Calidad de Datos e Información

Tarea 2

Corrección

Estructura esperada de la Tarea 2

- Correcciones de la Tarea 1
  - Indicando brevemente los cambios realizados y/o introducidos
- Entradas/salidas de la etapa ST4
- Presentación de las Entradas de la etapa ST4
  - Reporte del análisis de requerimientos de usuarios (resumen)
  - Reporte del análisis de datos (resumen)
  - Reporte de problemas de CD
  - Modelo de Contexto
- Especificación de las Salidas de la etapa ST4
  - Reporte de problemas de CD priorizados
  - Modelo de Calidad de Datos
- Conclusiones
  - De la Tarea 1 (si fuera necesario) y de la Tarea 2

Errores identificados en la Corrección

- Correcciones de la Tarea 1
  - Los data at hand son presentados como metadatos
  - Solo corrigen componentes de contexto faltantes
  - No presenta descripción de los componentes de contexto o de los problemas de CD, solo sus identificadores
  - No presentan el modelo de Contexto completo, solo los corregidos/agregados
  - Incluyen al Modelo de Contexto elementos que no son componentes
    - Ej: Fuentes de datos
  - Metadatos de Calidad mal interpretados
    - Ej: “La frecuencia esperada de actualización es mensual” o “Última actualización hace 5 años”, no son metadatos de CD
  - No presentan las lista de problemas de CD completa, solo los corregidos/agregados
  - Incluyen conclusiones al inicio del informe

Errores identificados en la Corrección

- Entradas/salidas de la etapa ST4
  - No son consideradas
  - No respetan el orden de ejecución de las actividades

- Presentación de las Entradas de la etapa ST4
  - Los reportes «análisis de requerimientos de usuarios», «análisis de datos» y/o «problemas de CD» son referenciadas al informe de la Tarea 1
  - No son claramente presentadas

Errores identificados en la Corrección

- Especificación de las Salidas de la etapa ST4
  - Problemas de Calidad de Datos
    - No se priorizan, solo seleccionan algunos de ellos
  - Modelo de Calidad de Datos
    - No incluye los problemas y/o componentes de contexto que influencian a los conceptos de CD
    - Componentes de contexto que sugieren dimensiones incorrectas
      - Ej: "Cada libro debe tener al menos 1 autor" sugiere exactitud
    - Las relaciones entre el Contexto y el Modelo de CD incluye componentes incorrectos o no incluyen componentes
      - Ej: Suggested by = {BR1, BR3, Referencial}, Uses = {nombre de una API} o Uses = {}
    - Las métricas son muy específicas
      - Ej: Verificar que la fecha tenga formato dd/mm/AAAA
    - Las métricas son muy genéricas
      - Ej: Mide si una tupla es precisa respecto a un concepto
    - La descripción de la métrica no coincide con su granularidad
      - Ej: Desc.: “Mide la precisión de un dato”. Granularidad: Columna
    - La descripción del método no coincide con la granularidad de la métrica

Errores identificados en la Corrección

- Especificación de las Salidas de la etapa ST4
  - Modelo de Calidad de Datos
    - No abordan todas las granularidades
    - La granularidad no es genérica
      - Ej: Granularidad: Tabla Libros
    - La descripción del método no coincide con la dimensión ni factor medidos
      - Ej: Dimensión Exactitud, Factor Precisión, Desc. del método: “verifica el grado de completitud de un string respecto al formato de fecha YYYY-MM-DD”
    - Descripción del método no coincide con la descripción del método aplicado
    - El tipo de datos de entrada del método (tabla) no coincide con el "aplica a" del método aplicado (tupla)
    - Repiten N métodos aplicados, cuando un método puede aplicar a N atributos
    - Crean métricas para las agregaciones

- Conclusiones
  - Presentan conclusiones solo de la Tarea 1
  - Presentan un resumen de las tareas realizadas

