DATOS VINCULADOS PARA BIBLIOTECAS

Jan Hannemann y Jürgen Kett
Biblioteca Nacional Alemana
Frankfurt, Alemania

(Traducción de:
Silvia Rubio Martín
Biblioteca Nacional de España)

Resumen:

La Web Semántica en general y la iniciativa Linking Open Data en particular animan a las instituciones a publicar, compartir y vincular entre sí su información. Esto tiene un enorme potencial para las bibliotecas que pueden complementar sus datos vinculándolos a otros que proceden de fuentes externas.

Este artículo habla del primer servicio abierto de vinculación de datos de la Biblioteca Nacional Alemana, centrándose principalmente en los desafíos que surgieron durante la puesta en marcha del servicio. Sin embargo, el artículo también contiene, extraída de esta experiencia, la visión de la Biblioteca Nacional Alemana sobre el futuro del intercambio de datos en el marco de las bibliotecas y el gran potencial que supondría para éstas la creación de datos globalmente vinculados entre sí.

1. Introducción:

Hoy en día, las bibliotecas están muy aisladas en materia de intercambio de datos, ya que desde siempre la información recogida por las bibliotecas era sólo para las bibliotecas. De hecho, el proceso de intercambio y utilización conjunta de datos con instituciones no bibliotecarias se encuentra todavía en sus inicios. Las colaboraciones existentes son fundamentalmente entre bibliotecas y los datos bibliotecarios no están integrados todavía como parte de la web. Esto se debe principalmente a la escasez de vinculaciones entre las bases de datos bibliotecarias y los datos procedentes de otros ámbitos, pero también es debido a que los procesos actuales de catalogación y los formatos se siguen enfocando instintivamente dentro de unos usos totalmente tradicionales.

La Web Semántica y la iniciativa Linking Open Data animan a las instituciones a publicar, compartir y vincular entre sí su información utilizando la web. La visibilidad de los datos puede aumentar enormemente mediante su vinculación con otras fuentes de información y esto es relevante tanto para las instituciones sin ánimo de lucro como para las instituciones comerciales. Además, llegar a formar parte de la red de datos vinculados, o “nube semántica”, supondría para las bibliotecas poder responder mejor a las necesidades de sus usuarios, tales como el acceso continuo a la información en un formato que sea comprensible para los expertos no bibliotecarios. Por otro lado, trabajar dentro de esta red de conocimiento creciente de la nube semántica permitiría a las bibliotecas encontrar ayuda para resolver algunas de las tareas complejas con las que se tienen que enfrentar a la hora de mantener y optimizar sus bases de datos locales.

Ejemplos importantes de estas tareas complejas son: la detección de duplicados, la desambiguación, la individualización, la gestión de la calidad de los datos y su enriquecimiento... Y, yendo aun más lejos, este trabajo en red también facilitaría el camino para la creación de nuevos servicios que necesitaran datos procedentes de más de una institución.

Por su parte, el conjunto de datos vinculados también se beneficiaría de los esfuerzos realizados por las bibliotecas y por el resto de instituciones culturales patrimoniales. Los datos procedentes de éstas normalmente tienden a ser de muy alta calidad debido a que son profesionales experimentados los encargados de recoger, revisar y mantener esta información. Tal es así que las bibliotecas podrían llegar a ser un eje central muy necesario para el crecimiento de la web semántica.

Las bibliotecas ya se han dado cuenta de este potencial y muchas instituciones están planeando publicar sus datos como Linked Data. Sin embargo, en la práctica, este proceso es un desafío. Además de los obstáculos propios de la organización, la parte técnica de la web semántica que permite publicar y utilizar los datos puede suponer un problema para algunos tipos de instituciones culturales patrimoniales, como las bibliotecas, debido al limitado presupuesto que suelen manejar para Tecnologías de la Información.

El propósito de este artículo es hablar de los datos vinculados desde la perspectiva de las bibliotecas y de otras instituciones culturales patrimoniales, así como proporcionar un informe concreto sobre la experiencia de la Biblioteca Nacional Alemana en el establecimiento de este tipo de servicio.

