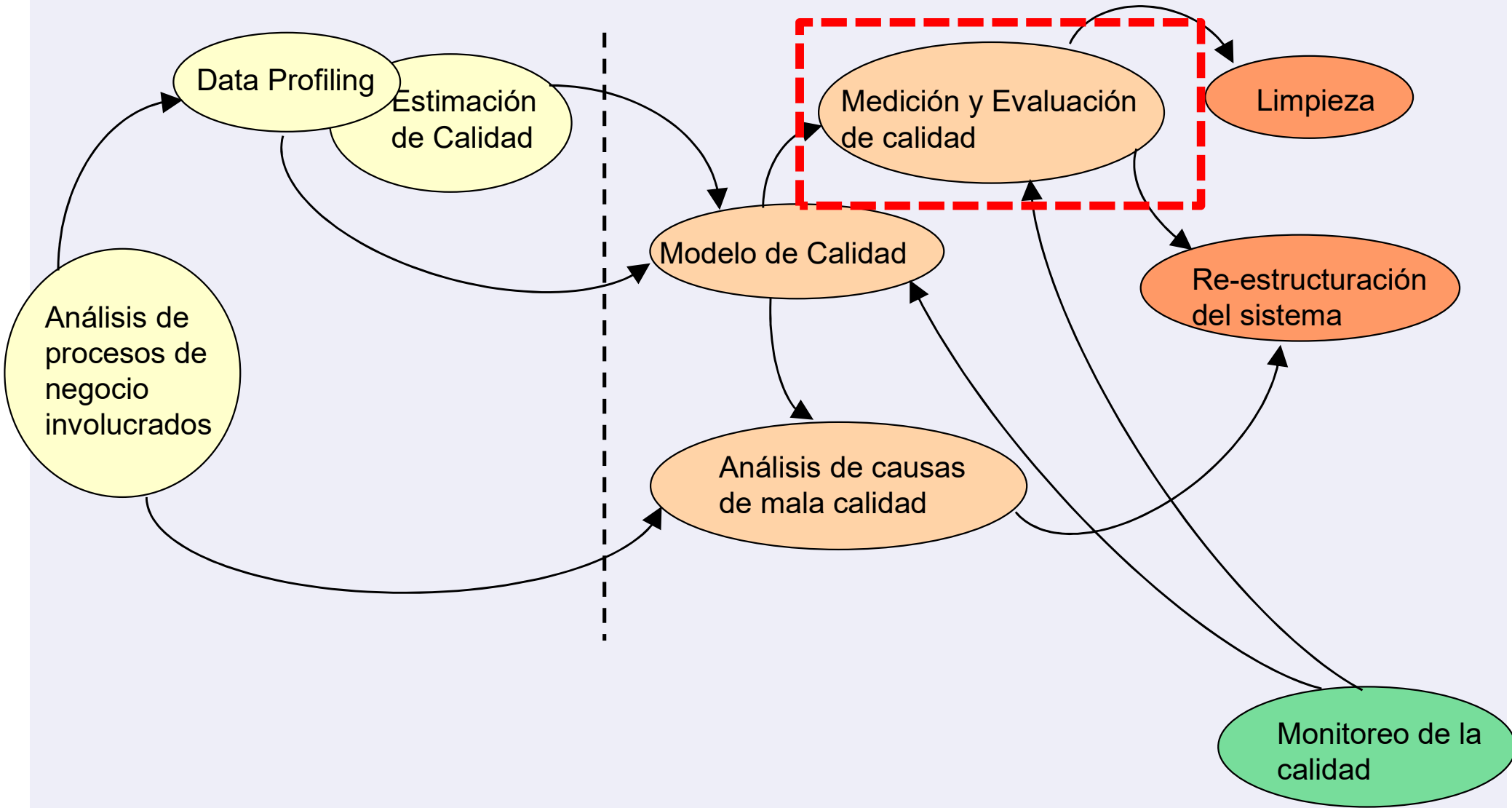

Calidad de Datos e Información

CaDQM

Phase 2 – DQ Assessment

Gestión de la calidad en SI



Medición de Calidad de Datos

- Para qué medimos?
 - Para poder brindar al usuario información acerca de la calidad de los datos que utiliza
 - Para poder mejorar la calidad de los datos
 - Para poder analizar el costo de mejorar la calidad
- Medición
 - Comparación cuantitativa entre una observación y un valor de referencia

