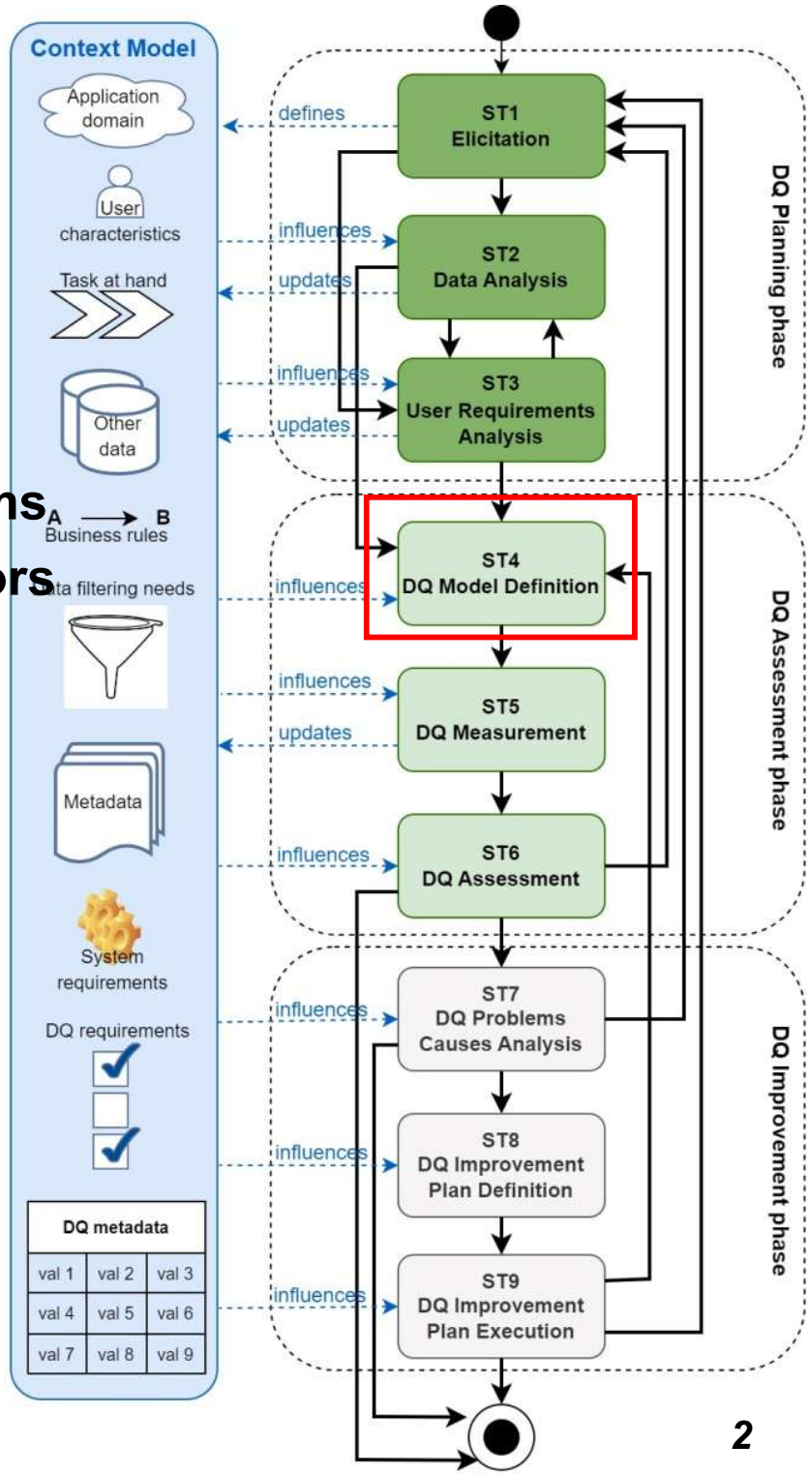
Calidad de Datos e Información

CaDQM

Phase 2 – DQ Assessment

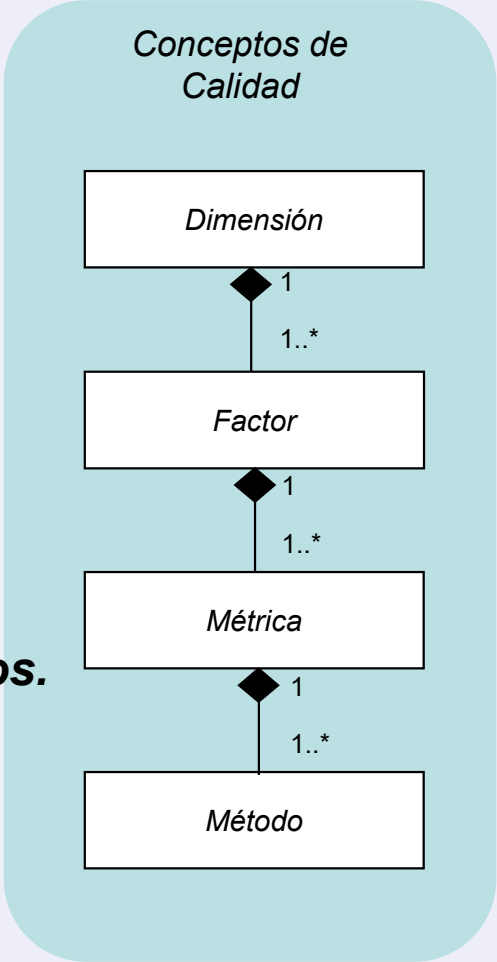
ST4 – Data Quality Model Definition

- ***Actividades***
  - **Prioritization and selection of DQ problems**
  - **Selection of DQ dimensions and DQ factors**
  - **Definition of DQ metrics**
  - Implementation of DQ methods



Actividades en ST4 – DQ Model definition

- ***Jerarquía de Conceptos de Calidad:***
  - ***Dimensiones:***
    - ***Facetas de la calidad a alto nivel.***
  - ***Factores:***
    - ***Aspectos particulares de las dimensiones.***
  - ***Métricas:***
    - ***Cada factor puede medirse con varias métricas.***
  - ***Métodos:***
    - ***Cada métrica puede implementarse con varios métodos.***



Métodos de CD

- *Los métodos de CD definen la forma de implementar las métricas de calidad*

- *Método de calidad - Especificación*
  - *Un nombre*
  - *Una descripción (cómo se realiza la medición)*
  - *Los tipos de los datos de entrada*
    - *tipo de datos esperado como parámetro de entrada*
  - *La tipos de los datos de salida*
    - *tipo de datos esperado como parámetro de salida*
  - *El algoritmo define la implementación de la métrica*
    - *proceso que implementa el método*

Ejemplos de Métodos de CD

- *La implementación del método depende:*
  - *de la aplicación en concreto y*
  - *de la estructura de la BD*

- *para medir el tiempo transcurrido desde la última actualización, se puede:*
  - *Usar timestamps de la BD*
  - *Acceder a los logs de actualización*
  - *Comparar versiones de la BD*
  - *…*