2. Visión: El grafo cultural global

El principal problema para los datos vinculados en web es asegurar su fiabilidad: ¿Son los datos correctos? ¿Existen procesos que garanticen la alta calidad de estos datos? ¿Quién es responsable de ellos? De igual importancia es la fiabilidad en el tiempo: ¿Un recurso es lo suficientemente estable como para ser citado o desaparecerá es algún momento? Estas cuestiones son de crucial importancia en el mundo de la investigación, en el que las citas son esenciales, y para los servicios de alto nivel que están basados en este tipo de datos.

Mientras que no es necesario asegurar el máximo grado de fiabilidad en todas las bases de datos para que éstas sean útiles, nosotros creemos que es fundamental proveer tanto de un núcleo estable como de un eje central de confianza al conjunto de datos vinculados en web, y pensamos también que las instituciones culturales patrimoniales se encuentran en una posición única para proporcionar parte de estos elementos esenciales ya que la conexión de sus bases de conocimiento locales podría conducir a un enorme grafo cultural mundial de información que fuera a la vez fiable y duradero.

Para explicar los diferentes grados de calidad de la información y de su fiabilidad, los datos del grafo cultural mundial deben organizarse sobre la base de un modelo de capas (ver Imagen 1). Cada capa está asociada a una política que se hace más estricta según nos vamos acercando al núcleo de fiabilidad de la base del conocimiento. Las entidades situadas en el núcleo son duraderas y por lo tanto suponen una cita fiable. Las descripciones de estas entidades deben ser versionadas para que cualquier cambio en sus

declaraciones o en su procedencia sea debidamente documentado. Se utilizan políticas transparentes para cada una de las capas, políticas que están basadas en una correcta documentación, en estándares asociados o en reglas de catalogación. Para garantizar su calidad y su persistencia, los datos del núcleo deben estar respaldados por una o más instituciones públicas de confianza.

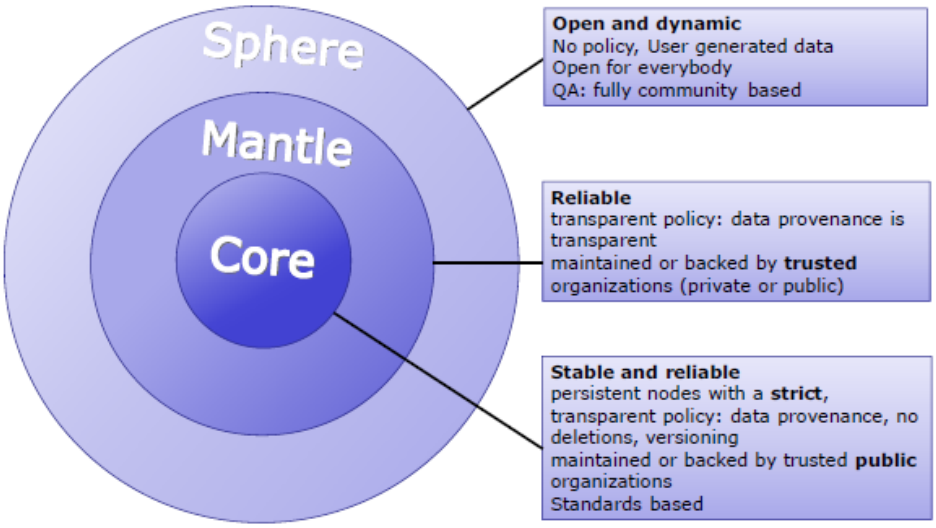

Figura 1: Modelo de capas de la fiabilidad y la durabilidad

Este modelo podría también incluir metadatos generados automáticamente, siempre y cuando el origen de los datos estuviera documentado. Es fundamental la posibilidad de que la información pueda moverse hacia el núcleo según vaya adaptándose a las reglas requeridas. De esta manera se consigue que el núcleo de información fiable, así como el valor de todo el conjunto de datos, vaya creciendo a lo largo del tiempo.

Las condiciones para hacer realidad este proyecto son prometedoras: las instituciones culturales patrimoniales ya utilizan estándares bien documentados y desarrollados de manera colaborativa, tales como MARC21 o RAK-WB, y también disponen ya de reglas para producir y mantener su información, aunque por supuesto necesiten adaptarse para lograr ser compatibles y formar parte así del intercambio general de datos en la web.