Medición de Calidad de Datos

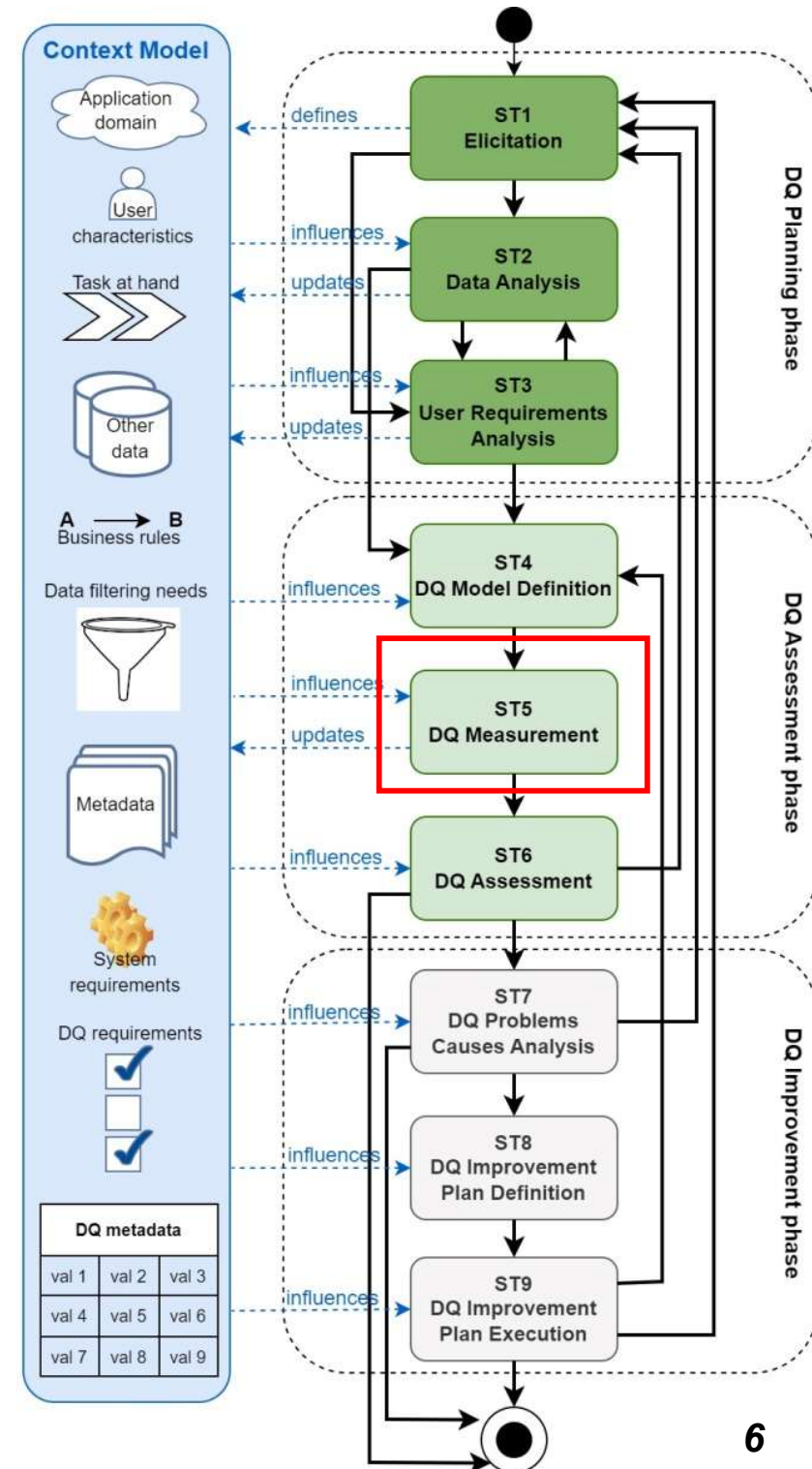
- La forma en que se mide la calidad de los datos es muy variable, dependiendo de:
 - La dimensión/factor/métrica de calidad
 - Sobre qué se va a ejecutar la medición
 - Sobre datos estructurados
 - Sobre datos semi-estructurados
 - Sobre datos no estructurados
 - Momento de la medición con respecto al uso de los datos
 - Off-line
 - On-line

Medición de Calidad de Datos

- Para realizar una medición, debe estar previamente definido:
 - Modelo de Calidad
 - Procedimiento de medición
 - Muestreo?
 - Implementación y ejecución de la medición
 - Modelo de datos para los valores de calidad obtenidos:
 - **Metadatos de Calidad**

