Integración de observaciones medioambientales: Solución inicial y retos futuros *

Manuel A. Regueiro, Sebastián Villarroya, Gabriel Sanmartín, and José R.R. Viqueira

Grupo de Gráficos por Computador e Ingeniería de Datos (COGRADE),
Instituto de Investigaciones Tecnológicas,
Universidade de Santiago de Compostela
Constantino Candeira S/N, Santiago de Compostela
manuelantonio.regueiro@usc.es
sebastian.villarroya@usc.es
gabriel.sanmartin@usc.es
jrr.viqueira@usc.es

Resumen. En este trabajo se presenta una solución inicial para la integración de fuentes de datos de observación en aplicaciones de tipo medioambiental. La solución se basa en la utilización de una arquitectura típica mediador/wrapper combinada con modelos de datos e interfaces estándar definidos en la iniciativa Sensor Web Enablement (SWE) del Open Geospatial Consortium (OGC). Los retos futuros van en la dirección de simplificar la incorporación de nuevas fuentes de datos en el sistema, simplificando el desarrollo de los wrappers y proporcionando mecanismos más avanzados para la definición de relaciones semánticas entre elementos locales y globales.

Palabras clave: Integración de datos, Sensor Web, Gestión de datos Medioambientales, Arquitecturas Mediador/Wrapper, Interoperabilidad Espacial

1 Introducción

En la actualidad existe una gran cantidad de datos generados diariamente por multitud de sensores relacionados con aplicaciones de gestión medioambiental. Tanto los modelos y formatos utilizados para el almacenamiento de estos datos como las interfaces implementadas para su acceso son completamente heterogéneas y en muchos casos basadas en tecnologías propietarias. Por otro lado, diversas iniciativas promulgadas por las distintas administraciones públicas incluyen como objetivo la creación de Infraestructuras de Datos Espaciales (IDEs) que faciliten el acceso a los datos arriba mencionados. Muchas de estas iniciativas están motivadas por la aparición de la directiva INSPIRE de la Comisión

* Este trabajo ha sido financiado por la Xunta de Galicia (ref. 09MDS034522PR) y el Ministerio de Ciencia e Innovación, Gobierno de España (ref. TIN2010-21246-c02-02)

A. Ruíz, L. Iribarne (Eds.): JISBD'2012, pp. 53-58, ISBN:978-84-15487-28-9.
Jornadas SISTEDES'2012, Almería 17-19 sept. 2012, Universidad de Almería.

Europea. La interoperabilidad entre los distintos componentes y servicios de estas IDEs se basa en el uso de estándares de modelos, lenguajes e interfaces de servicios web. En el caso de los datos de sensores en entornos medioambientales, se utilizan fundamentalmente los estándares propuestos por el Open Geospatial Consortium (OGC) en su iniciativa Sensor Web Enablement (SWE). El acceso a fuentes de datos de sensores de tipo heterogéneo a través de lenguajes e interfaces estandarizados es un caso específico del problema de integración de fuentes de datos heterogéneas estudiado ampliamente en el área de las bases de datos.

En este trabajo se presenta una solución inicial dada dentro del proyecto MeteoSIX, financiado por la Xunta de Galicia, para la integración de datos de observación de distintas fuentes disponibles en la agencia meteorológica gallega (MeteoGalicia), a través de la interfaz Sensor Observation Service (SOS) del OGC. Estas fuentes incluyen datos generados por estaciones meteorológicas y oceanográficas, datos de radar y radiosondaje. La solución propuesta utiliza una arquitectura clásica basada en el uso de un mediador y diversos wrappers. El modelo global utilizado para la integración de datos en el mediador está basado en el estándar Observations and Measurments (O&M) usado en la interfaz SOS del OGC. La solución utilizada para la transformación de las consultas sobre el modelo global a consultas sobre los modelos de las fuentes de datos locales se basa en una aproximación Global as View (GAV) sobre el lenguaje de consulta de observación proporcionado por el SOS.

El resto de este artículo se organiza de la siguiente manera. En la Sec. 2 se proporciona una breve descripción de los estándares SOS y O&M del OGC, así como algunas referencias a otros trabajos relacionados en el ámbito de la integración de datos. La solución propuesta y actualmente implementada en el proyecto MeteoSIX se proporciona en la Sec. 3. El artículo termina con una breve descripción de los retos futuros en el ámbito de la integración de datos mediante la interfaz SOS.