Errores graves a corregir

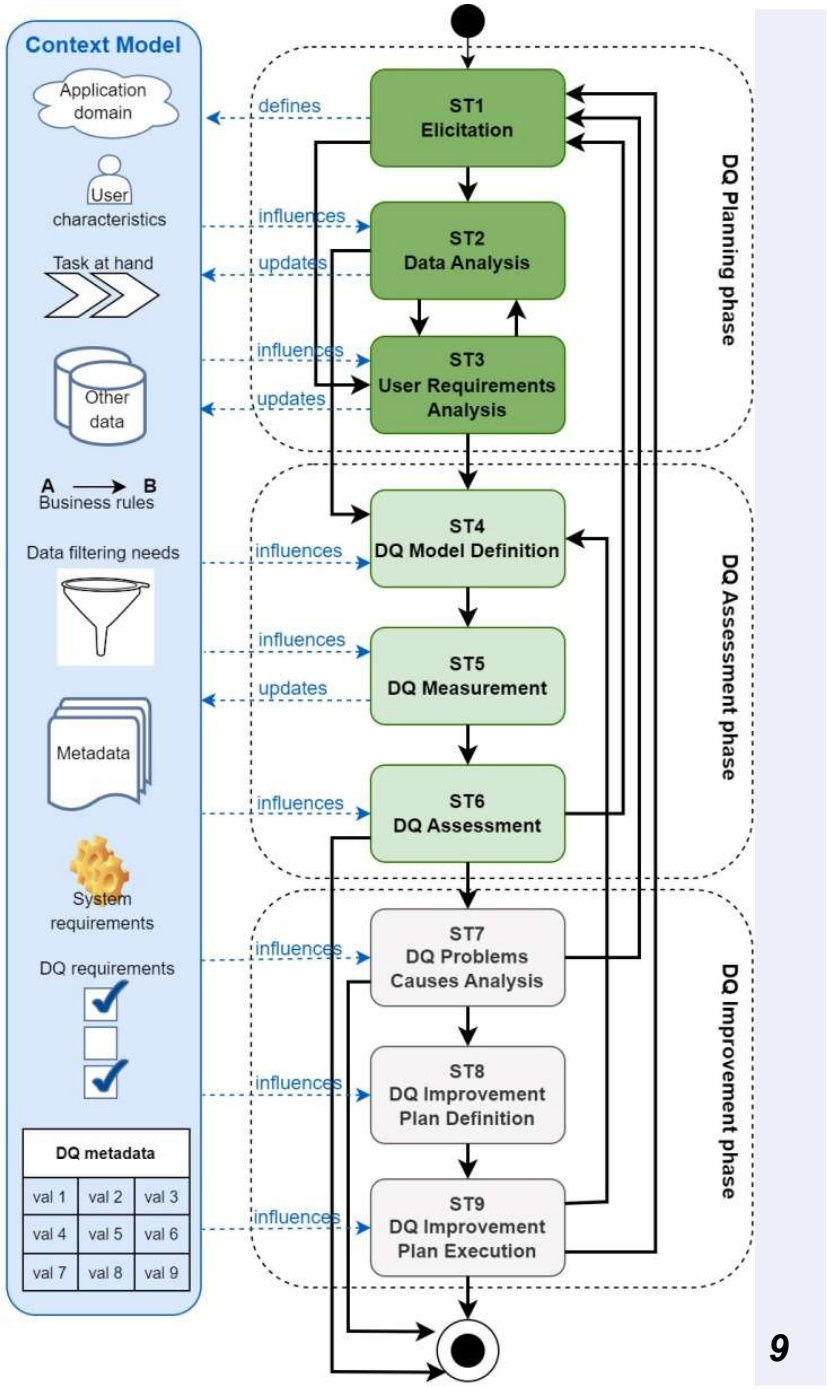
- Conceptos de CD con errores
  - Dimensiones, Factores, Métricas, Métodos y Métodos aplicados
- No hay coherencia entre los conceptos de CD
- El Modelo de CD no se relaciona con el Modelo de Contexto ni con los problemas de CD

