# Integración de Sistemas Información Geográficos

Mercedes Vitturini - Pablo Fillottrani*
{mvitturi, prf}@cs.uns.edu.ar

LISSI - Laboratorio de Investigación y Desarrollo
en Ingeniería de Software y Sistemas de Información
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur
Bahía Blanca - Argentina

## Resumen

Por varios años los Sistemas de Información Geográficos(SIG)[1] avanzaron independientemente definiendo sus propios modelos de representación que se ajustaran a un problema particular. Actualmente se necesita dar respuesta al requerimiento de integrar datos de origenes dispares y, en el caso de información con referencias geográficas, se añaden consideraciones particulares. La solución se encamina hacia definir modelos conceptuales mediadores claros y precisos que comuniquen infomación a personas, organizaciones y otros sistemas que deseen interactuar con el mismo. En este trabajo se recopilan y delinean los principales logros en conceptaulización y normalización de información geográfica.

## 1. Introducción

Los *Sistemas de Información Geográficos* surgieron en los '70 y evolucionaron hasta alcanzar un desarrollo, uso y aceptación considerables. El volumen y complejidad de los datos geográficos dio origen en los '90 a los *Sistemas de Manejo de Bases de Datos Espaciales* (SMBDE)[2], un tipo particular de Sistema de Manejo de Bases de Datos (SMBD) con capacidad para administrar información espacial, esto es, relacionada a una posición geográfica. Los SMBDE abstraen de la solución SIG los problemas de representación y gestión de información con componentes geográficos.

Los SIG abarcan una amplia variedad de tipos de aplicaciones: análisis demográfico, gestión de servicios públicos, análisis de redes de comunicación, entre otros. Los usuarios de SIG incluyen empresas privadas, organismos estales o gubernamentales y organizaciones internacionales. En la medida que la tecnología lo permita, sus propietarios están interesados en crear y mantener

*Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC).

[1] Geographic Information System – GIS

[2] Spatial Database Management System – SDBMS

sus bases de datos, así como en compartir información. Los distintos ámbitos de aplicación, en general comparten las siguientes necesidades:

- El SIG necesita estar integrado con otras aplicaciones. Es importante que los datos del SIG se almacenen y organicen de forma tal que permitan el acceso distribuido.

- Es importante considerar la integración de los datos geográficos con otros datos como pueden ser: datos de tiempo real, imágenes o bases de datos corporativas.

- Se debe optar por la estructura de datos apropiada que permita obtener la clase de análisis adecuada al dominio y de esto depende el tipo de consultas que se podrán realizar al sistema.

La Web Semántica está relacionada con la noción de interoperabilidad semántica. Busca representar la información disponible en la Web de forma que pueda ser interpretada por máquinas. Esto permite usar la información más inteligentemente que simplemente para mostrarla e incluye compartir y reusar datos entre aplicaciones. En [Ege02] se define la necesidad una Web Semántica Geoespacial, basada en un marco que abarque multiples ontologías espaciales temáticas, así como una forma canonizada de especificar consultas geoespaciales. Relacionado con el concepto de interoperabilidad tamién está la noción de Infraestructura de Datos Espaciales (IDE)[3] fuera del alcance de este trabajo.

[3]Spatial Data Infraestructure SDI

# 2. Modelos de Datos Geográficos

Los *modelos de datos* son un recurso para representar objetos del mundo real en la computadora. Una práctica común aplicada por los SMBD (y los SMBDE) es definir modelos de datos con diferentes niveles de abstracción: conceptual, lógico y físico. El modelo conceptual básicamente sirve a los efectos de clasificar, identificar y representar los fenómenos del mundo que se están modelando. Este modelo se utiliza como medio de comunicación entre el equipo de desarrollo y los clientes o usuarios del sistema. Las herramientas más populares para diseñar modelos conceptuales de datos son el *Modelo Entidad−Relación* (MER) y en entornos orientados a objetos el Lenguaje de Modelado Unificado (UML), específicamente el *Diagrama de Clases del Dominio*. El modelo conceptual no es directamente "representable" en la computadora, por ejemplo, para el caso de elementos geográficos el modelo conceptual se abstrae de cómo representar conjuntos infinitos de puntos. El modelo lógico o modelo discreto se ocupa de la representación adecuada para el modelo conceptual. Los diseñadores de SIG's interactúan con estos dos primeros niveles de abstracción para diseñar su propio modelo de datos. El nivel físico por su parte se encarga de administrar el almacenamiento persistente de la información.

