CONTENTUS: HACIA LAS BIBLIOTECAS MULTIMEDIA SEMÁNTICAS

Jan Nandzik
Acosta Consult
E-mail: jn@acosta-consult.de
Andreas Heß
German National Library
E-mail: a.hess@d-nb.de
Jan Hannemann
German National Library
E-mail: j.hannemann@d-nb.de
Nicolas Flores-Herr
Acosta Consult
E-mail: nf@acosta-consult.de
Klaus Bossert
Acosta Consult
E-mail: kb@acosta-consult.de

Traducción: Cayetano Hernández Muñiz, Biblioteca Nacional de España

Reunión: 149. Information Technology, Cataloguing, Classification and Indexing with Knowledge Management

Resumen:

La cantidad, cada vez mayor, de contenidos y conocimientos publicados en línea hace posible que las bibliotecas complementen su propia información y presenten sus colecciones de forma novedosa. La información relacionada conceptualmente puede vincularse semánticamente de modo que los usuarios se beneficien de colecciones de datos más ricas y posibilidades de búsqueda novedosas, que saquen partido de las relaciones inherentes entre los materiales, los metadatos locales y las fuentes de información externas.

Este trabajo presenta posibles soluciones a los retos fundamentales de integrar fuentes de información heterogéneas y proporcionar métodos innovadores de búsquedas semánticas, tal como se desarrollan para las bibliotecas y archivos multimedia en el proyecto CONTENTUS.

Introducción

En Alemania unas 30000 instituciones culturales albergan una increíble riqueza de contenido multimedia almacenado en soportes tales como libros, imágenes, cintas y películas. Estas organizaciones de patrimonio cultural afrontan el desafío de proporcionar a los ciudadanos acceso por Internet al conocimiento contenido en sus vastas colecciones multimedia. Los novedosos servicios de búsqueda semántica multimedia serán el fundamento tecnológico futuro para que los usuarios accedan a colecciones digitales. Una condición sine qua non

para la aplicación de estas tecnologías es la *integración de metadatos bibliográficos, metadatos generados automáticamente y recursos de información externa* en una base del conocimiento. Este trabajo se centra en el desarrollo técnico y metodológico relacionado llevado a cabo por la Biblioteca Nacional Alemana y sus socios en el contexto del proyecto CONTENTUS.

Buscar en colecciones multimedia: desafíos afrontados por organizaciones de patrimonio cultural

Para apoyar las búsquedas semánticas multimedia, las colecciones multimedia de gran envergadura necesitan enriquecerse con metadatos descriptivos suficientes y apropiados. Sin embargo, hasta el momento la mayor parte de los recursos multimedia es anotada y catalogada manualmente por expertos en información. Puesto que la mayoría de los recursos multimedia, como los materiales audiovisuales, contiene mucha información, la generación de metadatos manuales es muy compleja, costosa y lleva mucho tiempo.

En la práctica, al indizar manualmente grandes ficheros multimedia no estructurados, o bien el recurso no puede describirse completamente, o bien sólo unos fragmentos del recurso pueden indizarse en detalle. Esta indeseable situación se encuentra comúnmente en organizaciones de patrimonio cultural y se debe a la falta de recursos humanos que sean capaces de hacer frente a colecciones multimedia de rápido crecimiento.
Con recursos relevantes e importantes en Internet (por ejemplo, Wikipedia y Geonames) y conjuntos de datos editados en colaboración como los ficheros de autoridad, la tarea de indizar recursos multimedia no debería limitarse a describir sus contenidos. Puesto que estos recursos externos tienen el potencial de enriquecer semánticamente las unidades de las colecciones, los recursos multimedia como conjunto o las entidades individuales (por ejemplo, personas, lugares, sucesos) que se hallan dentro de recursos individuales deberían conectarse semánticamente con conjuntos de datos relevantes de Internet para complementar los metadatos disponibles.

No obstante, indizar multimedia así como conectar entidades de unidades de colecciones multimedia a recursos externos relevantes es difícil y está lejos de ser una rutina diaria en organizaciones de patrimonio cultural –éste es uno de los retos abordados por el proyecto CONTENTUS.

La aspiración de CONTENTUS: bibliotecas multimedia de próxima generación

CONTENTUS es un proyecto de desarrollo tecnológico y de investigación dirigido por la Biblioteca Nacional Alemana bajo la tutela de la iniciativa de investigación THESEUS, financiada por el gobierno alemán. Proporciona muchas herramientas con soluciones para instituciones culturales y otros poseedores de información, que facilita una perfecta transición desde los datos digitales puros hacia un entorno de búsqueda semántica multimedia [Bossert, Flores-Herr, Hannemann, 2009].

El marco CONTENTUS y las metodologías y conceptos que se están desarrollando en el proyecto producen un sistema que ayuda a las instituciones

culturales a proporcionar a los usuarios finales acceso a colecciones multimedia a gran escala. Los usuarios finales se benefician de opciones de búsqueda innovadoras que son alimentadas por la abundancia de recursos multimedia y metadatos procedentes de fuentes diversas, incluyendo datos "tradicionales" compilados intelectualmente, información generada automáticamente y fuentes de Internet.

CONTENTUS trabaja en estrecha colaboración con ALEXANDRIA y Mediaglobe –proyectos de THESEUS que se centran en el desarrollo de la Web 3.0 y en las tecnologías de los medios- para incorporar aquellas comunidades que construyen y mejoran las redes de conocimiento semántico. Estas colaboraciones producirán finalmente redes de conocimiento abiertas donde los recursos multimedia de las instituciones culturales puedan conectarse con recursos de la web social: las *bibliotecas multimedia de próxima generación*.

