Enseñanzas adquiridas en el desarrollo de un portal para integrar datos utilizando esquemas de intercambio de datos

José Miguel Cuadra Morales
Instituto Nacional de Biodiversidad
Heredia, Costa Rica
jcuadra@inbio.ac.cr

Resumen

Este documento es una exposición de la experiencia obtenida durante el desarrollo de un Portal que integra datos sobre biodiversidad, de fuentes heterogéneas, mediante la aplicación de esquemas para la estandarización de dichos datos. El enfoque principal consta en describir el proceso de estandarización de los datos de los poseedores de la información y los pasos necesarios a seguir para crear un Portal que integre dichos datos provenientes de diferentes proveedores.

Palabras clave: Estandarización de Datos, Esquemas conceptuales, Integración de Datos

Abstract

This document is an exposition of the experience obtained during the development of a Portal which integrates biodiversity data, from heterogeneous sources, via the application of schemas for the standardization of such data. The main focus is to describe the process of data standardization from the information owners and the steps necessary for the creation of a Portal that integrates such data from different providers.

Keywords: Data standardization, Conceptual schemas, Data Integration

1 Introducción

En la actual era del conocimiento, las tecnologías de comunicación en Internet se han convertido en grandes aliadas para compartir información, hecho que es de gran importancia. Es por esto que los poseedores o proveedores de la información se han visto en la necesidad de implementar soluciones para poder exponer sus datos de una manera estandarizada y esquematizada. Dada esta premisa, se ha vuelto más sencilla la creación de herramientas que extraen datos de estos proveedores, y los muestran de una manera centralizada en una sola entidad, por ejemplo un portal.

Los beneficiados de esta integración, son aquellas personas que utilizan esta información directamente en sus actividades, sean biólogos, agrónomos o de cualquier otra rama profesional. La orientación de este documento es sesgada hacia la biología pero las soluciones planteadas podrían ser utilizadas en otras ciencias o actividades profesionales.

2 Integración de fuentes de datos heterogéneas

El problema de la heterogeneidad de los datos entre diferentes proveedores (poseedores de información) puede ser considerado como uno de los factores que más obstaculiza las labores de integración. En otras palabras, cada institución proveedora de datos tiene su propia forma de describir o estructurar un elemento de información dentro de una instancia de almacenamiento, por ejemplo, estructurar el nombre de un autor dentro de una base de datos. Cuando se reúnen a muchas instituciones que poseen datos interrelacionados, lo más probable es que no sigan una misma estructura a la hora de almacenar un elemento en común. Esta situación deja como única opción de integración, desde el punto de vista informático, conocer la tecnología y protocolos de comunicación subyacentes propios de cada institución para consultar dichos datos. Esto representa un gran consumo de tiempo y recurso humano y tecnológico para crear una herramienta que realice dicha integración.

Para poder llegar a un consenso en una estructura para compartir información interrelacionada, se ha impulsado la creación de esquemas conceptuales, cuya función principal es la de agrupar elementos básicos de información que describen a alguna entidad en particular, por ejemplo una especie o un espécimen. Típicamente, este esquema se define siguiendo el lenguaje eXtensible Markup Language (XML), debido a su creciente importancia en el intercambio de una gran variedad de datos en la Web.[1] Luego, se hace una asociación entre los elementos del esquema conceptual y los elementos de la fuente de datos. Este procedimiento puede repetirse para varios proveedores con el fin de que todos puedan compartir datos de una manera estándar y facilitar la integración realizada por otras herramientas.

Existen ciertas herramientas, que se instalan localmente en los proveedores de datos, para llevar a cabo las labores de exposición de dichos datos bajo un mismo esquema conceptual, no sin antes realizar la asociación entre el esquema conceptual y la fuente de datos. Esta asociación es realizada típicamente mediante una interfaz gráfica en donde el poseedor de la información asocia sus elementos de información con los elementos del esquema conceptual. La forma en que esta herramienta es consultada sigue un protocolo de comunicación, definido típicamente en el lenguaje XML. Actualmente se encuentran en uso algunos protocolos tales como el Distributed Generic Information Retrieval (DiGIR), TDWG Access Protocol for Information Retrieval (TAPIR), y el Biological Collection Access Service for Europe (BioCASE). [2] Estas herramientas son claves para la integración pues ellas se encargan de la comunicación subyacente con el proveedor. Típicamente, estas herramientas realizan consultas utilizando el lenguaje Structured Query Language (SQL) directamente al motor de la base de datos para extraer la información solicitada, y exponen la información extraída en el formato XML siguiendo las pautas definidas en el esquema conceptual.

