

DAWIS: Una Arquitectura de Integración Web para el Acceso Compartido a Archivos Digitales

Santiago Eibe², Carmen Costilla¹, Ernestina Menasalvas² y César Acuña¹

Grupo de Bases de Datos, Universidad Politécnica de Madrid

¹ DIT-ETSI Telecomunicación, UPM,
Ciudad Universitaria, 28040 Madrid, España
{costilla, cjacuna}@dit.upm.es

² DLSIIS Facultad de Informática, UPM,
Campus de Montegancedo, 28260 Boadilla del Monte, Madrid, España
{seibe, emenasalvas}@fi.upm.es

Resumen La Web es el canal de comunicación preferido por millones de usuarios. Como consecuencia, muchas de las actividades cotidianas (formativas, intercambios comerciales y financieros, consultas a las fuentes disponibles en la red y, en particular, a bibliotecas digitales) se producen a través de este medio. La utilización de fondos documentales tampoco es ajena a esta tendencia. De hecho, múltiples fondos están siendo digitalizados y existe una iniciativa clara (OAI) para integrar el acceso a estos archivos digitales. DAWIS¹ es un proyecto nacional cuyo principal objetivo consiste en el diseño de un entorno para automatizar el desarrollo de portales Web que provean el acceso integrado a diferentes archivos digitales. Una parte esencial de DAWIS es la definición de una arquitectura o modelo de referencia. En este artículo se presenta el diseño de una arquitectura de integración que soportará la integración flexible y dinámica de múltiples archivos digitales.

1. Introducción

La Web se está convirtiendo en el canal de comunicación preferente para millones de usuarios. Si bien la tecnología es todavía muy joven, de lo cual se derivan los habituales problemas, las ventajas que provee predominan con claridad. Entre estas destacan la interactividad con los usuarios y la capacidad para facilitar el acceso compartido a múltiples informaciones en todo momento y lugar. Esta ubicuidad no supone un coste adicional para el usuario final sino que se mantiene contenido en valores mínimos. Como consecuencia, asistimos a la proliferación de servicios y al desarrollo de la infraestructura necesaria para soportarlos. En [1], [2] y [3] se describen algunos servicios basados en Web así como las arquitecturas extensibles para proveer tales servicios. Las bibliotecas digitales son un ejemplo de tales sistemas, si bien es posible encontrar muchos otros en las cada día más habituales aplicaciones de los sistemas de información web (educación, negocio, comercio, correo, etc.). En estos sistemas de bibliotecas digitales la

¹ DAWIS es el acrónimo de Digital Archive Web Information System.

entidad básica toma la forma de objeto digital, contenedor de la información manejada por el sistema, en base al cual ésta es almacenada, accedida, distribuida y gestionada. Se establecen así mismo los esquemas de denominación en base a los cuales los objetos digitales son identificados y localizados en todo el sistema, globalmente. Este esquema es utilizado, por ejemplo, en Fedora [4], [5] donde se aplica la tecnología Web como medio para la integración de bibliotecas digitales. En este orden, OAI (Open Archives Initiative) [6] es una primera tentativa para la estandarización de un modelo de archivo abierto sobre cuya base sea posible el acceso compartido, distribuido e interoperable a los contenidos de archivos heterogéneos. Los resultados y actividades promovidas por el foro de discusión asociado con OAI [8] confirman la relevancia de esta área de investigación entre la comunidad científica.

