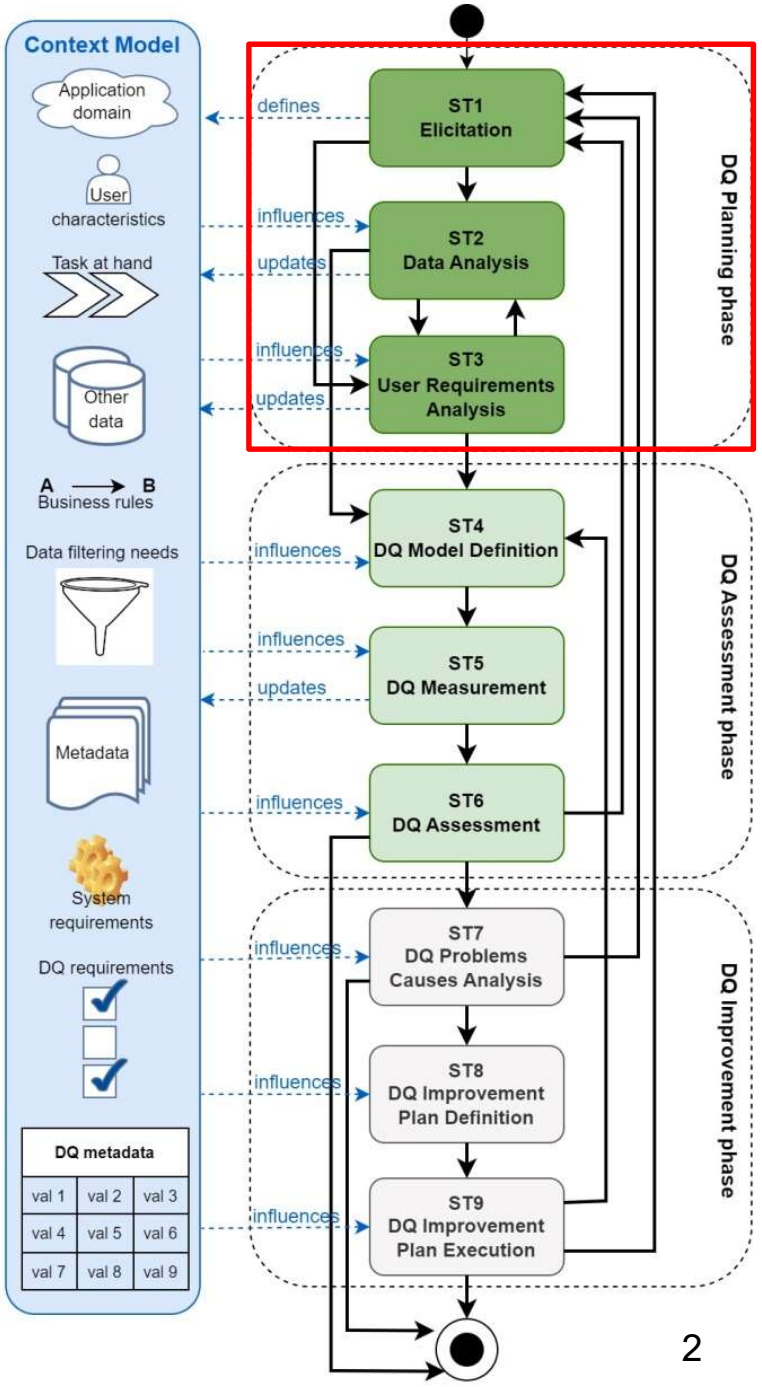
TAREA 1

CaDQM
Phase 1 – DQ Planning

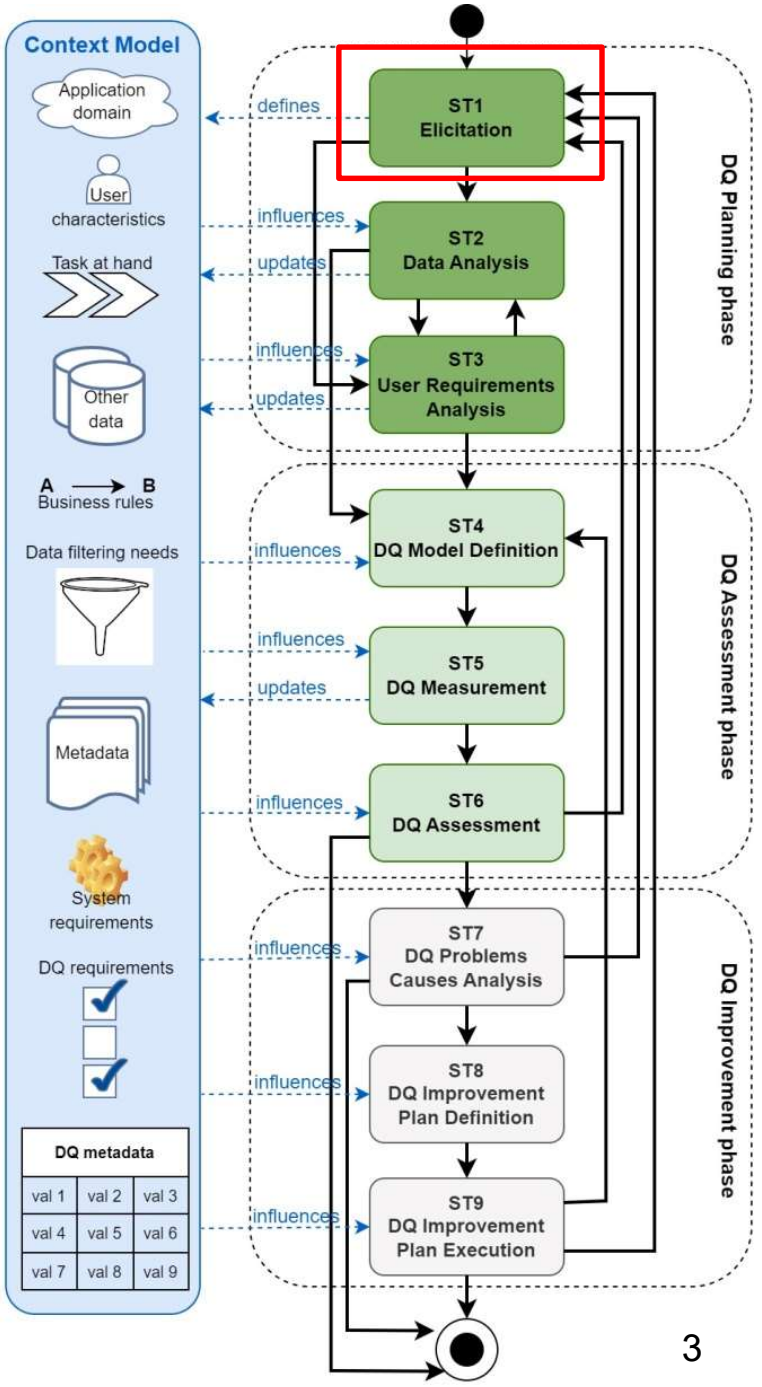
Phase 1 – DQ Planning

- Etapas
  - *ST1 – Elicitation*
  - *ST2 – Data Analysis*
  - *ST3 – User Requirements Analysis*



ST1 - Elicitation

- Actividades
  - *Selection of the Data at hand*
  - *Analysis of the Organization elements*
  - *Identification of DQ problems*
  - *Definition of the Context model*



Actividades en ST1 - Elicitation

- *Selection of the Data at hand*
  - *data at hand*: datos de NL, obtenidos al integrar:
    - *datasets books_data.csv* y *books_rating.csv* de L1
    - *books.csv, ratings.csv* y *users.csv* de L2