CONTENTUS pretende crear una infraestructura para organizaciones de patrimonio cultural que permita el procesamiento eficiente de grandes colecciones multimedia y su conexión con recursos de metadatos externos. Los pasos individuales forman una cadena de proceso, como muestra la **figura 1**.

Figura 1: La cadena de proceso de CONTENTUS

1.) **Digitalización**: todavía existen muchos archivos en forma analógica, lo que hace imposible un procesamiento automático de los recursos. Para tales archivos, la digitalización masiva es el primer paso para abrir los materiales a una búsqueda semántica completa.

2.) **Control de calidad**: es esencial que los análisis de calidad automatizados (por ejemplo, la comprobación de calidad por escaneado de la página de un libro) y los procedimientos de optimización vayan al mismo ritmo que la actualización de las máquinas de digitalización (por ejemplo, los robots de escaneado de libros). El objetivo aquí es mejorar la calidad de los materiales tanto para uso humano como para análisis del contenido (ver el siguiente paso).

3.) **Análisis del contenido**: las descripciones de metadatos manuales a menudo no son suficientes para permitir una búsqueda efectiva de objetos multimedia. Anotar extensamente el contenido textual, sonoro o audiovisual, por otro lado, por ejemplo transcribiendo discursos o indizando una emisión de noticias, supone enormes esfuerzos en términos de recursos humanos y financieros. En CONTENTUS, los servicios que analizan automáticamente recursos multimedia comunes

como imágenes, grabaciones musicales o vídeos son importantes para facilitar la generación de información relevante para la búsqueda.

4.) **Vinculación semántica**: para enriquecer los metadatos disponibles, la información generada automáticamente puede vincularse con metadatos bibliográficos y recursos de internet. Por ejemplo, el tema de un documental de noticias puede ser el autor de un libro, que puede a su vez enlazarse a un artículo de la Wikipedia o a una entrada de un fichero de autoridad. Además, a las entidades extraídas tales como lugares, personas, sucesos, etc., se les quita la ambigüedad (por ejemplo, para discernir automáticamente “apple” como una fruta de “apple” como una corporación) y se las conecta con la *Nube de Datos Abiertos Vinculados (Linked Open Data)*, es decir, la totalidad de las numerosas fuentes de información disponibles en la web como datos vinculados.

5.) **Redes de conocimiento abiertas**: en este paso, los recursos multimedia pueden ser enriquecidos más aún por comunidades externas con recursos externos.

6.) **Búsqueda semántica**: CONTENTUS ofrece a los usuarios finales una funcionalidad de búsqueda de multimedia innovadora al combinar búsquedas de textos, imágenes, contenido sonoro y audiovisual en una interfaz de usuario semántica y unificada.

En este trabajo nos centraremos en los retos de la integración de datos y los métodos para facilitar búsquedas semánticas.

Integración de datos

Uno de los desafíos clave en el proyecto CONTENTUS es la necesidad de integrar datos y metadatos de diversas fuentes. Normalmente, tales datos pueden comprender el producto de esfuerzos de digitalización, documentos de origen digital, las contribuciones de la comunidad de usuarios, etc. Puesto que pretendemos hacer el contenido multimedia procedente de diferentes fuentes accesible a través de una interfaz común, el desafío consiste en integrar y alinear los metadatos correspondientes. En oposición a los sistemas de catalogación tradicionales, enriquecemos nuestros datos con fuentes externas, ya que creemos que el usuario puede beneficiarse de la información adicional que contienen estas fuentes. Incluso si la calidad de los metadatos externos es a veces (pero no necesariamente) más baja de lo que puede esperarse normalmente de los metadatos creados por los bibliotecarios, pueden complementar los datos existentes si son más detallados o contienen aspectos que no se han considerado de otra manera. Por ejemplo, el catálogo del Archivo de Música Alemana (Deutsches Musikarchiv) –un archivo que alberga una colección fundamental de partituras y grabaciones sonoras en Alemania y sirve como centro de información bibliográfica relacionada con la música- no incluye las canciones o pistas individuales de una grabación. Al conectar los datos del catálogo con una base de datos de música, el usuario tiene acceso a información más detallada, como los listados de pistas.

Para cualquier servicio de información que integre datos de diferentes fuentes, tenemos que distinguir entre dos casos:

1.) Vinculación

2.) Integración

En el primer caso, los metadatos de fuentes diferentes no se almacenan en la misma base de datos, sino sólo aproximadamente asociados. A los metadatos de fuentes que no están bajo el control del proveedor del servicio, sólo se accede cuando se necesita. Este método tiene la ventaja de que los (meta-) datos que se presentan al usuario del servicio siempre están lo más actualizados posible, incluso si proceden de fuentes que no están bajo el control del proveedor del servicio. La desventaja es que no puede garantizarse la disponibilidad de las fuentes externas.

En el segundo caso, los metadatos de todas las fuentes se integran en la misma base de datos o almacén de ontologías por parte del proveedor del servicio. Así, la disponibilidad sólo depende del propio sistema del proveedor. Sin embargo, debe establecerse una política de actualización para asegurar que los datos que se presentan al usuario del servicio no estén demasiado desfasados. Otro problema que hay que considerar es el de las licencias, puesto que no sólo se accede a los datos de otras fuentes, sino que se copian.

