Incidencia de la calidad y semántica de datos en Sistemas de Información Federados

Raul Ruggia, Regina Motz
InCo - Facultad de Ingeniería - UDELAR

CLEI 2002

Advertencia ... esta charla presenta

- **... problemas y no soluciones.**
  - **Describe arquitecturas conocidas: *BD Federadas*.**
  - **Introduce "nuevos" contextos de aplicación.**
  - **Presenta las limitaciones de las arquitecturas con respecto a casos que surgen en los "nuevos" contextos.**
  - **Sugiere posibles líneas de trabajo para resolver los problemas planteados.**

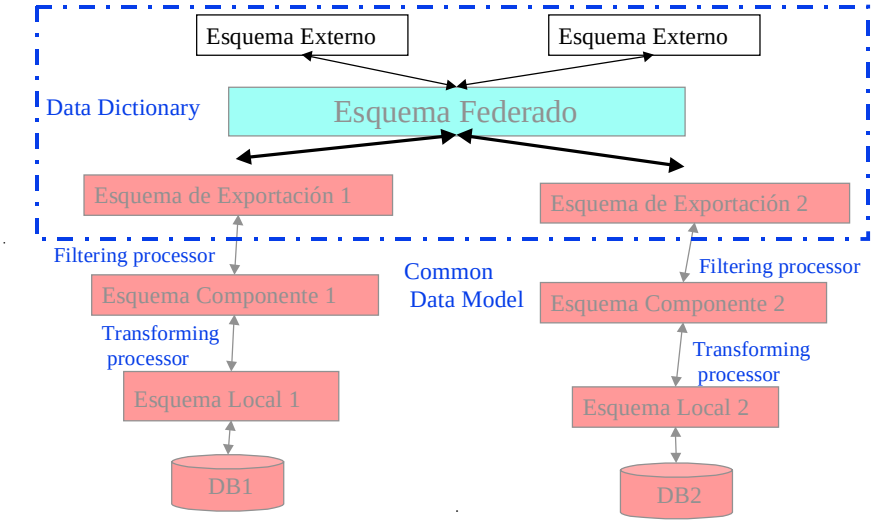
Contexto

- ´80: Distribución de datos (BDD).
- ´90: Datos distribuidos en distintas BD:
  - Bases de Datos Federadas
- ´2000: ….
  - Fuentes externas, muy numerosas y evolutivas (p.ej. Web).
  - Flexibilidad para manejar Escalabilidad y Evolución: Sistemas basados en Mediadores

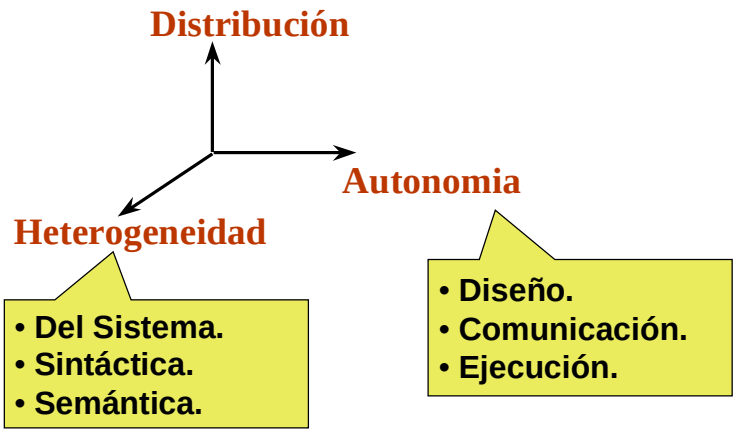
Contexto: Bases de Datos Federadas

- Principales Características
  - Construidas como una capa de integración sobre bases de datos ya existentes.
  - Soportan consultas distribuidas.
  - Respetan la autonomía de las BD Fuentes.
  - Manejan alto grado de heterogeneidad entre los componentes.