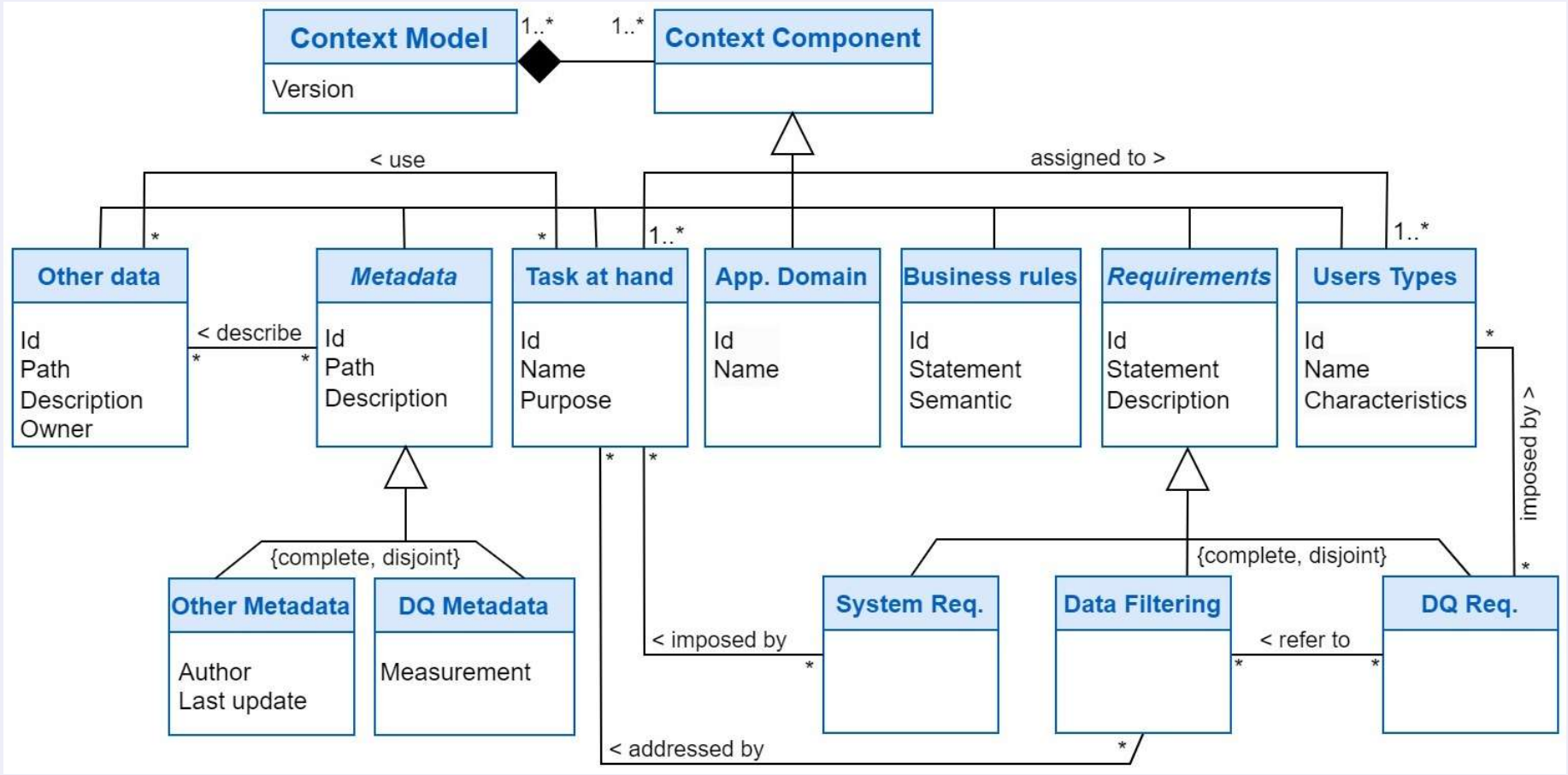
- *Analysis of the Organization elements*
  - Identificación y análisis de todos los elementos de la organización relacionados con los *data at hand*:
    - Análisis de la realidad planteada
    - Análisis de información identificada en la Web, relacionada con el dominio de aplicación
    - Descripción de los dataset integrados
    - Descripción del dataset obtenido en la integración (Por ejemplo, modelo relacional)
    - Identificación de usuarios, sus roles y tareas

Actividades en ST1 - Elicitation

- *Identification of DQ problems.*
  - Algunos problemas de CD:
    - Libros duplicados
    - Libros valorados que no forman parte de los datasets
    - Libros sin autores
    - Libros sin ISBN
    - Libros sin editores
    - Ratings con escalas diferentes
    - Campos con valores nulos
    - Fechas con distintos formatos
    - Etc, etc, etc….