## 2.1. Fenómenos Geográficos

Un modelo conceptual con capacidad para representar *información geográfica* (IG) debe capturar dos aspectos del problema. Por una parte debe proveer constructores para representar *atributos con valores de datos geográficos* que no pueden definirse directamente con tipos de datos

tradicionales. Por ejemplo si se desea mantener el atributo *curso* para objetos de tipo *río*. Además debe describir el *esquema de aplicación geográfico*. El esquema contiene los constructores para los objetos que en aplicaciones SIG se denominan *fenómenos*[4]. Un modelo conceptual que represente fenómenos geográficos necesita elementos para definir:

- *Las propiedades o atributos que describen el fenómeno*, incluyendo atributos descriptivos o geográficos. Los atributos geográficos representan las propiedades geográficas del fenómeno.
- *Relaciones entre fenómenos*, esto es el conjunto de asociaciones que existen entre dos o más fenómenos del mundo real.
- Aspectos relacionados con *el comportamiento del fenémeno*.

En la solución de un problema particular, cada emprendimiento SIG define sus propios modelos conceptuales de acuerdo a sus necesidades y restricciones.

### 2.2. Estándares sobre información geográfica

Inicialmente cada aplicación SIG no sólo definió sus modelos conceptuales, lo que es natural, sino también sus propios modelos de datos, formatos de almacenamiento y servicios. Todo esto llevó a serios problemas de interoperabilidad entre diferentes herramientas SIG y aún entre sistemas desarrollados con laA misma herramienta. En la solución de estos problemas trabajan desde la década del '90 en conjunto el Consorcio Open Gis (OGC) [OGC] y la Organización Internacional de Estándares (ISO) [211] a través del Comité Técnico ISO 211 (ISO/TC 211).

A los efectos de permitir interoperabilidad ambas organizaciones concensuaron en un conjunto de reglas que proveen los conceptos para definir los objetos geográficos del mundo real. Los resultados más importantes en estándares a nivel conceptual se encuentran en ISO 19107, 19109 y el Lenguaje de Marcado Geográfico (GML).[5]

## 3. Integración de Información Geográfica

A través de los años varios problemas se han enfrentado con la necesidad de compartir y reutilizar el conocimiento adquirido sobre un dominio. Compartir conocimiento incluye transferir el saber de una persona a otra, de una organización a otra, de un grupo a una persona, entre otros modelos de colaboración. La comunicación distingue dos entidades: el *enviador* y el *receptor*. Muchas veces ocurre que el receptor y el que envía son entidades arbitrarias que no comparten el mismo lenguaje, terminología, ni modelo mental, lo que hace difícil o imposible descifrar el mensaje. La respuesta a este problema está en estructurar el mensaje para que el receptor pueda entenderlo. Las ontologías potencialmente se ocupan de estos problemas. Una ontología es una especificación formal de los términos del dominio de una aplicación y las relaciones que existen entre ellos [Gru93]. La ontología define un vocabulario común entre varios usuarios que desean compartir información relativa a un dominio. Más interesante aún es que las definiciones de conceptos del dominio de aplicación y sus relaciones son

[4] feature

[5] Geographic Markup Languaje

potencialmente interpretables por agentes automaticos.

