Organizaciones Virtuales e Integración de Información

José Abásolo Prieto

Universidad de los Andes

Objetivo de la charla

- Mostrar que aunque la problemática de integración de información distribuida y heterogénea ya se ha trabajado extensivamente en lo que se llaman Sistemas Multi-Bases de Datos, las Organizaciones Virtuales le añaden un nuevo nivel de complejidad al problema, al manejar distribución a gran escala.
- Mostrar las estrategias, que a nivel investigación, se están utilizando para resolver los problemas propios de la integración en Organizaciones Virtuales.

Agenda

1. Necesidad y definición de Organizaciones Virtuales (OV).
2. Requerimientos, características y preparación de OV.
3. Ontologías para representación y manipulación de conocimiento.
4. Computación en malla y Data Grids.
5. Organizaciones Virtuales para intercambio de información en Dominios Específicos.
6. Estrategias.
7. Conclusiones

Motivación: Por qué Organizaciones Virtuales: Ejemplos [1]

1. Un consorcio industrial formado para crear un estudio de factibilidad para un nuevo avión supersónico, requiere hacer una simulación multidisciplinaria del aparato completo. Esta simulación involucra componentes de software propietario desarrollados por diferentes participantes, cada componente operando en los equipos de ese participante y teniendo acceso a datos aportados al consorcio por sus miembros.

Motivación: Por qué Organizaciones Virtuales (2)

2. Miles de físicos y otros científicos en cientos de laboratorios y universidades alrededor del mundo, se unen para diseñar, crear, operar y analizar los datos generados por un acelerador de partículas en el CERN, el laboratorio europeo de altas energías. Durante la fase de análisis, ellos comparten recursos de computación, almacenamiento y conectividad para poder analizar *petabytes* de datos.

Organización Virtual: Definición

Conjunto de organizaciones autónomas que colaboran entre sí, compartiendo recursos, trabajando en la obtención de un objetivo común [1].

Organizaciones Virtuales: Dominios de Aplicación

- Industria
- Ciencia
- Ingeniería

Organizaciones Virtuales: Requerimientos

- Esquemas para compartir altamente flexibles: de cliente-servidor a peer-to-peer
- Niveles de control sofisticados y precisos sobre cómo son usados los recursos compartidos
- Variedad de recursos a compartir: programas, archivos, datos, servidores, sensores, redes, etc.

Organizaciones Virtuales: Características

- Participantes suelen tener sus propias infraestructuras tecnológicas y sus propias políticas en cuanto a acceso, autorización y uso de la información.
- Datos almacenados en diferentes formatos y bajo diferentes esquemas en lo que se conoce como heterogeneidad sintáctica, semántica y de acceso.

Organizaciones Virtuales: Características (2)

- Alto número de fuentes de datos
- Altos volúmenes de datos.
- Alta distribución.
- Replicación de datos.

Organizaciones Virtuales: Preparación

Acciones típicas de preparación para una red de organizaciones que quieren cooperar incluyen:

- Desarrollo de procedimientos, estándares, procesos comunes y tecnologías de información y comunicaciones para soportar requerimientos hechos a la organización virtual.
- Crear y compartir conocimiento común.

Ontologías para representación de conocimiento: Definición

Ontología:

- Define los principales términos y relaciones que comprenden el vocabulario de un área, así como las reglas para combinar términos y relaciones para extender este vocabulario [2].
- Especificación explícita y formal de una conceptualización [3].

Ontologías para representación de conocimiento: Elementos

1. Clases.
2. Propiedades.
3. Jerarquía de herencia entre clases.
4. Jerarquía de herencia entre propiedades.
5. Restricciones para propiedades: dominio, rango, …
6. Instancias.

Requerimientos para Lenguajes Ontológicos

1. Sintaxis bien definida.
2. Semántica formal.
3. Conveniencia de expresión
4. Soporte de razonamiento eficiente.
5. Suficiente poder expresivo.

Semántica Formal

- Describe precisamente el significado del conocimiento.
- Permite razonar. Para conocimiento ontológico, algunos ejemplos de razonamiento son:

  - *Pertenencia a una clase*: Si x es una instancia de la clase C, y C es una subclase de D, entonces x es una instancia de D

Semántica Formal (2)

- *Equivalencia de clases*: Si la clase A es equivalente a la clase B, y B a C, entonces A es equivalente a C.
- *Clasificación*: Si hemos declarado que ciertas parejas propiedad-valor son una condición **suficiente** para pertenecer a una clase A, entonces si el individuo x satisface tales condiciones, podemos concluir que x es una instancia de A.