Definition of the Context model

Metamodelo de Contexto



Definition of the Context model

- Componentes de Contexto
  - Dominio de Aplicación (AD): Librería
  - Metadatos (M): descripción de los datasets obtenida en
    - [1] https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews/data?select=books_data.csv
    - [2] https://www.kaggle.com/datasets/saurabhbagchi/books-dataset/data
  - Usuarios:
    - U1, Administrador
    - U2, Publicista digital
    - U3, Analista de datos
  - Tareas:
    - U1: Administración y gestión de la librería
    - U2: Recomendación de libros y promoción de la librería
    - U3: Análisis de datos

Definition of the Context model

- Componentes de Contexto
  - Filtrado de datos:
    - DF1: Libros cuya publicación sea del año actual
    - DF2: top 3 de los libros con mayor score
    - DF3: libros editados por Wiley
  - Reglas de Negocio:
    - BR1: Cada libro debe tener un ISBN
    - BR2: Cada libro debe tener, al menos, 1 autor
    - BR3: Cada libro debe tener, al menos, 1 título
    - BR4: Cada libro debe tener, al menos, 1 editor
    - BR5: La librería debe ofrecer, al menos, 500 libros
    - BR6: La librería debe ofrecer, al menos, el 20% de los 100 mejores libro propuestos por *Goodreads*
  - Otro datos (OD): *Goodreads*
    https://www.goodreads.com/list/show/2681.Time_Magazine_s_All_Time_100_Novels

Definition of the Context model

- Componentes de Contexto
  - Requerimientos de CD:
    - DQR1: Es necesario que el sitio Web se actualice cada viernes
    - DQR2: Los libros con score mayor a 5 deben superar el 60%
    - DQR3: Más del 95% de ISBN no nulos
    - DQR4: Más del 95% de los títulos escritos correctamente
    - DQR5: Más del 95% de los nombres de autores escritos correctamente
    - DQR6: Los usuarios que califican deben ser mayores de 18 años
  - Requerimientos del sistema:
    - SR: los tiempos de respuesta del sitio Web no deben superar los 3 segundos

ST1 - Elicitation

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

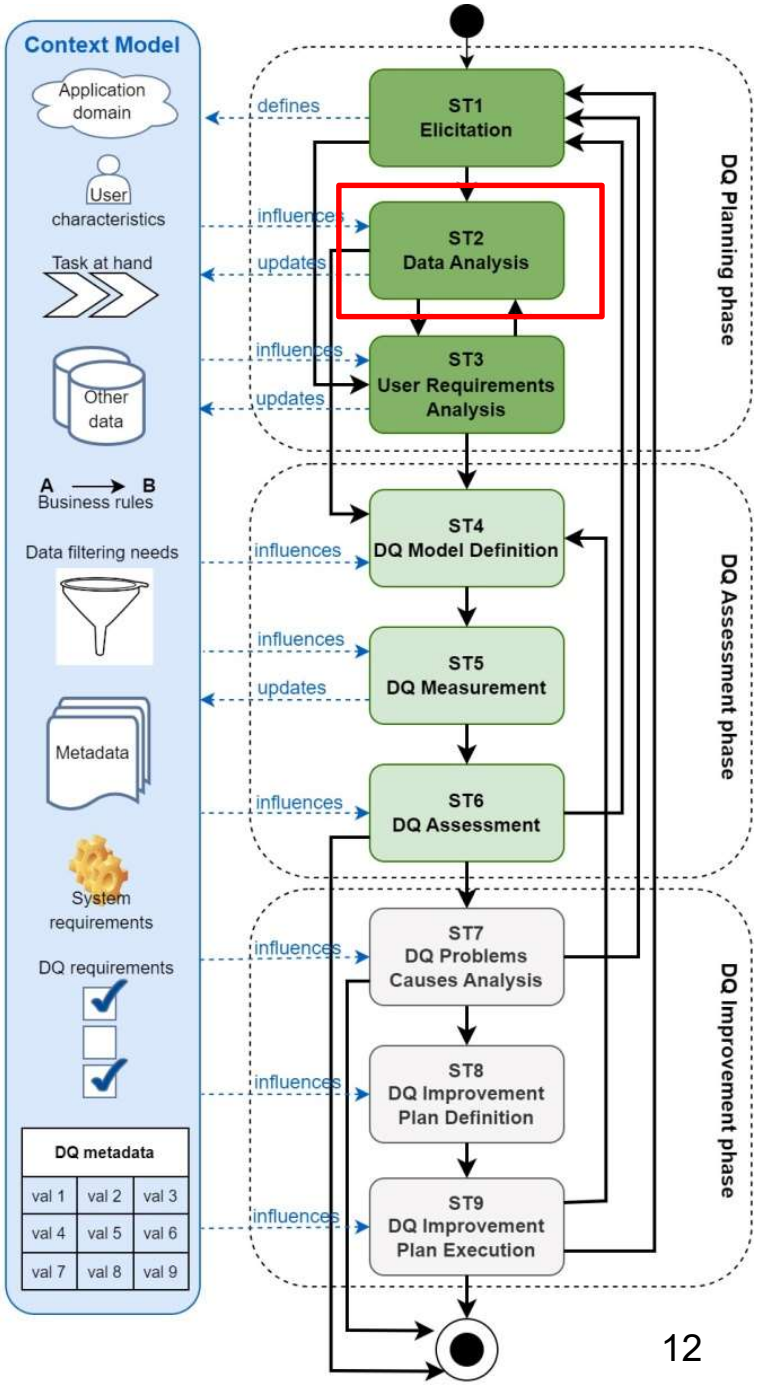
- Data at hand
  - Dataset de NL
- Reporte con problemas de CD
  - Listado de problemas de CD identificados en ST1.
- Modelo de Contexto
  - Listado de componentes de contexto identificados en ST1.