2 Trabajo Relacionado

El estándar O&M [6] del OGC define un modelo de datos y una codificación XML para la representación e intercambio de observaciones. En general, cada observación vincula un valor observado (distintos tipos de valores son posibles) con un instante de tiempo. Además, se recogen algunos metadatos que contextualizan el valor observado. Los más importantes son el procedimiento mediante el cual se ha obtenido el valor (*Process*), la propiedad observada (*Observed Property*) y la entidad sobre la cual se realiza la observación (*Feature of Interest - FOI*). Típicamente, un *Process* será un sensor para medición en local o remoto montado en alguna plataforma, estática o móvil. La propiedad más importante de la *FOI* es su localización geográfica. El estándar de servicio SOS [7] proporciona una interfaz para el acceso a datos y metadatos de observaciones. Las observaciones de un SOS se organizan en vistas llamadas *Offerings*. El conjunto de *Offerings* definidas en un SOS se puede obtener mediante la llamada a la operación *GetCapabilities* de su interfaz. La operación *GetObservation* obtiene

el conjunto de observaciones de una determinada *Offering* para una o varias propiedades observadas. Opcionalmente, el cliente puede filtrar el resultado especificando un conjunto de *Process*, un filtro temporal y un filtro espacial. Otras operaciones SOS permiten obtener datos adicionales de un determinado *Process* o FOI e insertar y borrar observaciones. En [2] se describen varios proyectos en los que se utilizan estándares SWE.

Dos grandes aproximaciones surgen para la integración de fuentes de datos de forma transparente al usuario. En un *Data Warehouse* los datos de varias fuentes se cargan en un modelo de datos común, proporcionando una solución eficaz en la que los datos pueden volverse obsoletos rápidamente. Una solución de este tipo para datos de observación se describe en [8]. Soluciones alternativas se basan en la integración virtual de las fuentes. Los problemas clave a resolver en este tipo de integracin virtual incluyen los siguientes [3]. En primer lugar es necesario definir un esquema global virtual en el que puedan integrarse todas las fuentes. En segundo lugar hay que especificar la relación entre el esquema global y cada uno de los esquemas locales. En este punto es importante la resolución de conflictos tanto sintácticos como semánticos. Dos grandes aproximaciones existen para esta tarea [3]: i) Global as View (GAV) en la que el modelo global se define mediante vistas sobre los esquemas locales, y ii) Local as View (LAV) en las que cada modelo local se define como una vista sobre el esquema global. Por último, se necesitan estratégias de procesamiento eficiente y optimización de consultas sobre fuentes distribuidas con capacidades heterogéneas. Trabajos relacionados existen en el ámbito de la integración de datos científicos [4] y en el modelado conceptual de datos de observación [1].

3 Solución Actual

En el diseño actual del sistema se han considerado cuatro fuentes de datos de observación disponibles en MeteoGalicia. La primera fuente almacena observaciones generadas por la red de más de 80 estaciones meteorológicas automáticas. La segunda fuente da acceso a las observaciones generadas por una pequeña red de estaciones oceanográficas. Los modelos de datos de estas dos fuentes han sido diseñados en MeteoGalicia y están implementados en un gestor Microsoft SQL Server. La tercera fuente consta de secuencias temporales de archivos de radar. Cada archivo almacena varias imágenes de precipitación la misma extensión geográfica gallega. La ltima fuente almacena información de radiosondaje en un gestor PostgreSQL. El radiosondaje consiste en la suelta de globos equipados con un GPS y varios sensores que miden temperatura, humedad, temperatura del rocío, presión atmosférica, etc.

