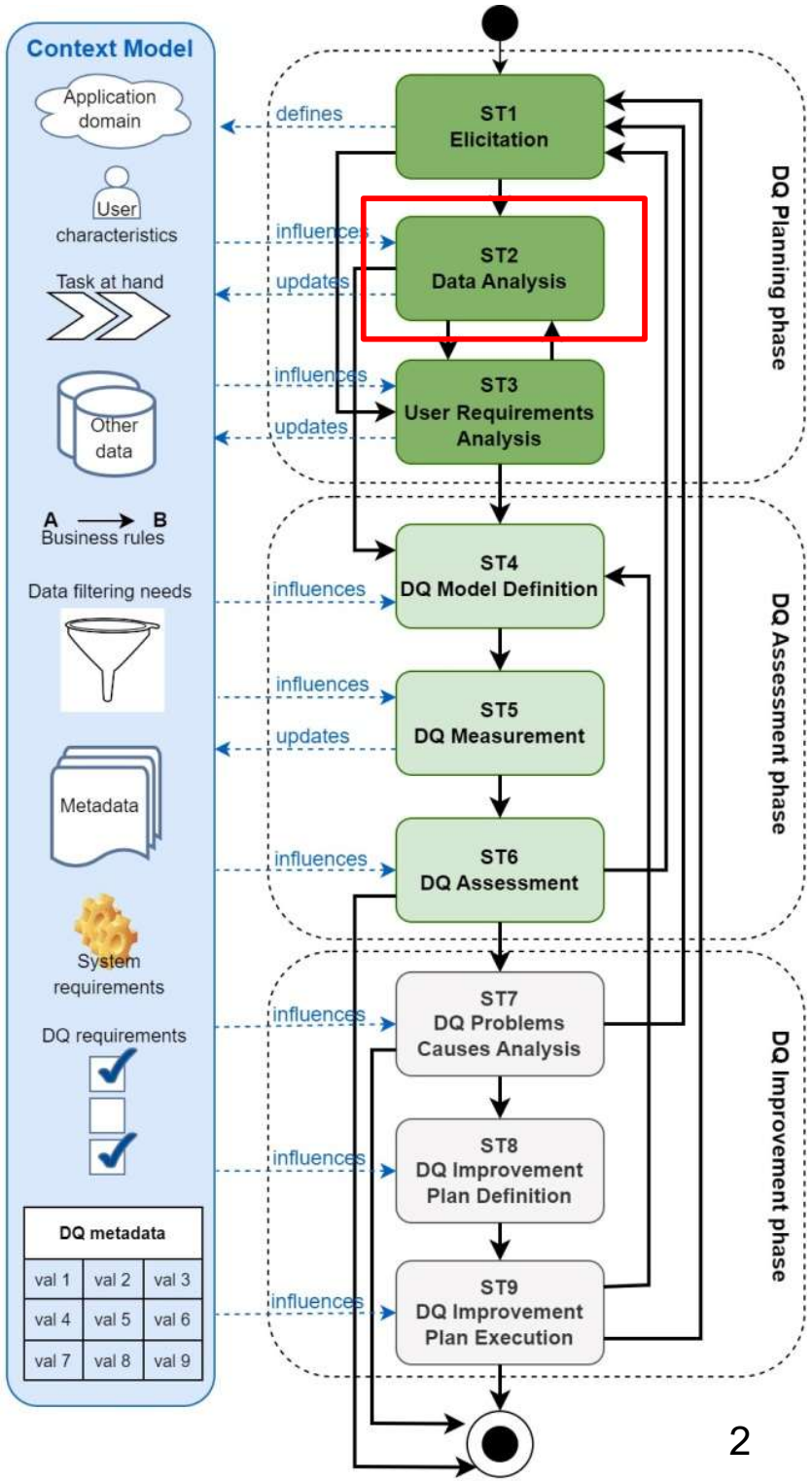
Calidad de Datos e Información

CaDQM

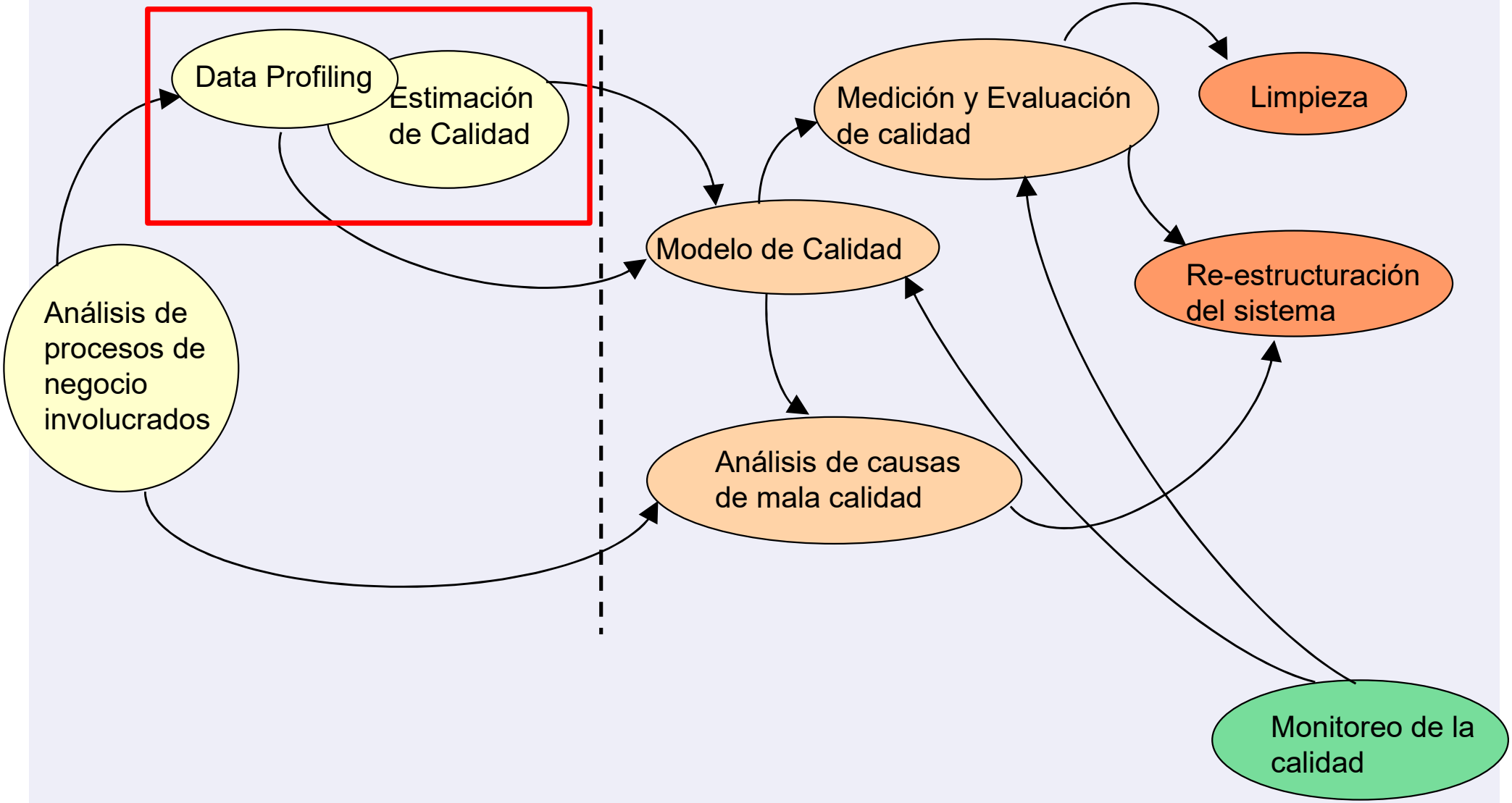
ST2 – Data Analysis

ST2 – Data Analysis

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



Gestión de la calidad en SI



Data Profiling

- *Data Profiling*
  - Primera aproximación al conocimiento sobre los datos del SI / dataset que queremos evaluar
    - Su estructura (si los metadatos son consistentes con los datos)
    - Sus relaciones
    - Su volumen
    - Sus problemas y frecuencia de los mismos
    - Patrones que se cumplen
  - Algunas técnicas, como
    - Estadísticas básicas
    - Análisis de metadatos
    - Análisis de patrones
    - Detección automática de “foreign keys”

Data Profiling

- *Data Profiling*
  - Análisis de atributos solapados de diferentes tablas
    - Redundancias, claves foráneas
  - Valores faltantes o erróneos
    - Cardinalidad actual vs. cardinalidad esperada (cant. clientes)
    - Frecuencia de valores nulos, maximo/minimo, etc.
  - Duplicados
    - Número de tuplas vs. cardinalidad del dominio del atributo
  - Claves difusas y dependencias funcionales difusas
    - Restricciones de integridad que no están explícitamente definidas pero que son satisfechas en la mayoría de los casos (un atributo que es clave, dependencias funcionales)

