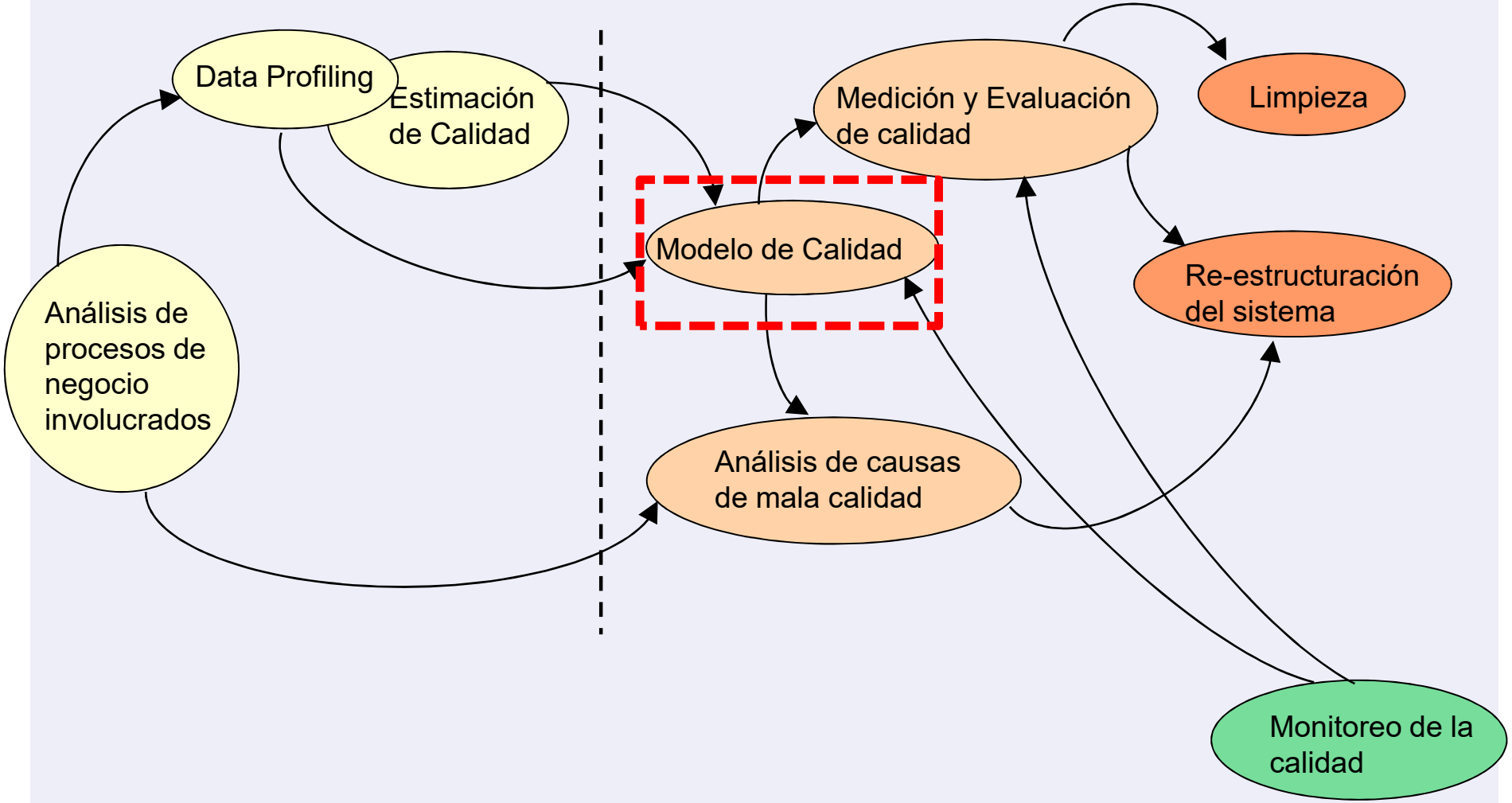
Calidad de Datos e Información

Evaluación de Calidad

Evaluación de Calidad

- Pasos
  - Construir modelo de calidad de datos
  - Diseñar base de metadatos de calidad
  - Implementar métricas especificadas en el modelo de calidad
  - Ejecutar medición
  - Comparar resultados con requerimientos de calidad
  - Definir pasos a seguir

Gestión de la calidad en SI



Modelo de Calidad de datos

- Define:
  - qué características de calidad se manejan
  - sobre qué datos aplican
  - cómo se miden esas características
- Para cada conjunto de datos se define un modelo de calidad particular
- Guía toda la gestión de la calidad de los datos

Modelo de Calidad

- Para poder tratar la calidad de un SI es necesario definir un modelo de calidad adecuado a las necesidades y prioridades de los consumidores de los datos en ese SI

- Se debe determinar y especificar
  - Dimensiones y factores a medir
  - Métricas y datos o grupos de datos donde éstas se aplican
  - Métodos de medición
  - Agregaciones

- Para determinar las dimensiones y factores de calidad, y los datos a medir, existen distintos métodos de trabajo

Especificación Modelo de Calidad

- Organizado por Dimensiones

Dimensión	Factor	Métrica	Agregación	Métrica Instanciada	Método Medición
DIM 1	F1	M1F1	Ag1	M1F1-dato1	….
				M1F1-dato2	….
					….
				M1F1-dato3	….
		M2F1	Ag2	M2F1-dato1	….
	F2	MF2	Ag3	MF2-dato4	….
	F3	MF3	Ag4	MF3-dato2	….
				MF3-dato3	….
DIM2	F4	….	….	….	….
	….	….	….	….	….