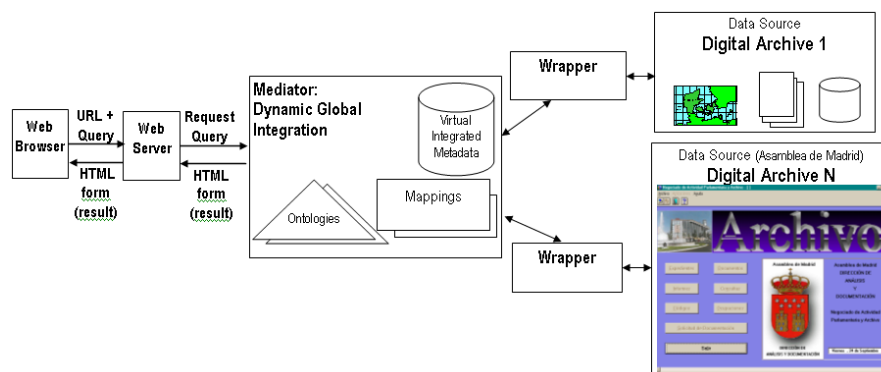


Figura 1. DAWIS: Primera Propuesta Arquitectural

El principal objetivo de DAWIS (Digital Archive Web Information System) [9] es automatizar tanto como sea posible el acceso integrado a diferentes archivos digitales a través de portales Web. DAWIS es un proyecto continuación de otro previo en el cual se abordó la digitalización y automatización del archivo de la Asamblea de Madrid. El sistema resultante de esta investigación (ver sección 2) está funcionando exitosamente desde 1999. El reto es ahora la integración vía Web de diferentes archivos digitales aprovechando el conocimiento y experiencia acumulados durante la puesta en funcionamiento del archivo de la Asamblea de Madrid. A tal efecto se ha producido un modelo arquitectónico [10] (ver figura 1) conforme a J2EE que soporta el acceso (sólo lectura) integrado a diferentes archivos digitales (en adelante, DA).

En este artículo se presenta la aproximación arquitectónica diseñada para DAWIS. En esencia, ésta consiste en una arquitectura basada en Web que posibilite la integración virtual, flexible y dinámica de múltiples archivos digitales. El resto del artículo se organiza como sigue. La sección 2 presenta los trabajos más relacionados con esta área de investigación. En la sección 3 se expone la

arquitectura y sus componentes esenciales. Finalmente, la sección 4 presenta las conclusiones y las líneas futuras de investigación.

2. Resultados y Precedentes Relacionados

Durante la pasada década la comunidad científica ha dedicado un generoso esfuerzo a la investigación en bibliotecas digitales. Los archivos digitales se consideran un tópico de investigación incluido en dicha área. De hecho, el elevado número de proyectos de investigación en desarrollo relacionados con documentos digitales, y la Web en general, evidencia el interés actual suscitado por este área. Debido a las limitaciones de espacio, no es posible incorporar todos los trabajos relacionados limitándonos a aquellos más relacionados con el planteamiento en este artículo. Estos son SGP, OAI y Fedora. Los resultados obtenidos en estos proyectos son de capital importancia para entender el alcance de la propuesta descrita en la sección 3.



Figura 2. Asamblea de Madrid. Sistema de Gestión del Archivo Parlamentario

El proyecto SGP produjo el diseño e implementación del sistema SGP (1998-2000), actualmente comercializado como SIAP por la empresa CRC Information Technologies [11]. SIAP [12] es el sistema de información parlamentaria en funcionamiento en la Asamblea de Madrid [13] desde el año 2000 hasta la actualidad. Para este artículo, una parte especialmente importante de SIAP es la relativa al Sistema de Gestión del Archivo Parlamentario, uno de los pioneros en España en la digitalización completa de su fondo documental. La figura 2 ilustra el aspecto de este Sistema de Gestión del Archivo.

El segundo de los proyectos señalados corresponde con la iniciativa OAI-PMH (ver figura 3). El objetivo del consorcio OAI es el desarrollo y promoción de estándares que posibiliten el acceso compartido y distribuido a los contenidos de diferentes archivos. En particular, PMH (Protocol Metadata Harvesting) provee un marco para la interoperabilidad entre diferentes fuentes de datos, independientemente de las aplicaciones que los utilizan. Este protocolo se basa en la recopilación y compartición de metadatos. En el informe OAI D3.1 [8] se