Arquitectura de referencia 5 Niveles



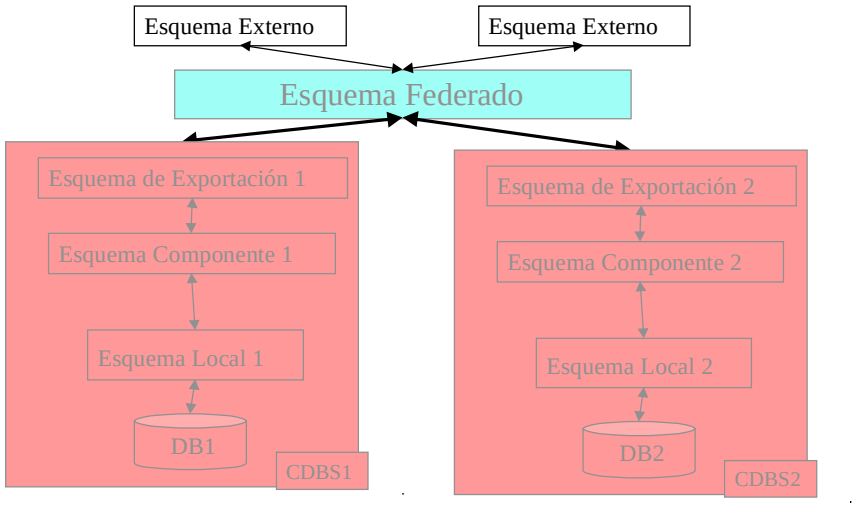
Aspectos Principales



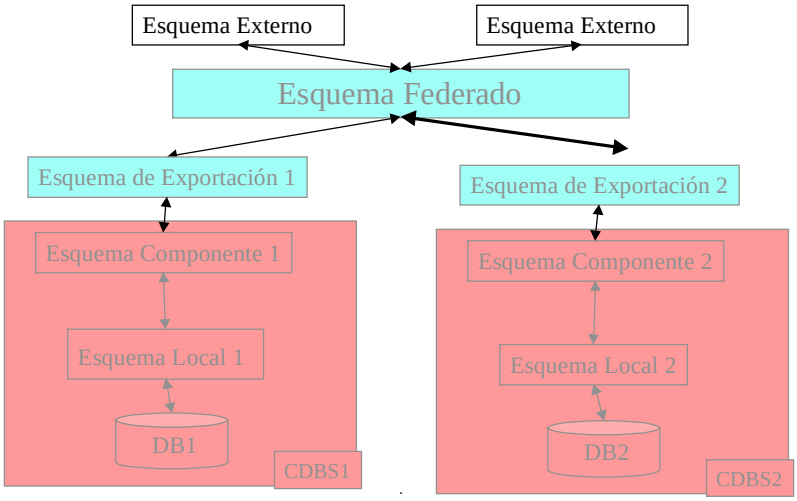
Semántica de los datos

- **Wood (1985):**
  - Significado y uso de los datos.
- **Sheth (1995):**
  - Mapeo entre el objeto modelado, representado y/o almacenado en un sistema de información y el objeto del mundo real que él representa.

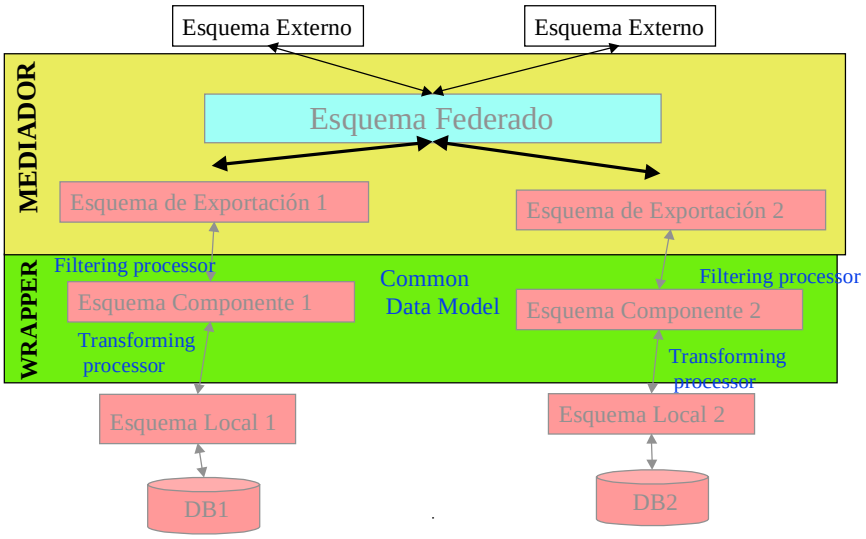
Arq. 5 Niveles FDBS (tight)