CaDQM

Repaso

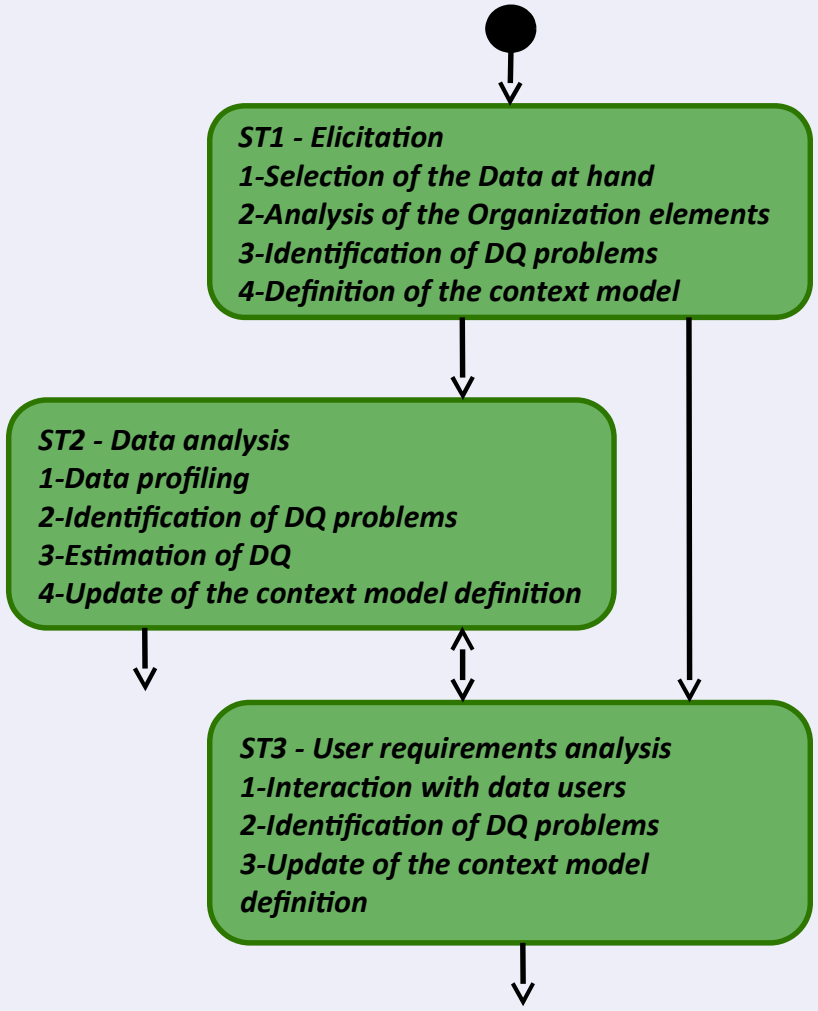
CaDQM

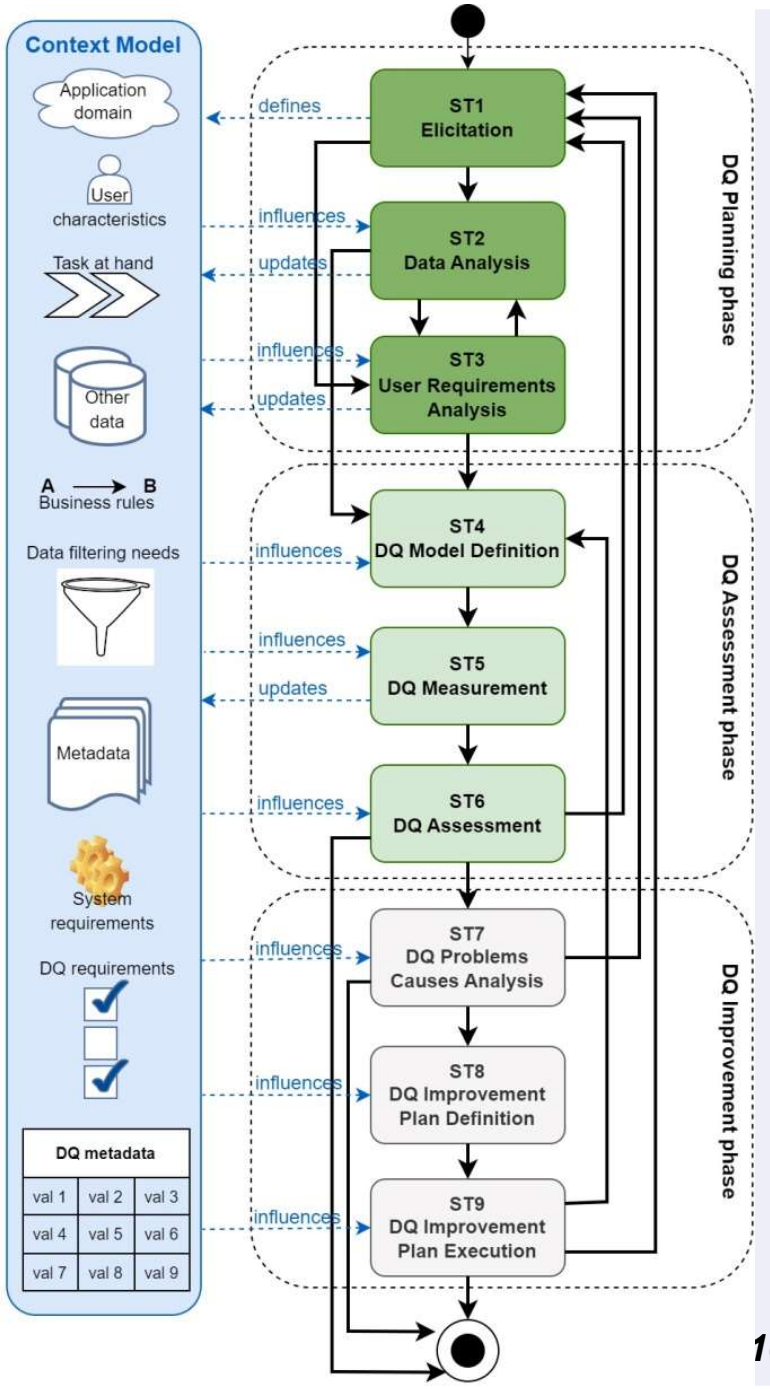
Context-aware Data Quality Methodology



CaDQM

Phase 1: DQ Planning





ST1 - Elicitation

Entradas	Salidas
	Data at hand
	Reporte con problemas de CD
	Modelo de Contexto

- Data at hand
  - descripción de los datos
  - breve explicación de las razones para evaluar la calidad (por ejemplo: gran cantidad de problemas de CD reportados, verificación de nuevas reglas de negocios, introducción de nuevos procesos de negocios o simplemente monitoreo de la calidad)

- Reporte con problemas de CD
  - Incluye cualquier problema de CD identificado en esta etapa.

- Modelo de Contexto
  - Incluye todos los componentes de contexto identificados en esta etapa, agrupados por tipo de componente.