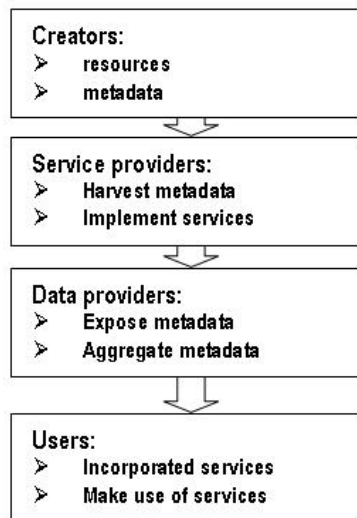


Figura 3. OAI-PMH

expone el modelo de compartición e intercambio básico para proveer servicios de acceso, almacenamiento y compartición de datos y metadatos.

Finalmente, resta mencionar el proyecto FEDORA (Flexible Extensible Digital Object and Repository Architecture) desarrollado por las Universidades de Cornell y Virginia. El resultado es un prototipo que provee la integración de la gestión necesaria para acceder a diferentes fuentes de datos tales como texto SGML, imágenes, vídeo, audio, información geográfica y demográfica, etc. Esto es posible gracias a que FEDORA construye una infraestructura con una arquitectura abierta y que soporta un conjunto extensible de servicios para la distribución de información digital como es el caso de las bibliotecas digitales. Los mecanismos de integración virtual flexible y escalable que aquí se describen pueden ser aplicados al diseño de DAWIS de cara a maximizar el rendimiento y la eficiencia en la arquitectura propuesta.

3. Propuesta de Arquitectura para DAWIS

En esta sección se presenta la arquitectura propuesta para DAWIS, como el siguiente paso a la primera aproximación representada en la figura 1. En primer término se describen las abstracciones básicas que definen la arquitectura del sistema y, seguidamente, los elementos que resultan de las mismas. Finalmente, se expone la arquitectura en su conjunto.

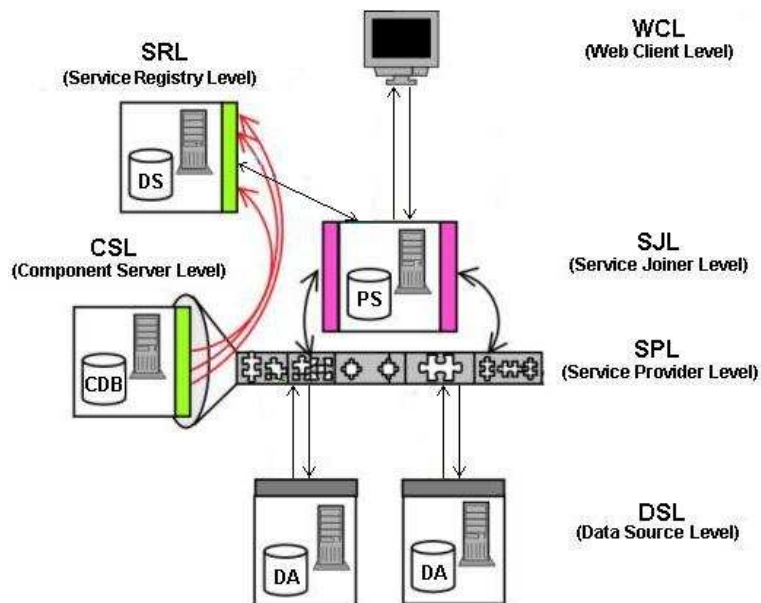


Figura 4. Propuesta de Arquitectura para DAWIS

3.1. Conceptos Básicos

La figura 4 muestra gráficamente la arquitectura propuesta. Como puede verse, es una arquitectura orientada a la Web y conforme con J2EE. En ella se distinguen seis niveles principales:

- WCL: cliente Web (Web Client Level)
- SRL: servicio de registro (Service Registry Level)
- CSL: servidor de componentes (Component Server Level)
- SJL: unificador de servicios (Service Joiner Level)
- SPL: proveedor de servicios (Service Provider Level)
- DSL: fuentes de datos (Data Source Level)

Estos niveles, descritos a continuación, son la base estructural que permite agrupar convenientemente los componentes requeridos por un sistema abierto, interoperable, flexible y sujeto a estándares (OAI y DCMI [7]) y normas archivísticas. Todos los niveles tienen como característica común que son niveles de servicio. Por consiguiente, la arquitectura propuesta es una arquitectura orientada al servicio.