En ambos escenarios el desafío técnico más importante es encontrar una correspondencia entre los diferentes esquemas. Esto puede hacerse

1.) manual/intelectualmente o

2.) (semi-) automáticamente.

En CONTENTUS, usamos ambos métodos de correspondencia dependiendo de las fuentes de datos. Debería señalarse que las fuentes mismas de metadatos podrían generarse intelectual o automáticamente. Por ejemplo, usamos algoritmos de extracción de información automáticos para encontrar y quitarles la ambigüedad a personas, organizaciones, lugares y materias de textos escaneados, pero también incorporamos datos de ficheros de autoridad generados intelectualmente.

En nuestro actual sistema, integramos las siguentes fuentes de metadatos:

- Biblioteca Nacional Alemana: ficheros de autoridad y datos del catálogo
- Wikipedia: retratros de personas (en proyecto: información contextual adicional para personas y lugares)
- MusicBrainz: listados de pistas de CDs

- Extraídos automáticamente: personas, organizaciones, lugares y materias del texto y audio, semejanza entre las pistas de música

La información de autoridad y del catálogo sirve como referencia. La correspondencia entre la Wikipedia y el fichero de autoridad la mantienen manualmente voluntarios. Esta correspondencia ya se usa en la Wikipedia alemana y el sistema de catálogo de la Biblioteca Nacional Alemana. La correspondencia entre los listados de pistas de MusicBrainz y el fichero de autoridad en la actualidad también se genera manualmente.

La figura 2 ilustra la integración de datos de fuentes diferentes en CONTENTUS. Los metadatos almacenados en CONTENTUS pueden verse como organizados en una red, conectando autores y obras así como información adicional, por ejemplo sobre lugares, materias o épocas relacionados.

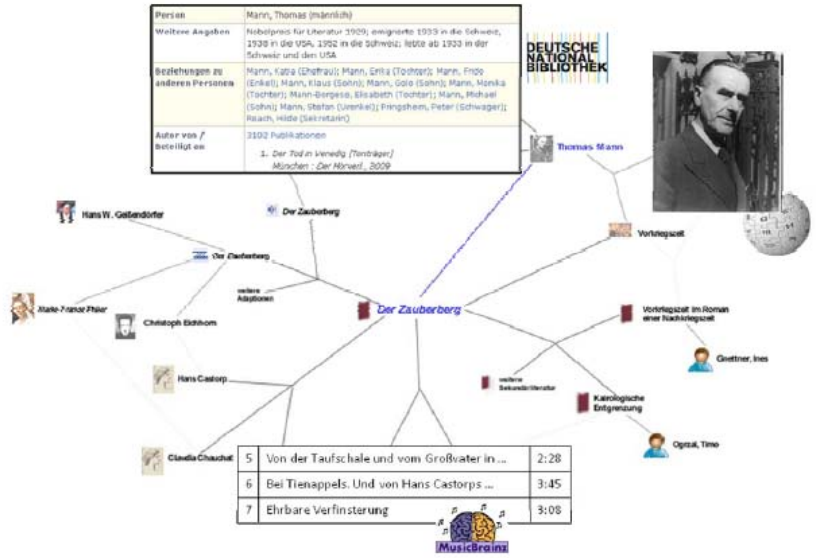

Figure 2: Integration of data from different sources regarding the work “Der Zauberberg” from German author Thomas Mann

Figura 2: Integración de datos de fuentes diferentes sobre la obra “Der Zauberberg” del autor alemán Thomas Mann

Usar URIs

Según los principios de los Datos Abiertos Vinculados (Linked Open Data) se recomienda usar URIs dereferenciables y persistentes como identificadores. Los URIs permiten unir diferentes fuentes de información referentes a entidades de interés (personas, lugares, organizaciones, etc.).Como tales, son fundamentales en nuestros esfuerzos de alinear conjuntos de datos e integrar información de fuentes de datos heterogéneas en la base del conocimiento de CONTENTUS.

Desambiguar personas

Para mejorar la funcionalidad de búsqueda es importante vincular materiales tales como documentos textuales, con otra información, como datos de ficheros de autoridad. Esto se está implantando actualmente para personas, entidades corporativas, lugares, etc. Uno de los retos aquí consiste en determinar a qué persona nos estamos realmente refiriendo, si existe más de una persona con el mismo nombre. Asimismo, los nombres de persona a veces no se distinguen a primera vista de otras palabras que aparecen en un diccionario. En CONTENTUS, no sólo se vinculan a ficheros de autoridad los autores de documentos u otras personas mencionadas en los metadatos, sino también las personas mencionadas en el texto extractado de los documentos mismos. Esto requiere el uso de algoritmos automáticos para extraer nombres de personas así como para quitarles la ambigüedad.

El método usado en CONTENTUS es el de Pilz y Paaß [Pilz y Paaß, 2009]. Para quitar la ambigüedad a un nombre de persona, se hace una comparación entre el contexto en que aparece el nombre y un documento de referencia donde se conoce la verdadera identidad de la persona. En nuestro ejemplo concreto el texto se compara con las entradas de la persona en cuestión en la Wikipedia. La persona, cuyo artículo de la Wikipedia es muy similar al texto relevante del documento, entonces es identificada como la persona mencionada en el documento original.