Especificación Modelo de Calidad

- Organizado por Datos
  - Primero:

Dimensión	**Factor**	**Métrica**	**Agregación**
DIM 1	F1	M1F1	Ag1
		M2F1	Ag2
	F2	MF2	Ag3
	F3	MF3	Ag4
DIM2	F4	….	….
	….	….	….

Especificación Modelo de Calidad

- Organizado por Datos
  - Segundo:

DATO: Tabla.atributo

Dimensión	Factor	Métrica Instanciada	Método de Medición
DIM 1	F1	M1F1-dato	….
		M2F1-dato	….
	F2	MF2-dato	….
	F3	MF3-dato	….
DIM2	F4	….	….
	….	….	….

Modelo de Calidad

- Métodos
  - Relevamiento **problemas y requerimientos de usuarios**
    - 2 opciones:
      - A partir de los **datos prioritarios** defino dimensiones-factores-métricas.
      - Primero defino **dimensiones-factores**, luego selecciono datos, luego defino métricas.
  - A partir de **perfil de los datos** (Data Profiling)
    - Sin intervención de usuarios finales.
    - El perfil de los datos me da las nociones de por dónde están las fallas de calidad. A partir de eso defino las dimensiones-factores-métricas.

Otras propuestas existentes

- Cada organización debe determinar los factores a medir y desarrollar métricas y métodos apropiados.
  - Pero las empresas no se plantean directamente las métricas
- Análisis top-down de la calidad
  - Se empieza por identificar los problemas de calidad y luego se determinan las métricas apropiadas para cuantificarlos
    - Ejemplo:

Quiero reducir la cantidad de cartas que no llegan a mis clientes

Necesitaría saber cuántos rechazos se deben a errores sintácticos en los nombres de las calles

Voy a medir el porcentaje de direcciones de clientes que no figuran en la guía de calles

Goal-question-metric (GQM)

- GQM es un paradigma de diseño de sistemas de información [Basili94].
- GQM propone tres niveles de abstracción:
  - Nivel conceptual: **GOALS**
    - Se definen objetivos de calidad de alto nivel que **apuntan a resolver problemas de calidad de la organización**.
    - Ej. *Reducir la cantidad de cartas que no llegan a mis clientes*
  - Nivel operacional: **QUESTIONS**
    - Cada objetivo se descompone en un conjunto de preguntas que **caracterizan la manera de alcanzar los objetivos**.
    - La idea es descomponer sucesivamente los objetivos hasta llegar a preguntas simples, cada una asociada directamente a un factor de calidad.
    - Ej. *¿Cuántos rechazos se deben a errores sintácticos en los nombres de las calles?*
  - Nivel cuantitativo: **METRICS**
    - Se define un conjunto de métricas para cada pregunta, **para responderla de una forma cuantitativa**.
    - E.g. *Porcentaje de direcciones de clientes que no figuran en la guía de calles*

Ejemplos de *goals* y *questions*

- Goal 1:
  - ***Mejorar la calidad de datos de localización de los estudiantes*** *(teléfono, dirección, etc.)*

	Questions	**Factores de calidad**
1	¿Las direcciones de los estudiantes son las correctas?	Exactitud semántica
2	¿Sus direcciones están bien escritas?	Exactitud sintáctica
3	¿Sus teléfonos son números válidos?	Exactitud sintáctica
4	¿Tenemos direcciones precisas?	Precisión
5	¿Sus direcciones están al día?	Actualidad
6	¿Tenemos las direcciones de todos los estudiantes?	Cobertura
…		

Ejemplo de aplicación de GQM

- Dominio biológico (Proyecto InCo – Instituto Pasteur)

Goal: Select studies with high accuracy in phenotype data

GOAL	**Purpose: Select a study** **Quality Dimension: Accuracy** **Measurable Object: phenotype variables metadata** **Stakeholder: meta-analyst**
Question	- How standardized the possible variable values are?
Metrics	- Boolean: true if the variable type is enumerated, false if it is free text. - Percentage of possible values of the variable that are mapped to SnomedCT

DQ Goals GWAS

G1	Select studies with high accuracy in phenotype variables (or clinical history forms).
G2	Select studies with high accuracy in genotype variables (or SNPs).
G3	Select studies with high accuracy in phenotype data (or phenotype variable values).
G4	Select studies with high accuracy in genotype data (or genotype variable values).
G5	Select studies with high completeness in phenotype data, with respect to target meta-analysis.
G6	Select studies with high completeness in genotype data.
G7	Select studies with high believability.
G8	Select studies with high compatibility for combination in Meta-analysis.