3.2. Elementos Básicos

Los elementos participantes en esta arquitectura se diferencian en cuatro categorías. Esta clasificación es bastante conforme con la de OAI y, aunque en menor grado, con la de Fedora. Los elementos son los siguientes:

1. **Recursos**, entendidos como cualquier objeto que puede ser descrito. Puede corresponder con un recurso hardware o virtual (un servidor de datos, datos simples o incluso un servicio).
2. **Paths a recursos o conectores**, todos los elementos desde los cuales se pueden propagar descripciones de recursos. Es importante remarcar que un recurso puede ser a la vez un path o conector, esto es, en el caso en que sea el propio recurso quien propaga descripciones de sí mismo. Adicionalmente, puede haber paths o conectores especializados, cuya única función sea la de diseminar descripciones de recursos.
3. **Descripciones**, esto es, conjuntos de metadatos. Aunque la estructura y formato de tales metadatos no está todavía determinada, como punto de partida se tomará el modelo propuesto por OAI. Esencialmente, un descriptor define un conjunto de métodos a través de los cuales es posible acceder (por invocación de los mismos) a un recurso.
4. **Contenido**, es decir, la información del archivo propiamente dicha.

A modo de ejemplo, los paths a recursos permitirían acceder a diferentes contenidos albergados en el archivo digital con formatos y características heterogéneas. Otro ejemplo posible sería el caso en el cual dichos paths proveen las definiciones de métodos requeridos en determinadas circunstancias especiales. Finalmente, en el caso más general, los paths proveerán las necesarias descripciones para soportar la interoperabilidad entre diversas fuentes. De esto último derivan las siguientes consideraciones adicionales:

- En primer lugar, resulta evidente la necesidad de proporcionar descripciones relativas a las propias interacciones. En caso contrario, para situaciones sofisticadas como, por ejemplo, la transformación de datos de un modelo a otro para la integración semántica de contenidos heterogéneos no sería posible.
- Segundo, tales descripciones resultan esenciales para el soporte de la semántica y de la conducta en un sistema de integración de archivos digitales a través de la Web. Es un objetivo básico del diseño de DAWIS que estas descripciones sirvan para definir no sólo las operaciones disponibles en cada archivo digital local sino también las operaciones de más alto nivel que resultan de la integración, disponibles globalmente, en el sistema como un todo. Así, las operaciones y funciones del archivo global no son estrictamente la suma de las operaciones y funciones de cada archivo independiente sino que la reunión deberá suponer un aumento de funcionalidad.

El párrafo anterior pone de manifiesto la importancia de considerar descripciones relativas a la lógica de funcionamiento, en general. En particular, el orden de ejecución de las operaciones es importante; pues si no es el correcto, entonces el resultado global no será el esperado. A menudo, las descripciones relativas a la lógica de funcionamiento son independientes de los contenidos (objetos digitales) implicados en las interacciones correspondientes. Este tipo de descripciones son clave para un buen funcionamiento global del sistema, por lo que es esencial identificar y definir precisamente las interacciones elementales que se aplican a los diferentes objetos y/o recursos en el sistema. De hecho, consideramos que ésta es la piedra angular en el diseño de DAWIS.

3.3. Arquitectura

En esta sección se expone cómo se aplicarán los conceptos y elementos descritos en la sección precedente en la arquitectura propuesta. La descripción sigue un orden ascendente, desde las fuentes de datos al nivel del usuario final.