### 3.1. Integración en Bases de Datos

Las aplicaciones que trabajan con bases de datos utilizan modelos de representación estructurados para su información (esquemas relacionales, XML y DTDs). En general existe más de una representación posible para un mismo problema y es así, que si se quiere integrar dos o más aplicaciones se necesitan resolver heterogeneidades en lo que respecta a los esquemas o modelos y a los datos en sí mismos. Las investigaciones para obtener sistemas de integración se enfocaron a resolver estos dos problemas, con propuestas alternativas [DH05, GMPQ$^+$04, IFF$^+$99] y distinos resultados.

En general se puede decir que las propuestas en integración semántica apuntan a traducir datos de distintas fuentes conforme a un esquema destino. Una interface uniforme, denominda *esquema mediador* se utiliza para integrar múltiples fuentes de datos. La figura 1 ilustra un ejemplo de aplicación en el ámbito inmobiliario. En el ejemplo, el sistema de integración de datos permite a los usuarios buscar propiedades en diferentes bases de datos inmobiliarias. La consulta del usuario se realiza sobre el esquema mediador. El sistema usa el conjunto de traducciones semánticas entre el esquema mediador y los esquemas locales para trasladar la consulta a los esquemas fuentes. La misma se ejecuta usando un programa seguro (wrapper) en las fuentes de datos. Finalmente los resultados parciales se combinan y se retorna un única respuesta a la consulta del usuario.

Otro problema en la construcción de sistemas de integración de datos es definir *reglas de coincidencia semántica*. Las distintas fuentes de datos además pueden compartir parte de la información, esto es, no siempre el problema es integrar datos que genuinamente corresponden a conjuntos disjuntos [BM03]. La integración en estos casos trae un problema adicional que es el de detectar y eliminar elementos duplicados en las respuestas particulares antes de presentar el resultado final.

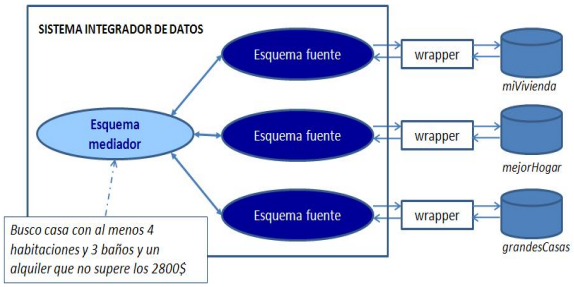


Figura 1: Sistema Integrador de Datos

### 3.2. Integración de Fenómenos Geográficos

La representación de IG distinguen los componentes espaciales y temáticos. Las variables temáticas se clasifican en *cualitativas* o *cuantitativas* según los valores que asumen. Los temas cuantitativos tienen rango de valores numéricos, que pueden ser contínuos (temperatura, precipitaciones) o discretos (población, número de especies). Otra característica de las variables temáticas cuantitativas es que se pueden agrupar en intérvalos. Así, por ejemplo la temperatura se podría clasificar por rango en "caluroso", "templado" o "fresco" asociando a cada valor con intervalos. Para el caso de variables temáticas cualitativas, existe un rango de valores posibles nominal discreto, cada valor típicamente se vincula a un término. El problema de distribución del uso de la tierra utiliza variables cualitativas.

Para definir interoperabilidad semántica relacionada con IG una restricción común es limitarse a modelos con una única variable temática sea esta cuantitativa o cualitativa y evitar así agregar la complejidad extra de heterogeniedad estructural. La interoperabilidad temática se maneja de distinta manera dependiendo si se requiere integrar datos cuantitativos o cualitativos y en particular la interoperabilidad semántica de temas cualitativos es más compleja.

### 3.3. Uso de ontologías en integración de SIG

La *heterogeneidad semántica* [Ter06] se refiere al hecho que distintos agentes representan fenómenos del mundo de diferente manera. En el caso de SIG, estas categorías corresponden a conceptos temáticos, y en consecuencia la semántica está principalmente relacionada con el componente temático de la IG que se esta modelando. Las soluciones al problema de integración semántica se relacionan con la definición de ontologías que provean la especificación formal de los esquemas de representación de cada SIG.