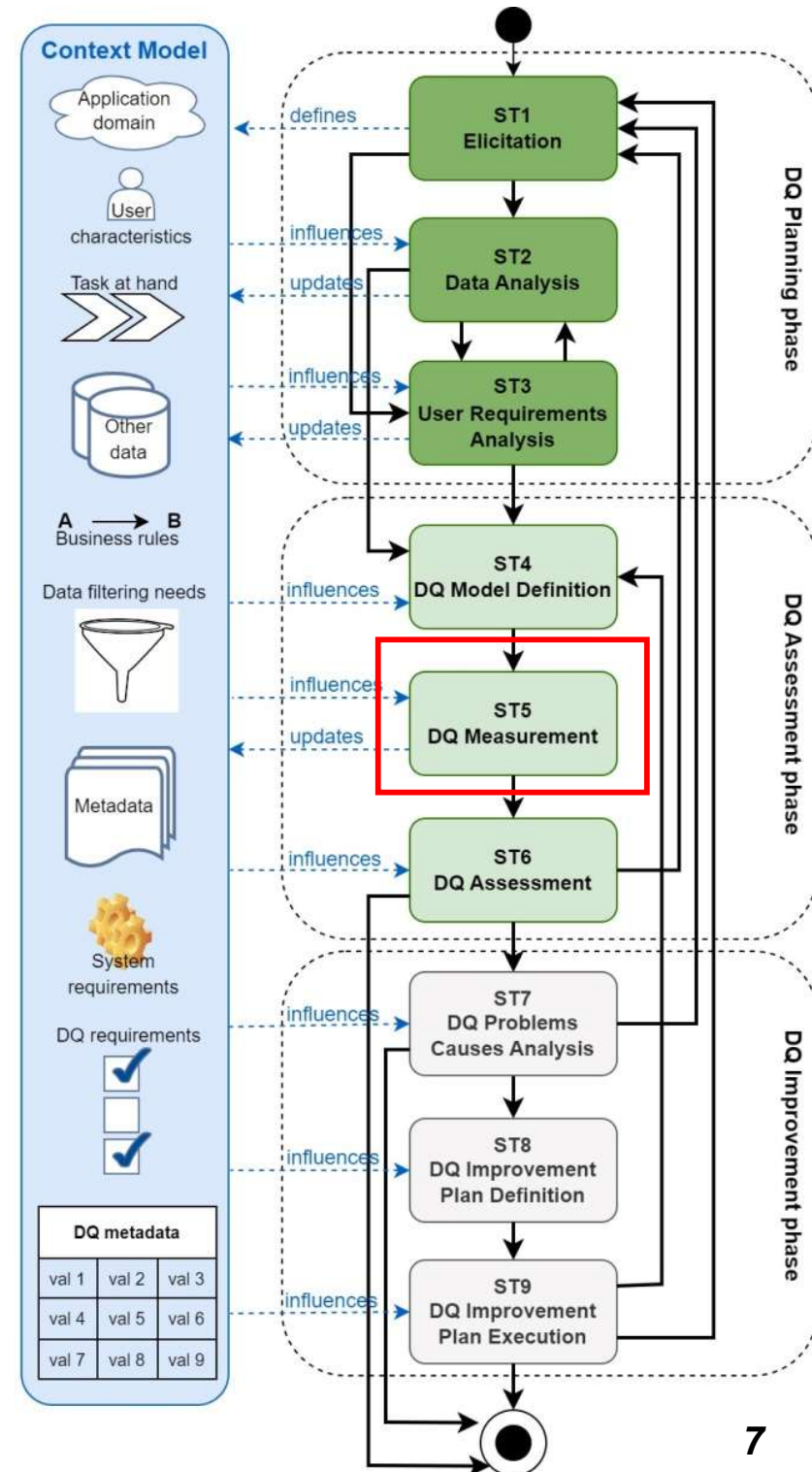
ST5 – DQ Measurement

- Actividades
 - Design of the DQ metadata database
 - Execution of the DQ measurement
 - Storage of measurement results in the DB
 - Update of the context model definition



ST5 – DQ Measurement

- Actividades
 - Design of the DQ metadata database
 - Execution of the DQ measurement
 - Storage of measurement results in the DB
 - Update of the context model definition



Design of the DQ metadata database

- Diseño de la BD donde se almacenan los resultados de las mediciones y evaluaciones:
 - estructurada o semi-estructurada
 - estrechamente relacionada con el Modelo de Contexto y el modelo de CD contextual
- Para cada dato cuya calidad se mide, se guardan:
 - dimensión, el factor, la métrica y el método de DQ aplicados
 - este diseño también debe considerar los resultados cualitativos de la evaluación
- **Recomendación:** la BD de metadatos de calidad y la BD donde se almacenan los datos evaluados deben ser totalmente independientes

Metadatos de Calidad

- Utilizamos **modelos de datos** para representar **datos**

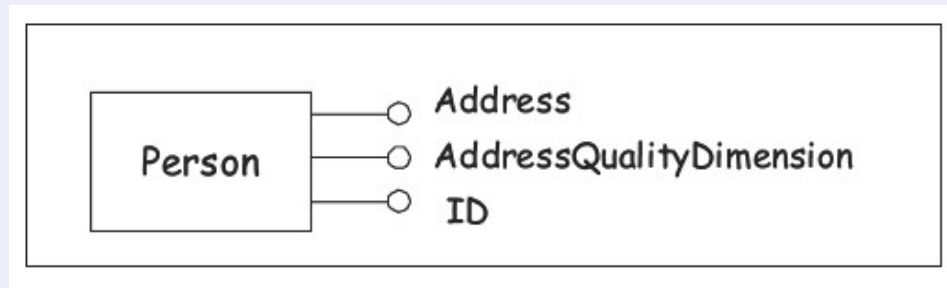
- Queremos además poder representar sus **dimensiones de calidad y sus medidas de calidad**. A esto le llamamos **METADATOS DE CALIDAD**.
 - Se proponen extensiones a los modelos tradicionales para bd, para representar y manejar aspectos relacionados con las dimensiones de calidad.
 - Se enriquecen los modelos convencionales con elementos para representar y analizar la calidad de los datos.

Metadatos de Calidad

- Modelos de datos
 - Modelado conceptual
 - Extensión del MER para representar los metadatos de CD
 - Modelado lógico
 - Esquema en Modelo Relacional para almacenar los metadatos de CD