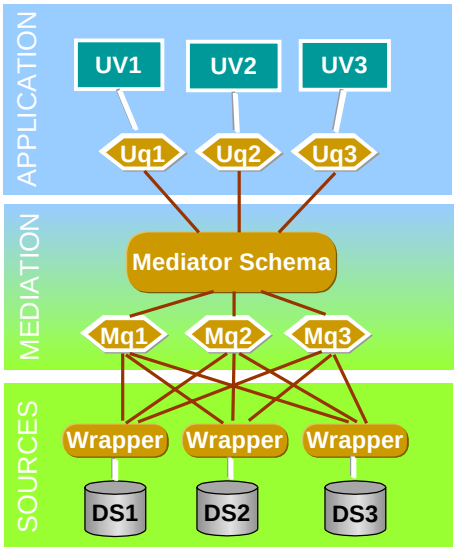
Arq. 5 Niveles FDBS (loose)



Arq. 5 Niveles - Basada en Mediadores



Arquitectura basada en Mediadores



- **Wrappers:**
  - Resuelven heterogenidad sintáctica y de sistema de Fuentes de Datos.
- **Mediator:**
  - Realiza la integración.
  - Incluyen esquema, equivalente federado.
  - Datos virtuales o materializados.
- **User Views:**
  - Datos usados en aplics.

Arquitectura basada en Mediadores

- **Principales aspectos a resolver:**
  - Diseñar *Mediator Schema*.
  - Definir Mappings, entre:
    - » Vistas del usuario y Mediator Schema.
    - » Mediator Schema y BD Fuente.
  - Mecanismo de resolución de consultas.
    - » Según relación entre BD Fuentes y Mediator Schema.
  - Definir *Mediation Queries*:
    - » Vinculan Mediator Schema y BDs Fuente.

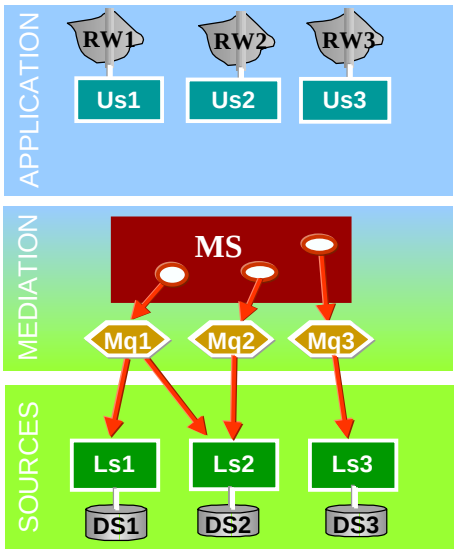
Resolución de Consultas Mediadas

- **Entrada:**
  - Consulta sobre el Mediator Schema.
- **Pasos:**
  - **Plan de la consulta:**
    - » Define una consulta a partir de un conjunto de candidatas equivalentes.
    - » Utiliza el conjunto de fuentes y consultas que las relacionan con el Mediator.
  - **Plan de ejecución:**
    - » Optimización y recolección de resultados
  - **Integración de resultados:**
    - » Remover redundancia y resolver conflictos de datos.

Plan de la Consulta

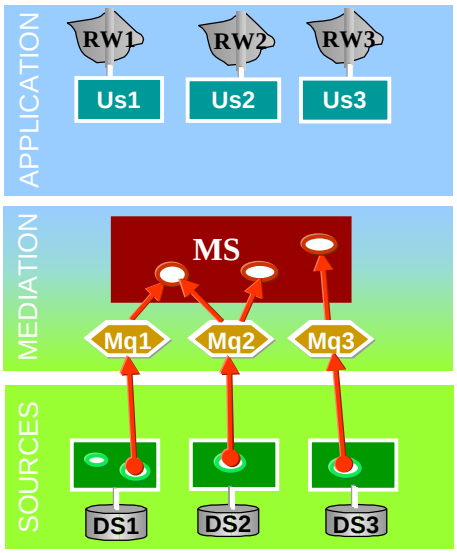
- **Basado en descripción de las fuentes respecto al esquema global.**
  - Global as View (GAV) (enfoque Top-down):
    - » Mediator schema es una vista de las Fuentes.
    - » [I-Manifold, DWQ, …]
  - Local as View (LAV) (enfoque Bottom-up):
    - » Fuentes son vistas del Mediator schema.
    - » [Tsimmis, SIMS,…]

Global as View (GAV)



- Cada objeto del Mediator Schema se define como una consulta sobre varias BD Fuentes.
- Consultas de usuario se reescriben en función de las BD Fuentes.

