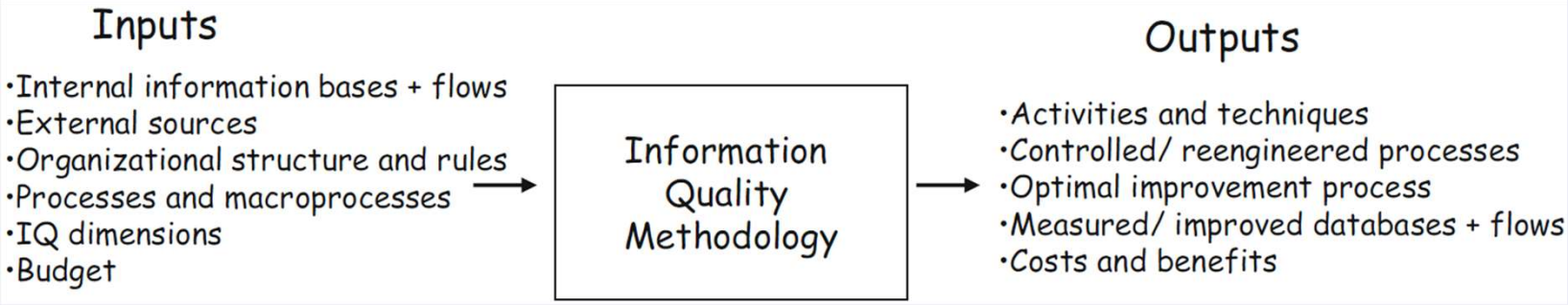
Calidad de Datos e Información

Metodologías

Introducción

- Qué es una metodología de Calidad de Datos (CD)?
  - *A DQ methodology provides a set of guidelines and techniques that, from input information that describes a given application context, defines a rational process to assess and improve the quality of data* [1].
  - Entradas y salidas de una metodología de CD para medición y mejora [2]:



Introducción

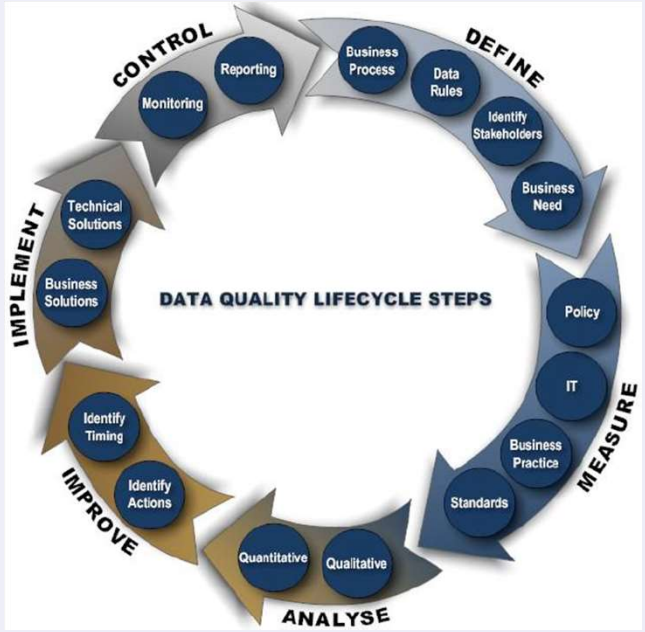
- Cómo se clasifican las metodologías? [2]:
  - Información vs. Proceso:
    - Información: se basan exclusivamente en el uso de fuentes de datos para mejorar la CD.
    - Proceso: los procesos de producción de datos se analizan y modifican para identificar y eliminar las causas de los problemas de CD.
  - Evaluación vs. Mejora:
    - Las actividades de evaluación y mejora están fuertemente relacionadas.
    - Ambas son posibles a partir de las mediciones de CD.

Introducción

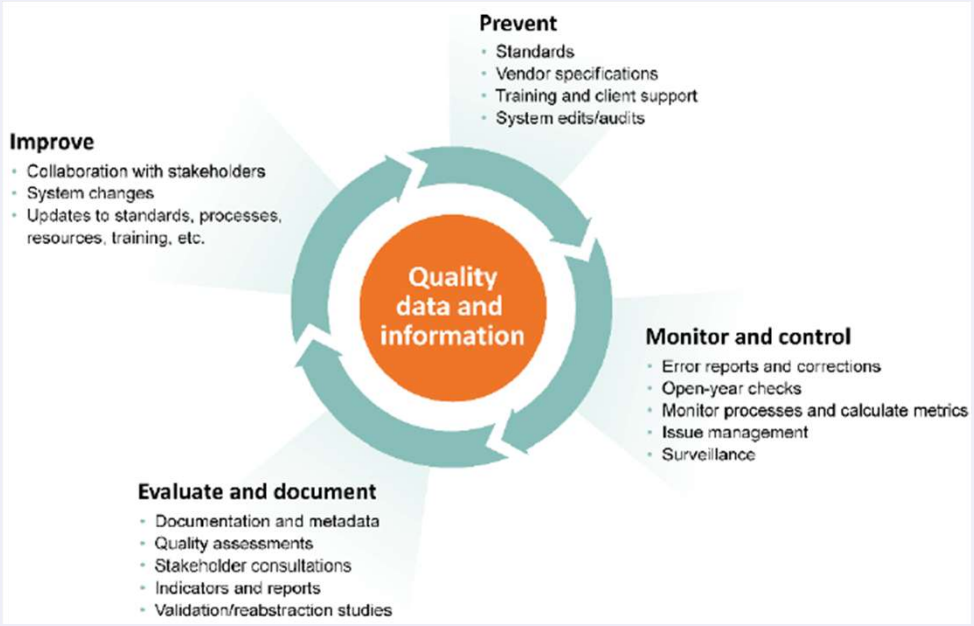
- Cómo se clasifican las metodologías? [2]:
  - Propósito general vs. Propósito específico:
    - Propósito general: cubre un amplio espectro de fases, dimensiones de CD y actividades.
    - Propósito específico: se centra en una actividad particular (ej. medición), en un dominio de información específico (el. un censo, un registro de direcciones de personas) o en dominios de aplicación específicos (ej. biología).
  - Intra-organizacional vs. Inter-organizacional:
    - Intra-organizacional: las actividades de CD conciernen a una organización específica, o a un sector específico de la organización, o incluso a un proceso o base de información específica.
    - Inter-organizacional: se trata de un grupo de organizaciones (ej. un conjunto de instituciones públicas).

Introducción

- De qué se compone una metodología?
  - Se compone de varias etapas, organizadas en fases [1], y cada etapa implica un conjunto de actividades.
  - Ejemplos:



Data Quality Framework.
Nueva Zelanda. 2008



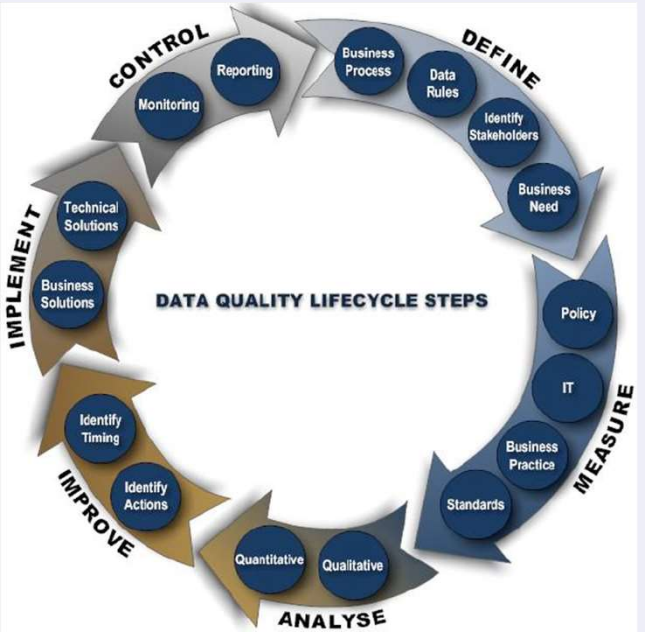
Information Quality Framework.
Canadian Institute for Health Information. 2017

Fases y etapas de una metodología

- Fases comunes y básicas de las metodologías de CD:
  - Evaluación y mejora [1].