Modelo de Contexto en ST1

Comp. de CTX.	Usuarios			
	Todos	**U1**	**U2**	**U3**
Dominio de Aplicación	AD			
Tareas		T1	T2	T3
Reglas de negocio	BR1, BR2, BR3, BR4, BR5, BR6			
Req. Sistema				
Req. CD	SR		DQR1, DQR2, DQR6	DQR3, DQR4
Filtrado de datos		DF1, DF2, DF3		
Metadatos	M			
Metadatos de CD	---			
Otros datos	OD			

ST2 – Data Analysis

- Actividades
  - *Data profiling*
  - *Identification of DQ problems*
  - *Estimation of DQ*
  - *Update of the context model definition*



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*
  - resultados obtenidos en el análisis de datos
  - componentes de contexto considerados en el análisis de datos
- Reporte con problemas de CD
  - Listado de problemas de CD identificados en ST2, **incluyendo** los problemas reportados en ST1/ST3
- Modelo de Contexto
  - Listado de componentes de contexto identificados en ST2, **incluyendo** los componentes de contexto identificados en ST1/ST3

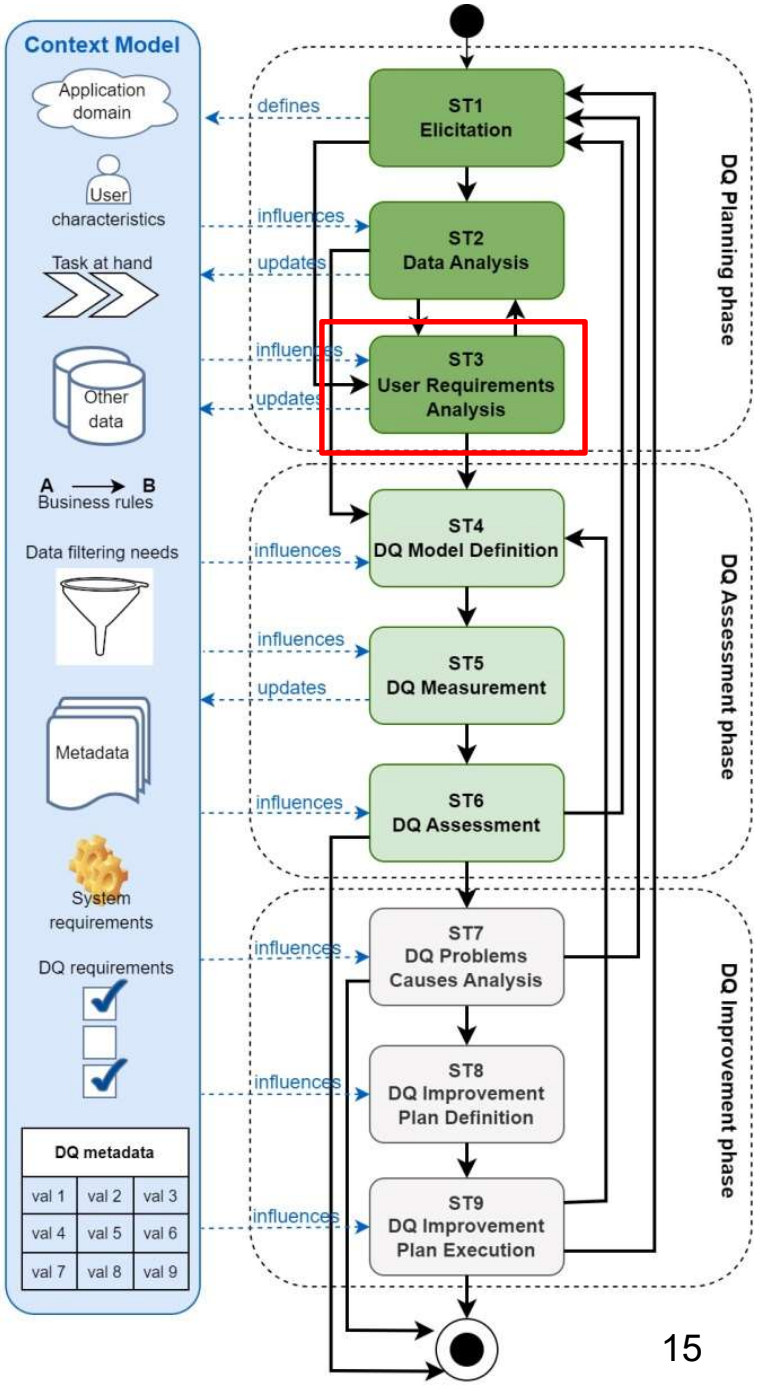
Modelo de Contexto actualizado en ST2

Comp. de CTX.	Usuarios			
	Todos	**U1**	**U2**	**U3**
Dominio de Aplicación	AD			
Tareas		T1	T2	T3
Reglas de negocio	BR1, BR2, BR3, BR4, BR5, BR6, BR7			
Req. Sistema				
Req. CD	SR		DQR1, DQR2	DQR3, DQR4
Filtrado de datos		DF1, DF2, DF3		
Metadatos	M			
Metadatos de CD	---			
Otros datos	OD			

Nota: Una buena práctica es diferenciar con color los componentes de contexto identificados en ST2.

ST3 – User requirements analysis

- Actividades
  - *Interaction with data users*
  - *Identification of DQ problems*
  - *Update of the context model definition*



ST3 – User requirements analysis

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

- Reporte del análisis de requerimientos de usuarios
  - descripción de las herramientas y técnicas utilizadas en el análisis de requerimientos de usuarios