ST2 – Data Analysis

Entradas	Salidas
Data at hand	Reporte del análisis de datos
Reporte con problemas de CD	Reporte con problemas de CD
Modelo de Contexto	Modelo de Contexto

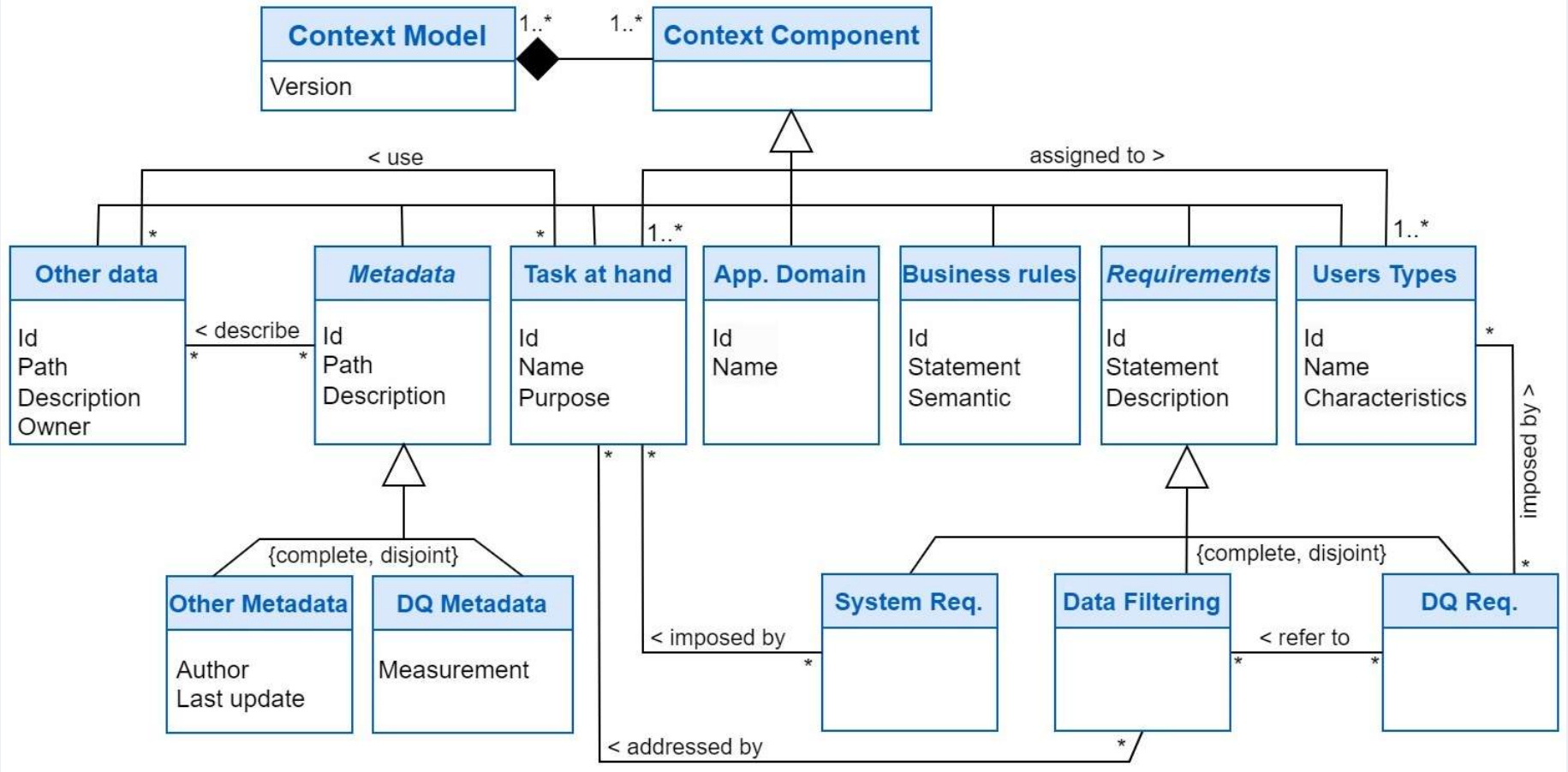
- Reporte del análisis de datos
  - descripción de las herramientas y técnicas utilizadas en la actividad de data profiling
  - componentes de contexto considerados en el análisis de datos
  - estado de CD que surge de la estimación inicial, es decir, los resultados del data profiling
- Reporte con problemas de CD
  - Incluye los problemas reportados en ST1/ST3 y todos los problemas de CD identificados en esta etapa
- Modelo de Contexto
  - Incluye los componentes de contexto identificados en ST1/ST3 y todos los componentes de contexto identificados en esta etapa, agrupados por tipo de componente

ST3 – User requirements analysis

Entradas	Salidas
Data at hand	Reporte de análisis de requerimientos de usuarios
Reporte con problemas de CD	Reporte con problemas de CD
Modelo de Contexto	Modelo de Contexto

- Reporte de análisis de requerimientos de usuarios
  - descripción de las herramientas y técnicas utilizadas en el análisis de requerimientos de usuarios
  - componentes de contexto considerados en el análisis
- Reporte con problemas de CD
  - Incluye a los problemas reportados en ST1/ST2, todos los problemas de CD identificados en esta etapa.
- Modelo de Contexto
  - Incluye a los componentes de contexto identificados en ST1/ST2 y todos los componentes de contexto identificados en esta etapa, agrupados por tipo de componente.

