OSHI y WebOSHI: Sistema para compartir y reusar ontologías de fuentes documentales históricas

A. Calvo
M. A. Serrano–Tenllado
J. Sánchez–Jurado
A. Maroto
J. A. Romero del Castillo

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio C2. 14071 – Córdoba

Resumen Se presenta un sistema local (OSHI) y remoto (WebOSHI) para crear y compartir ontologías en el dominio de las fuentes documentales históricas. El sistema que se propone permite a un grupo de investigadores obtener las conceptualizaciones y los interfaces de usuario desde un servidor central. Con ello podrá, bien de forma local o usando la red de internet, transcribir documentos históricos a XML que se ajusten a los modelos conceptuales propuestos. Asimismo, el sistema permite que el usuario pueda crear sus propias conceptualizaciones reusando estructuras previamente definidas. Finalmente el sistema dispone de un indexador y buscador de información sobre los documentos registrados por los diferentes investigadores.

1. Introducción

El trabajo de investigación en historia se caracteriza por la abundante información documental que se maneja de muy variada tipología. Las fuentes documentales históricas son, para muchos investigadores, parte esencial de su sistema de información, a partir de las cuales darán solución a los problemas históricos planteados. Recientes trabajos han mostrado la viabilidad del uso de estructuras reusables de conocimiento (ontologías) para su aplicación en la investigación histórica [CSTR01] [STCR01]. Para llevar a cabo de forma efectiva las propuestas presentadas es necesario construir sistemas informáticos que cumplan un conjunto de requisitos y que solucionen los problemas planteados.

1.1. Planteamiento del Problema

El trabajo con las fuentes documentales históricas presenta gran cantidad de problemas. Entre otros destacamos los siguientes:

1. Problemas en relación a los usuarios.

   a) A veces la investigación es llevada a cabo por un grupo de investigadores que aportan información parcial al sistema o explotan individualmente dicha información. Esto supone que cada investigador trabaja con sus propias fichas y estructuras, siendo difícil contrastar y explotar de forma conjunta la información que cada uno posee de forma individual.
   b) Los investigadores no pueden acceder eficientemente a la información del resto del grupo.
2. Problemas en relación a la conceptualización de los documentos y sus estructuras de información.
   a) No siempre el historiador separa la información que le proporcionan las fuentes documentales de la información que refleja la realidad histórica que trata de reconstruir.
   b) A veces se trata de investigadores en formación que no disponen de conocimientos sobre la estructura interna de los documentos, por lo que al realizar los resúmenes pueden dejar de recoger datos importantes.
   c) Las estructuras de conocimiento que representan a los documentos son, en la mayoría de los casos, rígidas, de forma que ante la necesidad de representar nuevos tipos de fuentes documentales, no contempladas a priori, resulta difícil y tediosa su creación.
3. Problemas en relación con el registro y entrada de información.
   a) A veces los sistemas de registro de información disponen de una gran variedad de campos y datos, de forma que el trabajo de entrada de datos es largo y tedioso, al estar éste distribuido a lo largo de una amplia gama de formularios, subformularios y campos. Ante esto, el historiador termina por abandonar este sistema de trabajo y pasa a registrar sólo resúmenes en lenguaje natural.
   b) Generalmente, el tiempo que se tarda en registrar la información siguiendo el método de formularios y subformularios es muy superior al tiempo de registro mediante resúmenes en lenguaje natural.
   c) El diseño de formularios no siempre es fácil, sobre todo si se trata de una amplia variedad de documentos históricos.
   d) El uso de formularios, subformularios, áreas de texto, campos, etc., distribuidos, hace que el historiador pierda la idea de la estructura de los documentos.
4. Problemas en relación al origen de la información.
   a) La transcripción y resumen de la información de las fuentes documentales se realiza de diferentes modos y entornos. En ocasiones el historiador trabaja en el archivo con los documentos originales sin conexión a red y de forma local.
   b) En otras ocasiones el historiador trabaja con fotocopias o microfilm de forma local o conectado a red.
   c) En otras ocasiones trabaja con documentos digitalizados dentro del propio computador o accediendo a la red.
5. Problemas en relación a la explotación de información.
   a) La información registrada no siempre es fácil de trasladar a otras formas de representación y se requieren programas especiales para lograrlo.
   b) No siempre quedan claramente separadas las estructuras de información que se sintetizan e infieren a partir de un conjunto de descripciones obtenidas de las fuentes documentales, de la información de las propias fuentes.