  - resultados obtenidos en el análisis de requerimientos de usuarios
  - componentes de contexto considerados en el análisis de datos
- Reporte con problemas de CD
  - Listado de problemas de CD identificados en ST3, **incluyendo** los problemas reportados en ST1/ST2
- Modelo de Contexto
  - Listado de componentes de contexto identificados en ST3, **incluyendo** los componentes de contexto identificados en ST1/ST2

Modelo de Contexto actualizado en ST3

Comp. de CTX.	Usuarios			
	Todos	**U1**	**U2**	**U3**
Dominio de Aplicación	AD			
Tareas		T1	T2	T3
Reglas de negocio	BR1, BR2, BR3, BR4, BR5, BR6, BR7			
Req. Sistema				
Req. CD	SR		DQR1, DQR2	DQR3, DQR4
Filtrado de datos		DF1, DF2, DF3		
Metadatos	M			
Metadatos de CD	---			
Otros datos	OD			

Nota: Una buena práctica es diferenciar con color los componentes de contexto identificados en ST3.

A tener en cuenta

- La integración no forma parte de la metodología
- Describir el dataset integrado
- Identificar problemas de calidad concretos
- Identificar todos los componentes de contexto
  - no incluyen componentes, como por ejemplo el dominio de aplicación
  - confunden tareas y problemas con requerimientos

- Respetar el Metamodelo de Contexto
  - agregan componentes que no propone el Metamodelo de Contexto

- Seguir la metodología estrictamente:
  - aplicando las actividades de cada una de las etapas (no superficialmente)
  - ST2 no solo implica verificar reglas de negocio y requerimientos de CD
  - completando las etapas que se abordan (ej: si se plantean preguntas en ST3, entonces se abordan)
  - presentando las entradas/salidas de cada etapa de forma completa (ej: el contexto se actualiza en cada etapa)

- Todo lo que se pide en la tarea se debe realizar
  - Conclusiones: siempre al final y no forman parte de la metodología, sino de la Tarea.