Metamodelo de Contexto



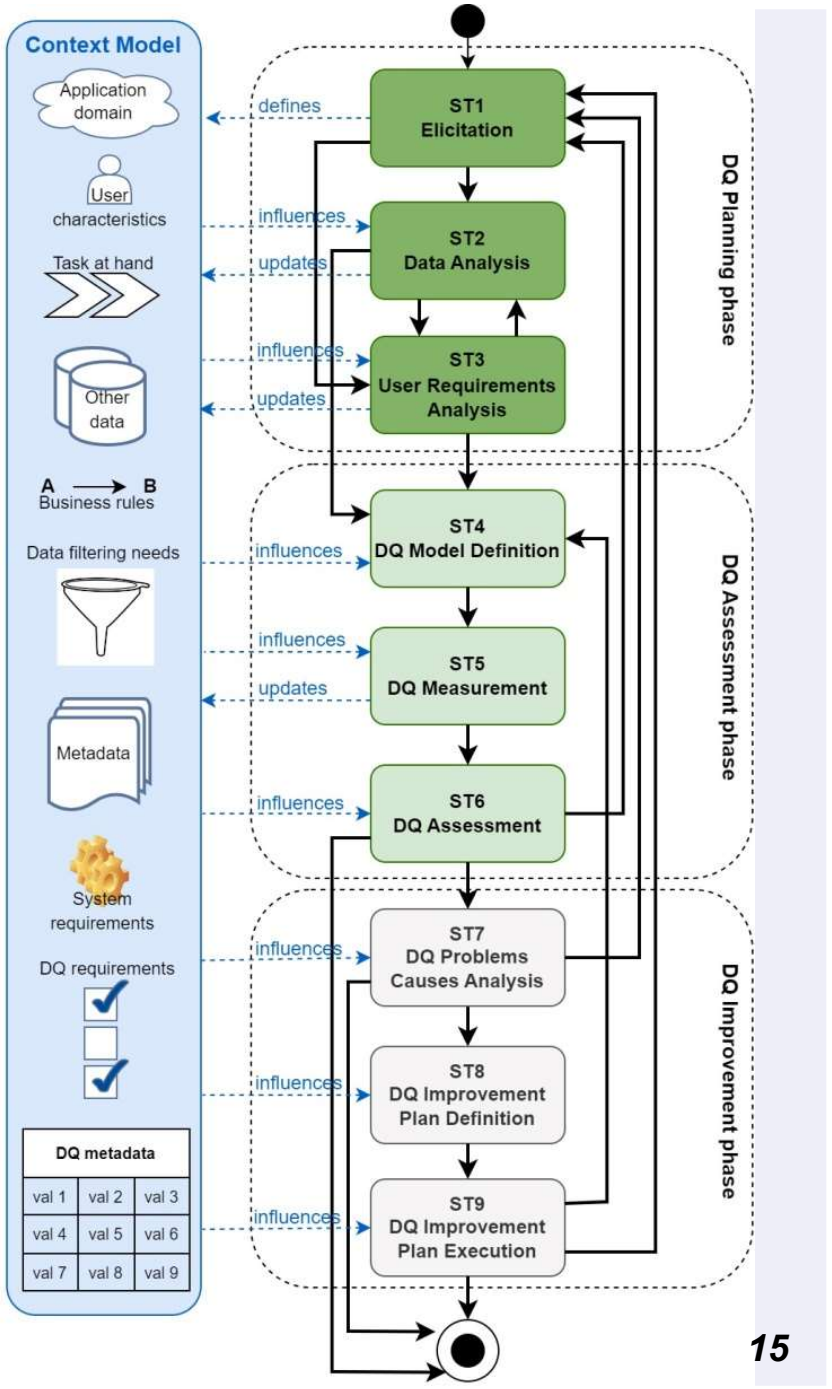
CaDQM

Phase 2: DQ Assessment

ST4 - DQ model definition
1-Prioritization and selection of DQ Problems
2-Selection of DQ dimensions and DQ factors
3-Definition of DQ metrics
4-Implementation of DQ methods

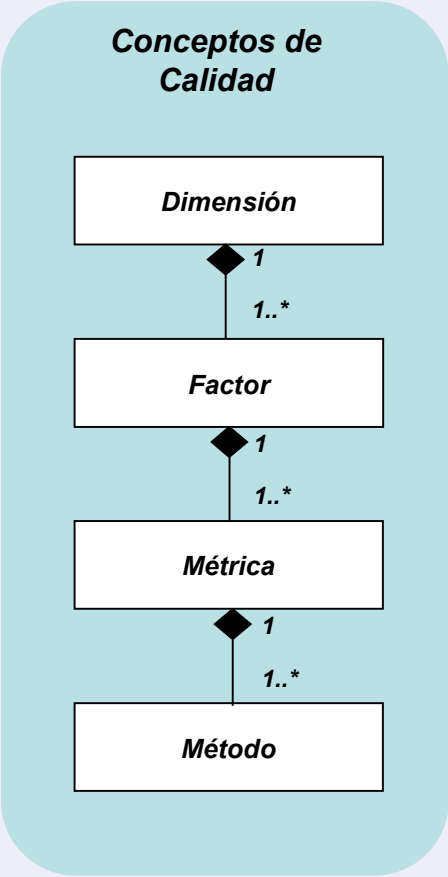
ST5 - DQ measurement
1-Design of the DQ metadata database
2-Execution of the DQ measurement
3-Storage of the DQ measurement results
4-Update of the context model definition