- Etapas básicas de evaluación:
  - Análisis (ej. datos, requerimientos de CD)
  - Identificación de áreas críticas
  - Modelado de procesos
  - Medición de la calidad, etc.

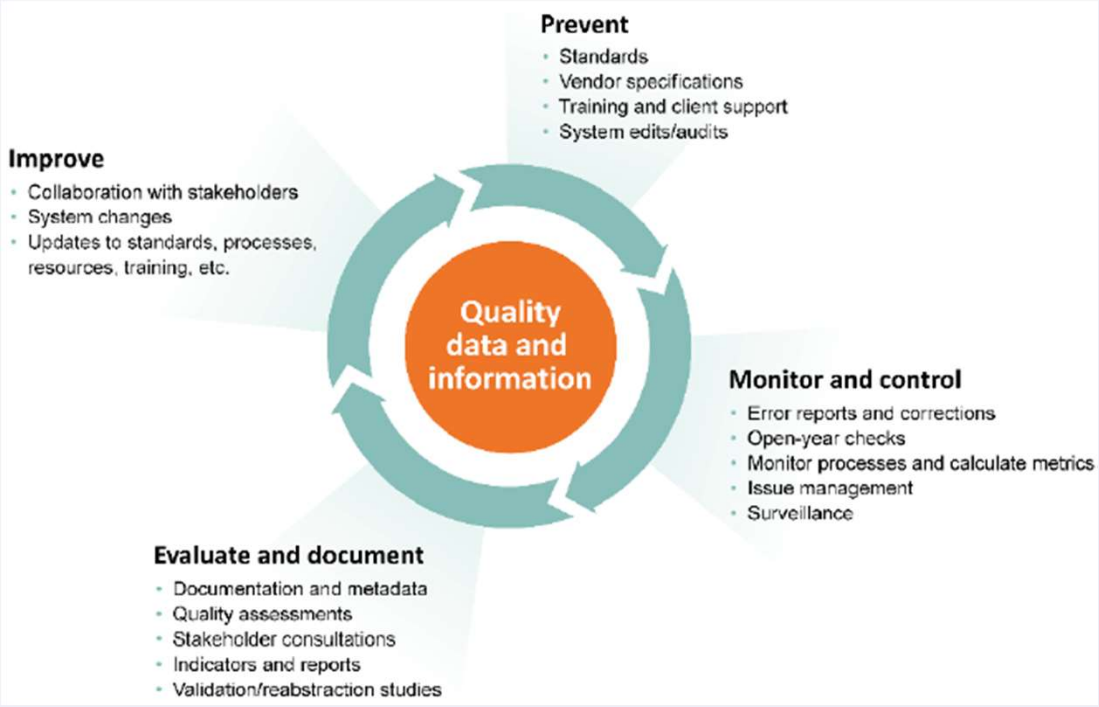
- Etapas básicas de mejora:
  - Evaluación de costos
  - Asignación de responsabilidades
  - Identificación de las causas de los errores.
  - Selección de estrategias y técnicas.
  - Gestión de mejora, etc.



Ejemplo: Data Quality Framework. Nueva Zelanda. 2008

Actividades abordadas por una metodología

- En las etapas de una metodología se realizan actividades de gestión de calidad de datos.



Ejemplo: Information Quality Framework.
Canadian Institute for Health Information. 2017

Actividades abordadas por una metodología

- Las metodologías tienden a centrarse en un subconjunto de actividades de calidad de datos [2].

- Las distintas actividades de gestión, abordadas por las metodologías del estado del arte, proponen una clasificación:

  - Metodologías completas: dan soporte a la fase de evaluación y de mejora, y abordan cuestiones técnicas y económicas.

  - Metodologías de auditoría: centradas en la fase de evaluación y brindan apoyo limitado a la fase de mejora.

  - Metodologías operativas: centradas en cuestiones técnicas de las fases de evaluación y mejora, no abordan cuestiones económicas.

  - Metodologías económicas: centradas en la evaluación de costos.

Metodologías y contexto de datos

- Dedicadas a dominios específicos, como salud [8], linked data [5,9], gobierno electrónico [6], toma de decisiones [4], gobernanza de datos [3] y arquitecturas orientadas a servicios [10].

- La mayoría de las actividades de gestión de calidad de datos están influenciadas por el contexto de los datos [2, 3, 4, 5, 6].

- Muy pocas metodologías consideran el contexto de los datos, cuando lo hacen, el contexto se aborda sólo en pocas etapas, y generalmente se trata de las etapas iniciales [7].

- No se especifica cómo usar el contexto, es un contexto muy simple (ej.: definido sólo por la tarea que usa los datos) o es estático (se define inicialmente y no se actualiza a lo largo de toda la metodología).

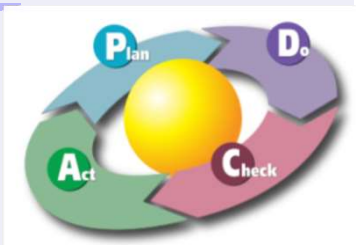
Metodologías y contexto de datos

- El contexto de los datos tiene un impacto significativo en toda la metodología [3]:
  - Análisis de los datos
  - Selección de las dimensiones de calidad
  - Estrategias de mejora, etc.

- Las dimensiones de CD dependen en gran medida del contexto, y su relevancia e importancia pueden variar entre organizaciones y tipos de datos [3].

- Retomando la definición:
  - *A DQ methodology provides a set of guidelines and techniques that, from input information* ***that describes a given application context****, defines a rational process to assess and improve the quality of data* [1].

Metodologías dependientes del contexto



- Metodologías basadas en un modelo de decisión:

  ciclo Shewhart-Deming Plan-Do-Check-Adjust (PDCA) [11]
  - **Total Data Quality Management** *(TDQM)* [12]:
    definir CD (plan), medir CD (do), analizar CD (check) y mejorar CD (adjust).
  - **Estándar ISO 8000-61:2016** [13] modelo de referencia para la gestión de CD: analiza requerimientos (plan), implementa un procesamiento de datos (do), monitorea y mide la CD, e informa los resultados (check) y toma acciones para mejorar (act).
  - **DQ framework for the estonian public sector** [21] usa el ciclo OPDCA [11]: Observe (la situación existente y la necesidad de mejora), Plan (el nivel de madurez objetivo de CD y el plan de mejora), Do (implementa el plan de mejora), Check (verifica si se alcanzaron los valores objetivo de CD) y Adjust (establece un nuevo estado de CD y mejorar el proceso de gestión de CD, si es necesario).

- Los 3 trabajos utilizan implícitamente el contexto en las etapas iniciales.

- El contexto está limitado a la identificación de requerimientos de CD [6, 12, 13] y reglas de negocio [6].