c) Al explotar la información de las fuentes documentales y lograr estructuras de síntesis, no siempre están garantizadas la consistencia de las mismas con la información registrada en las fuentes.

6. Problemas en relación a la presentación de la información.
   a) Al registrar la información de las fuentes documentales, hay una cierta tendencia a mezclar aspectos de presentación de la información con el contenido de la misma.
   b) Las fuentes documentales, por lo general, quedan sólo accesibles en el sistema local de computación, teniendo difícil acceso desde el exterior.

Para resolver esta problemática se han desarrollado los sistemas OSHI (**O**ntology in **S**ources **HI**story) y WebOSHI. El primero trabaja de manera local y el segundo trabaja en modo red de acuerdo con el modelo cliente–servidor.

1.2. Antecedentes

El desarrollo de sistemas para la creación de ontologías lleva varios años en estudio. Gómez-Pérez hace una revisión de los trabajos de los últimos años [BGP99]. En la actualidad existen varios sistemas que ofrecen la posibilidad de editar, construir y recuperar ontologías en varios dominios. Los sistemas Ontolingua Server, OILed, WebODE, Protègè2000 son algunos de estos sistemas.

En la bibliografía consultada son escasas las referencias a los sistemas creados para la transcripción, mantenimiento y recuperación de ontologías y documentos sobre fuentes documentales históricas. Son abundantes las referencias a la creación de sistemas de bases de datos relacionales, e incluso orientadas a objetos, pero siempre para resolver problemas puntuales y no bajo el punto de vista de estructuras reusables y compartidas.

El sistema KLEIO [Tha87] [Tha91] es uno de los pocos sistemas desarrollados para el tratamiento de fuentes documentales históricas. En la actualidad existe una versión que permite trabajar con una estación de tratamiento de imágenes de documentos pero no aborda la caracterización estructural de los documentos desde la perspectiva de la reusabilidad.

1.3. Objetivos

Los objetivos básicos de este trabajo consisten en presentar las características de un sistema software que resuelva los problemas planteados para la transcripción, registro, recuperación y explotación de fuentes documentales históricas bajo las orientaciones de la ingeniería del conocimiento y, más específicamente, bajo el punto de vista de la ingeniería de ontologías en su versión local y remota.

2. Modelo

Para llevar a cabo esta tarea de compartir estructuras, hay que establecer cuales son los elementos de que se compone el sistema y cómo se coordinan entre si. La figura 1a muestra un esquema del mismo. Se compone de un conjunto de tipos de documentos que permiten la gestión del sistema: esquemas conceptuales de los documentos en xml–Schema, esquemas de formularios, transformaciones XSLT, documentos XML e imágenes.

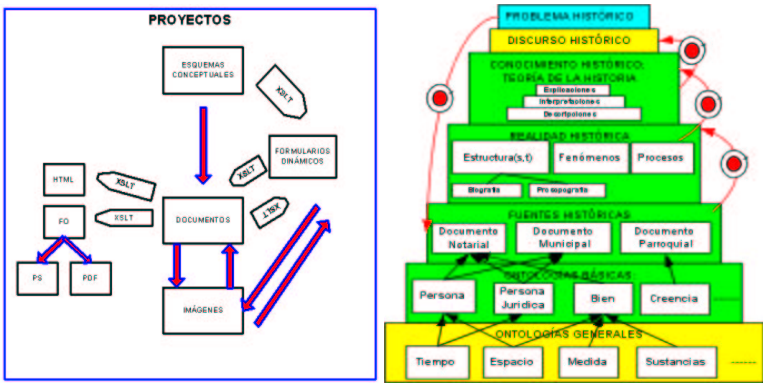

Figura 1. a) Elementos del modelo. b) Edificio Cognitivo.

