Proyecto SICO: Sistemas de Información en un entorno Cooperativo

Regina Motz, Raúl Ruggia, Jorge Abin, Adriana Marotta, Verónika Peralta, Fernando Carpani

Instituto de Computación - Facultad de Ingeniería
Universidad de la República – Uruguay

rmotz@fing.edu.uy

Resumen. Este trabajo presenta las líneas de investigación desarrolladas dentro del proyecto Sico referentes a la elaboración de un entorno cooperativo para sistemas de información. El objetivo principal del proyecto es que múltiples sistemas de información sean capaces de trabajar en forma cooperativa combinando sus datos y funcionalidades. Tradicionalmente estos sistemas se construyen al modo de una capa sobre sistemas ya existentes que pueden ser desde bases de datos, datos semi-estructurados o sistemas legados. Los mayores problemas para su construcción residen en que sus componentes son sistemas autónomos, distribuidos y heterogéneos. La propuesta central del proyecto es la de proveer un entorno único donde los diversos problemas a nivel de integración de datos e integración de funcionalidades suministradas por sistemas legados o por Web Services puedan ser tratados, razonados y comparadas entre sí. Es en este entorno donde las soluciones alternativas en base a los diversos aspectos de calidad que inciden en la construcción de sistemas cooperativos y en las mejores estrategias para adaptar sistemas legados pueden manejarse en forma semi-automática.

1 Introducción

Actualmente existe un enorme volumen de fuentes de información y servicios disponibles a través de Internet. Es común que no sea posible resolver una consulta usando una única fuente y sea necesario integrar conocimientos provenientes desde varias fuentes. Un ejemplo de esta realidad son los diferentes servicios de salud. Por ejemplo, los sistemas de información genética son sistemas federados, donde se almacena información concerniente a marcas genéticas, patologías asociadas a mutaciones genéticas, patrones de adn, etc. Por otra parte los sistemas de historias clínicas comunmente utilizados en los hospitales, administran información sobre los actos médicos, exámenes clínicos, patologías, etc. Se sabe que existen ciertas patologías que son de transmisión genética y por tanto son predecibles en aquellos pacientes cuyos antecedentes familiares presentan mutaciones genéticas que corresponden a dichas enfermedades. Algunas cooperaciones entre estos sistemas

pueden ser: (a) Se requiere la lista de pacientes y enfermedades a prevenir. (b) Se requiere la lista de pacientes a los que es conveniente realizar estudios genéticos en función de sus antecedentes médicos y del conocimiento existente en genética. Este ejemplo muestra claramente que el objetivo de la cooperación es el de incrementar el valor de la información cuando múltiples sistemas de información (*Sistemas Participantes*) son capaces de trabajar en forma cooperativa combinando sus datos y funcionalidades conformando un *Sistema Integrado.* La principal característica de estos sistemas integrados es que se construyen al modo de una capa sobre sistemas ya existentes que pueden ser desde bases de datos, datos semi-estructurados o sistemas legados (por ejemplo programas Cobol). El mayor problema para la construcción de estos sistemas es que sus componentes son sistemas autónomos, distribuidos y heterogéneos. Para atacar este problema el desafío del proyecto Sico es desarrollar un entorno donde sea posible que distintas fuentes de información y servicios (como sistemas legados o web services) puedan colaborar entre sí intercambiando e integrando no solamente datos sino también funcionalidades, teniendo en cuenta además factores de calidad de los sistemas participantes en relación al sistema integrado. Para estos problemas existen numerosos trabajos sobre integración de datos, ver por ejemplo [Abiteboul97, Bayardo et al. 97, Chen et al 98, CGM99, MP03]. La integración de funcionalidades suministradas por sistemas legados y Web Services, ha motivado varias conferencias en el área y múltiples trabajos de investigación. Varios de dichos trabajos abordan temas estrechamente relacionados con esta problemática, como ser integridad transaccional [SDTL03, TMR00], integración de datos con fuentes heterogéneas y Web Services [SAOB02, DKOS], orquestación y coreografías de software [BBQS03].