Metodologías dependientes del contexto

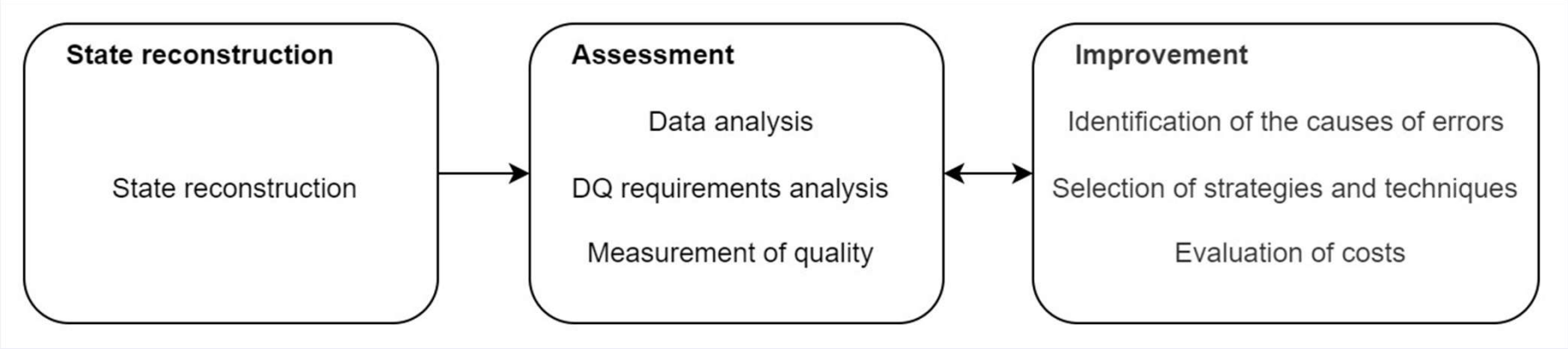
- Batini [1] y Cichy [3] comparan 13 y 12 metodologías (6 en común), respectivamente:
  - abordan principalmente 2 fases, evaluación y mejora de la CD.
  - sólo 3 metodologías proponen una fase previa donde se puede identificar y/o definir el contexto de los datos.
    - Comprehensive Data Quality (CDQ) [1, 14],
    - Heterogeneous Data Quality Methodology (HDQM) [15],
    - Hybrid Information Quality Management (HIQM) [16].

- Ninguno menciona el contexto explícitamente, recopilan cierta información contextual en las etapas iniciales.

- Información del contexto no se actualiza en etapas posteriores y se reduce a requerimientos de calidad [14, 15, 16] y reglas de negocio [14].

Comprehensive Data Quality

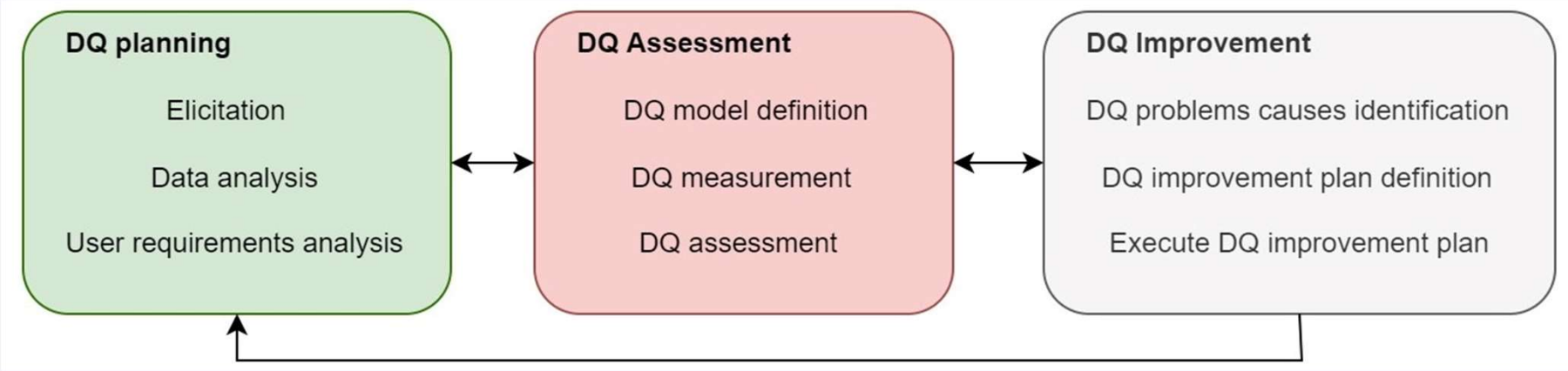
- *Comprehensive Data Quality Methodology* (CDQ) es una metodología completa propuesta por Batini [1,14]:
  - Considera la fase de evaluación (*Assessment*) y mejora (*Improvement*), abordando cuestiones técnicas y económicas.



  - Incluye una fase inicial llamada reconstrucción del estado (*state reconstruction*).
    - Donde se identifica la información contextual.
    - Se ejecuta una sola vez.

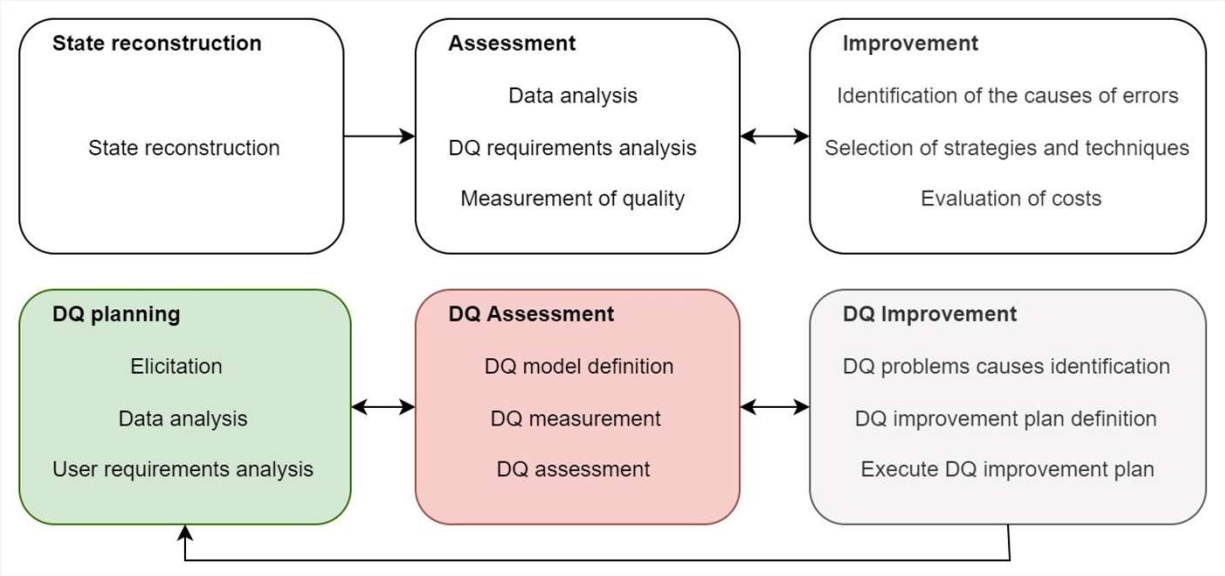
Context-aware Data Quality Management

- *Context-aware Data Quality Management Methodology* (CaDQM):
  - Está inspirada en la metodología CDQ.
  - Tiene 3 fases: *DQ planning, DQ assessment* y *DQ improvement*
  - Considera el contexto de los datos, explícitamente, en cada una de sus fases.



  - La mayoría de los componentes del contexto son identificados y definidos en la fase *DQ planning*.
  - El contexto de los datos puede ser actualizado en la fase *DQ assessment*.
  - Todas las fases de CaDQM son influenciadas por el contexto de los datos.