2.1. Esquemas conceptuales

Los esquemas XML–Schema son los encargados de conceptualizar los documentos y las estructuras descriptivas internas que en ellos aparecen. El lenguaje XML–Schema permite definir un conjunto de tipos simples y complejos de estructuras y restricciones sobre las mismas, capaces de conceptualizar de forma clara y eficiente los diferentes tipos de documentos. La posibilidad de establecer espacios de nombres, importar e incluir unos esquemas conceptuales en otros, etc., permite aplicar los principios de reusabilidad en los que estamos basando nuestro desarrollo [STCR01]. Estos esquemas son importantes y hacen posible comprobar si el documento que se crea a partir de ellos cumple con sus especificaciones. Se ha creado un conjunto de esquemas que corresponden a las ontologías identificadas en el sistema: generales (espacio, tiempo, unidades de medida, etc.), básicas (personas físicas, jurídicas, instituciones, bienes, etc.), ontologías de los documentos (notariales, parroquiales, hacendísticos, municipales, etc.), ontologías sobre documentos de síntesis(biografía, prosopografía, etc.) que corresponden con las plantas del edificio cognitivo [STCR01] propuesto en el modelo de la figura 1b.

La aplicación WebOshi permite a los usuarios, en función del grupo a que pertenecen, la gestión de estos esquemas conceptuales, es decir, usar los esquemas, descargarlos, crearlos, modificarlos, incorporarlos y borrarlos del sistema. Para la creación y modificación se ha utilizado la aplicación XML–Spy.

2.2. Formularios dinámicos en DFRML

Aún cuando es posible crear instancias de documentos directamente en XML, usando editores convencionales de texto o programas específicos como XML–Spy, se ha desarrollado un lenguaje para especificar formularios dinámicos que se adapten a los esquemas conceptuales de los documentos [SJ02]. El uso de este lenguaje y el componente OCX desarrollado a partir de él, permite crear el interface de entrada de información dinámico que presenta grandes ventajas. Como se observa en la figura 2, la representación de la estructura del documento se presenta de manera esquemática,

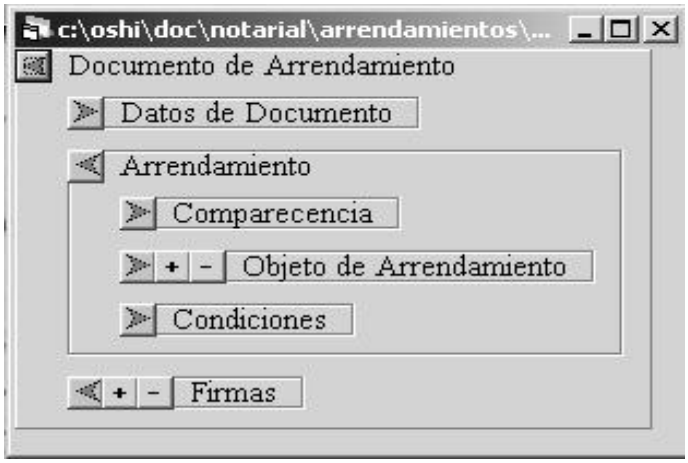

Figura 2. Interface general dinámico.

de forma que el usuario pueda observar en cualquier momento el esquema básico del mismo.

El elemento correspondiente a objeto de compraventa que se recoge en este documento, debe ser un elemento múltiple, y además, permitir seleccionar entre todos los posibles tipos de bienes definidos. La figura 3a muestra cómo seleccionar uno de los diferentes bienes. El interface se adaptará al bien seleccionado. La figura 3b muestra un ejemplo en el que en un documento notarial de compraventa se registran más de un bien. Cada bien contiene internamente la estructura correspondiente que se ha definido conceptualmente.