Dentro de la comunidad bibliotecaria alemana, la práctica del intercambio de información y el mantenimiento cooperativo de bases de datos centralizadas ha dado lugar a unos flujos de trabajo que conforman una base excelente para la creación de conjuntos de datos estables en la web. Un buen ejemplo de esto es el mantenimiento cooperativo de los registros de autoridad alemanes de personas, entidades y materias (PND, GKD y SWD, respectivamente). La buena definición de los procesos de eliminación, de actualización y de fusión de duplicados, junto con un esquema de identificadores ya establecido, garantizan un alto nivel de estabilidad. Todo esto nos llevó a elegir estos datos, precisamente, para establecer nuestro primer servicio de datos vinculados.

Publicar nuestras bases de conocimiento locales como datos vinculados es un paso necesario hacia la visión de futuro que acabamos de presentar. De ahora en adelante,

nuestra intención es poner de relieve los retos que conlleva este paso necesario y todas las experiencias que hemos llevado acabo en este terreno.

3. Los retos para establecer el servicio de vinculación de datos

A pesar de la opinión generalizada, el establecimiento de un servicio de datos vinculados no es algo trivial. Las instituciones culturales patrimoniales interesadas en formar parte de la "nube semántica", se ven obligadas a tratar con una serie de retos que se pueden agrupar en las siguientes categorías:

3.1 Retos técnicos

Para establecer un servicio de vinculación de datos, se requiere de una amplia infraestructura. Generalmente, esta infraestructura comprende un medio para el almacenamiento de datos (normalmente un triplestore o una base de datos), un servidor web, y un conmutador que interprete las peticiones de información entrantes, las transforme en preguntas lanzadas a los datos almacenados y sea capaz de devolver los resultados de esa búsqueda.

Debido a la relativa novedad del movimiento que promueve la vinculación de datos, las tecnologías disponibles están todavía poco desarrolladas y existe una información muy pobre sobre ellas. En particular, las instituciones que son nuevas en esto de la vinculación de datos no suelen tener del todo claro cuál es la opción tecnológica que mejor se adapta a sus necesidades.

3.2 Retos conceptuales

Otra cuestión esencial es la modelización de datos. Existen docenas de ontologías establecidas, más o menos amplias, entre las que elegir, cada una con sus ventajas y sus inconvenientes. Un aspecto importante a tener en cuenta a este respecto es la definición de sus propiedades individuales, que pueden permitir o no la modelización de los datos. Si no se encuentra ninguna ontología que pueda responder a nuestras necesidades será necesario mezclar varias y/o ampliarlas añadiendo propiedades personalizadas.

Hay información que es particularmente difícil de modelar, como las declaraciones de las declaraciones. Ejemplo de esto es la especificación de la procedencia de determinada declaración o de los procesos y reglas que se han aplicado a la hora de crear los datos. El último punto tiene especial importancia sobre todo para los datos producidos a partir de algoritmos automatizados, pero también es fundamental para documentar las reglas de catalogación manual y los estándares que se han utilizado. Existen varias aproximaciones para tratar este problema: las relaciones N-ary[1], las anotaciones OWL 2 axiom[2], la reificación[3], la ampliación personalizada de la ontología. Cada una de estas aproximaciones conlleva un conjunto de ventajas y de inconvenientes y no está establecido un protocolo de buenas prácticas en esta área, ni siquiera existe consenso dentro de la comunidad sobre qué solución se adaptaría mejor a las necesidades de las bibliotecas. En definitiva, es necesario llevar a cabo muchas más experiencias en el ámbito bibliotecario que utilicen todas estas tecnologías para la modelización de datos.