El primer paso para la integración de las fuentes anteriores es la definición de un modelo de datos global para el mediador. Este modelo está basado en los estándares O&M y SOS y su esquema relacional se muestra en la Fig. 1. Cada relación $\mathbf{Off_i\_Obs}$ almacena las observaciones de la *Offering* i definida en el sistema. De cada observación se almacena su instante de tiempo (time), *Process* (procId), *Observed Property* (propId), valor observado (result), y datos de su *FOI*

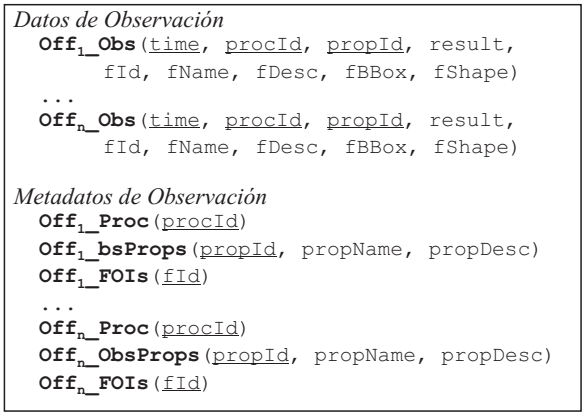

Fig. 1. Esquema Global.

que incluyen un identificador (fId, posiblemente nulo) y una geometría (fShape). Estos datos se utilizan para poder responder a las peticiones *getObservation*. Además de las observaciones, para cada *Offering* se almacenan los metadatos que la definen (procesos, propiedades observadas y FOIs). Estos metadatos se utilizan para poder responder a las peticiones *getCapabilities*.

Para cada fuente de datos se desarrolla un wrapper que implementa la interfaz SOS. Por lo tanto, cada una de las fuentes de datos tendrá el mismo modelo de datos que el usado como modelo global (ver Fig. 1). De acuerdo con lo anterior, el wrapper de cada fuente de datos S_i proporcionará para cada una de sus *Offerings* Off_j las relaciones $\mathbf{S_i\_Off_j\_Obs}$, $\mathbf{S_i\_Off_j\_Proc}$, $\mathbf{S_i\_Off_j\_ObsProps}$, $\mathbf{S_i\_Off_j\_FOIs}$, que dan acceso, respectivamente, a las observaciones, procesos, propiedades observadas y FOIs de la *Offering*. Además, para llevar a cabo el proceso de integración es necesario especificar la correspondencia entre cada proceso, propiedad observada y FOI local y su respectivo elemento global. Para conseguir esto, cada fuente de datos S_i debe de proporcionar las funciones *S_i_globalProc(procId)*, *S_i_globalProp(propId)* y *S_i_globalFOI(fId)*, que devuelven, respectivamente, el identificador global de proceso, propiedad observada y FOI para el correspondiente identificador local. Estas funciones proporcionan información semántica necesaria para la integración (ver "glue knowledge" en [4]). En la implementación actual de estas funciones, se asume que cada proceso y FOI de cada fuente de datos se incorpora al sistema global como un elemento distinto [1]. Por el contrario, varias propiedades observadas de varias fuentes de datos pueden incorporarse al sistema como una única propiedad observada global.

Para definir cada *Offering* Off_i del modelo global se especifican sus relaciones $\mathbf{Off_i\_Proc}$, $\mathbf{Off_i\_Prop}$ y $\mathbf{Off_i\_FOI}$. Además, se define un filtro temporal y un

[1] En concreto, el identificador local del elemento se concatena al identificador de la fuente de datos para obtener su identificador global

```
(1)Off_i_Obs = {time, procId, propId, result, fId, ..., fShape |
(2)         Off_i_Proc(procId) ∧  Off_i_Prop(propId, ...) ∧
(3)         temporalFilter_i(time) ∧
(4)         (spatialFilter_i(fShape) ∨ Off_i_FOI(fId)) ∧
(5)           (
(6)             (
(7)              S_1_Off_1_Obs(time, S_1_procId, S_1_propId, result,
(8)                            S_1_fId, ..., fShape) ∧
(9)              S_1_globalProc(S_1_procId) = procId ∧
(10)             S_1_globalProp(S_1_propId) = propId ∧
(11)             S_1_globalFOI(S_1_fId) = fId
(12)            )
(13)            ∨  . . .  ∨
(14)            (
(15)             S_n_Off_m_Obs(time, S_n_procId, S_1_propId, result,
(16)                           S_n_fId, ..., fShape) ∧
(17)             S_n_globalProc(S_n_procId) = procId ∧
(18)             S_n_globalProp(S_n_propId) = propId ∧
(19)             S_n_globalFOI(S_n_fId) = fId
(20)            )
(21)          )
(22)       }
```