Varios métodos de medición

Método de CD aplicado

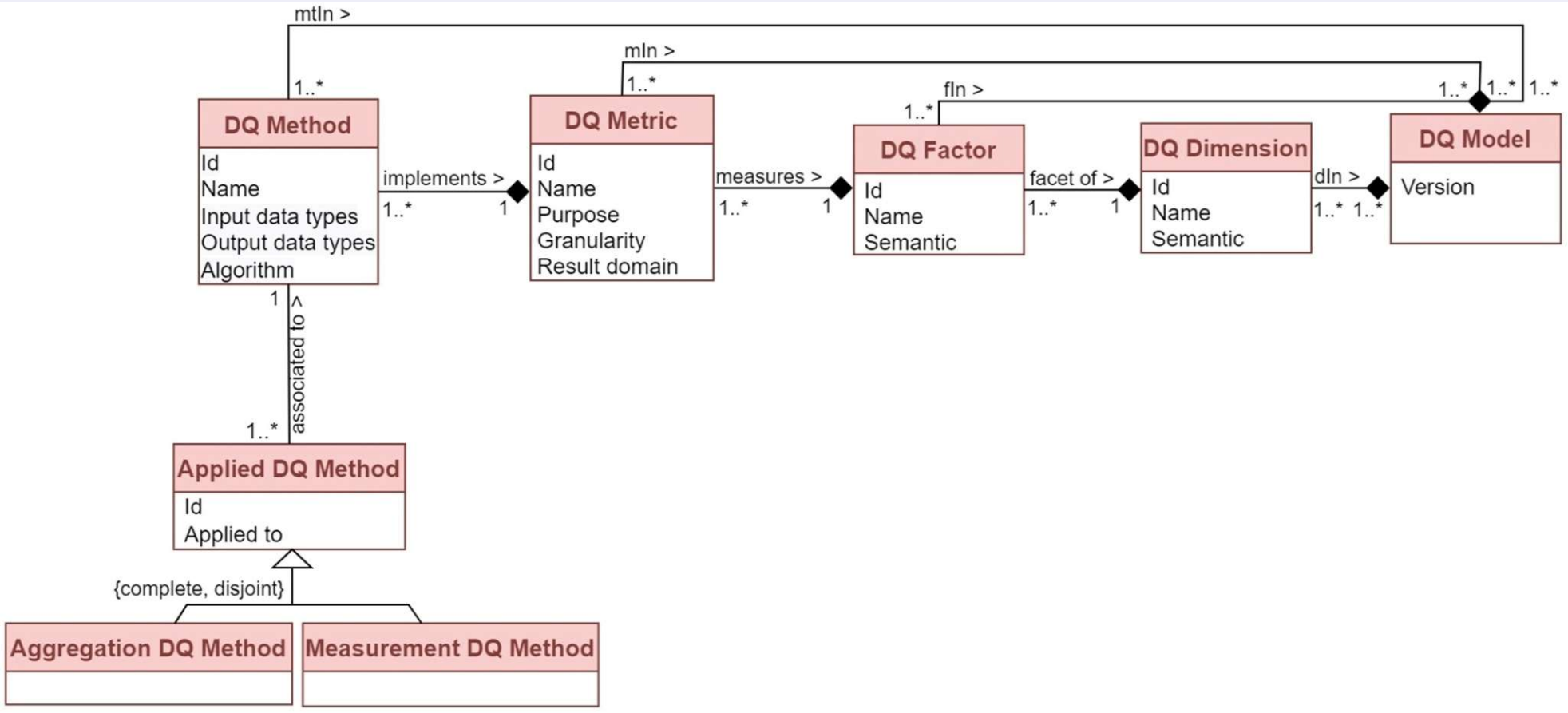
- *Define la aplicación de una métrica a un dato o conjunto de datos en particular.*

- *Método de calidad aplicado - Especificación*
  - *Un tipo: Tipo de método aplicado: medición O agregación*
  - *Una descripción*
  - *Un conjunto de atributos: Atributos del esquema de datos sobre los que se aplica el método*

- *Cada método de calidad está asociado a un conjunto de Métodos aplicados.*

- *Encargado de tomar una serie de medidas (correspondientes al método de una métrica específica), para una BD concreta.*

Metamodelo de Calidad de Datos



Ejemplos

Métrica 1: ExacSemantica-Bool	
Descripción:	***Mide si un dato existe en la realidad.***
Granularidad:	***Celda***
Dominio del Resultado:	***{0,1}***

- ***Para la métrica, ExacSemantica-Bool, proponemos:***
  - ***2 métodos llamados MET1 y MET2, para medir la exactitud semántica de valores que representan direcciones.***

Métodos de CD

- **Métrica 1 implementada con MET1:**

MET1: ExacSemantica-Bool-Google	
Descripción:	***Implementa la métrica ExacSemantica-Bool usando Google Maps***
Tipos de datos de entrada	***String***
Tipos de datos de salida	***Boolean***
Proceso	***Consultar Google Maps***

- **Método aplicado de MET1:**

MET1_ap: ExacSemantica-Bool-Google_ap	
Tipo	***Medición***
Descripción	***Busca en Google Maps una dirección dada***
Aplicado a	***Atributo «direccion»***

Métodos de CD

- **Métrica 1 implementada con MET2:**

MET2: ExacSemantica-Bool-Guía	
Descripción:	***Implementa la métrica ExacSemantica-Bool usando las guías de Antel***
Tipos de datos de entrada	***String***
Tipos de datos de salida	***Boolean***
Proceso	***Consultar Guías de direcciones de Antel***

- **Método aplicado de MET2:**

MET2_ap: ExacSemantica-Bool-Guía_ap	
Tipo	***Medición***
Descripción	***Busca en las guías de Antel una dirección dada***
Aplicado a	***Atributo «direccion»***