Modelo de Calidad GWAS

Goal	Dimension	Metric
G1	**Accuracy**	- Boolean: true if the variable is mapped to a SNOMED CT concept, false if not.
		- Matching degree assigned by bio-medical expert when mapping to SNOMED CT.
		- *Specificity Degree*
		- *Descendants Diversity*
G2	**Accuracy**	- Platform-study population adjustment.
G3	**Accuracy**	- Boolean: true if the variable type is enumerated, false if it is free text.
		- Percentage of possible values of the variable that are mapped to SNOMED CT
G4	**Accuracy**	- *HWE disequilibrium*
G5	**Completeness**	- Percentage of target variables that are present in the clinical history.
		- Percentage of target variables that correspond to some variable in the clinical history through SNOMED CT.
G6	**Completeness**	- For each genotype variable, null values ratio
G7	**Believability**	- Follow-up studies results.
		- Laboratory reputation obtained from a ranking.
G8	**Compatibility**	- For studies *S1* and *S2*, where *Vars(S)* are the variable names of study *S*: $\frac{\lvert Vars(S_1) \cap Vars(S_2) \rvert}{\lvert Vars(S_1) \cup Vars(S_2) \rvert}$
		- *Inter-study distance*, calculated according to the mapping of the variables to SNOMED CT concepts. In addition, variables are weighted taking into account target meta-analysis variables.
		- Percentage of variables that are present in both studies and have the same set of possible values.
		- *Platform compatibility degree (based on LD)*
		- *Population stratification*

Propuesta [Otto et. Al, ICIQ 2009]

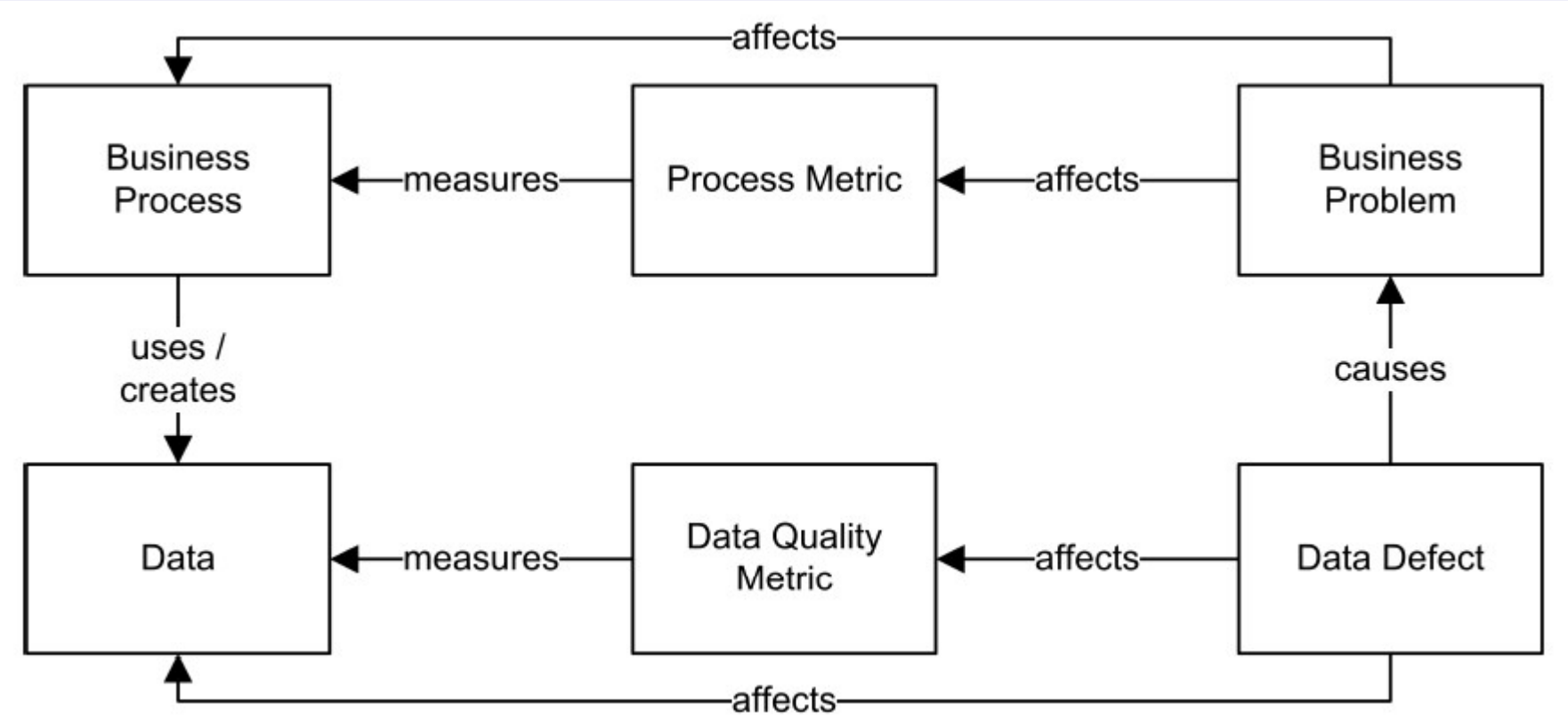
- Otra propuesta para definición de métricas de calidad para un sistema de información dado:

  "Identification of Business Oriented Data Quality Metrics"

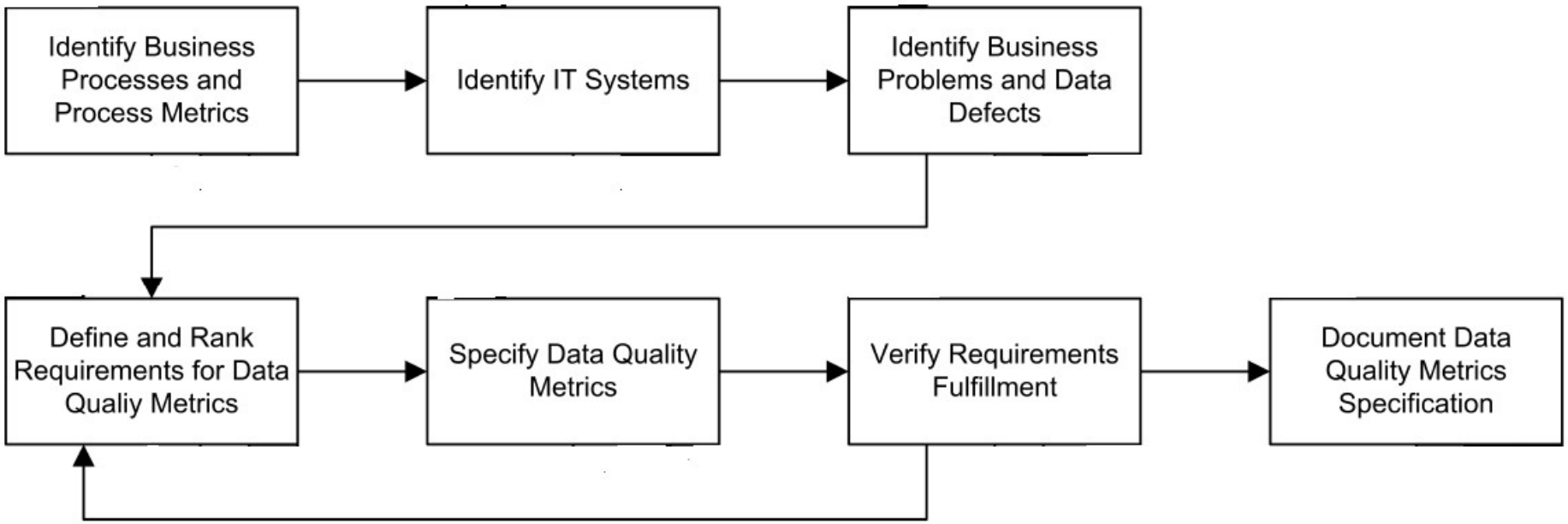
  - Problemas en los datos impactan performance de los procesos de negocio
  - Proponen considerar relación entre métricas de calidad de datos y métricas de procesos de negocio

  - **¿Cómo identificar las métricas de calidad de datos que son relevantes para sus métricas de procesos de negocio?**

Propuesta [Otto et. Al, ICIQ 2009]