Fig. 2. Definición de una *Offering* en el esquema global.

filtro espacial para la *Offering*. En base a estas definiciones las observaciones de Off_i se obtienen de las observaciones de las *Offerings* locales $\mathbf{S_i\_Off_j\_Obs}$ tal y como se especifica en la expresión de cálculo relacional de dominios de la Fig. 2. De forma muy breve, en la línea (2) se comprueba que la observación es de algún proceso y propiedad observada especificados para Off_i. Las líneas (3-4) chequean los filtros temporal y espacial. Por último, las líneas (7-19) verifican la existencia de alguna observación en alguna fuente de datos cuyos parámetros coincidan con los definidos globalmente para Off_i, haciendo uso de las funciones que especifican el enlace semántico entre elementos locales y globales.

No es difícil verificar que la consulta de la Fig. 2 se puede descomponer fácilmente en un conjunto de peticiones *GetObservation* a las distintas fuentes. Los filtros espacial y temporal pueden ser delegados en el wrapper de cada fuente, dependiendo de las capacidades de filtrado que implemente declaradas en su respuesta *GetCapabilities*.

4 Conclusiones y Retos Futuros

La solución actual asume la especificación por parte de las fuentes de datos de ciertas correspondencias semánticas con elementos del modelo global. Además, la implementación de los wrappers es excesivamente compleja debido a la transfor-

mación de modelos que deben de realizar. Líneas futuras a explorar para mejorar los problemas anteriores incluyen las siguientes:

- Reducir la funcionalidad de los wrappers limitándolos a transformar las consultas y datos a un lenguaje común. Una posible solución sería el uso de la interfaz Web Feature Service (WFS) del OGC que proporciona funcionalidad limitada de consulta sobre cualquier modelo de datos con información geográfica.
- Explorar la posibilidad de incorporar una solución LAV para la integración.
- Explorar el uso de ontologías para guiar el proceso de integración a nivel semántico [5]. Se trataría de simplificar la tarea del usuario a la hora de especificar las relaciones semánticas entre los elementos locales y globales. En la actualidad existen ya algunas ontologías conformes con el modelo O&M que se podrían usar [9].

References

1. Shawn Bowers, Joshua S. Madin, and Mark P. Schildhauer. A conceptual modeling framework for expressing observational data semantics. In *Proceedings of the 27th International Conference on Conceptual Modeling*, ER '08, pages 41–54, Berlin, Heidelberg, 2008. Springer-Verlag.
2. Helen Conover, Gregoire Berthiau, Mike Botts, H. Michael Goodman, Xiang Li, Yue Lu, Manil Maskey, Kathryn Regner, and Bradley Zavodsky. Using sensor web protocols for environmental data acquisition and management. *Ecological Informatics*, pages 32–41, 2010.
3. Alon Y. Levy. *Logic-based techniques in data integration*, pages 575–595. Kluwer Academic Publishers, Norwell, MA, USA, 2000.
4. Bertram Ludäscher, Amarnath Gupta, and Maryann E. Martone. A model-based mediator system for scientific data management. In *Bioinformatics: Managing Scientific Data*, pages 335–370. Morgan Kaufmann, 2003.
5. E. Mena, A. Illarramendi, V. Kashyap, and A. Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *International journal on Distributed And Parallel Databases (DAPD)*, 8(2):223–272, April 2000.
6. OGC. Observations and measurements part 1 - observation schema. Open Geospatial Consortium (OGC). Retrieved January 2011 from: http://www.opengeospatial.org, 2007.
7. OGC. Sensor observation service. Open Geospatial Consortium (OGC). Retrieved January 2011 from: http://www.opengeospatial.org, 2007.
8. José Ramon Rios Viqueira, José Varela, Joaquín A. Triñanes, and José Manuel Cotos. A sensor observation service based on ogc specifications for a meteorological sdi in galicia. In *ER Workshops*, pages 43–52, 2010.
9. W3C Semantic Sensor Network Incubator Group. Semantic sensor network xg final report. World Wide Web Consortium (W3C). Retrieved April 2012 from: http://www.w3.org/2005/Incubator/ssn/XGR-ssn/, 2011.