Local as View (LAV)



- Cada objeto de las BD Fuentes se define como una consulta sobre el Mediator Schema.
- Consultas de usuario se reescriben en función de las BD Fuentes.

GAV vs LAV

- ↗ **Encare natural y simple.**
  - ✓ *calidad del diseño depende de la definición del Mediator Schema.*
- ↗ **Facil de implementar en casos fijos y medianos o pequeños.**
  - ✓ *Re-escritura de consultas similar a reformulación de consultas con vistas.*
- ↘ **Poco flexible en evolución.**
  - ✓ *rediseño del Mediator Schema frente a evoluciones de BD.*

- ↗ **Mayor modularidad y evolutividad.**
  - – *Cambia Fuente = cambia 1 query.*
  - – *Agrega Fuente = agrega 1 query.*
- ↗ **Escalabilidad con aumento de BD Fuentes.**

Tratamiento complejo de consultas.

Limitaciones del modelo (1)

- **"Nuevo" contexto:**
  - – Abundantes fuentes de datos.
    - » BDs en áreas especializadas:
      - Biología Molecular (p.ej. 511 en InfoBIOGEN).
      - Cartografía.
      - Medio Ambiente.
    - » Información en Web.
      - Información semi-estructurada y multimedia.
      - Contactos.
  - – Fuentes heterogéneas.
    - » Interfaces generalmente no estructuradas, sin QL.
    - » Diversas estructuras: registros, texto, multimedia.

Limitaciones del modelo (2)

- **Características de las fuentes de dato**
  - Datos incompletos y solapados.
    - » Semántica dificil de descubrir
  - Calidad de datos poco
    - » Confiabilidad, A… …eutud, etc.
  - Accesibilid…
    - » Disponibili… …rmance.
  - Muy evolutiva… y con distintas frecuencias.
  - Muy autónomas.
    - » Dificultades para capturar los cambios.

Características NO corregibles por el cliente

Limitaciones del modelo (3)

- **Consecuencias.**
  - Cliente debe resolver los problemas anteriores:
    - » Elegir fuentes de datos según calidad y contenido.
    - » Realizar mappings.
    - » Realizar consultas con mappings complejos.
    - » Integrar resultados.
  - Aumenta complejidad con cantidad de fuentes.

➡ **No es escalable.**

➡ **No es generalizable.**

Encare a los problemas ...

- **Objetivos para el SI Federado:**
  - Caracterice con precisión fuentes de datos.
  - Seleccione de Fuentes de Datos con propiedades de calidad requerida (información y accesibilidad).
  - Integre datos heterogeneos teniendo en cuenta su semántica.
  - Procese consultas teniendo en cuenta la semántica de los datos.
  - Construya consultas de mediador que retornen información de calidad requerida.

Incorporando calidad y semántica (1)

- **Caracterizar con precisión fuentes de datos.**

- Seleccione de Fuentes de Datos con propiedades de calidad requerida (información y accesibilidad).
- Integración datos heterogeneos teniendo en cuenta su semántica.
- Procese consultas teniendo en cuenta la semántica de los datos.
- Construya consultas de mediador que retornen información de calidad requerida.

- **Asociar descripciones sobre semántica de datos.**
  - En base a descripciones de conceptos normalizadas.
- **Asociar propiedades de calidad de la Fuente.**
  - Propiedades de la Info.
    - » Confiabilidad, Actualización,
  - Propiedades de la Fuente.
    - » Disponibilidad, ...