Diferencias entre CDQ y CaDQM

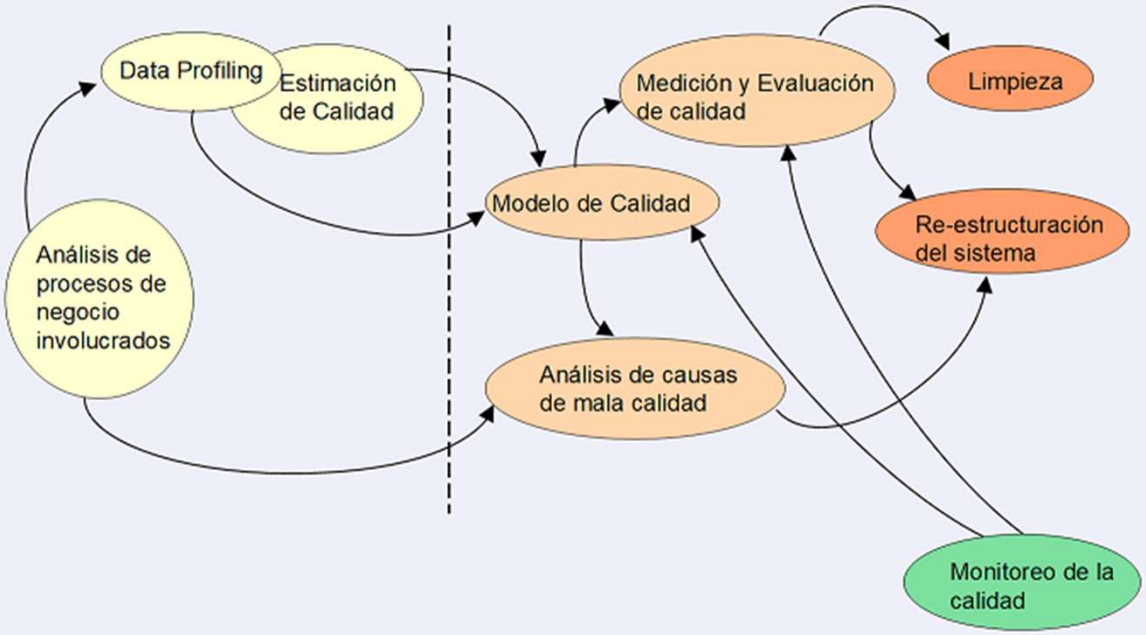


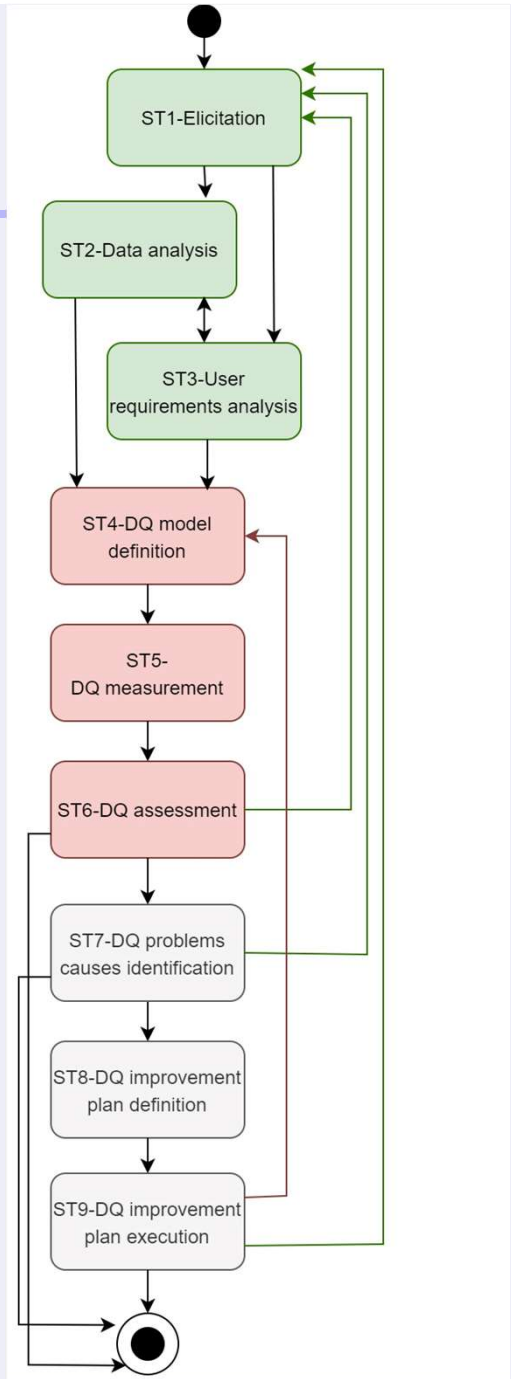
- CDQ ejecuta la fase inicial una sola vez, mientras que en CaDQM las fases definen un ciclo que se ejecutan tantas veces como sea necesario.

- CDQ recopila información contextual sólo en la fase inicial, mientras que CaDQM la identifica en la fase inicial y la actualiza en la segunda fase.

- CDQ no especifica cómo usar la información contextual, mientras que CaDQM describe explícitamente la influencia y el uso del contexto en cada fase.