[1] http://www.w3.org/TR/swbp-n-aryRelations/
[2] http://www.w3.org/TR/owl2-syntax/#Axioms
[3] http://en.wikipedia.org/wiki/Reification_(computer_science)

Otro tema es la especificación de las URIs. Normalmente, las organizaciones no bibliotecarias que publican su información poseen una única base de datos sin ningún tipo de identificador público ni ningún flujo de intercambio de información. Por este motivo, introducir nuevas URIs para las entidades y sus descripciones no es un problema para ellas. Las bibliotecas, por su parte, ya utilizan gran cantidad de identificadores públicos para sus datos y para las entidades que describen y están acostumbradas a un masivo flujo de información. Nosotros creemos que lo mejor no es separar la vinculación de datos de los procedimientos tradicionales de intercambio de información, sino unir ambos mundos. Un esquema adecuado de identificadores debería funcionar para cualquier tipo de flujo de información e incluso los identificadores podrían no variar dependiendo del protocolo de intercambio que haya sido utilizado para la representación de los datos (por ejemplo, RDF versus MARC21).

3.3 Retos legales

En lo que respecta a los temas legales, son importantes sobre todo dos cuestiones: los derechos de publicación, y la licencia de los datos vinculados. Las instituciones culturales patrimoniales, como las bibliotecas, tienden a menudo a recopilar sus datos en colaboración con otras instituciones. En esos casos, debe estar claramente establecido qué datos pueden ser publicados como datos vinculados y cuáles no, por ejemplo, para hacerlos accesibles públicamente.

El asunto de la privacidad atañe tanto a los registros bibliográficos como a los registros de autoridad. Por ejemplo, si una biblioteca recopila información sobre autores, es posible que esos autores no quieran que sus datos personales (fecha de nacimiento, lugar de nacimiento o del domicilio, su relación con determinadas entidades, etc) se hagan públicos.

El tema de las licencias también puede suponer un problema. Los términos de uso de los datos vinculados deben decidirse en una etapa temprana del proyecto puesto que los controles legales llevan tiempo. Por otro lado, distinguir entre la utilización comercial y no comercial de los datos puede ayudar a agilizar los trámites ya que los no comerciales pueden ofrecerse de manera gratuita.

3.4 Retos generales

Una cuestión común que afecta a las tres categorías que se acaban de mencionar es la falta de informes que recojan las experiencias en el establecimiento de servicios de vinculación de datos (algo que este artículo intenta solucionar). Las instrucciones paso a paso son muy raras y las que existen suelen dejar muchas preguntas sin contestar. Además, nos encontramos con una gran variedad de protocolos de buenas prácticas que ofrecen una muy necesaria flexibilidad a las instituciones que quieren llegar a formar parte de la web de datos vinculados, pero que, por otro lado, no son capaces, o no tanto como los estándares estrictos, de dar unas directrices de trabajo.

4. El Servicio de Datos Vinculados de la DNB

Este apartado pretende describir nuestra experiencia en el establecimiento del primer servicio de datos vinculados. El proyecto de la DNB se ha mantenido intencionadamente en un ámbito reducido para evitar el desbordamiento de trabajo y el establecimiento de objetivos poco realistas. Por esta misma razón, decidimos al principio del proyecto focalizar nuestro trabajo en partes seleccionadas de nuestra base

de datos en vez de intentar conseguir una solución para la totalidad de nuestra información.

Para el desarrollo técnico y conceptual necesario, elegimos un método iterativo con ciclos de 1 ó 2 meses que dio lugar a un más elaborado modelo de trabajo continuo. Esto nos permitió mantener una comunicación constante con la comunidad y con los usuarios potenciales cuyas opiniones fuimos introduciendo en el diseño final del servicio.

4.1 Realización técnica

La arquitectura del sistema se muestra en la Imagen 2. Toda la información sobre nuestros registros se encuentra almacenada en una base de datos central que es constantemente actualizada y ampliada. Para hacer que parte de estos datos estén disponibles como datos vinculados, un componente llamado RdfExporter extrae los metadatos de nuestro sistema central de catalogación[4], los convierte a RDF y los almacena en un Jena TDB[5] RDF Store. En una segunda etapa, estos datos se enriquecen con referencias a fuentes externas. La transformación de los datos requeridos es llevada a cabo por un conjunto de módulos de conversión (uno por cada tipo de información, por ejemplo, hay módulos separados para encabezamientos de personas, de entidades y de materias) proporcionados por nuestro Servicio Central de Conversión.