Correspondencias entre las instancias y los esquemas

Crear correspondencias entre conjuntos de datos heterogéneos es uno de los problemas más antiguos en ciencias de la información. Hemos de distinguir entre dos partes: la correspondencia de instancias u objetos y la correspondencia de estructuras de datos. El problema de hacer correspondencias de instancias automáticamente es equivalente a la detección de duplicados. Es común usar la métrica de distancia entre cadenas, como la bien conocida métrica de Levenshtein [Levenshtein, 1965] o la de Jaro-Winkler [Winkler, 1999], y/o medidas de similitud fonética como Soundex [Russell, 1918] como un medio para detectar instancias idénticas. En la práctica, los algoritmos de correspondencia modernos usan una combinación de métricas; véase, por ejemplo [Johnston y Kushmerick, 2004].

El problema de hacer correspondencias de esquemas automáticamente se ha tratado en la investigación y la literatura desde que aparecieron las bases de datos [Melnik et al., 2002] y se ha vuelto a tratar para la correspondencia del esquema XML, y recientemente también para la correspondencia de ontologías [Shvaiko et al., 2009; Hess, 2006]. Los algoritmos normalmente usan las similitudes estructuradas y léxicas y algunos también aprovechan los casos en que se sabe que las instancias están representadas en ambos esquemas.

Correspondencias de localizaciones

Podríamos confiar en las correspondencias generadas intelectualmente entre esquemas o instancias en algunos casos (véase arriba), porque fueron creadas en colaboración –en el caso de las correspondencias de la Wikipedia-, o porque fueron fáciles de crear –en el caso de las correspondencias de

MusicBrainz. Sin embargo, para correspondencias más grandes es crucial tener algoritmos de correspondencia automáticos razonablemente exactos.

Para el futuro desarrollo de la búsqueda semántica en CONTENTUS, pensamos incluir controles gráficos novedosos (véase la próxima sección). Con el fin de poder mostrar información geográfica sobre localizaciones que, por ejemplo, se hallan en el texto completo de los documentos de los medios o que se conectan por medio de metadatos. El objetivo es incluir correspondencias a una base de datos geográficos como GeoNames.

Pensamos usar una combinación de heurística y métrica de similitud para lograr esta tarea. En el fichero de autoridad que sirve como base para las correspondencias, normalmente está disponible la información sobre el país y (si existe) el estado federal o la provincia en que se localiza una ciudad. Esta información puede aprovecharse para quitar la ambigüedad, si el nombre de una ciudad no es único (e.g., París de Texas, EE.UU. vs. París de Francia). Se han usado con éxito métodos similares para quitar ambigüedad a otra información de ficheros de autoridad en el contexto del primer proyecto de datos abiertos vinculados de la Biblioteca Nacional Alemana [Hannemann et al., 2010].

Búsqueda y navegación:
El motor de búsqueda desarrollado en el proyecto CONTENTUS combina dos fuentes de información: un índice tradicional de texto completo de transcripciones de OCR y audio, así como información semántica mantenida en una ontología. Los medios subyacentes a esta “Búsqueda Semántica Multi-Media” (BSMM) comprenden material audiovisual y sonoro, soportes impresos escaneados y documentos textuales de origen digital.

La búsqueda de CONTENTUS pretende dar acceso a todas estas fuentes de información a través de una interfaz unificada. En consecuencia, los principales retos de diseño para la interfaz de usuario (IU) fueron:

1.) La combinación transparente de diferentes fuentes de información

2.) La integración perfecta de información multimedia y metadatos asociados

3.) Un acceso amigable a las características de búsqueda semánticas

Usar la información semántica para las búsquedas supone tres ventajas principales sobre los motores de búsqueda tradicionales:

1.) Los usuarios pueden ojear la información de forma exploratoria siguiendo los enlaces semánticos entre los materiales y las fuentes de información

2.) A los individuos y a las palabras clave se les puede quitar la ambigüedad por su significado

3.) Las relaciones entre los resultados de la búsqueda se hacen evidentes

El planteamiento de CONTENTUS para el diseño de la IU

Para que los usuarios finales utilicen plenamente las fuentes de metadatos integradas y los materiales diferentes, es esencial proporcionar una interfaz de búsqueda que sea intuitiva y suministre nuevas capacidades de búsqueda semántica. El proyecto CONTENTUS ha producido dos prototipos que funcionan, basados en la web, de su Búsqueda Semántica Multi-Media. Hemos recogido la reacción del usuario frente a la facilidad de uso de los dos prototipos desde 2008 mediante demostraciones en ferias comerciales como la Feria del Libro de Frankfurt de 2008 y 2009, y la feria de la International Broadcasting Conference (IBC) en Amsterdam en 2009.

Antes de la fase de diseño del tercer demostrador (actualmente en desarrollo), mantuvimos dos sesiones de prototipos de examen (véase e.g. [Maass, 2008]) en el Institut für Rundfunktechnik en Munich en 2010 para confirmar la reacción previa (positiva) de los visitantes de la feria comercial, esta vez con usuarios de un entorno archivístico y bibliotecario.

Decidimos no confrontar al grupo de usuarios de prueba con los prototipos existentes de nuestro motor de búsqueda basado en la web. En su lugar, en un primer paso presentamos a los participantes un conjunto de tareas de búsqueda predefinidas y les pedimos ideas sobre cómo una IU podría resolverlas más fácilmente. En un segundo paso mostramos a nuestro grupo de prueba un conjunto de elementos de control para obtener comentarios sobre cómo los entendían y qué clase de interacción esperaban.