Métrica de Calidad

Métrica 2: Densidad-Grado	
Descripción:	***Mide el grado de densidad de un conjunto de datos.***
Granularidad:	***Columna***
Dominio del Resultado:	***[0..1]***

- ***Para la métrica, Densidad-Grado, proponemos:***
  - ***1 método llamado MET3, para medir la densidad de un atributo.***

Métodos de CD

- **Métrica 1 implementada con MET3:**

MET3: Densidad-Grado-Contar	
Descripción:	***Cuenta la cantidad de valores NULL en un conjunto de datos***
Tipos de datos de entrada	*Atributo?*
Tipos de datos de salida	***Integer***
Proceso	***Contar cantidad de valores NULL***

- **Método aplicado de MET3:**

MET3_ap: Densidad-Grado-Contar_ap	
Tipo	***Medición***
Descripción	***Recibe el nombre de un atributo y cuenta la cantidad de valores NULL en la columna dada***
Aplicado a	***Atributos: «ciudad», «direccion», «telefono»***

Ejercicio 12

- ***Considerando los ejercicios realizados anteriormente en clase (Ej. 1 al 5), para algunos de los factores de Exactitud identificados:***

  - ***Dar una** métrica de calidad **de** Exactitud.*

  - ***Definir un** método de calidad **y** aplicar **dicho método para algún dato o para algunos datos en particular.***

Ejercicios 13

- ***Considerando una tabla cualquiera, especificar una** métrica**, un** método de medición **para dicha métrica y el** método aplicado **para uno de los siguientes factores:***
  - ***Cobertura***
  - ***Actualidad***
  - ***Integridad intra-relación***

Influencia del Contexto

- ***La selección de las dimensiones y sus respectivos factores se ve influenciada por los componentes del contexto.***
  - ***Ejemplo: la existencia de reglas de negocio en el contexto de los datos sugiere el análisis de la consistencia de los mismos.***

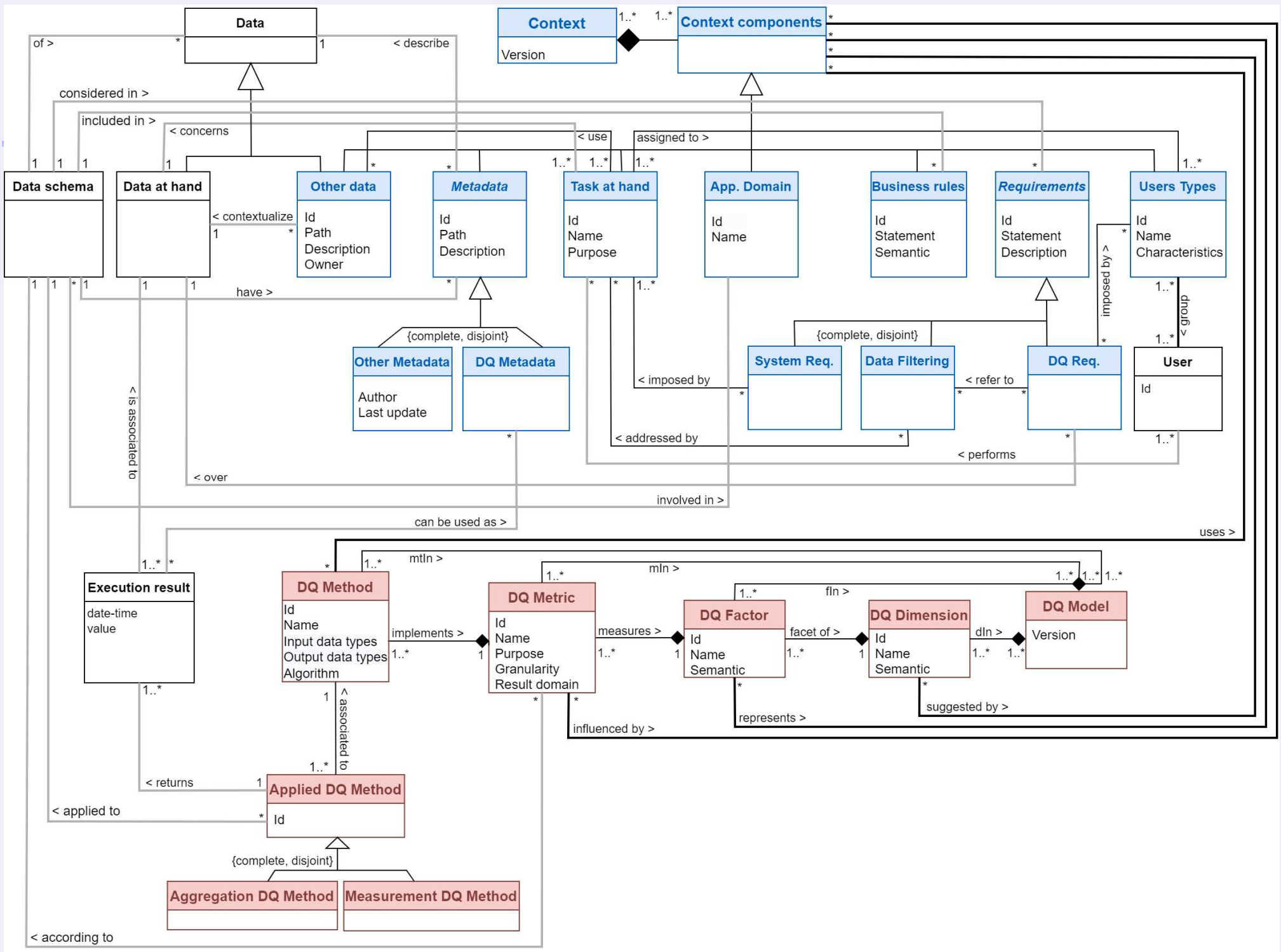
- ***Los componentes de contexto condicionan y apoyan la definición de las métricas y los métodos de calidad.***
  - ***Ejemplo: la granularidad de las métricas podría establecerse en función de requerimientos de CD (ej.: req. de CD para una tabla relacional o para una tupla).***