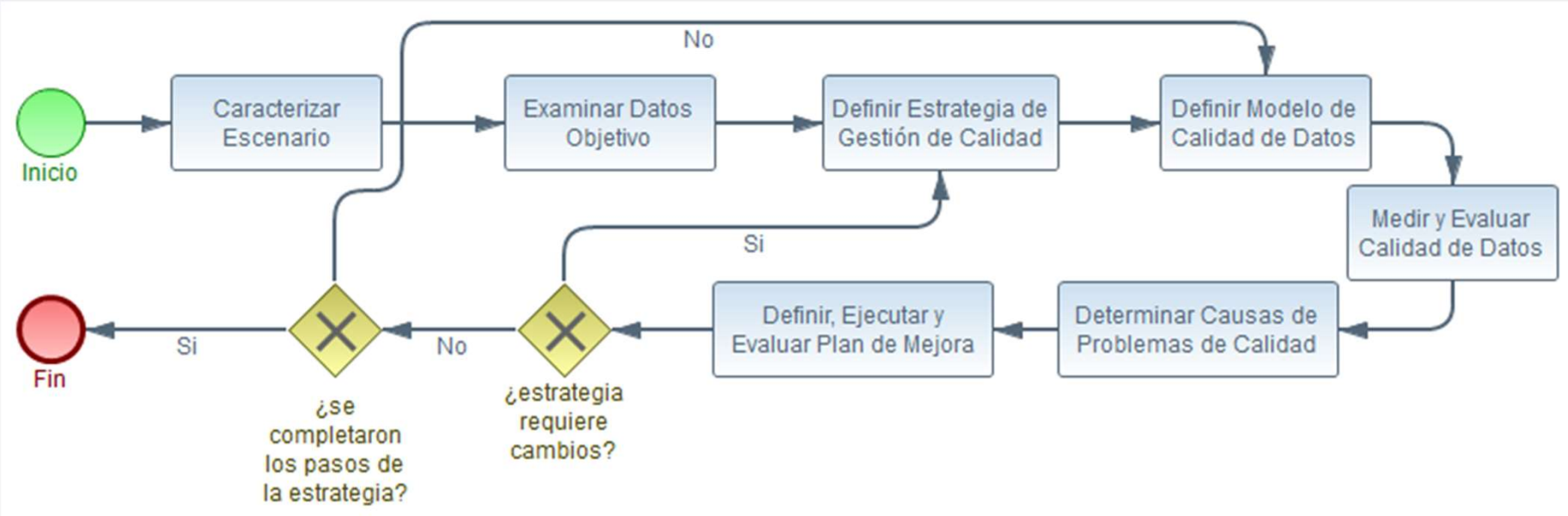


Propuesta [Otto et. Al, ICIQ 2009]