Lenguajes para definir ontologías

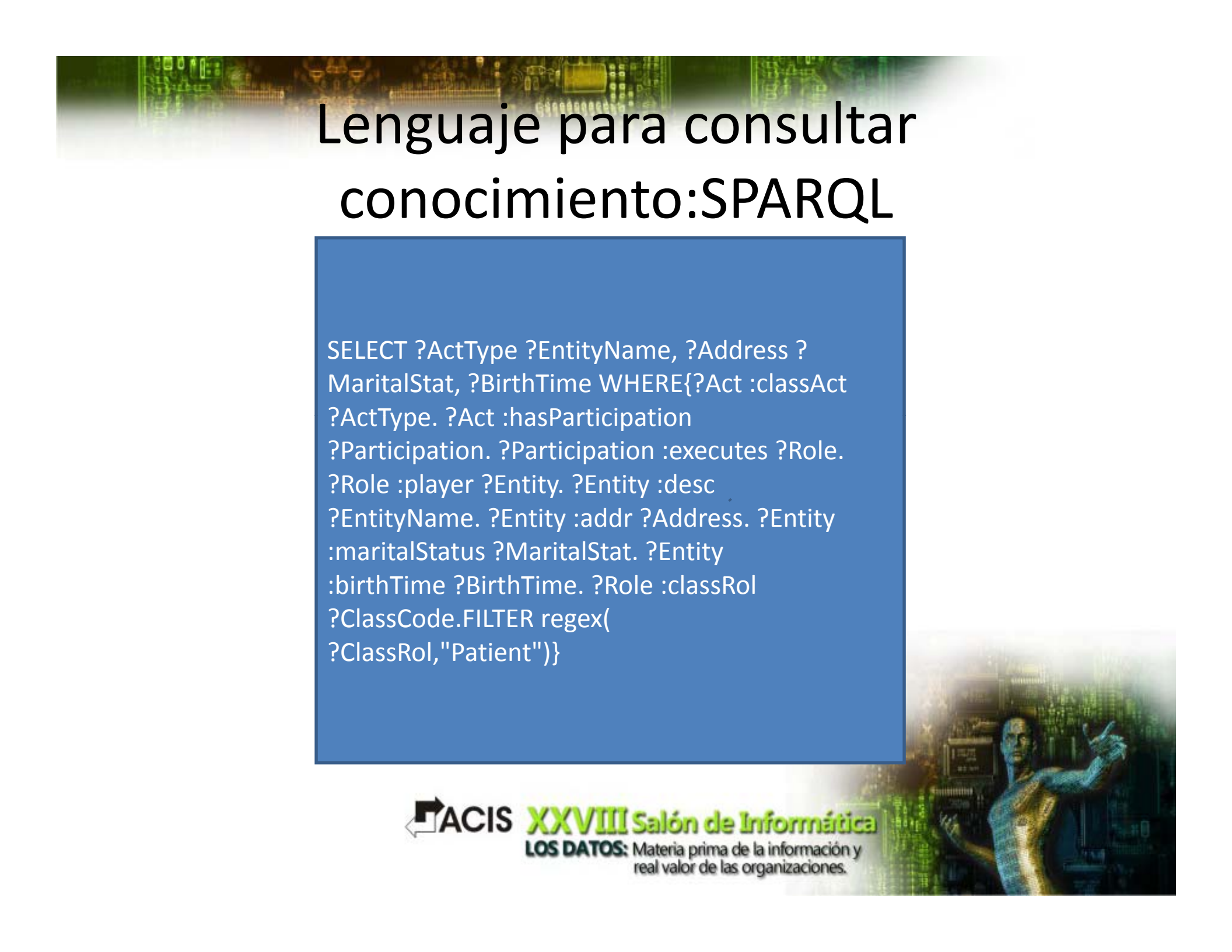
Existen lenguajes para expresar las ontologías, con mayor o menor riqueza semántica y con más o menos formalismo matemático.

Ejemplos:

- Diagramas de clase UML
- Esquemas de Bases de Datos
- Tesauros
- Taxonomías
- RDF
- OWL

Lenguaje para consultar conocimiento:SPARQL

```
SELECT ?ActType ?EntityName, ?Address ?
MaritalStat, ?BirthTime WHERE{?Act :classAct
?ActType. ?Act :hasParticipation
?Participation. ?Participation :executes ?Role.
?Role :player ?Entity. ?Entity :desc
?EntityName. ?Entity :addr ?Address. ?Entity
:maritalStatus ?MaritalStat. ?Entity
:birthTime ?BirthTime. ?Role :classRol
?ClassCode.FILTER regex(
?ClassRol,"Patient")}
```



Grid Computing y Data Grids

- *Grid Computing*: Paradigma de computación en Internet que propone agregar recursos de cómputo, almacenamiento y comunicaciones, geográficamente distribuidos y heterogéneos, para ofrecer acceso unificado y seguro a sus capacidades combinadas [ ].
- *Data Grids*: Proveen servicios que ayudan a los usuarios a descubrir, transferir y manipular grandes conjuntos de datos almacenados en repositorios distribuidos, y también a crear y administrar copias de esos datos.