Influencia del Contexto

- *En el caso de los métodos de CD:*
  - *sus tipos de datos de entrada/salida podrían configurarse de manera diferente dependiendo de la tarea en cuestión.*
    - *Ejemplo: una tarea requiere códigos alfanuméricos, mientras que otra tarea requiere códigos numéricos.*
  - *podrían incluir (en su algoritmo) condiciones dadas por umbrales de calidad requeridos por los requerimientos de CD.*
    - *Ejemplo:*
      - *el 90% de los nombres deben ser sintácticamente correctos*
      - *los datos son recientes si se registraron después del mediodía de hoy*

Influencia del Contexto

- *Otros ejemplos:*
  - *Metadatos de CD obtenidos en mediciones preliminares*
    - *Para definir métricas de calidad agregadas*
  - *Otros metadatos, como la fecha de creación de los datos*
    - *Para analizar la frescura de los datos*
  - *Los tipos de usuarios también influyen en el modelo de CD*
    - *imponen los requerimientos de CD*
    - *nuevos tipos de usuarios podrían sugerir nuevos requerimientos de CD*
  - *Dominios de aplicación definen modelos de calidad diferentes*
    - *cada dominio tiene objetivos diferentes*


Data
of >
< describe
Context
Version
Context components
considered in >
included in >
< concerns
< use
assigned to >
Data schema
Data at hand
< contextualize
Other data
Id
Path
Description
Owner
Metadata
Id
Path
Description
Task at hand
Id
Name
Purpose
App. Domain
Id
Name
Business rules
Id
Statement
Semantic
Requirements
Id
Statement
Description
Users Types
Id
Name
Characteristics
imposed by >
< group
have >
{complete, disjoint}
Other Metadata
Author
Last update
DQ Metadata
{complete, disjoint}
System Req.
Data Filtering
DQ Req.
User
Id
< imposed by
< refer to
< addressed by
< performs
< over
involved in >
< is associated to
can be used as >
uses >
Execution result
date-time
value
DQ Method
Id
Name
Input data types
Output data types
Algorithm
mtIn >
mIn >
fIn >
implements >
DQ Metric
Id
Name
Purpose
Granularity
Result domain
measures >
DQ Factor
Id
Name
Semantic
facet of >
DQ Dimension
Id
Name
Semantic
dIn >
DQ Model
Version
< associated to
< returns
Applied DQ Method
Id
< applied to
represents >
suggested by >
influenced by >
{complete, disjoint}
Aggregation DQ Method
Measurement DQ Method
< according to