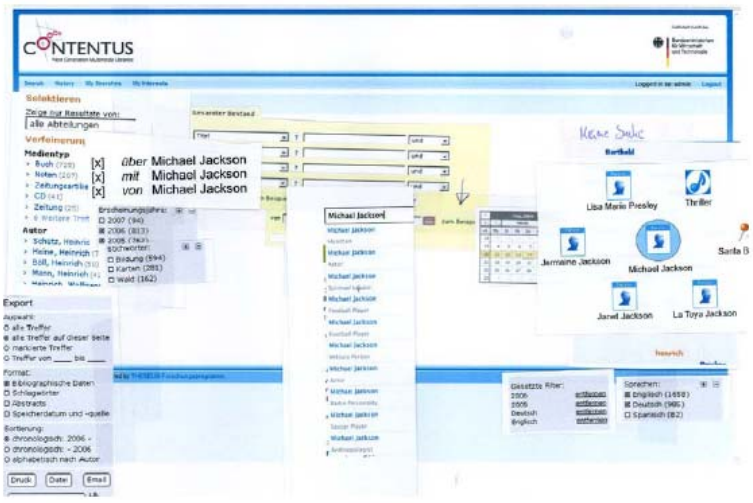

Figure 2: A CONTENTUS paper prototyping example. Users could freely select and arrange predefined controls in a second pass of our prototyping sessions in 2010

Figura 3: Un ejemplo de prototipo de examen de CONTENTUS. Los usuarios podían seleccionar y fijar libremente controles predefinidos en un segundo paso de nuestras sesiones de prototipo en 2010

Los resultados de nuestra prueba de usuarios muestran que el usuario medio realmente prefiere un clásico enfoque tipo Google para su entrada de búsqueda: una casilla de búsqueda y una representación textual de la lista de los resultados de búsqueda. Sin embargo, sospechamos que una razón para esta preferencia es que muchos usuarios no están familiarizados con los elementos de la interfaz de usuario más innovadores o inusuales y son así reacios a usarlos.

Ya que consideramos la posibilidad de exploración como una de las mayores ventajas de las interfaces de búsqueda asistida semánticamente, en consecuencia tuvimos que elegir una interfaz que animara a los usuarios a utilizar el “valor semántico añadido” y al mismo tiempo no los forzara en exceso ni los desorientara con posibilidades de interacción desconocidas. La mayoría de los usuarios prefirió una interfaz de búsqueda facetada para restringir sus listas de resultados iniciales con palabras clave sin ambigüedad, antes que un lenguaje de pregunta específico y antes que la eliminación de ambigüedad como una característica de búsqueda asistida.

Un caso de uso

La actual interfaz de usuario permite la siguiente interacción de muestra:

Un usuario busca libros escritos por un periodista llamado *Michael Jackson*. Por consiguiente introduce el término “Michael Jackson” en la aplicación de búsqueda. Sin embargo, Michael Jackson también es el nombre de un cantante y músico muy popular. De forma parecida a un motor de búsqueda convencional la BSMM en primer lugar recupera resultados de texto plano del índice de búsqueda, puesto que no hay manera de que la aplicación adivine a cuál de los dos individuos podría haberse referido el usuario.

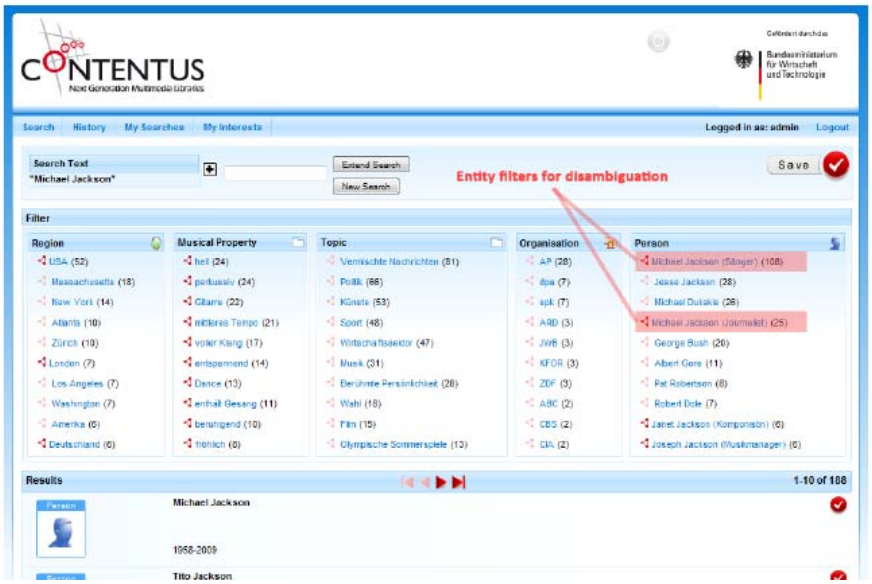

Figura 4: El conjunto de resultados de CONTENTUS para el término de búsqueda “Michael Jackson” antes de filtrar y quitar la ambigüedad de persona

Debido a la semejanza de los nombres la aplicación devuelve una lista de resultados que contiene una mezcla de resultados de búsqueda deseados y no deseados en todos los tipos de soportes. La mayor parte de éstos están relacionados con el artista Michael Jackson (y no con el periodista) y no son, así pues, de interés para el usuario.

Además de la lista de resultados de materiales la interfaz de búsqueda también proporciona varias listas de filtros dinámicos (facetas), que se generan automáticamente desde los conjuntos de resultados de la búsqueda. Éstas comprenden los conceptos y nombres de entidades más relevantes dentro del conjunto de resultados, y se compilan a partir de los metadatos del catálogo preparados intelectualmente y la información reconocida por los módulos automáticos de análisis de contenido de CONTENTUS.