Extensión del Modelo Entidad-Relación

- Una posible solución

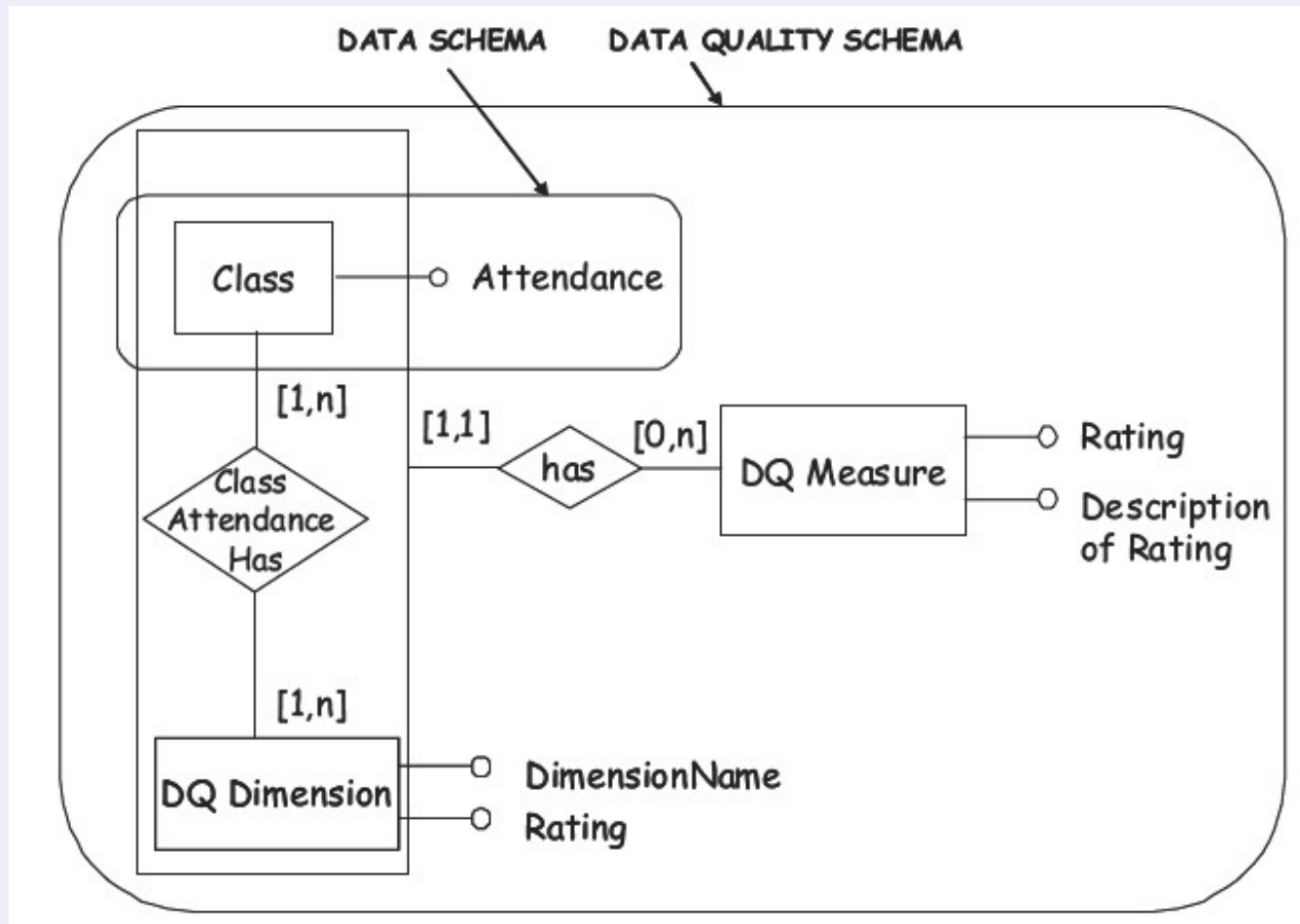


Desventajas?

◆ Otra solución

- Agregamos 2 nuevas entidades:
 - DQ dimension: representa cada dimensión y todos sus posibles valores
 - DQ measure: representa las mediciones