3 Portal integrador de las fuentes de datos

El portal se encarga de las labores de integración de los datos de todos los proveedores. El primer requisito para esta integración, es que el portal sepa cómo consultar al proveedor de datos. Para lograr esto, el portal debe conocer el protocolo de comunicación que está utilizando el proveedor, por ejemplo el DiGIR, TAPIR o BioCASE. De la misma manera, debe conocer el esquema conceptual del proveedor. Para el caso específico del Portal en cuestión, se manejan dos esquemas conceptuales; el Darwin Core [3], para la estandarización de datos de especimenes y el Plinian Core [4], para la estandarización de información de especies. Dado esto, la consulta física puede realizarse mediante una comunicación HTTP. Al momento de la respuesta, cada uno de los proveedores, envía una respuesta estándar al portal, lo que hace fácil la manipulación de los datos por parte de éste último.

La visualización de dichos datos se refiere más que todo a la forma en que el portal modela la respuesta de los proveedores para que sea desplegada al usuario. Típicamente, se le puede aplicar una conversión utilizando el lenguaje eXtensible Stylesheet Language (XSLT) [5] para que itere por todas las respuestas XML, y así realizar la transformación desde XML hacia cualquier otro formato. Con este esquema, existe una muy buena separación entre la capa de la presentación y la capa de datos, pues el diseñador del portal no debe preocuparse por la implementación subyacente de las fuentes de datos. De esta manera, siempre recibirá la información como se espera, aun cuando se agreguen nuevos proveedores para consultar, y solo deberá preocuparse por dicha transformación.

Debido a que los proveedores son entidades externas independientes del portal se deben tomar previsiones para no estar consultando dichos proveedores cada vez que un usuario ejecuta una solicitud, pues se podrían presentar problemas de conectividad o de latencia. Para resolver este problema se puede recurrir a la creación de un índice que almacene un subconjunto de los datos que está devolviendo el proveedor. Una "araña" se encargaría de hacer la consulta con los proveedores registrados en el portal, en horas de poco tráfico, para así indexar este subconjunto de datos. La intención es que, cuando el usuario ejecute una solicitud, el portal acceda al índice para devolver los datos almacenados en él, sin la necesidad de consultar físicamente a los proveedores. Sí sucediera que el usuario desea la totalidad de la información para alguna entidad, por ejemplo una especie, el portal podría realizar la consulta física pero advertiría al usuario de posibles fallas en la conexión. La experiencia a la hora de desarrollar un portal, dicta que se debe tomar muy en cuenta la inclusión de un mecanismo de detección de fallas al momento de realizar la conexión con el proveedor, y mostrar un mensaje al usuario el cual notifique que dicha comunicación falló por causas ajenas al portal.

4 Conclusiones y Recomendaciones

Para el desarrollo de un Portal que integre fuentes de datos heterogéneas sobre biodiversidad, es necesario que los poseedores de la información se enteren de la importancia de compartir sus datos de una manera estandarizada. Existen varias herramientas, principalmente desarrolladas por una persona con un alto conocimiento técnico, que le facilitan al proveedor la estandarización y exposición de sus datos. Al desarrollar este tipo de herramientas, se debe considerar diseñarlas de fácil uso para una persona con poca experiencia técnica, para que dicha persona se encargue de todo el proceso de instalación y configuración, lo que daría beneficios claros pues no habría necesidad de contar con una persona técnica para realizar dicho proceso. Dada esta premisa, el desarrollador del Portal, no tendría la necesidad de conocer las tecnologías de almacenamiento subyacentes de cada proveedor y solo se preocuparía de conocer los protocolos de comunicación y esquemas conceptuales que las herramientas instaladas en cada proveedor están utilizando. La integración resultaría relativamente sencilla, y el desarrollador tendría que preocuparse más por la implementación propia del portal y del diseño gráfico del mismo, que por cuestiones de instalación y configuración de proveedores. En resumen, para lograr una eficiente integración de datos en un portal, el proveedor de información, el desarrollador de las herramientas de integración y el desarrollador del portal deben colaborar muy de cerca para lograr el máximo beneficio.

Referencias

[1] Quin, Liam. *Extensible Markup Language (XML)*. [citado 30/04/2007], disponible en http://www.w3.org/XML

[2] Döring, M. and De Giovanni, R. *A unified protocol for search and retrieval of distributed data*. [citado 25/04/2007], disponible en http://www.cria.org.br/protocols/newprotocol.pdf

[3] Wieczorek, John. *Darwin Core* [citado 29/08/2007], disponible en http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome

[4] INBio, GBIF.ES. *Plinian Core* [citado 29/08/2007], disponible en http://www.pliniancore.org

[5] Clark, James. *XSL Transformations (XSLT)* [citado 29/08/2007], disponible en http://www.w3.org/TR/xslt