DSL (Data Source Level). DSL se refiere al nivel de los datos en cada archivo digital. Son las fuentes de datos, la parte *back-end*, situada en el nivel inferior de la arquitectura que almacenan la información (de carácter cultural, socioeconómico, administrativo, etc) o contenido informativo propiamente dicho. Este nivel es así mismo responsable de gestionar y procesar las consultas locales emitidas en casa caso, a la fuente de datos local correspondiente. Estos accesos incluyen tanto información estructurada (bases de datos relacionales tradicionales) como semiestructurada (fundamentalmente metadatos en documentos XML).

SPL (Service Provider Level). Posiblemente sea éste el nivel más importante y complejo de los existentes en la arquitectura propuesta. Como es sabido, un objeto digital es un recurso identificado globalmente (en un espacio de nombres) por su OID. Cada objeto digital encapsula un contenido que se asocia con una o más interacciones, ya sean relativas a operaciones o al comportamiento del objeto digital. Por su parte, los conectores o paths permiten asociar los métodos en cada objeto digital con la implementación de los mismos. Para esto, los conectores proveen las descripciones de los métodos y mecanismos que, en conjunto, constituyen la definición abstracta de los servicios ofrecidos por el archivo digital. Por tanto, desde este nivel se ofrecen los servicios (de ahí su nombre) en la forma de componentes básicos que los implementan. Es importante destacar que cualquier interacción con las fuentes de datos (nivel DSL) tiene lugar a través de algún componente del nivel SPL de modo que, en conjunto, actúa como un nivel de intermediación (mediator) entre las fuentes de datos y el unificador de servicios (nivel SJJ).

CSL (Component Server Level). Este nivel corresponde con una base de datos de componentes (CDB) donde se almacenan físicamente todos los componentes disponibles en el nivel SPL y registrados en el nivel SRL (descrito a continuación). Coincide con el habitual servidor de aplicaciones presente en cualquier arquitectura conforme con J2EE.

SRL (Service Registry Level). El conjunto de paths a recursos existentes en el sistema, así como las descripciones correspondientes a los métodos y/o mecanismos que manejan, se localizan en este nivel. Como consecuencia, ni los métodos ni los mecanismos que los implementan forman parte de los objetos digitales ni tampoco de los componentes que los constituyen. Como ya se ha mencionado anteriormente, estas descripciones se almacenan junto con los propios conectores en un servidor de directorio (DS en la figura 4). Como en el caso anterior, este servidor es también un elemento característico en cualquier solución conforme a J2EE.

SJJ (Service Joiner Level). Este nivel es responsable de proveer la funcionalidad necesaria para facilitar la reunión, composición y agregación dinámica de servicios. Más concisamente, éste es el punto del sistema desde el cual se invocan los métodos en los diferentes componentes que constituyen los objetos digitales.

Tales métodos se determinan a partir de paths en el SRL. Desde la perspectiva de usuario final, éste es el nivel que unifica y hace disponibles servicios de alto nivel sobre el archivo digital global. Define, por tanto, un contexto para el usuario del archivo donde aspectos como control de sesión, persistencia o tratamiento de las interdependencias entre servicios son clave. Obviamente, previo al acceso de cualquier servicio, es preceptivo resolver su identificación por lo que en el DS en el nivel SRL existirán descripciones análogas a las ya descritas que faciliten la resolución del servicio.

WCL (Web Client Level). Los clientes (usuarios o aplicaciones) que consultan al archivo digital integrado se sitúan en este nivel de la arquitectura. No hay ninguna limitación relativa al tipo de acceso desde el cliente, pudiendo ser -por tanto- vía SOAP, HTTP, FTP o por medio del correo electrónico.

