La gestión de objetos digitales: una aplicación para la e-Ciencia

Digital object management: an application for e-Science

Luis Zorita Vicente y Alicia López Medina

Resumen

El nuevo entorno de la e-ciencia, en red y altamente colaborativo, está generando un nuevo tipo de unidad de información de naturaleza compleja cuya gestión, acceso, almacenamiento y reutilización requieren un sistema capaz de tratar esa complejidad. Se propone un nuevo modelo de Objeto Digital que estructura la naturaleza compleja de la unidad de información en red y una arquitectura, FEDORA, capaz de integrar ese modelo de Objeto Digital. Partiendo de estos conceptos y arquitectura, se presentan algunas aplicaciones desarrolladas utilizando Fedora: creación de un repositorio de objetos digitales, un sistema para la creación de revistas digitales mediante query RDF y transformaciones xslt, un servicio web de búsqueda a texto completo (datos y assets textuales) mediante indexación con estándar Lucene, un servicio de generación de RSS en la búsqueda y un servicio web de diseminación mediante protocolo OAI-PMH.

Palabras clave: Objeto Digital, FEDORA.

Summary

The new online and highly collaborative e-science environment is creating a new type of complex information unit that requires a system capable of managing, accessing, storing and reusing this sort of complexity. A new Digital Object model has been proposed to structure the complex nature of the online information unit and an architecture, FEDORA, capable of integrating this Digital Object model. Based on these concepts and architecture, applications developed using Fedora have been presented: digital object repository creation, a system for creating digital magazines through RDF queries and xslt transformations, a full-text (data and textual assets) search web service by means of indexing with the Lucene standard, an RSS generation service in the search, and a dissemination web service by means of the OAI-PMH protocol.

Keywords: Digital Object, FEDORA

El nuevo entorno de la e-ciencia está generando un nuevo tipo de unidad de información de naturaleza compleja

1. Información en la e-Ciencia

La aplicación y el uso de los recursos computacionales y las tecnologías en red está permitiendo una forma de hacer ciencia altamente colaborativa

Según el libro blanco de la ciencia en España -2004-[1] "La e-Ciencia se entiende como el conjunto de actividades científicas desarrolladas mediante el uso de recursos distribuidos accesibles a través de Internet". La aplicación y el uso de los recursos computacionales y las tecnologías en red está permitiendo una forma de hacer ciencia altamente colaborativa, basada en la red y en grandes volúmenes de información heterogénea y distribuida.

Este nuevo entorno de investigación está provocando grandes transformaciones en los procesos de comunicación científica; una de estas transformaciones es la incorporación a estos procesos de nuevas unidades de información que incluyen ahora no sólo las revistas y sus unidades, los artículos, sino también datasets, simulaciones, software, contenidos dinámicos, representaciones, anotaciones, etc. Por otra parte, estas unidades de información (que denominaremos **data**), tienen una naturaleza compleja y dinámica, son objetos "vivos" que tienen diferentes ubicaciones en la red, se agregan y mantienen relaciones entre sí, y cuyas necesidades de almacenamiento, gestión, acceso, difusión y reutilización, en un entorno de trabajo en colaboración, requieren un sistema escalable y flexible, capaz de tratar y representar esa complejidad y adaptarse a los más que previsibles cambios tecnológicos.

En este contexto, se propone un nuevo modelo de Objeto Digital que tiene una uri asociada, capaz de representar cualquier data, soportando la agregación de varios data en un único Objeto digital, y capaz de expresar relaciones semánticas entre ellos.

Trataremos de demostrar cómo todo ello nos coloca en muy buena posición para apoyar los distintos flujos del proceso de investigación científica en la red: creación, obtención de versiones, reutilización, almacenamiento , comunicación y publicación.

2. El modelo de Objeto Digital (DO)

- Nuestro modelo de Objeto Digital es un contenedor capaz de "estructurar" los datos que contiene. Se trata, por tanto, de una agregador de componentes. Característica clave para funcionar en la red es que tanto el contenedor como sus componentes sean direccionables mediante URI's.
- Se basa en la idea de que las unidades de información pueden ser:
  a) heterogéneas
  b) complejas
  c) generadas dinámicamente
  d) y mantienen relaciones entre sí

Característica clave para funcionar en la red es que tanto el contenedor como sus componentes sean direccionables mediante URI's

2.1. DO Heterogéneo:

Ha de poder representar muchos tipos de **unidades de información:**

- objetos textuales,
- imágenes,
- libros electrónicos,
- objetos multimedia,
- datasets,
- metadatos
- y muchas otras entidades.