El proyecto Sico propone un nuevo entorno de integración basado en *metadatos* que utilizados tanto para la integración de datos como de funcionalidades y que utilicen aspectos de calidad de las fuentes. Cuando nos referimos a metadatos, consideramos los metadatos *descriptivos* que son externos a los datos de la fuente (por ejemplo, autor, fecha, etc.) y los metadatos *semánticos* que caracterizan el contenido de un documento [BR99]. Este último tipo de metadatos es sobre el cual se basan los extractores de información (wrappers) y los integradores (mediadores) para automatizar el proceso de integración semántica de datos. En nuestro enfoque los metadatos son extendidos para que describan también las funcionalidades a integrar de los sistemas de información. Adicionalmente, los metadatos de la arquitectura Sico contienen datos sobre la calidad de las fuentes. En relación al tema de calidad en sistemas de información se aprecia que es un tema con activa investigación en los últimos años [LSKW01, NLF99, HH02]. En particular, existen trabajos recientes que se concentran en el manejo de la calidad en sistemas cuya información proviene de múltiples fuentes. En [LSKW01] se propone una metodología para la evaluación y mejoramiento de la calidad en la información. En particular, el artículo presenta dimensiones de calidad, y clasificaciones de éstas en categorías. En [NLF99] se presenta un modelo de calidad en un sistema *multidatabase* que permite calcular valores de calidad para los posibles planes de una consulta. Para esto estudian como propagar factores de calidad desde las bases de datos fuentes, a través del árbol correspondiente a un plan de la consulta global. En [HH02] se presentan enfoques para manejo de calidad en Data Warehouse (DW), basados en metamodelos, y en [CPPS01] se presenta un conjunto de métricas de calidad para diseño “estrella” de

DWs y se les aplica un proceso de validación formal. Sin embargo, hasta donde es de nuestro conocimiento no existen trabajos que relacionen la calidad de los sistemas participantes con la calidad del sistema integrado cuando los participantes corresponden a sistemas de información heterogéneos, ejemplo bases de datos vrs. sistemas legados.

2 Arquitectura

El mayor propósito de la arquitectura de *Sico* es proveer un entorno unificado para cooperación entre diversas fuentes de información y servicios. Como se muestra en la Figura 1, se tienen tres módulos correspondientes a los temas de Metadatos Semánticos, Integrador y Metadatos Factores de Calidad. Cada módulo será tratado en forma independiente, teniendo en cuenta las relaciones entre ellos.
El módulo *Metadatos semánticos* contiene: especificación de los datos que exporta cada sistema participante y permisos de acceso (lectura, escritura, modificación, etc.) para el sistema integrado y para cada uno de los participantes, especificación de funcionalidades que exporta cada sistema participante y permisos de acceso para el sistema integrado y para cada uno de los participantes, especificación de los datos y funcionalidades que provee el sistema integrado y un diccionario semántico de los términos que utiliza cada sistema. En estos metadatos se debe tener en cuenta toda la información necesaria para describir cada participante a nivel estructural y semántico así como las relaciones entre los diferentes Sistemas. Debe contemplar también toda la información estructural y semántica del Sistema Integrado incluyendo las vistas del Sistema que se deben presentar a cada usuario. Estos metadatos se representan vía RDF y sobre RDF usando las primitivas del lenguaje de descripción de ontologías DAML+OIL se implementa la manipulación de los metadatos. Proyectos como On2broker [Dieter et al 99] y el de [Broekstra et al. 02] están dedicados a construir herramientas basadas en ontologías para gerenciar el conocimiento necesario para realizar integración semántica. Sin embargo, en esto sproyectos, hasta donde es de nuestro conocimiento, no tienen en cuenta los criterios de calidad del sistema integrado.

El módulo *Integrador* se ocupa fundamentalmente de la resolución de las operaciones solicitadas al sistema integrado, con el nivel de calidad requerido por el usuario y de acuerdo a las especificaciones y reglas definidas a nivel del sistema integrado. El modelo canónico para la integración es el *Schema Graph* [Motz02] que corresponde a una representación formal de grafo para ODMG [Catell93]. Se comunica con los usuarios mediante la capa de interfaz de usuario, con la cual intercambia solicitudes de operaciones y respuestas.Se observan dos tipos de operaciones diferentes: a) Operaciones de los sistemas participantes. Son aquellas operaciones que son procesadas en uno sólo de los sistemas participantes. b) Operaciones definidas a nivel del sistema integrado. Son aquellas operaciones que no existen como tales en los sistemas participantes cuya resolución puede involucrar una o más operaciones del tipo a). Conceptualmente, este componente está constituido por dos piezas de software básicas, el Procesador de Solicitudes y el Orquestador de Operaciones

El Procesador de Solicitudes tiene las siguientes responsabilidades: recibir las solicitudes de operaciones al sistema integrado, analizar la operación y descomponerla en las operaciones necesarias para su resolución. Este proceso puede tener más de un resultado posible ya que es factible que existan distintos sistemas que aporten la misma información. Este módulo deb también elegir la solución apropiada del conjunto de soluciones posibles (reformulaciones de la consulta sobre el esquema integrado), deberá elegirse aquella que responda a los requerimientos de calidad solicitados por el usuario. Finalmente deberá solicitar al Orquestador de Operaciones la ejecución de las operaciones requeridas según la lógica correspondiente a la solución elegida y esperar su respuesta.