La relevancia de las facetas no sólo se basa en su frecuencia en el conjunto de resultados, sino en la sumamente efectiva reducción del tamaño del conjunto de resultados –las facetas que aparecen en todos (o la mayor parte de) los resultados se omiten, ya que no ofrecen posibilidades de filtrado sustanciales.

Las facetas de filtrado se agrupan en un conjunto fijo de clases:

- Conceptos musicales
- Localizaciones
- Materias
- Organizaciones
- Personas

Ahora el usuario puede usar estas facetas de filtrado para restringir su búsqueda –internamente al término de búsqueda original se le añade el término correspondiente o entidad no ambigua con un “y” lógico. Cada faceta de filtrado figura como un icono de color que representa la procedencia de los datos (véase la figura 4) –esto permite distinguir entre personas no ambiguas contenidas en los ficheros de autoridad de las bibliotecas y entidades de nombre genérico encontradas por análisis estadísticos del material textual.

Como la mayor parte de los resultados de la búsqueda en nuestro ejemplo tienen conexión con el *artista* Michael Jackson, muchas de las materias y entidades también tienen relación con la música. Pero también vemos materias como “cerveza” y “whisky” que son comunes a las obras del *periodista* Michael Jackson. La lista de filtrado de personas muestra a ambos Michael Jackson en nuestra base de datos de personas, así como a personas relacionadas como los hermanos del cantante pop. Un clic en la faceta del periodista reduce el conjunto de resultados a todos los materiales relevantes para el usuario –ya no se muestran los resultados de la búsqueda relacionados con el cantante.

Curiosamente, las facetas de filtrado para el término de búsqueda *Michael Jackson* también muestran materias y organizaciones (como KFOR, la Fuerza de Kosovo) que nada tienen que ver con las dos personas más evidentes, el periodista y el artista. Mientras que algunos usuarios se desconcertaron por esto y descartaron estas entradas como ruido no relevante, muchos exploraron más y descubrieron un tercer Michael Jackson, un general de las fuerzas de la OTAN –un resultado no previsto pero, sin embargo, útil.

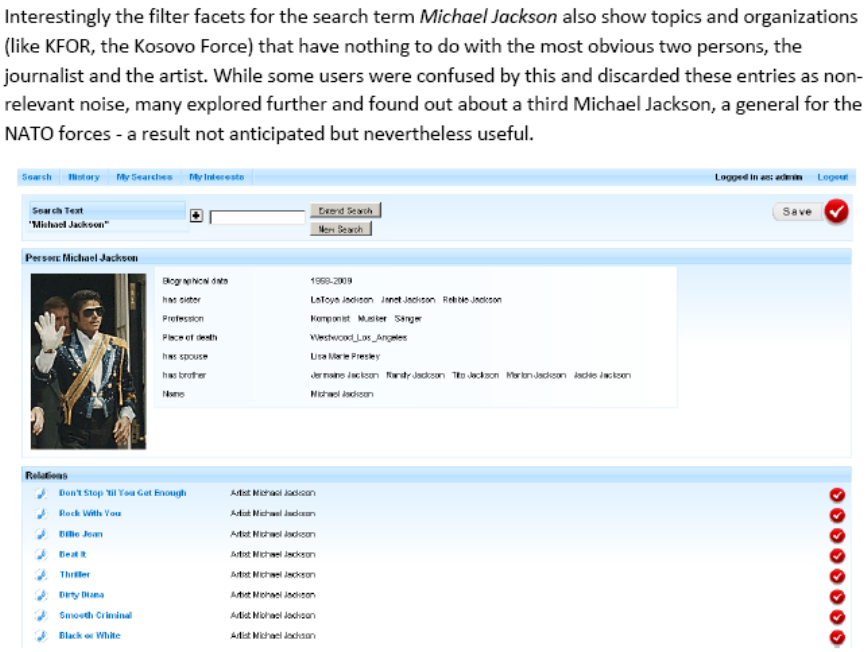

Figura 5: la página de entidad para el cantante Michael Jackson

Cada persona no ambigua tiene también una *página de entidad* que puede invocarse pulsando en la entrada de personas en la lista de resultados. Aquí, a los usuarios se les proporciona toda la información semántica asignada, como parientes de personas, sus obras como creadores, fechas y localizaciones de

nacimiento, etc. Las páginas de entidad se enriquecen además con imágenes, información bibliográfica y texto de la Wikipedia. La figura 5 muestra la página de entidad del *cantante* Michael Jackson.

Desde la página de entidad los usuarios pueden lanzar una nueva búsqueda pulsando en cualquiera de las entidades, materias, localizaciones, etc. conectadas, posibilitando así una verdadera experiencia de exploración semántica que combina perfectamente con el aspecto visual relativamente convencional de la interfaz.

Elementos novedosos de la interfaz de búsqueda

Nuestras pruebas de usuarios mostraron que *la interacción con representaciones gráficas de grafos semánticos* en la mayor parte de los casos no se entendió completamente o se consideró poco práctica. Mientras que los usuarios comprendieron el significado de una vista gráfica de las relaciones entre personas, no captaron la idea de interactuar con las visualizaciones.

Un *control cronológico interactivo* ha resultado ampliamente aceptable a la mayoría de nuestro grupo de prueba del prototipo. Restringir el conjunto de resultados señalando un marco de tiempo sobre el control pareció ser un modo intuitivo de buscar datos y esto se aplica de forma universal en la mayor parte de los dominios del conocimiento.