El lenguaje DFRML permite además la declaración de estructuras múltiples, la inclusión de tablas, la selección de opciones y el establecimiento de restricciones sobre los campos. Para simplificar la creación del interface, también permite en su definición la importación de otros formularios, la incorporación de entidades y la definición de clases. De esta forma se logra crear, de una manera rápida y sencilla, interfaces de nuevos documentos a partir de otros ya definidos. La posibilidad de poder usar clases, entidades e incluir ficheros, facilita también el mantenimiento de la aplicación. La figura 4 muestra el uso de las clases y la inclusión de ficheros para la descripción del formulario.

2.3. Transformaciones XSLT

Todos los ficheros que se usan en el presente proyecto, esquemas de documentos y entidades asociadas (xsd), formularios asociados a los documentos (frm), instancias de los esquemas conceptuales de los documentos notariales, parroquiales, municipales, etc., están basados en XML. Son por tanto susceptibles de aplicarles transformaciones y convertirlos en otros formatos en función de las necesidades. Por ejemplo, se han

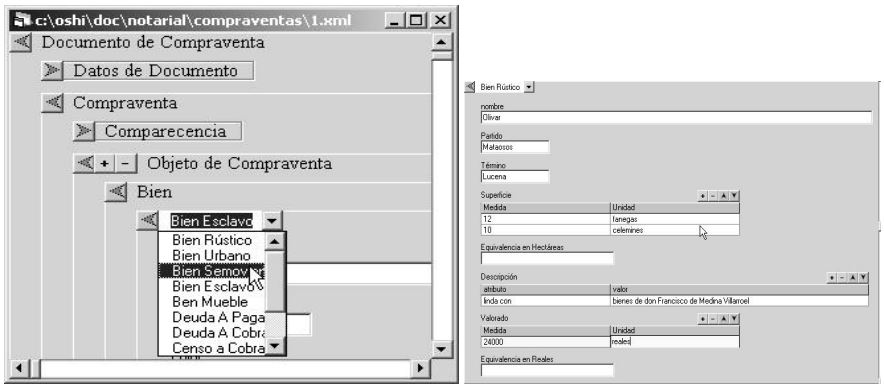

Figura 3. a) Objetos múltiples y selección de un tipo de bien. b) Expansión del interface.

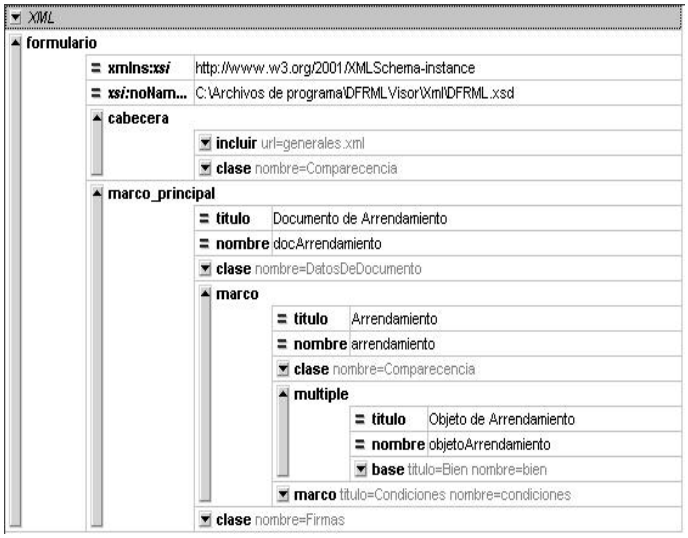

Figura 4. Declaración del interface en el lenguaje DFRML.

definido plantillas XSLT para presentar los documentos en HTML, FO[1], LATEX, etc. El sistema también contempla la creación, edición, descarga y carga de estos ficheros XSLT. Como en el caso anterior, el lenguaje XSLT permite la inclusión e importación de unos ficheros XSLT en otros, lo que facilita la creación y el mantenimiento de los mismos. Por otra parte, también es necesario definir hojas de transformación XSLT que trasladen, de manera transparente, la información de los documentos desde el lenguaje DFRML a XML y viceversa.