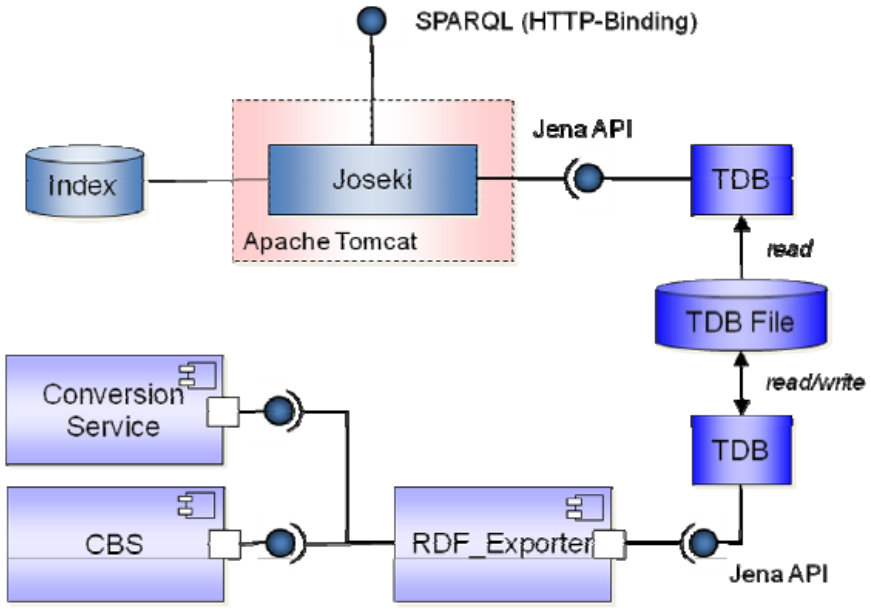

Imagen 2: Arquitectura actual del sistema

Elegimos en concreto estas tecnologías de almacenamiento y acceso a los datos porque las habíamos utilizado ya en un proyecto anterior. Sin embargo, hay que tener en cuenta que están todavía en desarrollo y, en consecuencia, la documentación sobre ellas carece de consistencia.

[4] OCLC CBS: http://www.oclc.org/cbs/default.htm
[5] http://openjena.org/TDB/

Aunque el sistema actual funciona lo más seguro es que tenga que ser reemplazado en un futuro. La escalabilidad, en particular, es un objetivo que necesitamos lograr. Por otro lado, la gran cantidad de datos manejados plantea otro desafío: la conversión completa de nuestras tres bases de datos tarda aproximadamente dos días en realizarse con un ordenador moderno, esto sin contar con la adición de los enlaces externos ni con la unión conjunta de todas las bases de datos enriquecidas. Además, el servidor Joseki tiene gran capacidad de memoria (muchos gigabytes, que aumentan según el tamaño de la base de datos) y no hemos encontrado todavía la forma de reducirla.

Otro problema es que el sistema actual requiere una iniciación manual de la transformación de los datos, problema que tendrá también que solventarse en el futuro. Lo ideal sería que la conversión de los datos y su enriquecimiento con enlaces externos se hiciera bajo demanda (por ejemplo, cuando se recibiera una petición) o que estuviera asociada a un mecanismo de actualización automática que se activara cada vez que la base de datos CBS cambiara en algún sentido.

4.2 Selección de datos

Una pregunta importante que debemos hacernos a la hora de crear un servicio de datos vinculados es qué datos van a ser publicados y con qué fuentes de información externas vamos a realizar los enlaces.

Nosotros decidimos centrarnos en el archivo de autoridades, más concretamente en 1.8 millones de autoridades de persona (del Archivo de Autoridades de Nombre PND), 160,000 autoridades de materia (del Archivo de Encabezamientos de Materia SWD) y 1.3 millones de autoridades de entidad (del Archivo de Autoridades de Entidad GKD). Esta selección se debió a la solicitud externa de dichos datos, a la decisión de movernos en un ámbito manejable, y a las propiedades de los datos en sí (ver el apartado 2):

- Los datos son utilizados ya por muchas organizaciones
- Los trabajos de mantenimiento de estos datos permiten su publicación y garantizan un alto grado de persistencia.
- Ya existe un esquema de identificadores establecido y aceptado para estos conjuntos de datos.
- Y en lo que respecta a los enlaces con fuentes externas, tuvimos la suerte de poder aprovechar los resultados de muchos proyectos de colaboración entre bibliotecas que se centraban en la alineación de datos. Podemos proporcionar enlaces con la Wikipedia alemana[6], y la BDpedia[7], con VIAF[8], con LCSH[9] y con RAMEAU[10].