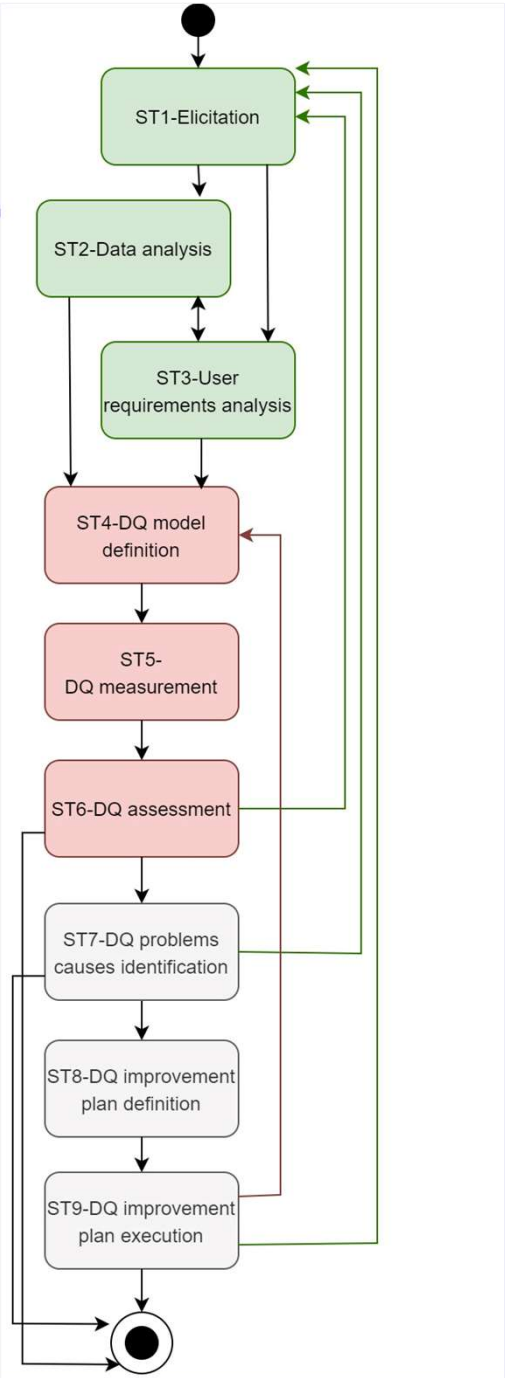
Etapas de CaDQM





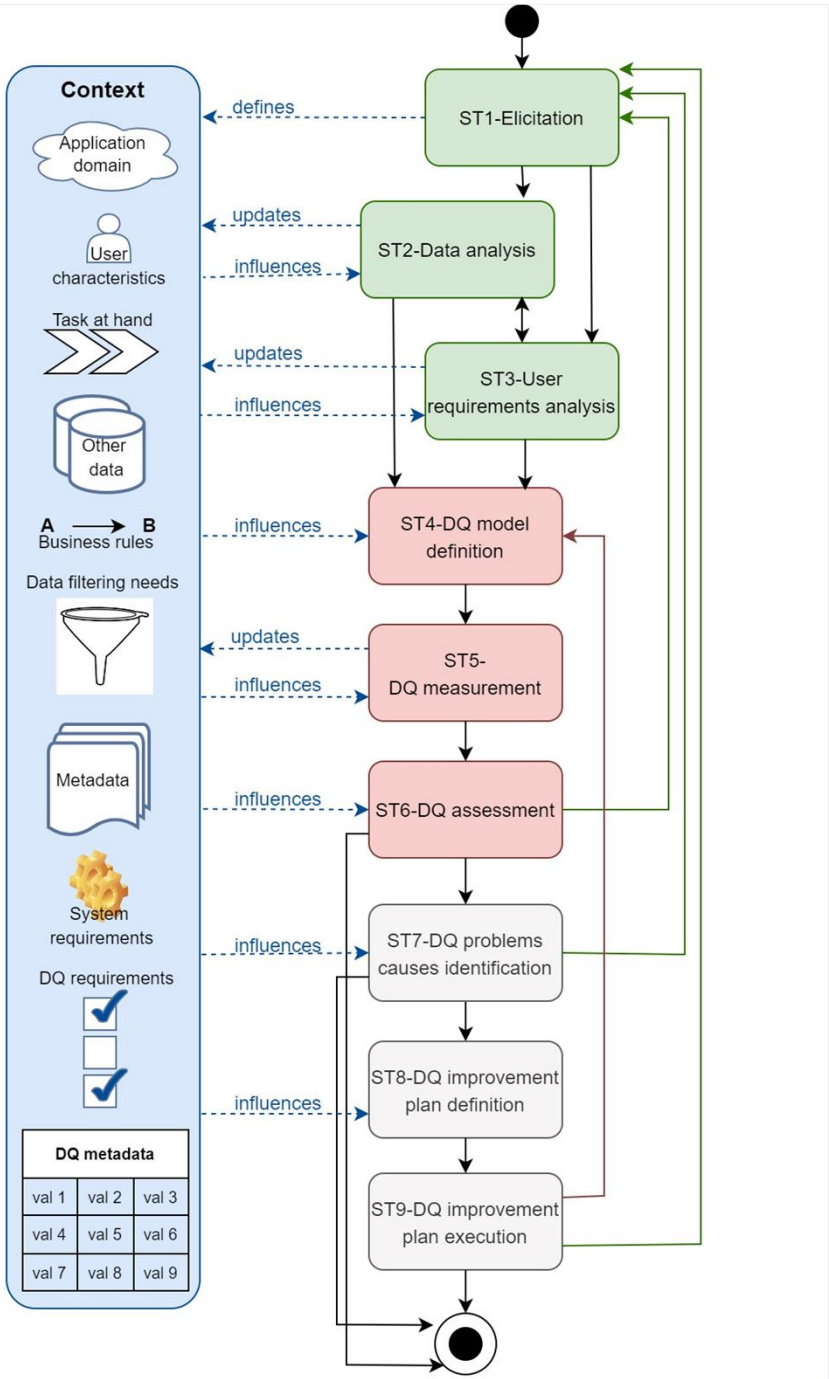
Etapas de CaDQM

- DQ Planning:
  - ST1: Elicitation
  - ST2: Data analysis
  - ST3: User requirements analysis
- DQ Assessment:
  - ST4: DQ model definition
  - ST5: DQ measurement
  - ST6: DQ assessment
- DQ Improvement:
  - ST7: DQ problems causes identification
  - ST8: DQ improvement plan definition
  - ST9: DQ improvement plan execution

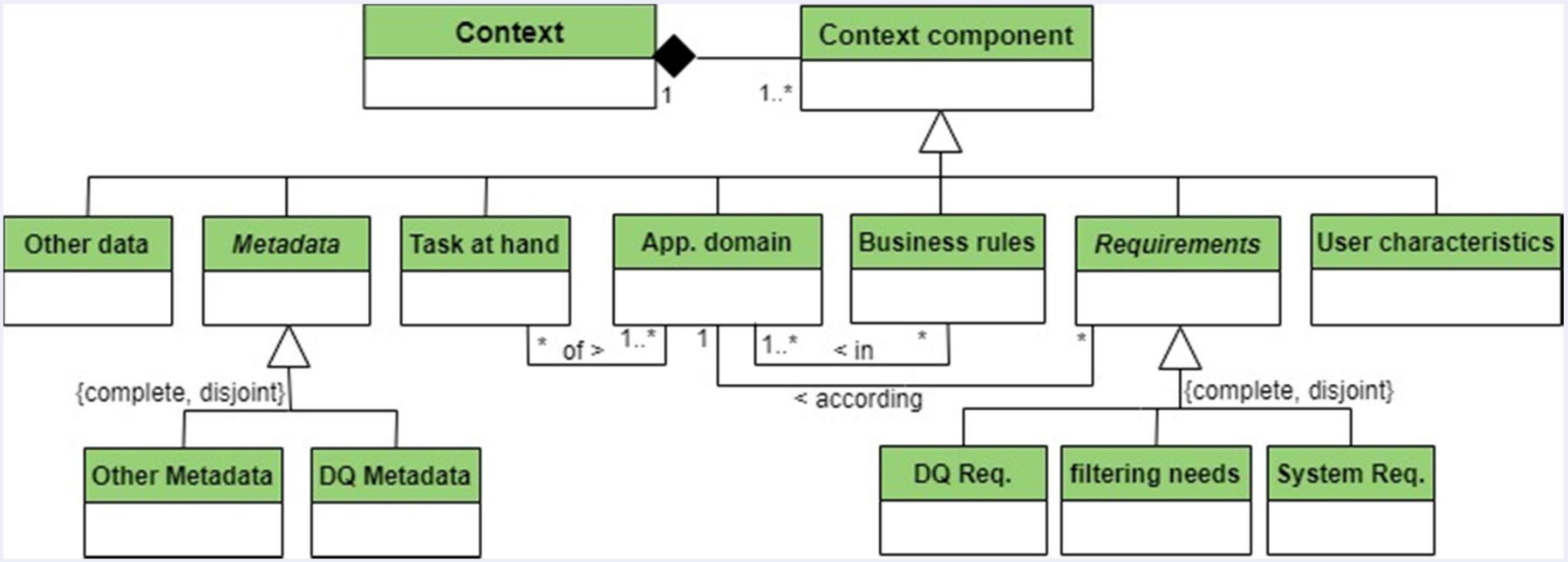


Contexto y CaDQM

- Definición, actualización e influencia del contexto de datos.