Especificación del Modelo de CD usando el Contexto de datos

set of DQ problem Ids	DQ Dimension	DQ Factor	DQ Metric	DQ Method	Applied DQ Method
	ID: Name: Description: Suggested by = {comp1, comp2, comp3}	ID: Name: Description: Represents = {comp1, comp2}	ID: Name: Description: Influenced by = {comp1} Granularity: Result domain:	ID: Name: Description: Uses = {comp1} Input data types: Output data types: Algorithm:	ID: Type: Description: Applied to:
					ID: Type: Description: Applied to:
				ID: Name: Description: Uses = {} Input data types: Output data types: Algorithm:	ID: Type: Description: Applied to:
			ID: Name: Description: Influenced by = {comp2} Granularity: Result domain:	ID: Name: Description: Uses = {comp2} Input data types: Output data types: Algorithm:	ID: Type: Description: Applied to:

Ejemplo

- ***Evaluación de la calidad de los datos registrados en un sitio Web para el registro de quejas de una institución pública.***

- ***Problemas de calidad de datos identificados***
  - ***P1: El 80% de los nombres de los ciudadanos registrados en la intendencia están abreviados o incompletos.***
  - ***P2: El 65% de las direcciones reportadas están mal escritas.***

Ejemplo: Modelo de Contexto

Context comp.	All users	User 1	User 2
Application domain	**AD:** e-government		
Business rules:	**BR1:** Las denuncias deben ser realizadas por ciudadanos mayores de edad.		
	BR2: cada queja debe incluir la dirección del lugar de Montevideo que se esta reportando en la queja.		
Other data:	**OD1:** Base de datos relacional de la Dirección de Identificación Civil		
	OD2: Datos abiertos con las Direcciones oficiales de Montevideo		
Users types:		**UC1:** ciudadanos	**UC2:** funcionarios públicos
Tasks at hand:		**T1:** registran quejas en el sitio Web de la institución	**T2:** gestionan las quejas de los ciudadanos
Data filtering:			**DF1:** Interesan los datos de los ciudadanos que registran quejas.
			DF2: Interesan consultar las direcciones que han sido reportadas en las quejas.
DQ requirements:			**RQ1:** Los nombres de los ciudadanos deben ser reales.
			RQ2: Los nombres de los ciudadanos deben estar bien escritos y completos.
			RQ3: Las direcciones deben deben estar bien escritas y deben contener calle, número y esquina.

Ejemplo: Modelo de CD

DQ Problems	DQ Dimension	DQ Factor	DQ Metric	DQ method	Applied DQ method
P1, P2	**ID:** Acc_D1 **Name:** Accuracy **Description:** Evalúa si un dato es correcto **Suggested by =** {RQ1, RQ2, RQ3}	**ID:** SyntAcc_F1 **Name:** Syntactic accuracy **Description:** Evalúa la correctitud sintáctica **Represents =** {RQ2, RQ3}	**ID:** SyntAcc_M1 **Nombre:** SynAcc dictionary check **Description:** Evalúa si un dato es sintácticamente correcto comparándolo con un diccionario. **Influenced by =** {RQ2, RQ3, DF1, DF2, OD1, OD2} **Granularity:** Celda **Result domain =** {0, 1}	**ID:** M1_checkValue **Uses =** {} **Name:** Check value **Description:** chequea si un dato está en una colección de datos. **Input data types:** String, Collection **Output data types:** Boolean **Algorithm:** Return isInCollection(dato, col)	**ID:** M1_checkValue_v1 **Type:** Measurement **AppliedTo:** "nombreCiudadano" **ID:** M1_checkValue_v2 **Type:** Measurement **AppliedTo:** "dirección"

ST4 – Data Quality Model Definition

Entradas	Salidas
Reporte del análisis de requerimientos de usuarios	Reporte de problemas de CD priorizados
Reporte del análisis de datos	Modelo de Calidad de Datos
Reporte de problemas de CD	
Modelo de Contexto	

- *Reporte de problemas de CD priorizados*
  - *Incluye a todos los problemas de CD identificados en las etapas ST1, ST2 y ST3. A cada uno de estos problemas se les asigna una prioridad, por ejemplo: alta, media, baja.*

- *Modelo de Calidad de Datos*
  - *El modelo de CD incluye todas las dimensiones y factores que surgen del Modelo de Contexto y que son guiadas por la lista de problemas de CD priorizados.*
  - *Además, métricas y métodos de CD son definidos en función del Modelo de Contexto y los problemas priorizados.*