Algunos usuarios también propusieron *facetas de filtrado con jerarquías* (de hiperónimos o hipónimos como planta -> flor -> rosa), así que probaremos su facilidad de uso en un prototipo futuro del mismo modo que ya tenemos en orden jerárquico partes de los encabezamientos de materia del fichero de autoridad.

Muchas de las personas de la prueba han evaluado positivamente *un mapa interactivo*, una visualización gráfica de las localizaciones en los resultados de la búsqueda. Los usuarios podrán limitar sus resultados marcando un área geográfica sobre el mapa.

Resultados de la prueba de la interfaz

Nuestras pruebas de usuarios han mostrado que:

- Las características de la búsqueda semántica ayudan mucho a reducir el esfuerzo de localizar combinaciones relevantes en grandes archivos multimedia.

- Es crucial que los usuarios entiendan *cómo y por qué* cualquier hito de la búsqueda llegó al conjunto de resultados. De lo contrario, la base semántica puede ser confusa, sobre todo si incluimos conexiones adicionales, como parientes de una persona, en el conjunto de resultados.

- Los usuarios son reacios a usar visualizaciones novedosas como entrada de búsqueda única. Esperan una casilla de búsqueda tradicional, pero aceptan visualizaciones interactivas como herramienta para refinar la búsqueda.

- Una búsqueda de exploración se usa sobre todo como un paso secundario después de introducir una o más palabras clave. Ningún usuario propuso la exploración pura como método preferido para responder a nuestras preguntas, pero todos, por otro lado, reaccionaron favorablemente a las facilidades de exploración ofrecidas por nuestros prototipos, especialmente las páginas de entidad.

Adiciones proyectadas para la IU

Durante el marco temporal del proyecto añadiremos al menos las siguientes funcionalidades a nuestra interfaz:

- *Funciones de las entidades en las facetas de filtrado*: nuestros usuarios de la prueba hicieron hincapié en la necesidad de poder diferenciar entre filtrado para, por ejemplo materiales escritos por la persona A y para materiales que tuvieran como materia a la persona A.

- *Explicación mejorada de resultados y facetas*: la lista de resultados debería reflejar por qué cualquier elemento ha salido en el conjunto de resultados, sobre todo los resultados que sólo tengan una relación semántica indirecta con el término de búsqueda.

- *Control de visualización cronológica interactivo*: los usuarios deberían poder acotar su conjunto de resultados marcando un intervalo de tiempo sobre la visualización de modo que sólo se muestren los resultados dentro de ese marco temporal.

- *Mapa interactivo*: los usuarios deberían poder restringir sus resultados a un área libremente seleccionable sobre el mapa.

Conclusión

No sólo en Alemania, sino también en la Unión Europea, se están haciendo enormes esfuerzos para asegurar la disponibilidad del patrimonio cultural digitalizado, ya sea para su archivo a largo plazo o para la creación de bibliotecas digitales como *Europeana* o la *Deutsche Digitale Bibliothek*. Por tanto, cada vez más bibliotecas y archivos se enfrentan al reto de integrar recursos de diversos proyectos de digitalización, metadatos locales y conjuntos de datos externos. Por desgracia, aún faltan herramientas que faciliten un suministro sencillo aunque extenso de los recursos y metadatos para sistemas y catálogos de bibliotecas.

Por otro lado, hay una fuerte demanda de los usuarios de bibliotecas y archivos para acceder a los medios digitales en un contexto organizado de modo equitativo a través de todos los tipos de materiales. El contenido de audio y vídeo, según los hábitos actuales de uso de los medios, se espera que se

integre directamente en los objetos de información y, por tanto, debería también formar parte de los motores de búsqueda en bibliotecas y archivos. Esta demanda a menudo conduce a la necesidad de integrar herramientas de terceros y fuentes de datos. CONTENTUS está desarrollando tecnologías y conceptos que afrontarán estos desafíos y, de forma significativa, simplificarán la producción, provisión y uso de colecciones de materiales digitales.

Con los dos sistemas de demostración basados en la web ya desarrollados, hemos mostrado que esta clase de agregación y presentación de materiales y metadatos es factible y –de forma más importante- también valiosa para los usuarios de bibliotecas y otros archivos. El acceso y descubrimiento del conocimiento por medio de búsquedas asistidas semánticamente sin duda crecerá en importancia y creemos que los resultados de CONTENTUS pueden ser importantes pilares para los sistemas bibliotecarios digitales de próxima generación.

Lecciones aprendidas

Podrían establecerse unos pocos principios orientativos que han demostrado ser útiles para lograr el objetivo del proyecto. Las lecciones que hemos aprendido hasta ahora:

· *El diseño modular es esencial*. Puesto que no todas las bibliotecas son iguales en términos de sus necesidades, algunas podrían no requerir técnicas de digitalización, y otras puede que ya ofrezcan interfaces de búsqueda y simplemente requieran tecnologías para generar y combinar metadatos para sus recursos. Por consiguiente, el diseño de las soluciones de CONTENTUS es intencionadamente modular. Para cada uno de los diferentes pasos del proceso (véase la introducción), existen soluciones independientes que las instituciones interesadas pueden usar de forma individual o conjunta.