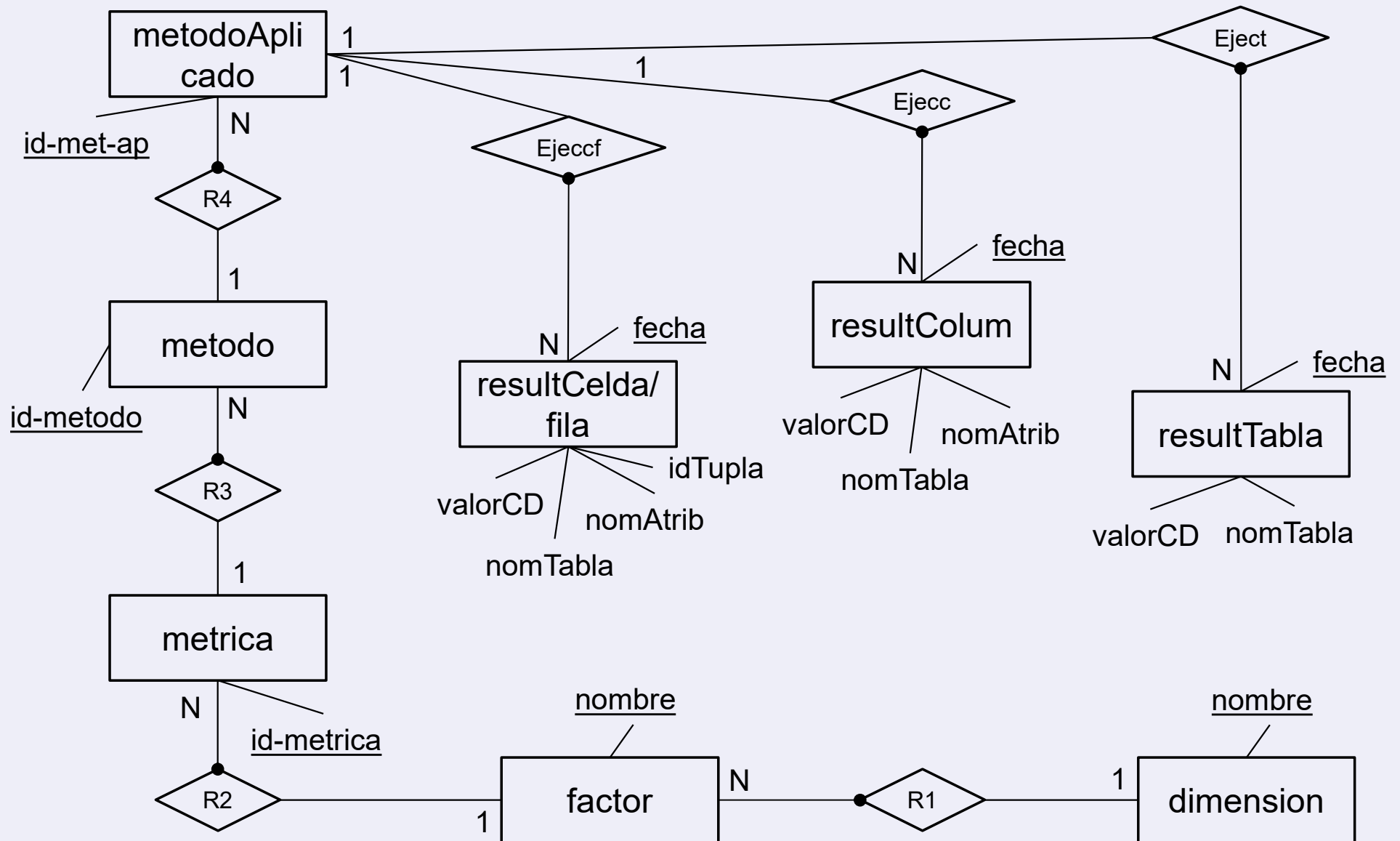
Extensión del Modelo Entidad-Relación



Qué cosas falta representar?

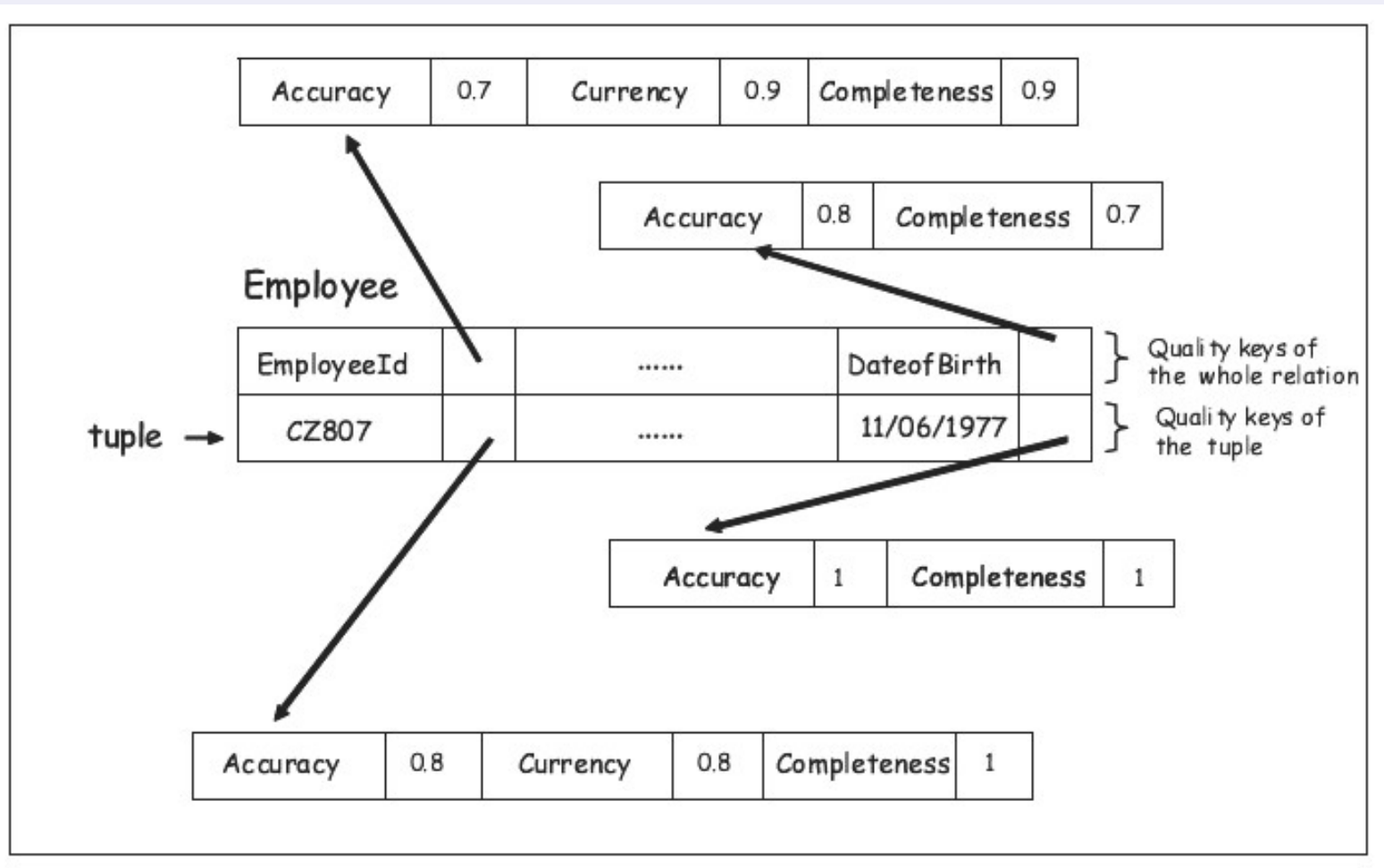
Qué construcciones habría que agregarle?