Incorporando calidad y semántica (2)

- Caracterizar con precisión fuentes de datos.
- **Seleccionar Fuentes de Datos con propiedades de calidad requerida.**
- Integración datos heterogeneos teniendo en cuenta su semántica.
- Procese consultas teniendo en cuenta la semántica de los datos.
- Construya consultas de mediador que retornen información de calidad requerida.

- **Expresar requerimientos sobre Propiedades de Calidad.**
- **Propagar valores de calidad a través de la arquitectura.**
- **Comparar valores de calidad resultantes con los requeridos.**
- **Acceder a Fuentes con datos útiles s/semántica.**

Incorporando calidad y semántica (3)

- Caracterizar con precisión fuentes de datos.
- Seleccionar Fuentes de Datos con propiedades de calidad requerida.
- **Integrar datos heterogeneos teniendo en cuenta su semántica.**
- Procese consultas teniendo en cuenta la semántica de los datos.
- Construya consultas de mediador que retornen información de calidad requerida.

- **Usar descripciones semánticas en la integración de esquemas.**
- **Usar descripciones semánticas en la resolución de conflictos entre datos.**

Incorporando calidad y semántica (4)

- Caracterizar con precisión fuentes de datos.
- Seleccionar Fuentes de Datos con propiedades de calidad requerida.
- Integrar datos heterogeneos teniendo en cuenta su semántica.
- **Procesar consultas teniendo en cuenta la semántica de los datos.**
- Construya consultas de mediador que retornen información de calidad requerida.

- **Aplicar mappings "semánticos" en la construcción de las Mediation Queries.**
- **Aplicar información sobre semántica del contenido de Fuentes.**

Incorporando calidad y semántica (5)

- Caracterizar con precisión fuentes de datos.
- Seleccionar Fuentes de Datos con propiedades de calidad requerida.
- Integrar datos heterogeneos teniendo en cuenta su semántica.
- Procese consultas teniendo en cuenta la semántica de los datos.
- **Construir consultas de mediador que retornen información de calidad requerida.**

- **Calcular valores de calidad resultantes de las Mediation Queries.**
- **Comparar valores de calidad resultantes con los requeridos.**

Incorporando calidad y semántica (6)

- **Objetivos (revisados):**
  - **Dados:**
    - » Requerimientos de Calidad en User Views.
    - » Descripciones de Calidad y Semántica en Fuentes de Datos.
  - **Se desea obtener:**
    - » Diseño de las componentes de forma de que se cumplan los requerimientos de calidad.
      - Mediator Schema y Mediator Queries.

Trabajos existentes

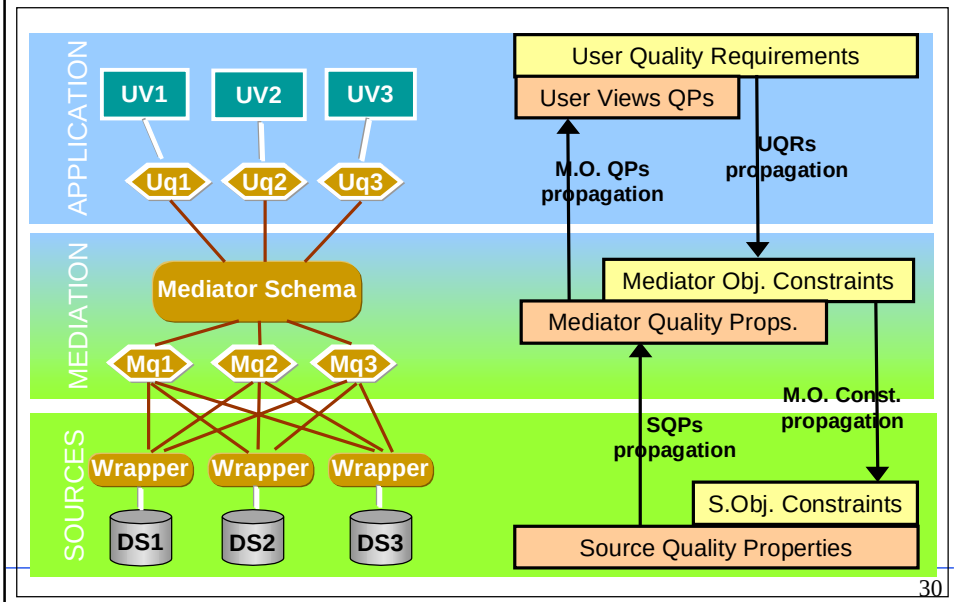
- **Selección de Fuentes de Datos.**
  - F. Naumann et al:
    - » [2001] From Databases to Information Systems Information Quality Makes the Difference.
    - » [1999] Quality-driven Integration of Heterogeneous Information
    - » [1998] Quality-driven Source Selection using Data Envelopment Analysis
  - [G. Mihaila et al. 01] Using Quality of Data Metadata four Source Selection and Ranking.
- **Propiedades de Calidad de Información.**