3. Organización interna de la información

3.1. Ontologías

Para organizar internamente la información se han definido varias ontologías: notariales, parroquiales, síntesis, etc. Cada una de ellas contiene la definición de un conjunto de documentos que han sido conceptualizados. Así, dentro de los documentos notariales se han conceptulalizado documentos de arrendamiento, compraventa, dote, testamento, etc.; dentro de los parroquiales se han definido las actas de bautismo, matrimonio, defunción, etc. y dentro de la síntesis se han creado documentos para caracterizar la biografía y la prosopografía. Esta catalogación de grupos de documentos a los que hemos denominado ontologías (notarial, parroquial, de síntesis, etc.) se define, así misma, dentro de una ontología sobre las ontologías. El sistema contempla un directorio de configuración que incluye un fichero de ontologías en el que se declaran todas las ontologías establecidas en el mismo. La figura 5 muestra la estructura del fichero. Para cada ontología se especifica un conjunto de propiedades (fecha de creación, autor, versión, etc.). Cada ontología contiene las especificaciones del conjunto de documentos que contiene la ontología. Para cada documento se especifica, entre otras características, la ubicación de los ficheros xml–schema, las transformaciones xslt, o el directorio donde se almacenan las instancias de los documentos. OSHI tiene un menú dedicado al mantenimiento de este fichero de ontologías. La figura 5 muestra la estructura del documento ontologías y las opciones disponibles en el sistema.

3.2. Imágenes

Otro de los elementos de información de OSHI son las imágenes. El sistema permite el almacenamiento y recuperación de imágenes digitales que pueden registrarse de forma individual o asociadas a documentos como se observa en la figura 6a.

3.3. Documentos

El elemento de información básico de OSHI es el documento. Como se ha indicado, en cada ontología se han definido un conjunto de documentos caracterizados por una estructura propia interna. Un elemento común a todos es la descripción de sus datos básicos.

[1] El formato FO, permite mediante el uso de programas apropiados, la transformación de los documentos a formatos pdf, ps, etc.

Figura 5. Ontologias.

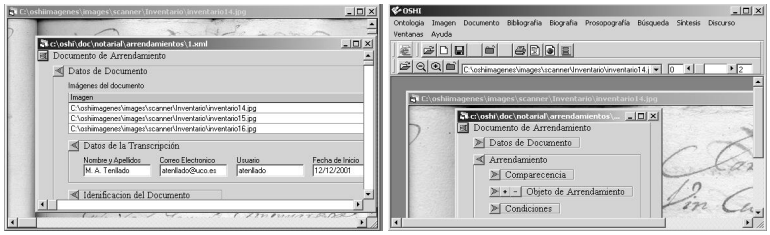

Figura 6. *a*) Documentos e imágenes asociadas. *b* Aspecto general.

4. Modelo funcional

La aplicación WebOSHI está destinada a gestionar la información de un grupo de usuarios que deseen compartir modelos conceptuales, documentos e imágenes. Es la aplicación que desempeñará el papel de servidor. Por su parte la aplicación OSHI, puede trabajar de forma autónoma o como cliente de la aplicación WebOSHI.