MER Metadatos de CD - Ejemplo



Extensión del Modelo Relacional

- Basado en atributos



Ejercicio – Metadatos de Calidad

- Base de Datos
 - Clientes (ci, nom, dir, tel, fnac, sexo, categoria)
 - Productos (cod, pres, desc, prov, cant-stock)
 - Ventas (ci, cod, pres, fecha, cantidad, importe, sucursal)
- Parte del Modelo de Calidad de Datos:

Dimensión	Factor	Métrica	Método aplicado
Exactitud	Exactitud Sintáctica	M1: VerifFormato gran: celda tipo-res: {0,1}	Clientes.ci Clientes.sexo Productos.prov Ventas.importe
	Precision	M2: CantDecim gran: columna tipo-res: {0,1}	Productos.cant-stock Ventas.importe
Compleitud	Cobertura	M3: CoberturaRef gran: tabla tipo-res: [0,1]	Clientes

Metadatos en BDs documentales (json)

Colección de películas

```
{
  "_id": 1,
  "titulo": "Isla de Perros",
  "puntaje": {
    "imdb": 7.9,
    "rottenTomatoes": 90,
    "metacritic": 82
  },
  "elenco": [
    {
      "nombre": "Bryan Cranston",
      "personaje": "Chief"
    },
    {
      "nombre": "Koyu Rankin",
      "personaje": "Atari"
    },
    {
      "nombre": "Edward Norton",
      "personaje": null
    }
  ]
}
```

```
{
  "_id": 2,
  "titulo": "Memento",
  "puntaje": {
    "imdb": 8.4,
    "rottenTomatoes": 92,
    "metacritic": null
  },
  "genero": "Comedia",
  "director": "C. Nolan"
}
```

```
{
  "_id": 3,
  "titulo": "Amores Perros",
  "puntaje": {
    "imdb": 8.1
  },
  "genero": null,
  "director": null
}
```

Mediciones

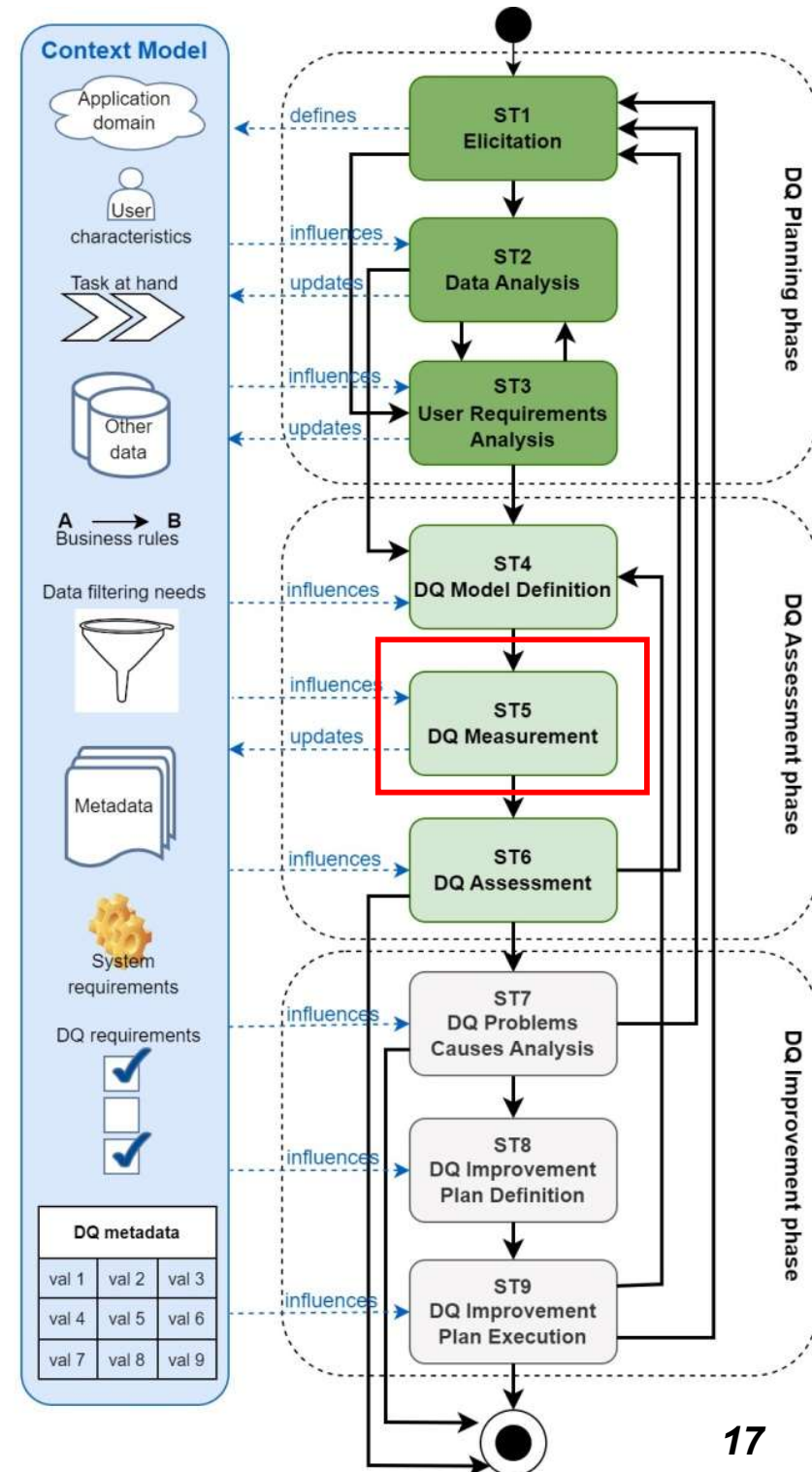
```
{
  "target": {
    "collection": "movies",
    "id": "movie_1"
  },
  "value": true,
  "execution_id": "ejecucion_1"
}
```

```
{
  "target": {
    "collection": "movies",
    "id": "movie_2"
  },
  "value": false,
  "execution_id": "ejecucion_1"
}
```

```
{
  "target": {
    "collection": "directors"
  },
  "value": 0.9,
  "execution_id": "ejecucion_2"
}
```