El Procesador de Solicitudes tiene las siguientes responsabilidades: recibir las solicitudes de operaciones al sistema integrado, analizar la operación y descomponerla en las operaciones necesarias para su resolución. Este proceso puede tener más de un resultado posible ya que es factible que existan distintos sistemas que aporten la misma información. Este módulo deb también elegir la solución apropiada del conjunto de soluciones posibles (reformulaciones de la consulta sobre el esquema integrado), deberá elegirse aquella que responda a los requerimientos de calidad solicitados por el usuario. Finalmente deberá solicitar al Orquestador de Operaciones la ejecución de las operaciones requeridas según la lógica correspondiente a la solución elegida y esperar su respuesta.

El Orquestador de Operaciones tiene las siguientes responsabilidades: ejecutar la secuencia de operaciones en base a la lógica establecida, resolviendo el flujo de ejecución, pasaje de parámetros, temporización de operaciones (sincronismo, asincronismo, time out, etc.), identificar los procesos complementarios tales como "logging", verificación de derechos de uso, mecanismos de "tunneling", etc. que puedan ser requeridos por la operaciones participantes e invocarlos conforme a la lógica del proceso de la operación solicitada al orquestador, manejar las condiciones de excepcionalidad que puedan presentarse durante la ejecución de una operación y retornar el resultado al Procesador de Solicitudes.

Metadatos de Factores de Calidad. Para poder trabajar en el sistema integrado con factores de calidad específicos es necesario estudiar el impacto que los factores de calidad existentes en los sistemas participantes tienen sobre el sistema integrado. Para esto es necesario resolver los siguientes problemas para cada factor de calidad: (a) obtención del valor (o los valores) en cada sistema participante, (b) propagación de los valores de calidad hacia el sistema integrado, (c) combinación del factor considerado con otros factores de calidad, (d) posibles consecuencias de los valores de los factores en el diseño del sistema integrado.

(a)La obtención de los factores de calidad de los sistemas participantes puede tener una complejidad muy variada. En algunos casos el sistema participante puede publicar los datos necesarios, en otros casos éstos pueden ser estimados, pero en ciertos sistemas deben construirse programas a medida de consulta y estimación. Por ejemplo, el factor *tiempo de respuesta* de un cierto sistema puede ser dado por él mismo, puede ser estimado o puede ser tan variable que obligue a calcularlo periódicamente mediante consultas de prueba. (b) El problema de propagación de los valores de calidad hacia el sistema integrado se refiere a cómo se transforman los valores de calidad que presentan los sistemas participantes (o la información obtenida

a partir de éstos) para obtener los valores de calidad del sistema integrado. Por ejemplo, si se sabe que cierta información en un sistema participante tiene determinado valor para el factor de calidad *frescura*, se puede deducir el valor de *frescura* de esa información en el sistema integrado, sumándole el lapso de tiempo transcurrido entre la obtención de la información del participante y la consulta al sistema integrado. (c) También es necesario estudiar cómo pueden combinarse los diferentes factores de calidad. Algunas combinaciones pueden dar como resultado un valor que corresponde a un nuevo factor de calidad.(d) La consideración de los valores de los factores de calidad de los sistemas participantes y su propagación al sistema integrado puede incidir en el diseño del sistema integrado en muchos aspectos, como por ejemplo, selección de participantes cuando cierta información puede ser obtenida de más de un participante, selección del plan de una consulta a realizarse sobre una base de datos, determinación de frecuencias de carga de datos desde los participantes al sistema integrado, etc.