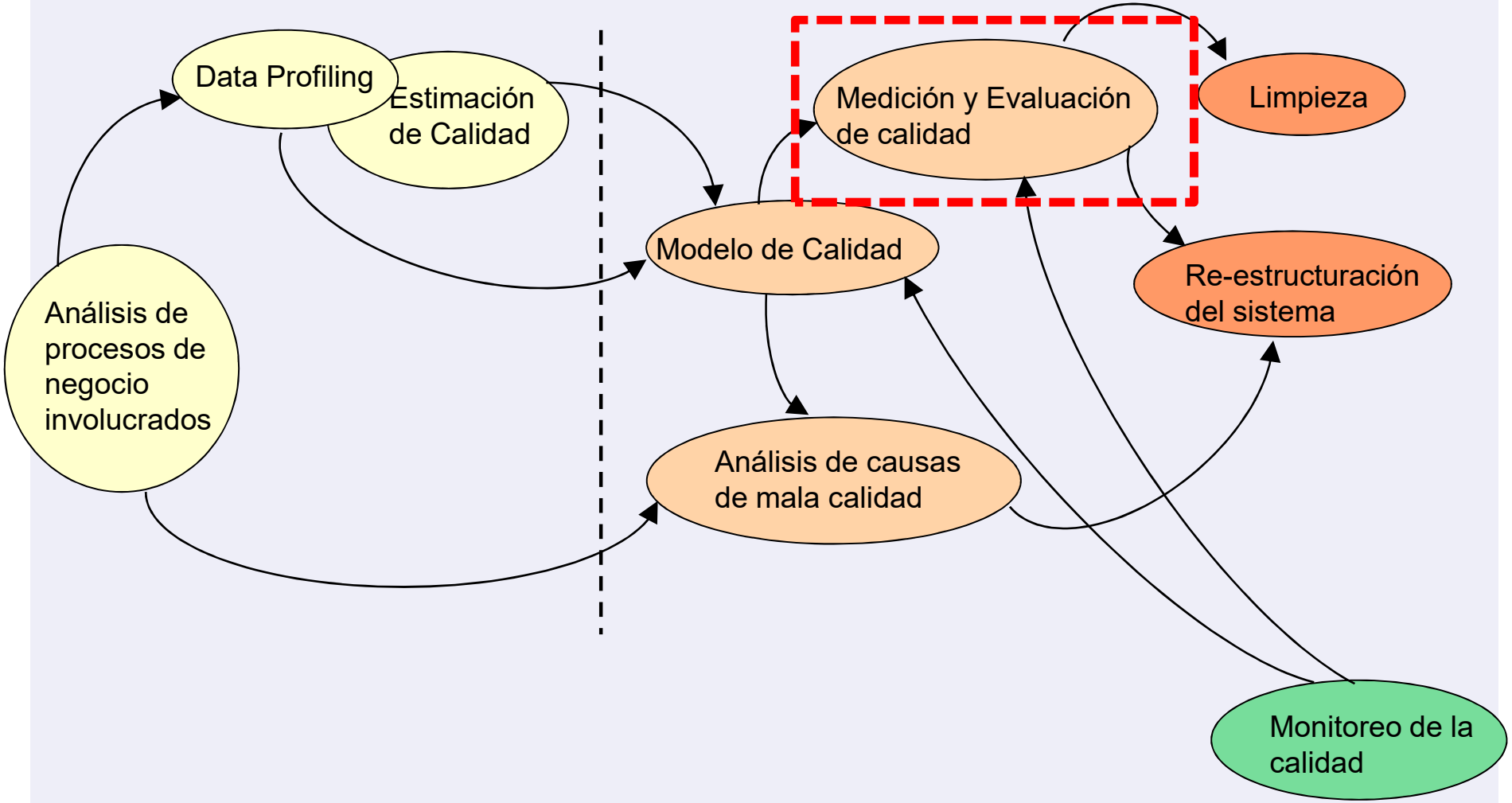


Proyecto Inco - AGESIC

- Proceso de Gestión de Calidad de Datos:



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 se le entregan
  - 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 de calidad
  - La métrica elegida
  - 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
  - Agregaciones que se van a calcular
  - Modelo de datos para los valores de calidad obtenidos: **Metadatos de Calidad**

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.

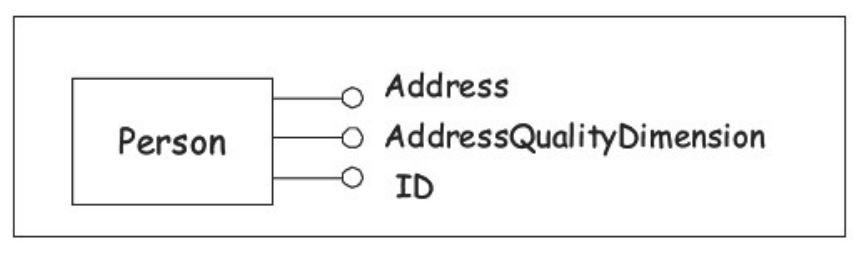
  – Se enriquecen los modelos convencionales con elementos para representar y analizar la calidad de los datos.

Metadatos de Calidad

- Modelos de datos
  - Modelado conceptual
    - Extensión del MER
  - Modelado lógico
    - Extensiones del Modelo Relacional
    - Extensión del modelo XML
- Modelo de proceso
  - Modelo para el proceso de la producción de información
    - IP-MAP