Data Profiling

- "*Profiling*" con SQL

```
SELECT MIN(edad), MAX(edad), COUNT(DISTINCT edad)
FROM Empleados;

SELECT ciudad, COUNT(*) AS cant
FROM Clientes
GROUP BY ciudad ORDER BY cant;

SELECT COUNT(DISTINCT C1.ciudad)
FROM Clientes C1, Clientes C2
WHERE C1.ciudad = C2.ciudad AND
        C1.pais <> C2.pais;
```

Data Profiling

- "*Profiling*" con SQL

Estudiantes (ci-est, nombre, email, telefono, direccion, fnac)
Creo que el email no se repite casi nunca.

```
SELECT COUNT(DISTINCT E1.email)
FROM Estudiantes E1
WHERE E1.email in
        (SELECT E2.email
        FROM Estudiante E2
        GROUP BY E2.email
        HAVING COUNT(*) > 1)
```

Data Profiling

- "*Profiling*" con SQL

Actividades (ci-est, tipo-act, fecha, carrera, asignatura, instituto)

asignatura, carrera → instituto ?

```
SELECT DISTINCT A1.asignatura, A1.carrera
FROM Actividades A1, Actividades A2
WHERE A1.asignatura = A2.asignatura and
        A1.carrera = A2.carrera and
        A1.instituto <> A2.instituto
```

Data Profiling

- *Data Profiling*: Algunas herramientas

Empresa	**Productos**
Ataccama	DQ Analyzer, Data Quality Center, DQ Issue Tracker, DQ Dashboard
Datactics	Data Quality Platform, Data Quality Manager, Master Record Manager
DataMentors	DataFuse, ValiData, NetEffect
Human Inference	HIquality Suite, HIquality Name Worldwide, HIquality Identify, HIquality Data Improver, DataCleaner
IBM	InfoSphere Information Analyzer, InfoSphere QualityStage, InfoSphere Discovery
Informatica	Data Explorer, Data Quality, Identity Resolution, AddressDoctor
Information Builders/iWay	iWay Data Quality Center
Innovative Systems	i/Lytics Data Quality, i/Lytics Data Profiling, i/Lytics ProfilerPlus, FinScan
Oracle	Oracle Enterprise Dat Quality, Oracle Enterprise Data Quality for Product Data
Pitney Bowes Software	Spectrum Technology Platform
RedPoint (DataLever)	RedPoint Data Management
SAP	Data Quality Management, Information Steward, Data Services
SAS/DataFlux	Data Management Platform
Talend	Talend Open Studio for Data Quality, Talend Enterprise Data Quality
Trillium Software	Trillium Software System, TS Discovery, TS Insight, Trillium Software On-Demand
Uniserv	Data Quality (DQ) Explorer, DQ Batch Suite, DQ Real-Time Suite, DQ Real-Time Services, DQ Monitor
Melissa Data	Contact Zone
Datiris	Datiris Profiler
CloverETL	Address Doctor
Microsoft	Data Quality Services

Data Profiling

- *Data Profiling*: Ejemplos de Herramientas
  - DataCleaner
    - https://datacleaner.github.io/
  - Python Pandas Profiling
    - https://pypi.org/project/pandas-profiling/
  - Talend Open Studio
    - www.talend.com

ST2 – Data Analysis

- *Identification of DQ problems*
  - Actividad introducida en ST1
  - Nuevos problemas de calidad de datos se identifican a partir de los resultados del *data Profiling*.
  - Ejemplos de problemas de CD son datos incompletos, duplicados, obsoletos, inconsistentes o incumplimiento de las reglas de negocio.

- *Estimation of DQ*
  - Con los problemas de CD y los resultados del *data profiling*, obtenemos una estimación de CD.
  - Indicador del volumen de trabajo para evaluar y mejorar la CD.
  - Permite priorizar los problemas de CD que servirán de base para la definición del modelo de CD.

ST2 – Data Analysis

- *Update of the context model definition*
  - Identificación de nuevos componentes de contexto que surgen de los resultados de la actividad de *data profiling*.
  - Por ejemplo:
    - problemas de CD identificados en esta etapa pueden determinar nuevos requerimientos de CD y reglas de negocio.
    - metadatos, como detalles sobre los atributos de los *data at hand*
    - otros datos mediante el descubrimiento de relaciones con los *data at hand*.

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

Bibliografía

- Felix Naumann. Data Profiling Revisited. ACM SIGMOD Record 42.4 (2014): 40-49 https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2013/profiling_vision.pdf

- Ziawasch Abedjan, Lukasz Golab, Felix Naumann. Data Profiling. SIGMOD 2017 Tutorial. https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2017/SIGMOD_2017_Tutorial_Data_Profiling.pdf