  [Wang, Strong, et al. 00] AIMQ: A Metodology for Information Quality Assessment.

Lineas de trabajo

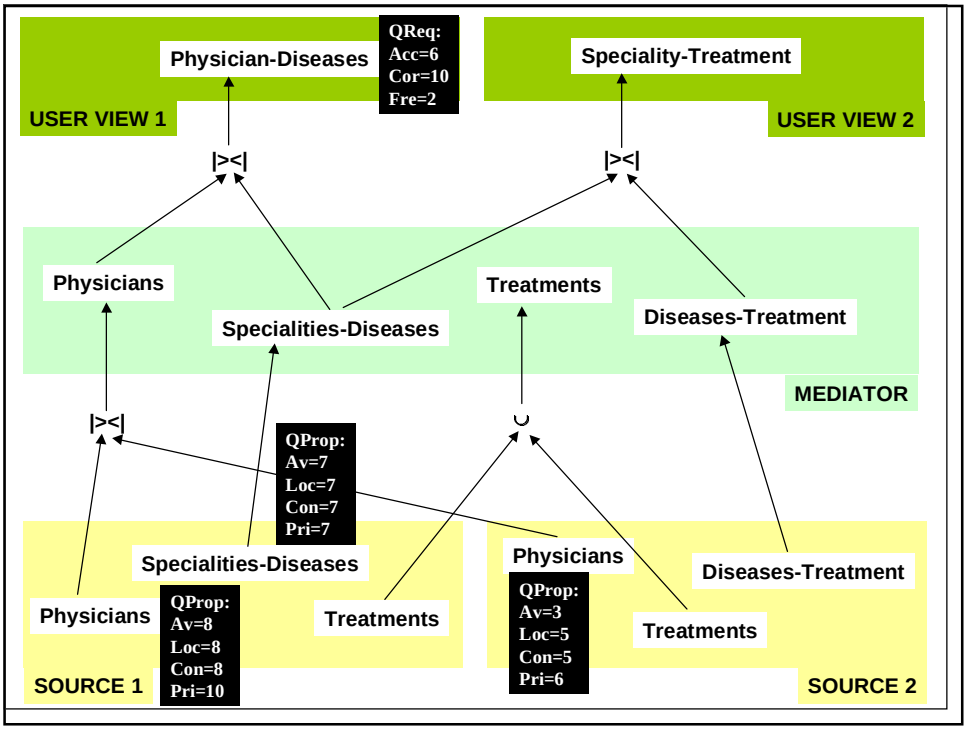
- **Especificar Semántica de datos.**
  - – Según estándares de Dominios de Aplicación.
  - – Tratable en la Integración y procesamiento de consultas.
- **Propiedades de Calidad.**
  - – Definición precisa y comparable de propiedades
  - – Especificación de Requerimientos y Valores.
    - » Integrar Requerimientos de usuarios.
  - – Propagación a través de operaciones.
  - – Evaluación de Valores vs. Requerimientos.

Quality Propagation and Evaluation



Lineas de trabajo

- **Generar Mediator Queries.**
  - Seleccionar BDs Fuentes.
  - Aplicar operadores según resultados de calidad deseados.
    - » Propagación de propiedades de calidad a través de operadores.
- **Diseñar Mediator Schema.**
  - De forma de cumplir con requerimientos de calidad.
  - Implica resolver los problemas anteriores.



Referencias

- M. Bouzeghoub. *Heterogeneous Data Sources Integration and Evolution*. DEXA 2002.
- A. Marotta. *Quality Management in MSIS*. Reporte Técnico 2002.

Muchas gracias