4.1. OSHI

La figura 6b muestra el aspecto general de la aplicación OSHI. Se compone de ocho módulos esenciales: Ontologías, Imágenes, Documentos, Bibliografía, Biografía, Prosopografía, Discurso Histórico y Búsqueda. El módulo ontologías gestiona el empaquetado, la instalación y desinstalación de las ontologías en comunicación con la aplicación WebOshi. El módulo imágenes gestiona todo lo relacionado con las imágenes digitales: abrir, almacenar y editar las imágenes, así como diversas operaciones sobre ellas. El módulo documentos gestiona la transcripción de documentos. El usuario selecciona la ontología y el tipo de documento que desea transcribir y OSHI le presenta la estructura del documento para iniciar la transcripción. Para cada estructura interna del documento y para cada documento se ha implementado una axiomática que lleva a cabo inferencias para la creación y mantenimiento de estructuras de síntesis, por ejemplo la biografía. Se han establecido diversas categorías de axiomas: axiomas que se aplican a la descripción de personas, axiomas que tienen en cuenta el papel que cada persona desempeña en el documento, axiomas de contradicción, etc.

4.2. WebOshi

WebOSHI presenta unos módulos similares a los de la aplicación OSHI, añadiendo aspectos relacionados con la gestión de usuarios. Por ello, uno de los módulos está reservado para el administrador del sistema, donde se realizarán las tareas de altas, bajas, y especificación de privilegios a los usuarios. El módulo de conceptualización permite gestionar todas las tareas relacionadas con el mantenimiento y distribución de esquemas conceptuales de los documentos y sus entidades descriptivas básicas, la conceptualización de los formularios asociados con las clases y entidades que en ellos se hayan definido y la gestión de las transformaciones XSLT para la transformación de los documentos. El módulo de documentos permite gestionar los documentos (envío y recepción de transcripciones) de acuerdo con los esquemas conceptuales definidos. Esto es, crear, modificar, visualizar, descargar, subir, validar, buscar, etc. Sobre el conjunto de documentos que se va almacenando se lleva a cabo un proceso de indexación para favorecer su búsqueda. El módulo de imágenes gestiona el archivo de imágenes digitalizadas.

5. Explotación del sistema de información: Búsqueda y síntesis de estructuras

Se están obteniendo los primeros resultados sobre la explotación del sistema de información. Este módulo permite la síntesis de estructuras a partir de la información

registrada en los documentos. Inicialmente se han caracterizando dos estructuras fundamentales: la biografía y la prosopografía. El objetivo último es que el sistema pueda crear y rellenar las partes internas de estas estructuras generales a partir de los datos registrados en las fuentes documentales históricas, teniendo en cuenta toda una axiomática que se está definiendo a tal fin. Se necesitarán axiomas de identificación y reconocimiento, axiomas de detección de contradicciones, axiomas de síntesis, etc. En cualquier caso el sistema deberá proporcionar una síntesis e integración de la información existente en las fuentes eliminando redundancias y realizando inferencias.

Referencias

[BGP99] R. Benjamins and A. Gómez-Pérez. Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem solving Methods. In *Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends*, Estocolmo, 1999.

[CSTR01] A. Calvo, M. A. Serrano-Tenllado, and J. A. Romero. La ingeniería del conocimiento en la investigación de la historia. In *La representación y organización del conocimiento: metodologías, modelos y aplicaciones: Actas del V Congreso Isko-España*, Alcalá, April 25–27 2001. Universidad de Alcalá de Henares. Servicio de Publicaciones.

[SJ02] J. Sánchez-Jurado. Dfrml. uso de lenguajes marcados en la construcción interfaces dinámicos. In *Proyecto Fin de Carrrera*, Córdoba, Marzo 2002. Universidad de Córdoba.

[STCR01] M. A. Serrano-Tenllado, A. Calvo, and J. A. Romero. La representación de documentos históricos: Aplicación de esquemas y xml para modelar y representar la información y el conocimiento descrito en las fuentes de protocolos notariales. In *La representación y organización del conocimiento: metodologías, modelos y aplicaciones: Actas del V Congreso Isko-España*. Universidad de Alcalá de Henares. Servicio de Publicaciones, 8–11 de Abril 2001.

[Tha87] M. Thaller. *Methods and Techniques of historical computations*, pages 147–156. P. Denley, Manchester, 1987.

[Tha91] M. Thaller. The historical workstation project. *Computers and the Humanities*, (25), 1991.