ST6 - DQ assessment
1-Definition of the DQ assessment approaches
2-Execution of the DQ assessment approaches
3-Store DQ assessment results

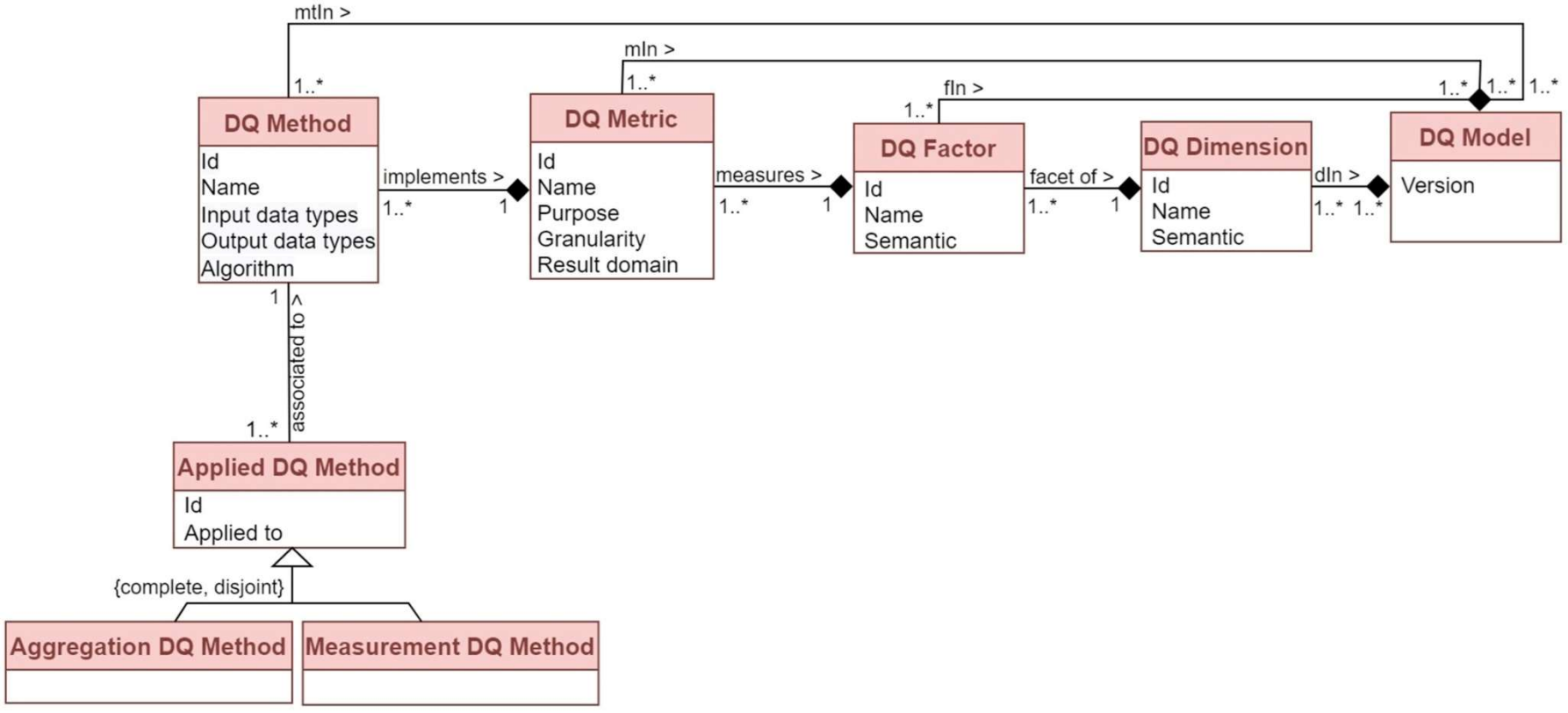


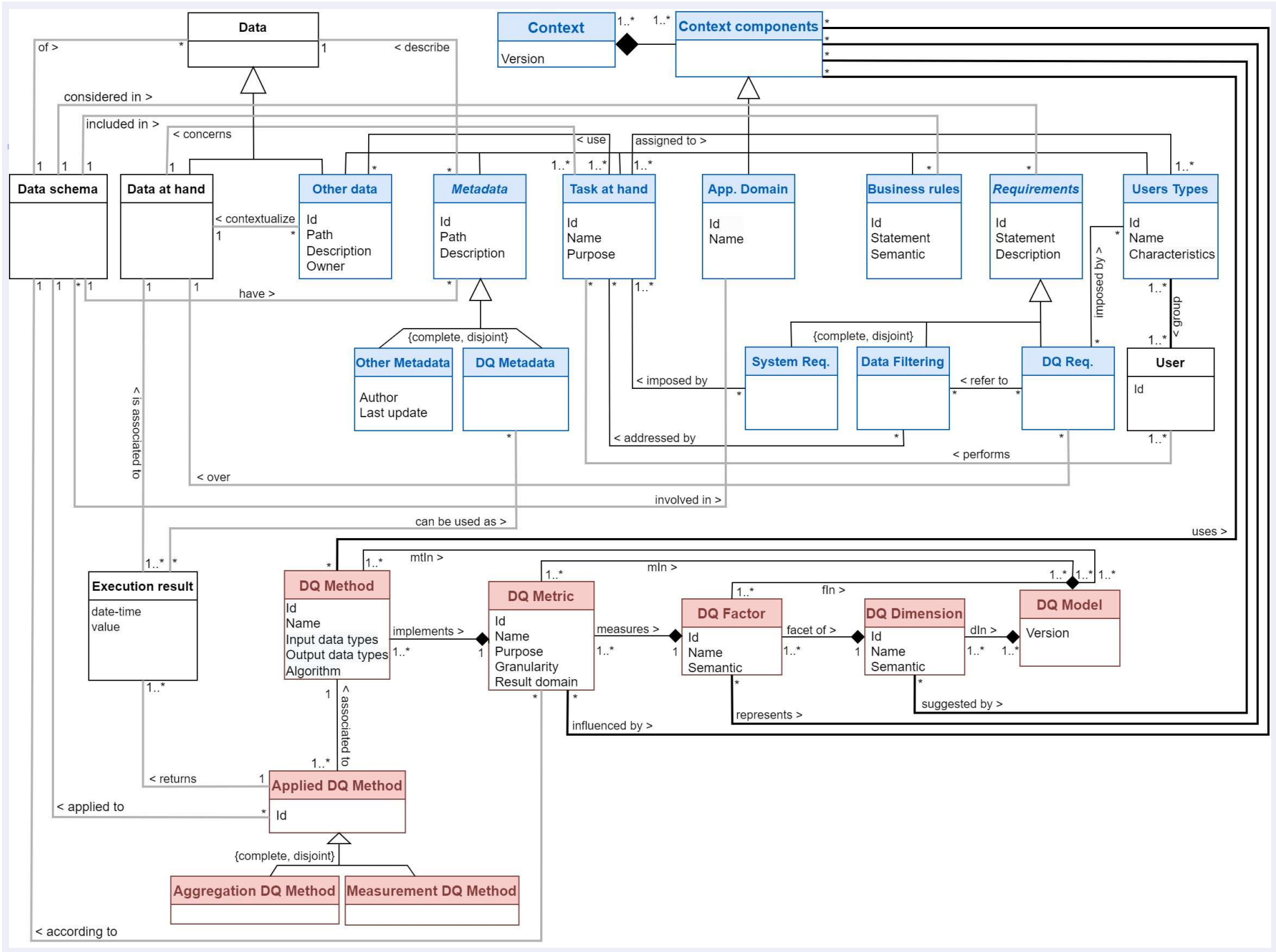
Actividades en ST4 – DQ Model definition

- 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.



Metamodelo de Calidad de Datos




Data
of >
< describe
considered in >
included in >
< concerns
< use
assigned to >
Context
Version
Context components
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
implements >
DQ Metric
Id
Name
Purpose
Granularity
Result domain
mtIn >
mIn >
fIn >
measures >
DQ Factor
Id
Name
Semantic
facet of >
DQ Dimension
Id
Name
Semantic
dIn >
DQ Model
Version
represents >
suggested by >
influenced by >
< associated to
< returns
Applied DQ Method
Id
< applied to
{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 = {}	ID: Name: Description: Represents = {}	ID: Name: Description: Influenced by = {} Granularity: Result domain:	ID: Name: Description: Uses = {} 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 = {} Granularity: Result domain:	ID: Name: Description: Uses = {} Input data types: Output data types: Algorithm:	ID: Type: Description: Applied to:

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	

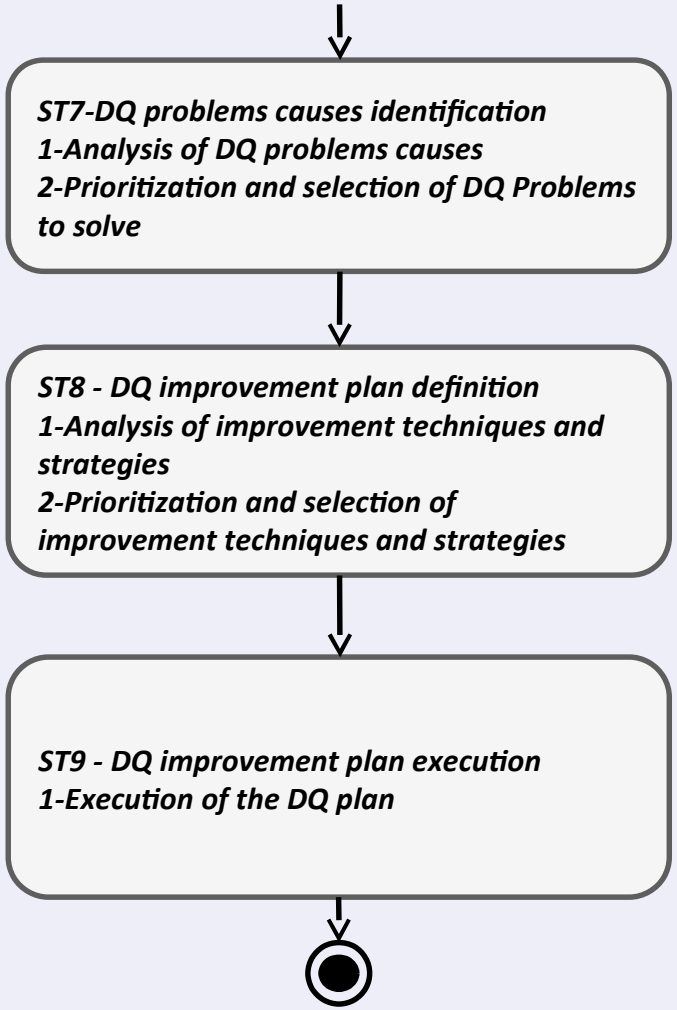
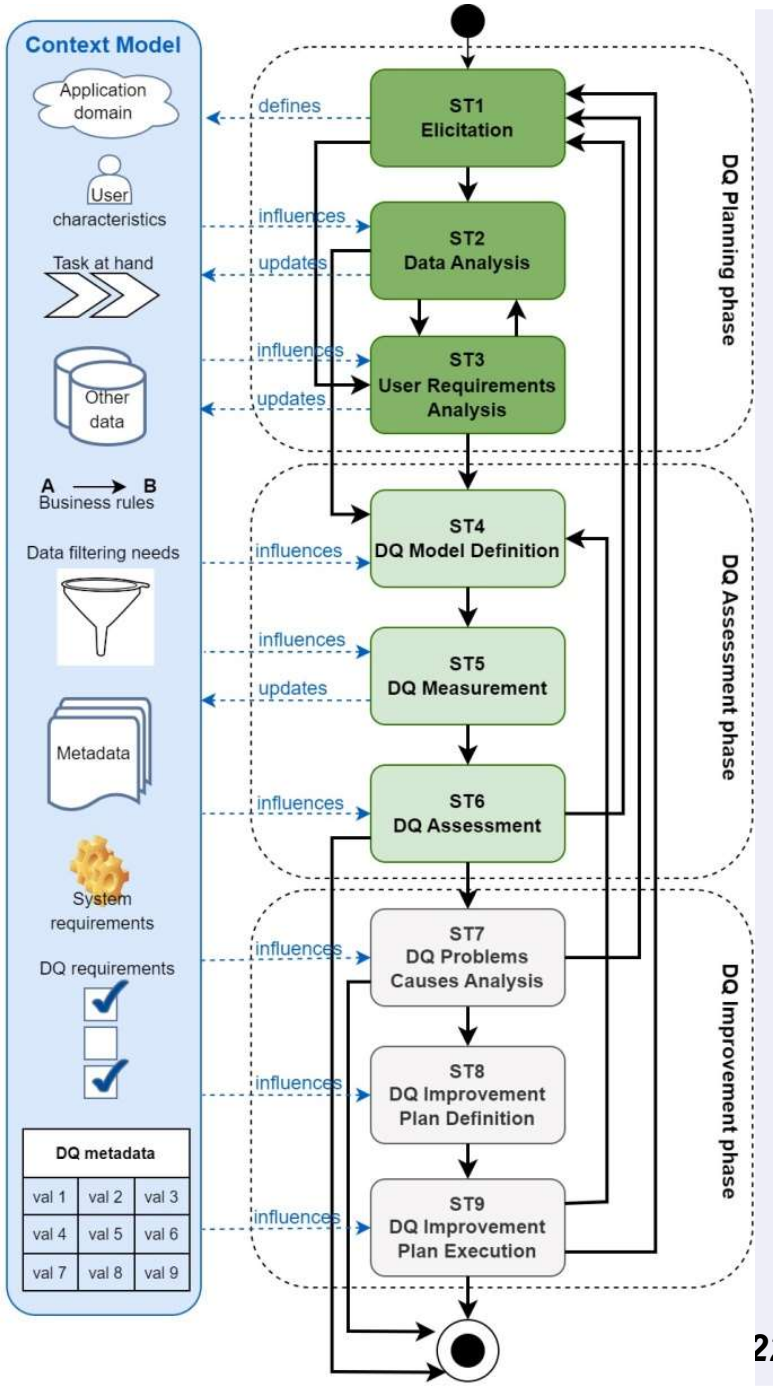
- ***Reporte de problemas de CD priorizados***
  - ***Incluye a todos los problemas de CD identificados en las etapas ST1, ST2 y ST3. A cada uno de estos problemas se les asigna una prioridad, por ejemplo: alta, media, baja.***

- ***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.***

ST5 – DQ Measurement

Entradas	Salidas
Reporte de problemas de CD priorizados	Especificación de la BD de metadatos de CD
Modelo de CD contextual	Reporte de medición de la CD
	Modelo de Contexto

- Especificación de la BD de metadatos de CD:
  - La información sobre las entidades de la BD y sus relaciones. Si la base de datos es estructurada o semiestructurada el esquema asociado
- Reporte de medición de la CD:
  - Este informe especifica los sistemas, herramientas y lenguajes de programación, entre otros, utilizados para llevar a cabo la medición de DQ.
- Modelo de Contexto:
  - Modelo de Contexto completo. Esto implica todos los componentes de contexto identificados en ST1, ST2, ST3 y ST5.

CaDQM

Phase 3: DQ Improvement