ST5 – DQ Measurement

- Actividades
 - Design of the DQ metadata database
 - Execution of the DQ measurement
 - Storage of measurement results in the DB
 - Update of the context model definition

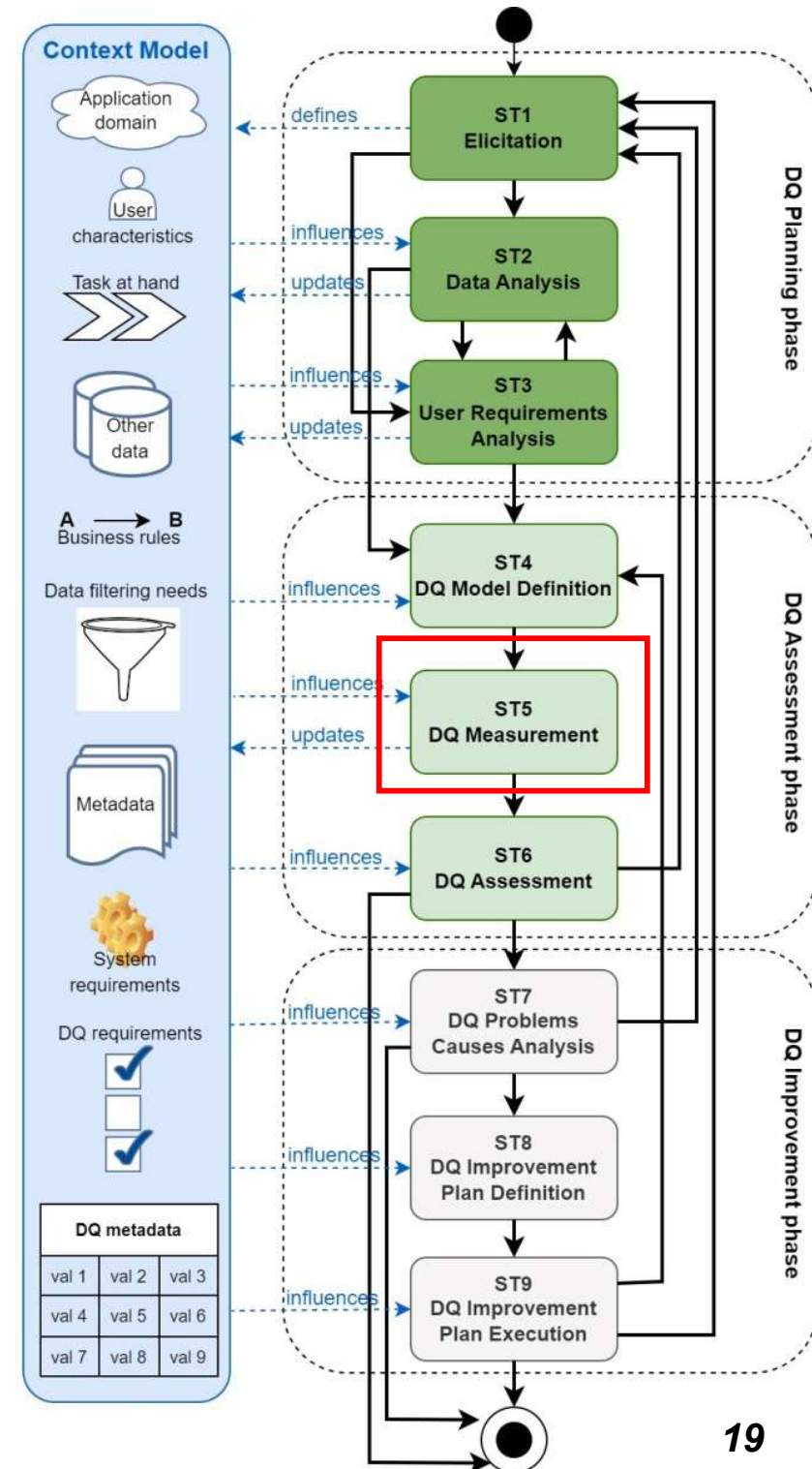


Execution of the DQ measurement

- La medición de CD se realiza mediante la ejecución de los métodos implementados para cada métrica.
- Los valores de calidad resultantes constituyen **Metadatos de Calidad**.