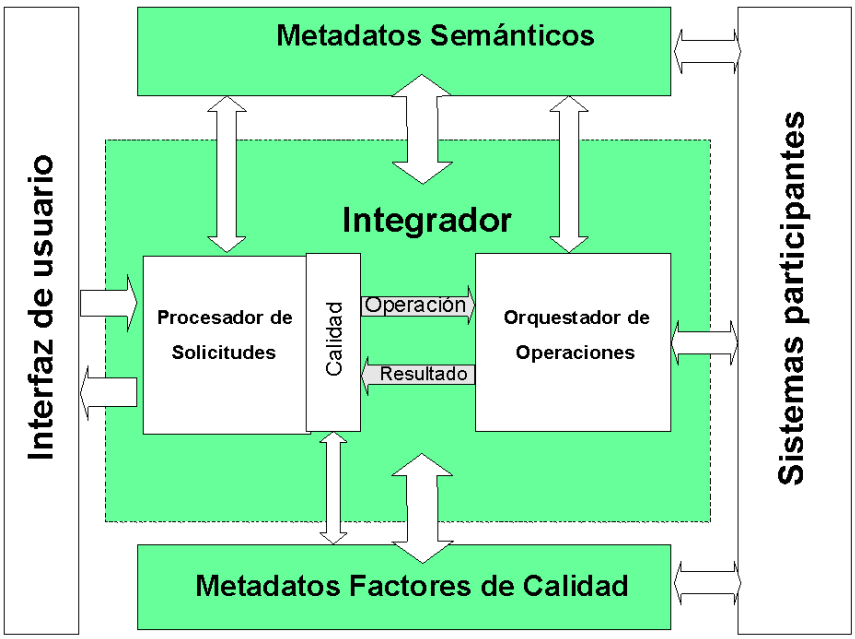

Fig. 1. Arquitectura Proyecto *Sico*

3 Conclusiones

El trabajo presentado en este reporte está en desarrollo. Hemos desarrollado un prototipo para integrar y mantener Schemas Graphs y actualmente estamos definiendo completamente la ontología dentro del proyecto *Development of a Metadata-based DSS Environment: applied to the Ocean Resources area*. Founded by DINACYT – URUGUAY (Fondo "Clemente Estable" Project). El siguiente paso es implementar un prototipo que integre basado en los metados semánticos y en los metadatos de los factores de calidad.

4 Bibliografía

[Abiteboul97] Serge Abiteboul, *"Querying Semi-Structured Data*", ICDT pages 1-18,1997.

[Bayardo et. al 97] R. J. Bayardo et al. "*InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments*" ACM SIGMOD International Conference on Management of Data, vol 26, N 2, ACM Press, New York", pages 195--206, 1997.

[BR99] R. Baeza-Yates and B. Ribeiro-Neto. "*Modern Information Retrieval*". Addison-Wesley, Wokingham, UK, 1999.

[Broekstra et al. 02] Jeen Broekstra, Michel Klein, Dieter Fensel, Stefan Decker, Frank van Harmelen, and Ian Horrocks. "*Enabling knowledge representation on the Web by extending RDF Schema."* Computer Networks, 39(5):609-634, August 2002.

BBQS03] Boualem Benatallah, Quan Z. Sheng, Marlon Dumas. "*The Self-Serv Environment for Web Services Composition*", IEEE Internet Computing, January.February 2003.

[Cattell93] Atwood, Barry, Duhl, Eastman, Ferran, Jordan, Loomis, Wade. *"The Object Database Standard: ODMG - 93."* R.G. G. Cattell.

[Chen et al. 98] Chen Li, Ramana Yerneni, Vasilis Vassalos, Hector Garcia-Molina, Yannis Papakonstantinou, Jeffrey Ullman,Murty Valiveti. "*Capability Based Mediation in TSIMMIS"*. SIGMOD 98 Demo, Seattle, June 1998.

[CGM99] Chen-Chuan K. Chang, Hector Garcia -Molina "*Mind Your Vocabulary: Query mapping Across Heteerogeneous Information Sources*" ACM SIGMOD 1999 (12pp)

[CPPS01] C. Calero, M. Piattini, C. Pascual, M. A. Serrano. *"Towards Data Warehouse Quality Metrics"*. DMDW 2001: 2

[Dieter et al 99] Dieter Fensel et al. "*On2broker: Semantic-based access to information sources at the {WWW}"* In WebNet (1), pages 366-371, 1999.

[DKOS99] David Konopnicki and Oded Shmueli, **"***A Comprehensive Framework for Querying and Integrating WWWData and Services*" Fourth IECIS International Conference on Cooperative Information Systems, 1999.

[HH02] M. Helfert, C. Herrmann. "Proactive Data Quality Management for Data Warehouse Systems". DMDW 2002: 97-106

[LSKW01] Y. W. Lee, D. M. Strong, B. K. Kahn, R. Y. Wang. "AIMQ: A Methodology for Information Quality Assessment". Information & Management, Elsevier Science (North Holland).

[JQJ98] M. A. Jeusfeld, C. Quix, M. Jarke. Design and Analysis of Quality Information for Data Warehouses. ER 1998: 349-362

[Motz02] Regina Motz, "*Schema Evolution in ODMG*", Workshop Chileno de Bases de Datos, Copiapo, Chile, November 2002.

[MP 03] P. McBrien and A. Poulovassilis, "*Data integration by bi-directional schema transformation rules*". ICDE'03, March , 2003.

[NLF99] Felix Naumann, Ulf Leser, Johann Christoph Freytag. "*Quality-driven Integration of Heterogenous Information Systems*". VLDB 1999: 447-458

[SAOB02] S. Abiteboul, Omar Benjelloun, Tova Milo; "*Web services and data integration*". WISE 2002 .

[SDTL03] Sanjay Dalal, Sazi Temel ,Mark Little, MarkPotts, Jim Webber . "*Coodinating Business Transactions on the Web*". IEEE Internet Computing, February 2003.

[TMR00] Thomas Mikalsen, Isabelle Rouvellou, Stanley Sutton Jr., Stefan Tai, Mandy, Chessell, Catherine Griffin, David Vines, "*Transactional Business Procesess Servers: Definition and Requirements"* Business Object Component Workshop - OOPSlA2000