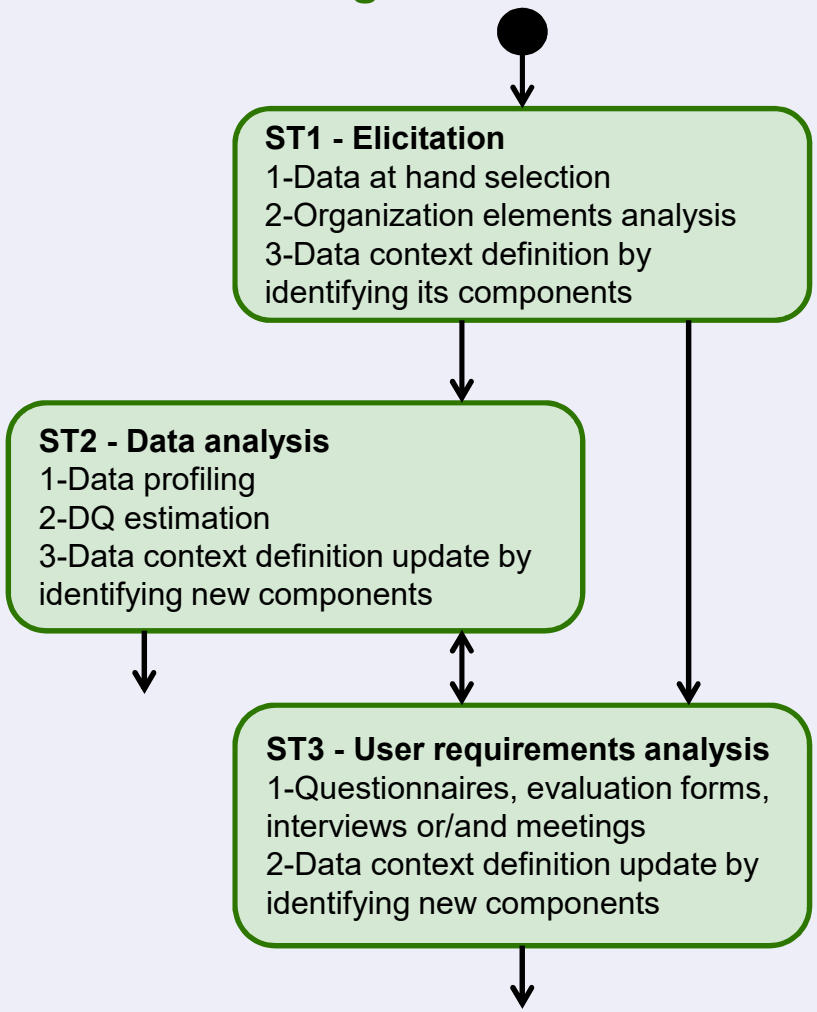


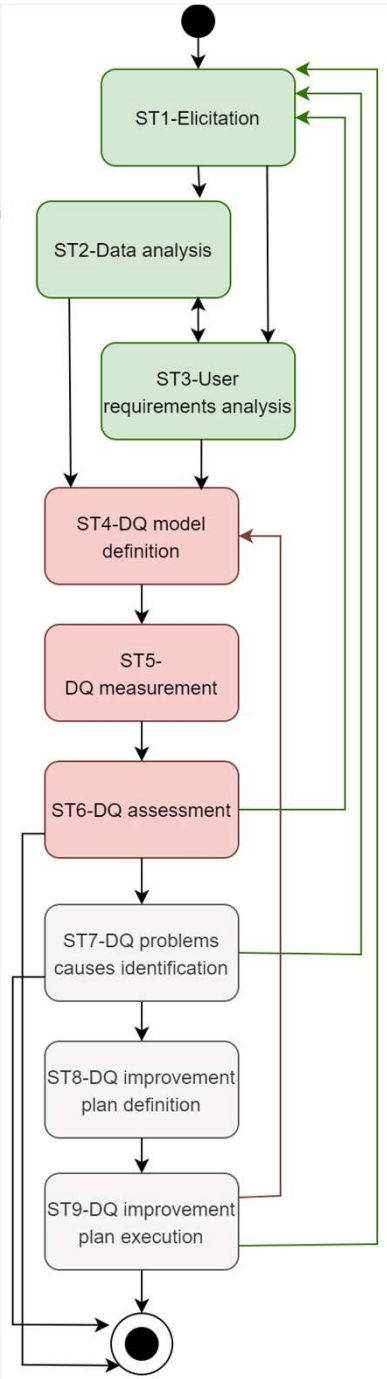
Contexto


Context
Context component
1
1..*
Other data
Metadata
Task at hand
App. domain
Business rules
Requirements
User characteristics
* of >
1..*
1
1..*
< in
*
*
< according
{complete, disjoint}
{complete, disjoint}
Other Metadata
DQ Metadata
DQ Req.
filtering needs
System Req.

Actividades de CaDQM

Fase 1: DQ Planning





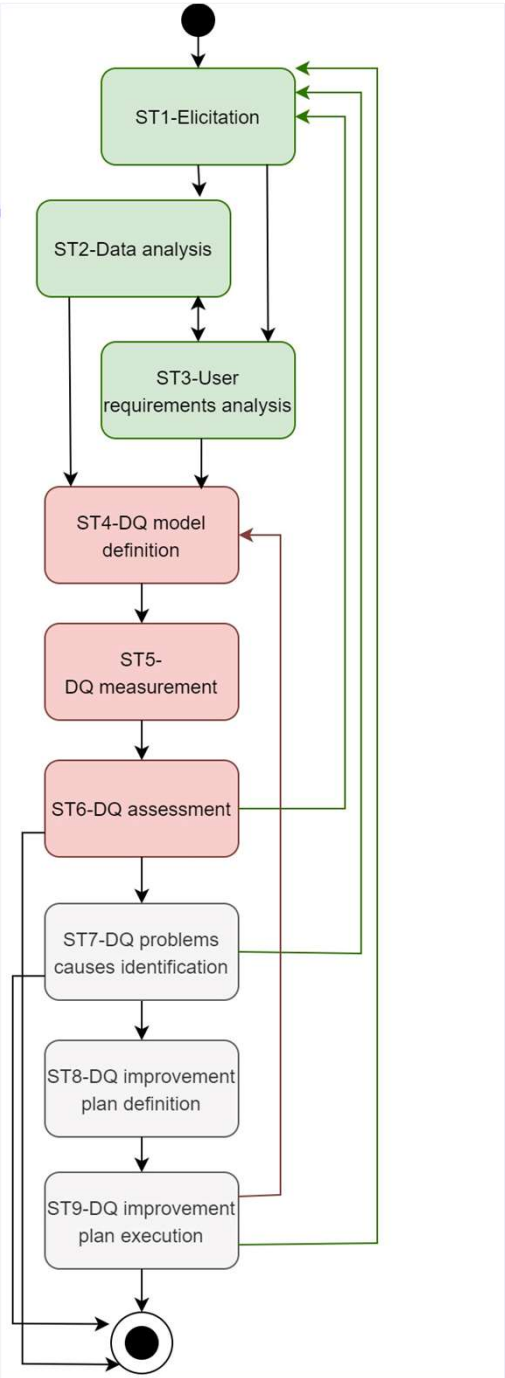
Actividades de CaDQM

Fase 2: DQ Assessment

ST4 - DQ model definition
1-Data context, DQ problems analysis
2-DQ dimensions selection
3-DQ metrics definition

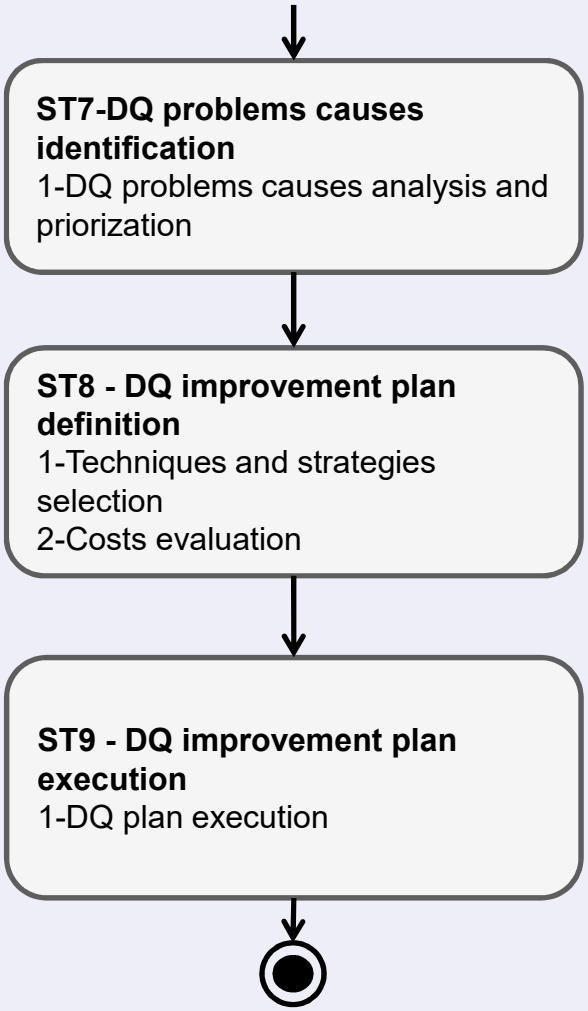
ST5 - DQ measurement
1-Database design for DQ metadata
2-DQ metrics execution, results storage
3-Data context definition update by identifying new components