OGSA: Open Grid Services Architecture

- Basada en el paradigma de *servicios web*: componentes auto-contenidos, sin estado, que usan mecanismos estandar parala representación e intercambio de datos.
- OGSA construye sobre propiedades de los *servicios web*, tales como definiciones de servicios usando XML y protocolos de comunicación estandar tales como SOAP, para crear *servicios grid*: interfases *web service* estandarizadas que ofrecen capacidades Grid.

GLOBUS Toolkit

- Conjunto de servicios y librerías con fuente y arquitectura abierta, que permiten la implementación de aplicaciones Grid.

OGSA-DAI: Open Grid Services Architecture, Data Access Integration

- Componente de la capa de datos de Globus Toolkit.
- Permite:

  - Saber cuáles son las fuentes de datos disponibles y qué se puede hacer con ellas.

  - Acceder a las fuentes de datos, encargándose de la traducción a las características particulares de la fuente.

Organizaciones Virtuales para intercambio de información en Dominios Específicos

Ejemplos:

- Organización Virtual en Salud.
- Organización Virtual en Educación.
- Organización virtual en Banca y Seguros.

Organizaciones Virtuales para intercambio de información en Dominios Específicos (2)

Características:

- Información fragmentada, distribuida, heterogénea y posiblemente replicada de manera independiente.
- Muchas fuentes.
- Se conoce información intensional (esquema)de las fuentes, mas no extensional (instancias).
- Para una entidad particular, no se sabe qué fuentes tienen información sobre ella.

Organizaciones Virtuales para intercambio de información en Dominios Específicos (3)

- Problemática parecida a la de los Sistemas Multi-Bases de Datos.
- Diferencia: distribución a gran escala, que hace ineficiente descartar fuentes utilizando únicamente información intensional.

Estrategias de Integración

1. Manejo de conceptualizaciones comunes (ontologías de referencia).
2. Equivalencias entre conceptos locales a una fuente de datos y conceptos comunes.
3. Definición de objetos virtuales para consumo de los usuarios de la organización virtual.
4. Generación de información extensional.

Manejo de conceptualizaciones comunes (ontologías de referencia).

- En el caso de una Organización Virtual en Salud, la ontología utilizada puede ser el RIM de HL7
  - ➔ Compleja para un usuario común
  - ➔ Necesidad de Objetos Virtuales

Generación de Información Extensional

Posibles Estrategias:

1. Fuente de datos referencial natural. Ejemplo: Bodega de Datos que integra información de protección social de los habitantes de un país.
2. Aprender de resultados de consultas anteriores.
3. Minería de datos a las fuentes: *Clustering* (Segmentación).
4. Conocimiento suministrado por especialista de la fuente.

Generación de Información Extensional (2)

Implantación estrategias:

1. Un solo lenguaje de representación para conceptos, equivalencias, clusters, etc. Puede ser OWL.
2. Índices sobre conocimiento.
3. Clausura transitiva: hacer inferencias cuando se agrega el conocimiento, no cuando se consulta.
4. Particiones.

Conclusiones

- Filtrar fuentes involucradas en una consulta es crítico en distribución a gran escala.
- Para filtrar, además de información intensional, se necesita información extensional.
- La minería de datos puede ayudar en la obtención de información extensional.
- Por eficiencia, se requiere de índices, agregados y particiones sobre la metadata.
- Las Bodegas de Datos podrían jugar un papel, si se las mira como fuente referencial.
- Los Data Grids tienen una problemática similar, y son una buena opción para soportar integración en organizaciones virtuales.
- Pasos: Primero lograr la funcionalidad, luego la eficiencia.

Bibliografía

[1] Foster, I., Kesselman, C., Tuecke, S.: *The anatomy of the grid: Enabling scalable virtual organizations*. Int. J. High Perform. Comput. Appl. 15, 200–222 (2001)

[2] Neches, R. et al : *Enabling technology from knowledge*. AI magazine, Vol 12, 1991.

[3] Antoniou, G., van Harmelen, F. : *A semantic web primer*. The MIT Press, 2004.

[4] Venugopal, S. et al : *A Taxonomy of data grids for distributed data sharing, management, and processing*. ACM CS, Vol. 38, March 2006, Article 3.