Finalmente, para una mejor comprensión de la propuesta, vamos a resumir la misma resaltando sus principales características: el nivel SPL alberga un número de objetos digitales constituídos por componentes elementales. Estos objetos digitales residen en el nivel CSL y acceden a los contenidos del archivo digital a partir de la interacción entre los niveles SPL y DSL. Físicamente, los componentes que constituyen los objetos digitales están almacenados en una CDB y son registrados en el DS (las páginas amarillas) del nivel SRL. Es importante resaltar que ningún componente existe en el sistema hasta que es registrado y publicado en el nivel SRL. Como consecuencia, se le asigna un URI. Adicionalmente, los métodos de los componentes de los objetos digitales son invocados desde el nivel SJL para componer servicios más complejos y elaborados. Esta invocación viene determinada por los paths a recursos o conectores, que son los encargados de proveer la descripciones necesarias (los paths también residen en el directorio). Por último, los usuarios o clientes vía Web solicitan servicios a través del nivel SJL que resuelve dinámicamente la solicitud estableciendo la sesión y demás controles necesarios.

4. Conclusiones

En este artículo se ha presentado la propuesta de arquitectura de integración para el proyecto DAWIS. Esta propuesta supone un primer paso en el camino a recorrer como parte de la investigación en este proyecto. Nuestra meta es producir una arquitectura para la integración de archivos digitales a través de la Web que reúna las características de eficiencia, interoperabilidad, extensibilidad y, lo que resulta fundamental, que sea orientada al servicio. Esto último es clave para el éxito del sistema final resultante, pues la calidad y adecuación de los servicios que se ofrezcan junto con el grado de automatización interna y la facilidad de uso determinan la utilidad y, en consecuencia, el éxito de la solución final.

Así mismo, la arquitectura propuesta es conforme a las actuales tendencias y estándares relativos a la interoperabilidad entre archivos digitales así como las arquitecturas de integración basadas en Web.

5. Reconocimientos

Este trabajo está parcialmente financiado por el Ministerio de Ciencia y Tecnología (MCYT- TIC2002-04050-02, proyecto que forma parte de DAWIS) y por la Comunidad de Madrid (07T/0056/2003 3, Proyecto EDAD-UPM).

Referencias

1. Mohan, C.: C. Dynamic e-Business. Trends in Web Services. IBM Almadén Research Center, 2002.
2. Shirky, C.: C. Web Services and Context Horizons. IEEE Computer, pp. 98-100, September 2002.
3. Kreger, H.: Web Services: Conceptual Architecture (WSCA 1.0) IBM Software Group, May 2001
4. The Mellon Fedora Project: Digital Library Architecture Meets XML and Web Services. Forthcoming European Conference on Research and Advanced Technology for Digital Libraries, Rome, Italy, September 2002.
5. Payette, S., Lagoze, C.: Flexible and Extensible Digital Object and Repository Architecture Second European Conference on Research and Advanced Technology for Digital Libraries, Heraklion, Crete, Greece, September 21-23, 1998, Springer, 1998, (Lecture notes in computer science; Vol. 1513).
6. Open Archives Initiative: <http://www.openarchives.org>
7. Dublin Core Metadata Initiative: <http://www.dublincore.org>
8. Carpenter, L. et al.: Open Archives Forum: Project Deliverable D3.1, Issue:1.0 November 2002, in http://www.oaforum.org/otherfiles/oaf_d31organizational1.pdf
9. Marcos E., Cáceres P., Caverro JM, Vela B., Costilla C., Eibe S., Menasalvas E. y Sáenz J.: DAWIS: Sistematización del Desarrollo de Portales para el Acceso Integrado a Archivos Digitales en la Web. Taller RedBD dentro de JISBD 2002, El Escorial, Madrid, Noviembre 2002
10. Costilla C., Eibe S., Menasalvas E., Sáenz J., Marcos E., Cáceres P., Caverro JM y Vela B. DAWIS: Enfoques preliminares sobre la arquitectura de referencia para la Integración de Archivos Digitales en Web. Taller RedBD dentro de JISBD 2002, El Escorial, Madrid, Noviembre 2002
11. CRCIT: <http://www.crcit.es>
12. SIAP: <http://www.crcit.es/siap>
13. Asamblea de Madrid: <http://www.asambleademadrid.es>