Extensión del MER

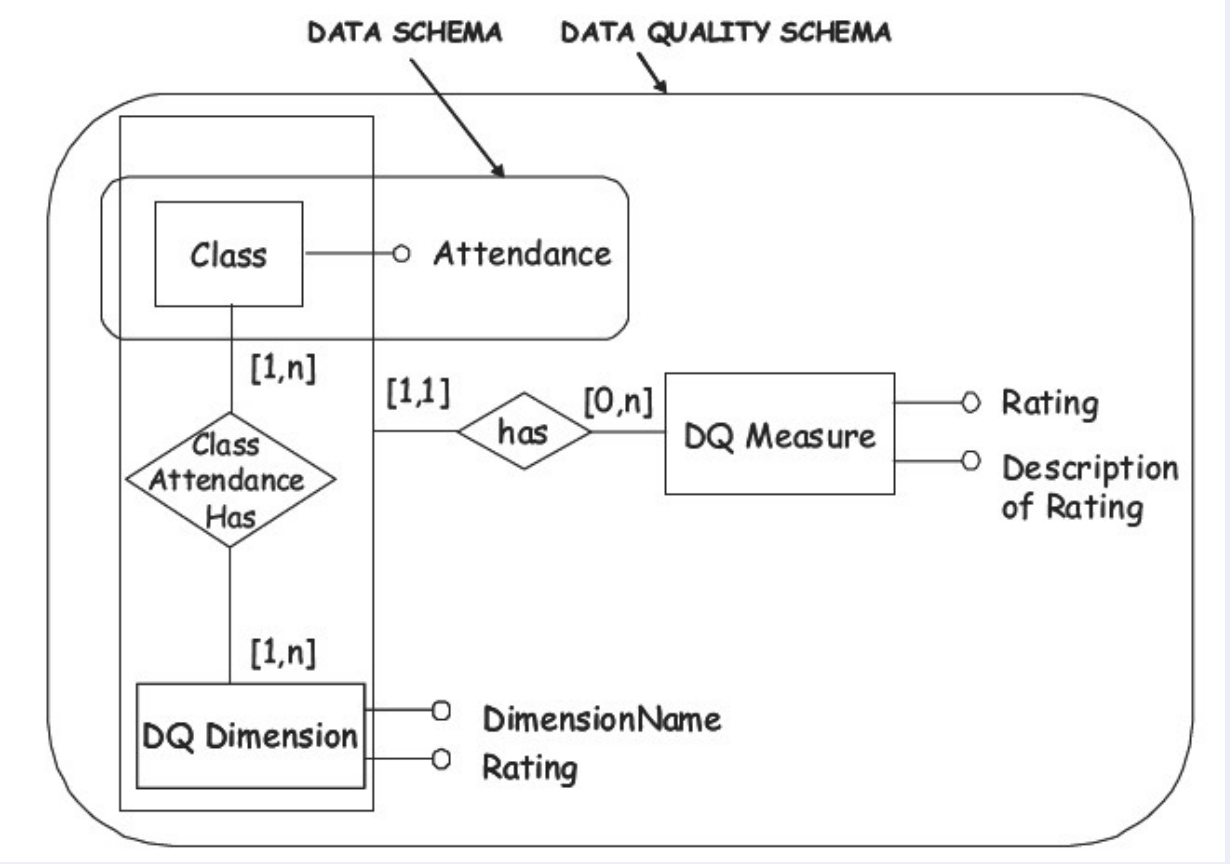
- Una posible solución



Desventajas?

- **Otra solución**
  - Agregamos 2 nuevas entidades:
    - Data quality dimension
      - Representa cada dimension y todos sus posibles valores
    - Data quality measure
      - Representa las mediciones

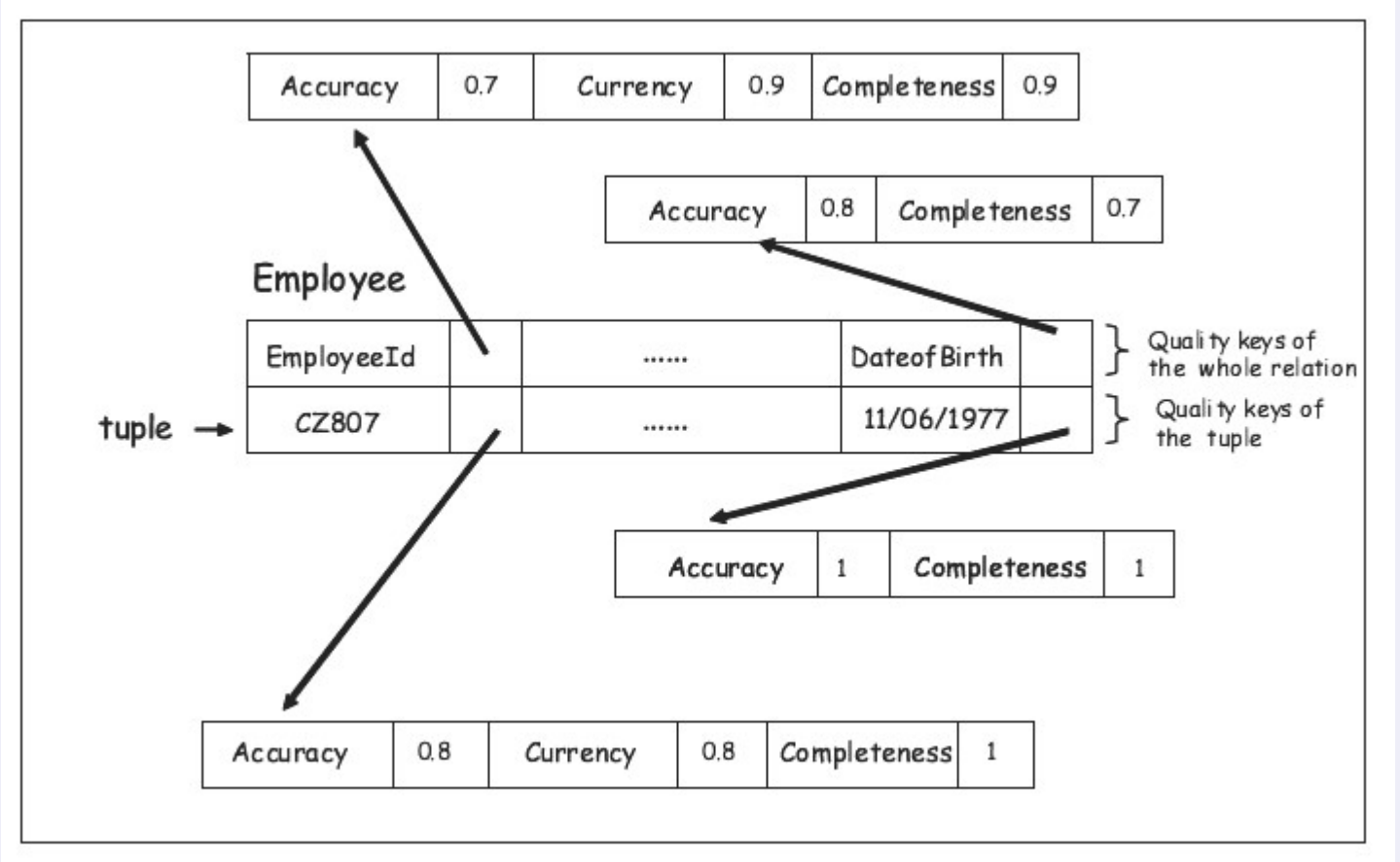
Extensión del MER



Qué cosas falta representar?
Qué construcciones habría que agregarle?