2.2. DO Complejo:

Ha de soportar la agregación en único DO de más de un componente de cualquiera de los tipos especificados anteriormente. Esos componentes pueden estar incluidos en el objeto o bien estar referenciados mediante URL.

2.3. DO dinámico:

Debe disponer de métodos asociados (otro tipo de DO) capaces de actuar sobre él. Por ejemplo, ofrecer la vista ampliada de una imagen o la tabla de contenidos del documento libro obtenida mediante su generación en tiempo de ejecución.

Los componentes pueden estar incluidos en el objeto o bien estar referenciados mediante URL

2.4. Soporta relaciones semánticas:

Ha de ser capaz de expresar esas tripletas (sujeto, verbo y predicado). Por ejemplo :

```
<rdf:description rdf:about="info:fedora/bibliuned:ETFSerieV-55C14806-F9E2-237B-99D3-045AD53B1069">
  <rel:isMemberOf rdf:resource="info:fedora/bibliuned:ETFSerieV2004"/>
  <rel:isMemberOf rdf:resource="info:fedora/bibliuned:Setarticulo"/>
 </rdf:description>
```

3. Estándares de metainformación

Este modelo de DO soporta cualquier esquema de metadatos que podamos crear; no obstante, creemos que, frente a la utilización de esquemas de metadatos creados para necesidades específicas, es más adecuada la combinación de estándares ya definidos y vigentes en la red aumentando de esta manera la interoperabilidad. Así utilizaremos DublinCore[2] con carácter general, LOM[3] para representar Objetos de aprendizaje, RDF[4] para representar relaciones semánticas, y estaremos atentos a las nuevas estructuras que defina el W3C[5] .

Otro tipo de estándares como OAI-PMH[6] (y posiblemente ORE[7] en el futuro) serán referencia cara a la difusión e interoperabilidad de los DO´s.
Creemos que esta apuesta por los estándares es imprescindible cara a las posibilidades futuras de conexión a esa red de la siguiente generación, donde los contornos de la información digital se difuminarán. Quien no esté en condiciones de conectarse al resto estará condenado a desaparecer, por muy buenos que sean sus valores específicos.

4. Fedora. Una arquitectura para DO's

FEDORA[8] (no confundir con el OS) es el middleware que hemos escogido como arquitectura capaz de integrar este modelo de objeto digital.
Sus características principales son:

- Es un middleware con arquitectura SOA desarrollado en Java
- Todas sus funciones están expuestas como servicios web.
  http://www.fedora.info/definitions/1/0/api/Fedora-API-A-LITE.wsdl
  http://www.fedora.info/definitions/1/0/api/Fedora-API-A.wsdl
  http://www.fedora.info/definitions/1/0/api/Fedora-API-M.wsdl
  http://www.fedora.info/definitions/1/0/api/Fedora-API-M-LITE.wsdl
- Montado sobre Tomcat como servlet
- Serializa los objetos digitales en XML con arreglo al siguiente esquema:
  http://www.fedora.info/definitions/1/0/foxml1-0.xsd
- Múltiples vistas de los DO mediante su asociación con métodos definidos en servicios web.
- Pueden gestionar el(los) asset(s) en forma local o remota
- Relaciones entre objetos digitales : arquitectura basada en RDF
  http://www.fedora.info/definitions/1/0/fedora-relsext-ontology.rdfs
- Metadatos sobre relaciones entre DO´s basados en RDF
  http://62.204.194.45:8080/fedora/get/bibliuned:ETFSerie1-F444A3E3-4230-C5FF-C492-BA3D011D7738/RELS-EXT
- Posibilidad de buscar en el repositorio como un grafo , esto es mediante las relaciones definidas en RELS-EXT(relaciones con otros objetos)
- Control de acceso utilizando el estándar XACML con diferentes niveles de granularidad:
  - de repositorio
  - del objeto
  - de componentes dentro de un objeto (datastreams)
  - Permite incorporar LDAP y Shibboleth[9] como sistemas de autenticación
  - Control y mantenimiento de las diferentes versiones de un objeto digital
  - Auditoría
  - Preservación

Quién no esté en condiciones de conectarse al resto estará condenado a desaparecer, por muy buenos que sean sus valores específicos

FEDORA es el middleware que hemos escogido como arquitectura capaz de integrar este modelo de objeto digital