Además de definir las ontologías se requieren servicios semánticos para integrar los datos de diferentes fuentes de información. El Consorcio de Conocimiento de la Web (KnowledgeWeb Consortium 2005) identifica tres niveles de hetegoneidad semántica, para cada uno de estos niveles puede ser necesario definir una ontología:

1. El *nivel sintáctico*, que se refiere al hecho que distintas ontologías pueden estar expresadas en diferentes lenguajes (OWL o KIF).

2. El *nivel terminlógico*, comprende las diferencias en la denominación de los conceptos. Por ejemplo sinónimos, diferencias relacionadas con el idioma (Español, Inglés, etc), polisemias (sobrecarga de un término o derivados (sufijos o prefijos).

3. El *nivel conceptual*, incluye las diferencias relacionadas con el contenido de la ontología. Existen dos tipos de diferencias conceptuales: diferencias *metafísicas* que se refieren a "cómo está dividido el mundo en piezas", esto es, que entidades, propiedades y relaciones están representadas en la ontología; y diferencias *epistémicas*, que se refieren a "cómo se entienden las entidades", esto es, que aserciones representan. En particular las diferencias metafísicas se subdividen en tres tipos: *cubrimientos*, distintas ontologías cubren diferentes porciones del mundo, *granularidad*, una ontología puede dar una descripción con distinto grado de detalle para los mismos conceptos que otra y *perspectiva* ontologías diferentes para el mismo mundo real pero con distintos puntos de vista, como es el caso típico del modelado desde diferentes disciplinas.

## 4. Conclusiones

La posibilidad real de contar con acceso a diversas fuentes de información heterogéneas distribuidas en la red motiva el crecimiento de la investigación en el área de integración de información. Particularmente en el dominio de SIG's combinar datos que provienen de multiples modelos de representación heterogéneos hace que la integración de IG no sea un problema trivial. Algunas propuestas de solución en este tema apuntan hacia abstraer y conceptualizar el dominio de IG a través de la definición de ontologías que sirvan como una herramienta semántica traductora a un lenguaje de interpretación común para

personas, organizaciones y otros sistemas que requieran interacción.

En este trabajo se presentan los lineamientos en investigación sobre modelos de representación de conocimiento e integración de IG que forman parte del proyecto investigación "Lenguajes e Inferencias para Representación del Conocimiento y Bases de Datos" de la Universidad Nacional del Sur. Como parte de las inventigaciones se recopilarán y organizarán los distintos problemas en integración de IG y las soluciones obtenidas hasta el momento enfocadas en el uso de ontologías. Los resultados de estos temas de investigación son objetivo de diversas tesis de grado de Licenciatura y Maestrías en Ciencias de la Computación.

## Referencias


[211] ISO/TC 211. Geographic information/geomatics. `http://www.isotc211.org/`.

[BM03] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48, New York, NY, USA, 2003. ACM.

[DH05] A. Doan and A. Halevy. Semantic integration research in database community: A brief survey, 2005.

[Ege02] M. Egenhofer. Toward the semantic geospatial web, 2002.

[GMPQ+04] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, V. Vassalos, and J Widom. The tsimmis approach to mediation: Data models and languages. *Journal of Intelligent Information Systems*, 8(2):117–132, 2004.

[Gru93] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 1993.

[IFF+99] Zachary G. Ives, Daniela Florescu, Marc Friedman, Alon Levy, and Daniel S. Weld. An adaptive query execution system for data integration. *SIGMOD Rec.*, 28(2):299–310, 1999.

[OGC] OGC. Open geospatial consortium, inc. `http://www.opengeospatial.org/`.

[Ter06] Navarrete Terrasa. Semantic integration of thematic geographic information in a multimedia context. *Doctorate in Computer Science and Communication Department of Technology. Universitat Pompeu Fabra*, 2006.