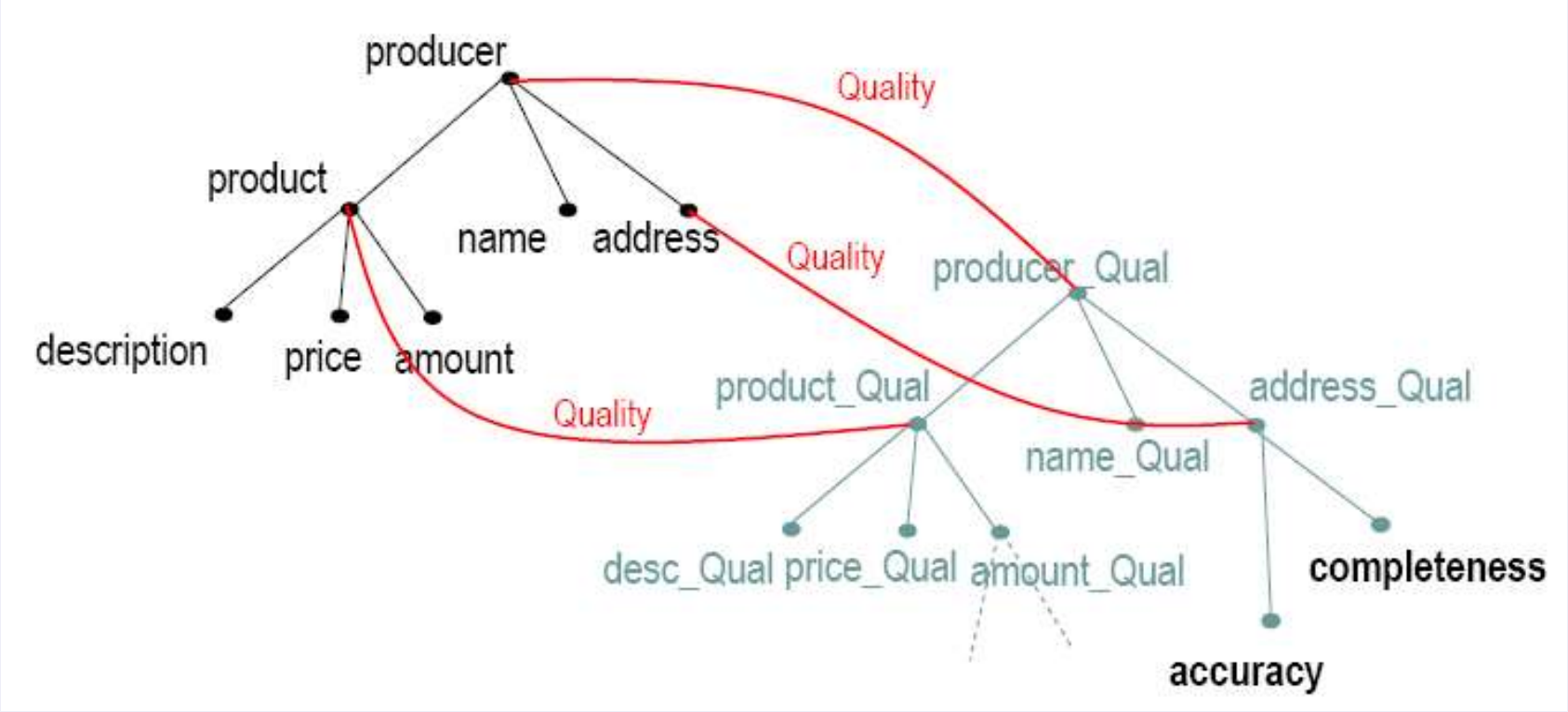
Extensiones del Modelo Relacional

- Basado en atributos



Extensión de XML

- *Data and Data Quality* (D^2Q)



Ejercicio – Metadatos 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 Gral	Métrica Inst. sobre
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
Completitud	Cobertura	M3: CoberturaRef gran: tabla tipo-res: [0,1]	Clientes

Evaluación de Calidad

- Requerimientos de calidad de datos
  - Umbrales establecidos por el usuario para cada dimensión/factor/métrica de calidad
  - 2 formas posibles de trabajar
    - Embeberlos en las métricas
      - 2 opciones
        - » Requerimientos fijos para todos los usuarios por igual
        - » “Parametrizar” las métricas
    - Comparar resultados de las mediciones con requerimientos de usuarios

Reqs. en las métricas

- Fijos
  - Ej.: Métrica de exactitud sintáctica – 0 o 1 según distancia a valor válido con respecto a un umbral fijo.
- Métricas “parametrizadas”
  - Concepto de ***Contexto de usuario***
    - Perfil, tarea, preferencias, requerimientos de calidad
  - Las métricas son ***dependientes del contexto***
  - Ej anterior: el umbral dependerá del usuario o del tipo de usuario

Comparar resultados con reqs.