5. Servicios desarrollados en torno a los objetos digitales

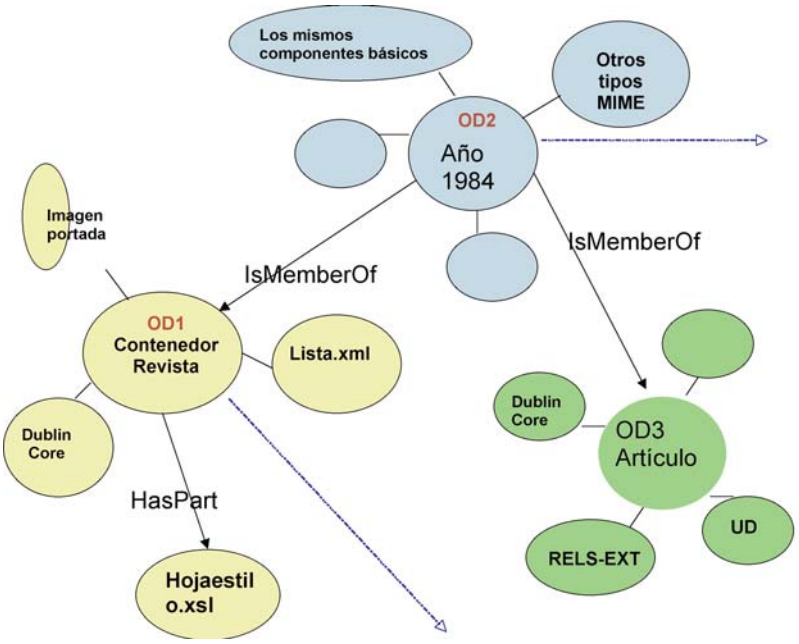
Partiendo de estos conceptos y arquitectura, vamos a indicar algunas aplicaciones desarrolladas utilizando Fedora:

- Creación de un repositorio conteniendo objetos digitales (locales y remotos): http://e-spacio.uned.es. En este momento disponemos de unos 11.500 DO's, procedentes mayoritariamente de la migración de datos ya existentes en BBDD.
- Creación mediante query RDF y transformaciones xslt de 4 revistas digitales basadas en la existencia de las bases de datos de tripletes y en transformaciones xslt:
  http://62.204.194.45:8080/fedora/get/bibliuned:revistaETF/demo:Collection/view/
- Servicio web de búsqueda a texto completo (datos y assets textuales) mediante indexación con

estándar Lucene. http://62.204.194.45:8080/fedoragsearch/rest incluyendo acceso directo a texto, imagen y video:
http://62.204.194.45:8080/fedoragsearch/rest?operation=gfindObjects&query=maeso&hitPageSize=

- Servicio de generación de RSS en la búsqueda, por ejemplo:
http://62.204.194.45:8080/fedoragsearch/rest?operation=gfindObjects&query=casa&hitPageStart=1&hitPageSize=10&indexName=DemoOnRepouned&restXslt=salidarss
- Servicio web de diseminación mediante protocolo OAI-PMH [10] .
http://www.madrimasd.org/informacionidi/e-ciencia/buscar-documentos/default.asp

FEDORA permite incorporar LDAP y Shibboleth como sistemas de autenticación



En este momento disponemos de unos 11.500 DO's procedentes mayoritariamente de la migración de datos ya existentes en BBDD

6. Por hacer

- Aprovechar la posibilidad de intercambiar información en formato XML para crear y desarrollar proyectos de colaboración en red.
- Estudiar y, si es posible, participar en la en el proyecto ORE[11] (Object Reuse and Exchange) orientado a modificar la forma en que se realiza el proceso de la comunicación y publicación científica.
- Estudiar mecanismos de reutilización de data en proyectos GRID-Europa.

7. Conclusiones

- Creemos que esta elección tecnológica nos permite un alto grado de independencia de nuestros contenidos respecto de la representación elegida.
- Está orientada a su despliegue en la web.

- Permite crear relaciones internas y externas entre unidades de información.
- Permite asociar data con información de una forma sencilla.
- Utiliza esquemas estándar y publicados en la red para representar sus contenidos lo cual facilita la posibilidad de federación de contenidos.

Referencias

[1] *http://www.fecyt.es/documentos/e-Ciencia.pdf*
[2] *http://dublincore.org/*
[3] *http://www.imsproject.org/metadata/*
[4] *http://www.w3.org/RDF/*
[5] *http://www.w3.org/*
[6] *http://www.openarchives.org/OAI/openarchivesprotocol.html*
[7] *http://www.openarchives.org/ore/*
[8] *http://www.fedora-commons.org/*
[9] *http://shibboleth.internet2.edu/*
[10] *http://www.openarchives.org/pmh/*
[11] *http://www.openarchives.org/ore/*

Luis Zorita Vicente
(lzorita@pas.uned.es)
Alicia López Medina

UNED

Se deben estudiar mecanismos de reutilización de data en proyectos GRID-Europa

Esta elección tecnológica permite un alto grado de independencia de nuestros contenidos respecto de la representación elegida