· *Las normas abiertas y las interfaces son importantes*. Para facilitar la susodicha integración de tecnologías de CONTENTUS, nos centramos en normas abiertas, interfaces y formatos de datos. Para los elementos de la búsqueda semántica multimedia en CONTENTUS, por ejemplo, empleamos una arquitectura orientada al servicio (AOS). La interacción de los diferentes módulos a través de los *servicios web* tiene en cuenta la necesidad de una infraestructura bibliotecaria moderna y ofrece máxima flexibilidad para la integración de diferentes fuentes de datos ya sea de forma interna o bien proporcionada por un proveedor de servicio externo. Para integrar fuentes de información externas, los formatos típicos de colecciones de datos asociados (XML/RDF) facilitan utilizar tales metadatos.

· *Los URIs valen para enlazar semánticamente recursos, conceptos y fuentes de información*. Véase más arriba “Usando los URIs”.

· *Los usuarios prefieren interfaces sencillas, bien estructuradas pero potentes*. Esto es verdad sobre todo cuando se trata de nuevas funcionalidades como las que proporcionan las búsquedas semánticas. Las representaciones gráficas, e.g., de conceptos y relaciones tienen que ser intuitivas y fáciles de usar, incluso a través de dominios del conocimiento. Se demandan mucho –

particularmente por parte de usuarios profesionales- unas extensas opciones de configuración personal para la interfaz del usuario.

Trabajo y perspectiva futuros

Actualmente, el proyecto está manteniendo su ciclo de publicación anual y alcanzando así rápidamente el tercer sistema de demostración basado en la web, que será presentado al público profesional en la exposición de la International Broadcaster Conference (IBC) en septiembre de 2010. El nuevo demostrador comprende una interfaz de usuario reelaborada así como un motor de facetas semánticas extendido, y un mejor manejo del contenido multimedia. A finales de año el nuevo demostrador de BSMM de CONTENTUS también se presentará en el centro estable de demostraciones del programa de investigación de THESEUS en Berlín y en eventos escogidos de la comunidad bibliotecaria y archivística.

Los esfuerzos de desarrollo posterior se concentrarán en extender las capacidades semánticas integrando un *visor semántico de los materiales* para permitir una interacción mejor con los nombres de entidades reconocidos por el sistema. Otro gran desafío será la extensión de las características de personalización y comunidad, que formará un modo novedoso de *aprovechamiento cooperativo de la información*. Hará posible que los usuarios interactúen exhaustivamente con los recursos de información, ya sea para uso personal o en cooperación con colaboradores o grupos de usuarios. Por último, integraremos fuentes de información más valiosas desde la nube de datos abiertos vinculados.

Una aspiración de CONTENTUS es demostrar sus conceptos de integración de metadatos y búsqueda asistida semánticamente en el contexto de una colección histórica real. Para este propósito se han digitalizado grandes partes del archivo del *Musikinformationszentrum* (MIZ) de la antigua República Democrática Alemana (RDA). Los diversos materiales de esta apartada colección se integrarán en el demostrador final de CONTENTUS, que estará disponible a comienzos de 2012. Creemos que este contenido es muy adecuado para mostrar las ventajas de nuestro sistema en un dominio del conocimiento específico y que llevará a una nueva comprensión de la vida musical en la antigua RDA.

Referencias

Bossert, Klaus y Nicholas Flores-Herr y Jan Hannemann. *CONTENTUS: Technologien für digitale Bibliotheken der nächsten Generation*. Dialog mit Bibliotheken, Bd. 21, p. 14-20. ISSN 0936-1138. German National Library, 2009

Hannemann, Jan y Jürgen Kett. *Linked Data for Libraries*. En: Proceedings of World Library and Information Congress: 76th IFLA General Conference and Assembly (IFLA 2010), Gothenburg, Sweden

Heß, Andreas, 2006. *An Iterative Algorithm for Ontology Mapping Capable of Using Training Data.* En: Proceedings of the 3rd European Semantic Web Conference (ESWC 2006), Budva, Montenegro

Johnston, Eddie y Nicholas Kushmerick, 2008. *Web Service aggregation with string distance ensembles and active probe selection*. Information Fusion 9(4): 481-500 (2008)

Levenshtein, Vladimir I., 1965. *Binary codes capable of correcting deletions, insertions, and reversals*. En: Doklady Akademii Nauk SSSR. 163, Nr. 4, 1965, S. 845–848 (In Russian. Traducción inglesa en: Soviet Physics Doklady, 10(8) S. 707–710, 1966)

Maaß, Christian y Elica Savova, 2008. *Paper Prototyping in der Softwareentwicklung*. En: Das Wirtschaftsstudium, 11/2008 (In German)

Melnik, Sergey y Hector Garcia-Molina y Erhard Rahm, 2002. *Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching*. En: Proceedings of the 18th International Conference on Data Engineering (ICDE), San Jose CA, USA

Pilz, Anja y Gerhard Paaß, 2009. *Named Entity Resolution Using Automatically Extracted Semantic Information*. En: Proceedings of workshop Lernen, Wissen, Adaptivität (LWA 2009), Darmstadt, Germany

Russell, Robert C., 1918. United States Patent 1261167, application filed Oct. 25, 1917, patented Apr. 2, 1918.

Shvaiko, Pavel y Jérôme Euzenat y Fausto Giunchiglia y Heiner Stuckenschmidt y Natasha Noy y Arnon Rosenthal (Editors), 2009. *Ontology Matching (OM-2009), Papers from the ISWC Workshop*. October 2009.

Winkler, W. E., 1999. *The state of record linkage and current research problems*. Statistics of Income Division, Internal Revenue Service Publication R99/04, 1999.