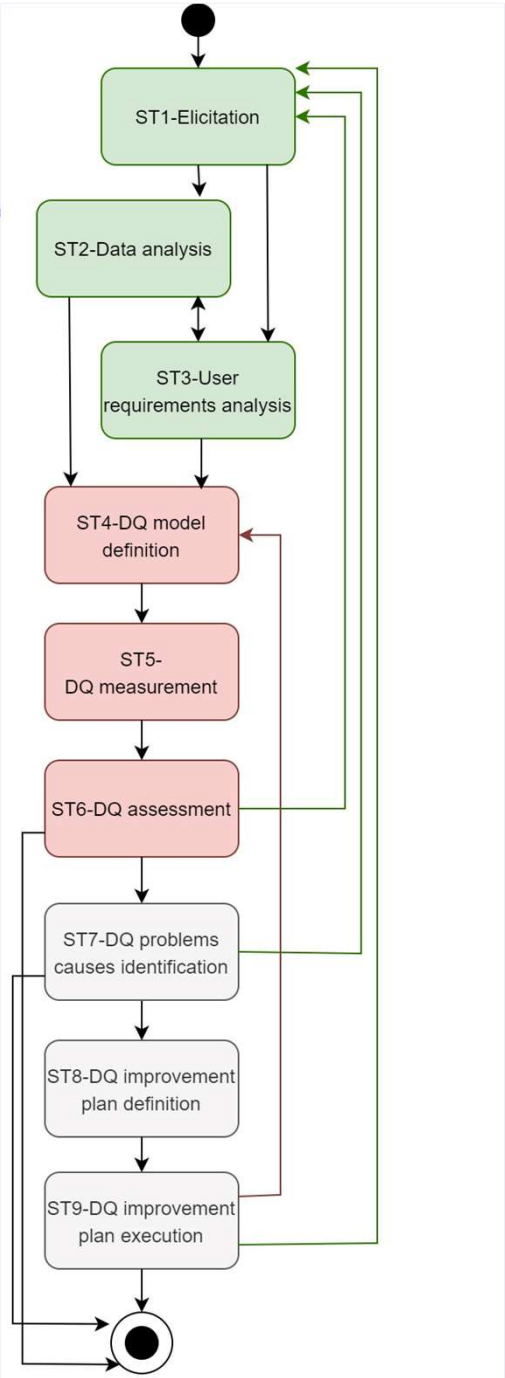
ST6 - DQ assessment
1-Data context analysis
2-DQ metadata assessment



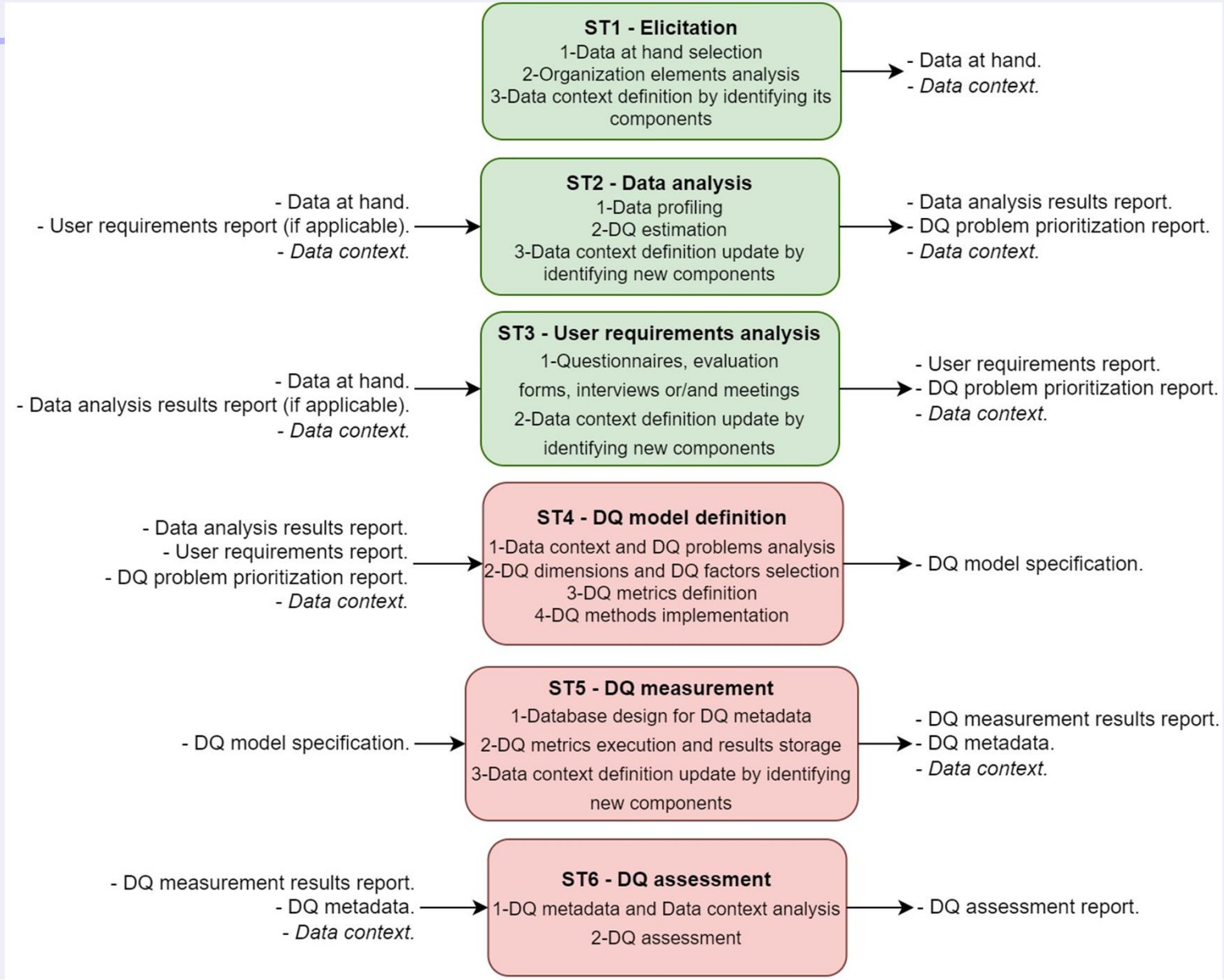
Actividades de CaDQM

Fase 3: DQ Improvement





Entradas y salidas



Caso de estudio

- Descripción de la Realidad

Preparación y análisis de datos médicos reales para la estratificación (ordenación) de pacientes con Esclerosis Lateral Amiotrófica (ELA).

El dataset (archivos CSV anonimizados) contiene datos demográficos, clínicos y biológicos sobre 1660 pacientes locales diagnosticados con ELA entre 1991 y 2022.

Un archivo con datos de pacientes (una línea por paciente) y otros archivos con resultados clínicos y biológicos.

Un científico de datos realizó un análisis de estratificación preliminar y preparación de datos:

- Sin metodologías de calidad de datos
- Mediante implementación de tareas de data profiling y ETL

Caso de estudio

Etapas	Componentes de Contexto	Enfoque basado en los datos
ST1 Elicitation	Dominio: Salud Características de Usuarios: U1-investigadores médicos, U2-científicos de datos Tarea: estratificación de ELA Reglas de Negocio: La variable ALSSFR solo toma valores entre 0 y 48 (0 ≤ ALSSFR ≤ 48). Necesidades de Filtrado: Los pacientes no diagnosticados son descartados	
ST2 Data Analysis	RQ1: el formato de fecha debe ser DD/MM/AAAA, RQ2: se descartan valores biológicos con muchos valores nulos, RQ3: cada paciente debe tener al menos 5 citas médicas, RQ4: los valores posibles para el diagnóstico son Espinal, Bulbar y Respiratorio, RQ5: las fechas de las citas deben ser crecientes, RQ6: las puntuaciones ALSSFR (referentes a ELA) deben estar disminuyendo, RQ7: los valores atípicos deben descartarse.	
ST3 User Requirements Analysis	No aplica	
ST4 DQ Model Definition	Dimensión: **Exactitud**. Factores: **exactitud sintáctica** (RQ4) y **precisión** (RQ1) Dimensión: **Completitud**. Factor: **densidad** (NF y RQ2) Dimensión: **Consistencia**. Factores: **integridad de dominio** (RN y RQ4), **integridad intra-relación** (R5 y R6), **integridad inter-relación** (RQ3)	

Ejercicio

Una base de datos (BD) almacena datos de una Universidad uruguaya y contiene las siguientes tablas:

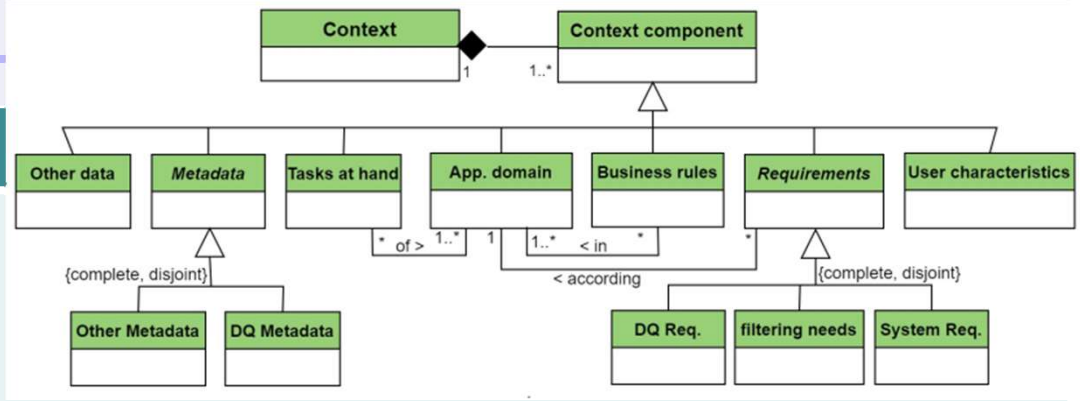
Estudiantes (CI, nombre, fechaNacimiento, dirección, teléfono, fechaIngreso, ultimaActualizacion)
Cursos (codigoCurso, nombre, cupo)
DictadoCurso (codigoCurso, codigoDocente, esResponsable)
Docentes(CI, fechaIngresoCargoActual, nombre, teléfono)
Evaluaciones(codigoCurso, CIDocente, CIEstudiante, fecha, nota)

Solo interesa evaluar la calidad de los datos de las primeras 3 tablas.

Los usuarios administrativos realizan análisis con los datos almacenados. Por otro lado, los estudiantes, actualizan y realizan consultas sobre sus datos y los datos de los cursos. La Universidad exige que todas las fechas tengan el formato DD/MM/AAAA. Además, las direcciones y teléfonos deben ser datos válidos de Uruguay. En particular, las direcciones no pueden ser abreviadas. Por otro lado, los docentes solo pueden ser responsables si tienen al menos 5 años en el cargo. Finalmente, se sabe que esta BD solo almacena datos de los cursos Lógica, Matemática Discreta, Álgebra, Física y Química. Cualquier otro curso sería un dato inválido.

Los usuarios administrativos plantearon que hay campos importantes para el análisis de datos. Por lo tanto, plantean que al menos el 95% de las direcciones deben ser no nulas. Además, la fecha de ingreso de los estudiantes debe tener al menos 17 años más que la fecha de nacimiento. Por otro lado, los usuarios administrativos necesitan que los estudiantes actualicen sus datos 1 vez al año y solo están interesados en analizar datos recolectados desde 2010.

Ejercicio 1



Etapas	Componentes de Contexto
ST1 Elicitation	
ST2 Data Analysis	------
ST3 User Requirements Analysis	------
ST4 DQ Model Definition	

Ejercicio

Comp. de CTX.	Etapa: ST1 - Elicitation
Dominio	D: Educación
Características de Usuarios	U1: Usuario administrativo U2: estudiantes
Tareas	T1: Análisis de datos, T2: Actualización de datos, T3: Consulta de datos.
Reglas de negocio	RN1: Todas las fechas con formato DD/MM/AAAA RN2: Direcciones y teléfonos datos válidos de Uruguay. RN3: Direcciones no pueden ser abreviadas. RN4: Docentes responsables si tienen al menos 5 años en el cargo. RN5: Nombre de cursos = Lógica, Matemática Discreta, Álgebra, Física o Química.
Req. Sistema	---
Req. CD	RQ1: Al menos el 95% de las direcciones deben ser no nulas. RQ2: La fecha de ingreso de los estudiantes debe tener al menos 17 años más que la fecha de nacimiento. RQ3: Los estudiantes deben actualizar sus datos 1 vez al año.
Necesidades de filtrado	NF1: Los usuarios administrativos solo están interesados en analizar datos recolectados desde 2010.
Metadatos	---
Metadatos de CD	---
Otros datos	OD1: Tabla «Docentes»

Ejercicio

Comp. de CTX. (Identificados en las etapas ST1, ST2 y ST3)	Etapa: ST4 – DQ Model Definition
D: Educación U1: Usuario administrativo U2: estudiantes T1: Análisis de datos, T2: Actualización de datos, T3: Consulta de datos. RN1: Todas las fechas con formato DD/MM/AAAA RN2: Direcciones y teléfonos datos válidos de Uruguay. RN3: Direcciones no pueden ser abreviadas. RN4: Docentes responsables si tienen al menos 5 años en el cargo. RN5: Nombre de cursos = Lógica, Matemática Discreta, Álgebra, Física o Química. RQ1: Al menos el 95% de las direcciones deben ser no nulas. RQ2: La fecha de ingreso de los estudiantes debe tener al menos 17 años más que la fecha de nacimiento. RQ3: Los estudiantes deben actualizar sus datos 1 vez al año. NF1: Los usuarios administrativos solo están interesados en analizar datos recolectados desde 2010. OD1: Tabla «Docentes»	Dimensión: **Exactitud**. Factores: **exactitud sintáctica** (RN1, RN3, RN5), **exactitud semántica** (RN2) Dimensión: **Frescura**. Factores: **actualidad** (RQ3) Dimensión: **Completitud**. Factores: **densidad** (RQ1) Dimensión: **Consistencia**. Factores: **integridad inter-relación** (RN4, OD1), **integridad intra-relación** (RQ2)

Bibliografía

[1] Batini, C., et al.: Methodologies for data quality assessment and improvement. CSUR 41(3), 1–52 (2009).

[2] Batini, C., Scannapieco, M.: Methodologies for information quality assessment and improvement. In: Data and information quality, pp. 353–402. Springer (2016).

[3] Cichy, C., Rass, S.: An overview of data quality frameworks. IEEE Access 7, 24634–24648 (2019).

[4] Günther, L.C., et al.: Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. Procedia Manufacturing 29, 583–591 (2019).

[5] Gürdür, D., et al.: Methodology for linked enterprise data quality assessment through information visualizations. JIII 15, 191–200 (2019).

[6] Tepandi, J., et al.: The data quality framework for the estonian public sector and its evaluation. In: TLDKS, vol. 10680, pp. 1–26. Springer (2017).

[7] Serra, F., Peralta, V., Marotta, A., & Marcel, P. (2023, August). Context-Aware Data Quality Management Methodology. In European Conference on Advances in Databases and Information Systems (pp. 245-255). Cham: Springer Nature Switzerland.

[8] Kerr, K., Norris, T.: The development of a healthcare data quality framework and strategy. In: ICIQ. pp. 218–233 (2004).

[9] Debattista, J., et al.: Luzzu—a methodology and framework for linked data quality assessment. JDIQ 8(1), 1–32 (2016).

Bibliografía

[10] Petkov, P., Helfert, M.: A methodology for analyzing and measuring semantic data quality in service oriented architectures. In: 14th International Conference on Computer Systems and Technologies. pp. 201–208 (2013)

[11] Foresight university: Shewhart-deming's learning and quality cycle. https://foresightguide.com/shewhart-and-deming/, accessed: October 2023.

[12] Wang, R.Y.: A product perspective on total data quality management. ACM 41(2), 58–65 (1998).

[13] Standard ISO 8000-61:2016. Data quality — part 61: Data quality management: Process reference model. Tech. rep. (2022).

[14] Batini, C., et al.: A comprehensive data quality methodology for web and structured data. In: ICDIM. pp. 448–456 (2007).

[15] Batini, C., et al.: A data quality methodology for heterogeneous data. IJDMS 3, 60–79 (2011).

[16] Cappiello, C., et al.: Hiqm: A methodology for information quality monitoring, measurement, and improvement. p. 339–351 (2006).