ST5 – DQ Measurement

- Actividades
 - Design of the DQ metadata database
 - Execution of the DQ measurement
 - Storage of measurement results in the DB
 - Update of the context model definition

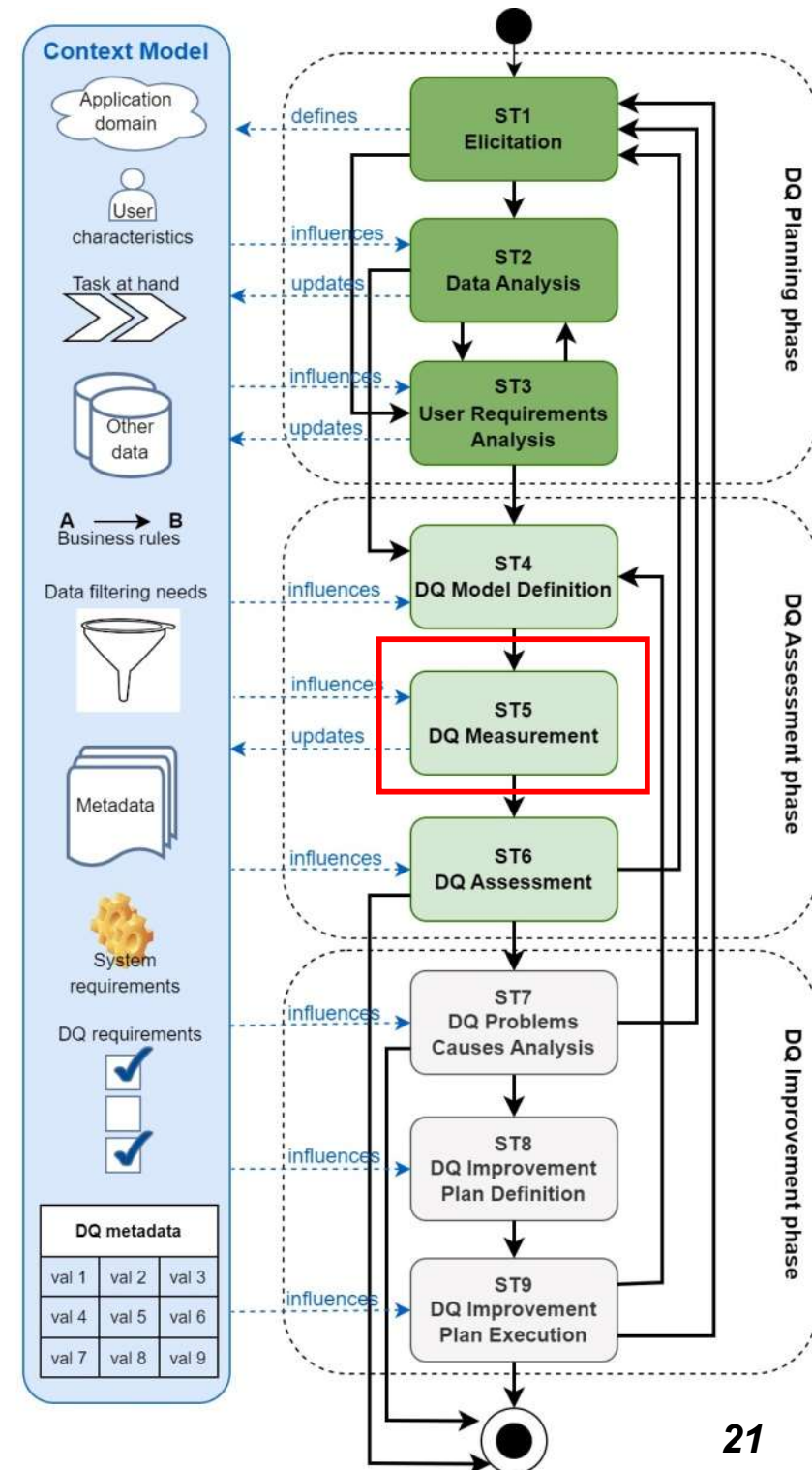


Storage of measurement results in the DB

- Cada valor de CD obtenido en la medición se almacena en la BD de metadatos de calidad.
- También se registran la dimensión, el factor, la métrica y el método de CD, si no se registraron previamente.

ST5 – DQ Measurement

- Actividades
 - Design of the DQ metadata database
 - Execution of the DQ measurement
 - Storage of measurement results in the DB
 - Update of the context model definition



Update of the Context Model definition

- Los **metadatos de CD** se agregan al Modelo de Contexto como **un nuevo componente de contexto**.
- Estos componentes completan la definición del Modelo de Contexto en la ejecución de CaDQM.
- Las mediciones de CD de ciertos datos pueden utilizarse para contextualizar otros datos:
 - Ejemplo: métricas que combinan subconjuntos de estos metadatos de DQ para obtener medidas combinadas.

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.