Nuestro catálogo bibliográfico, sin embargo, es mucho más extenso, razón por la cual decidimos abordarlo en un proyecto posterior.

4.3 Selección de la Ontología

Seleccionar ontologías sólidas como base para la modelización de datos es algo que la comunidad de la web semántica recomienda encarecidamente ya que facilita el

[6] http://de.wikipedia.org/wiki/Wikipedia:Hauptseite
[7] http://wiki.dbpedia.org/About
[8] http://viaf.org/
[9] http://authorities.loc.gov/
[10] http://rameau.bnf.fr/

intercambio de información. Como consecuencia de esto, nosotros quisimos introducir este punto también en nuestro proyecto.

Sin embargo, nos dimos cuenta de que, en la práctica, las ontologías existentes no eran del todo adecuadas para modelar nuestros datos. Sus propiedades individuales no se ajustaban a nuestro tipo de datos, por lo que ninguna ontología en concreto se consideró aceptable. En lugar de eso, tuvimos que determinar de forma meticulosa qué partes de qué ontologías cubrirían de forma conjunta la mayor parte de nuestros datos y, para las partes restantes, nosotros mismos definimos nuestras propiedades, con la intención de registrar en un futuro la ontología resultante.

La modelización de los datos para la representación de personas y entidades utiliza muchas ontologías existentes como el grupo de elementos de las RDA, el vocabulario de FOAF y el Vocabulario de Relaciones, siendo las RDA la base de la modelización debido al hecho de que cubre muy bien los elementos fundamentales de los Requisitos Funcionales para los Registros Bibliográficos (FRBR) (por ejemplo, personas y entidades). Además utilizamos propiedades creadas por la Biblioteca Nacional Alemana (el vocabulario Gemeinsame Normdatei (GND)) para complementar estas ontologías. En lo que respecta a los encabezamientos de materia, la modelización de datos se basa en el uso del Sistema de Organización Simple del Conocimiento (Simple Knowledge Organization System, SKOS) y en los elementos del Dublin Core, los cuales son complementados por las propiedades especiales GND.

En nuestro servicio de datos vinculados existe documentación detallada sobre todas las ontologías que tuvimos en cuenta durante el proceso y sobre las razones que nos llevaron a la selección final de determinados elementos de ontologías concretas[11].

4.4 Ejemplos

Los siguientes ejemplos ilustran el trabajo realizado en el proyecto:

- El autor alemán *Bertolt Brecht* (http://d-nb.info/gnd/118514768) XML/RDF
La representación se encuentra en: http://d-nb.info/gnd/118514768/about

- La representación XML/RDF de una entidad "*IFLA / Section of Public Libraries <The Hague>*" (http://d-nb.info/gnd/10352988-3) se encuentra en:
http://d-nb.info/gnd/10352988-3/about

- El encabezamiento de material para "*Führungskraft*" (En español: "*Ejecutivoe*") se encuentra aquí: http://d-nb.info/gnd/4071497-4 , y la representación asociada en XML/RDF se encuentra aquí: http://d-nb.info/gnd/4071497-4/about

5. Experiencias

Nosotros habíamos estado involucrados en actividades de la web semántica antes de decidirnos a establecer nuestro propio servicio de datos vinculados por lo que pensábamos que estábamos bien preparados para los retos que el proyecto pudiera plantear. Nuestra experiencia práctica, aunque siempre positiva, también nos hacía ver las dificultades y nos mostraba algunas expectativas equivocadas. Por ejemplo,

[11] wiki.d-nb.de/display/LDS

aprendimos que había que coger con pinzas la opinión generalizada dentro de la comunidad de la web semántica de que establecer un servicio como el nuestro era algo sencillo. Nuestros descubrimientos fueron los siguientes:

- *Crear un servicio no es algo trivial*. Las iniciativas en la vinculación de datos son de muy reciente desarrollo. Como consecuencia de esto, las soluciones (herramientas) del software necesarias no están todavía maduras, lo que significa, entre otras cosas, que la documentación existente carece de consistencia. Dentro de un servicio, es necesario organizar muchos componentes del software para que trabajen conjuntamente, lo que requiere de una gran pericia (ejemplo, ver Imagen 2). Los datos probablemente tendrán que ser transformados a un formato adecuado (RDF) y esta parte del trabajo no solo necesita una modelización adecuada de los datos y su transformación (un esfuerzo potencial considerable si la modelización de datos debe ajustarse estrechamente a los datos originales), sino que también necesita la creación de programas de conversión o de filtros de exportación. Para la visualización final de los datos deberá utilizarse la codificación UTF-8, aunque las bibliotecas sigan un sistema de codificación internacional diferente.
- *La modelización de los datos puede ser compleja*. Cuando se publica información en la web, lo más conveniente es utilizar ontologías registradas y por lo tanto ya existentes. Desafortunadamente, estas ontologías no siempre se ajustan a la representación de los datos de las bibliotecas concretas (ver apartado 3). Es particularmente importante el hecho de que la definición de las propiedades individuales puede variar considerablemente. Hay fundamentalmente dos caminos para solucionar este problema: coger las ontologías publicadas tal cual están o definir nuevas propiedades que se ajusten a los datos. La primera solución es más fácil, pero puede tergiversar la información, mientras que la segunda es mucho más compleja, pero representa los datos correctamente. No hay una respuesta sencilla a la pregunta de cuál es el camino mejor a seguir. Nosotros elegimos el modelo de intentar representar nuestra información lo más fielmente posible porque pensamos que de la otra forma pondríamos en peligro la calidad de nuestros datos.
- *La mentalidad del intercambio abierto de datos no existe en todas partes*. Incluso antes de la existencia de los datos vinculados, las bibliotecas intercambiaban y alineaban sus bases de datos. Los resultados de aquellos proyectos podrían suponer la fuente de información principal para la conexión posterior de los datos vinculados. Desgraciadamente, no todas las instituciones involucradas comparten esa mentalidad de intercambio abierto, y la propiedad compartida hace difícil la publicación de dichos resultados. En lo que se refiere a esto, nosotros hemos tenido tanto experiencias positivas como negativas y recomendamos a las bibliotecas que se encuentren en una situación similar que hablen largo y tendido con todas las partes involucradas antes de considerar la utilización de los resultados de esas colaboraciones.
- *Las prácticas más recomendables son vistas como normas*. La vinculación abierta de datos se basa fundamentalmente en la búsqueda de los procedimientos más adecuados, no en normas. Sin embargo, este modo de actuar tan pragmático no es visto como importante por parte de la comunidad de los datos vinculados. Las desviaciones de los estándares prefijados suelen ser criticadas, lo que puede producir que las instituciones que son nuevas en esto de la web semántica duden de sus decisiones, incluso aunque éstas tengan sentido dentro del marco de la

organización en cuestión. Las bibliotecas no deben desistir por ello, todo lo contrario, deben ver como una motivación el hecho de poder contribuir con su propia experiencia y conocimiento al desarrollo de la comunidad. Las directrices y las prácticas recomendadas deben valorarse sólo y exclusivamente dentro del marco de las necesidades de cada institución, especialmente en esta etapa tan temprana de la existencia de la nube semántica. Por ejemplo, nosotros hemos sido criticados por no ofrecer un SPARQL-endpoint. Aunque esta tecnología supone una adición muy útil, no está del todo claro por qué razón esto tiende a ser una obligación, sobre todo, porque hoy en día existen otras alternativas ya establecidas de búsqueda (por ejemplo, la Búsqueda y Recuperación vía URI[12] (SRU) y OpenSearch[13]) y de sincronización (por ejemplo, ORI[14]) de datos bibliotecarios.

- *Los usuarios siguen siendo en gran medida anónimos*. Para mejorar nuestro servicio, nos preguntamos a nosotros mismos, al inicio del proyecto, dos cuestiones fundamentales: ¿Quién utiliza nuestros datos? Y: ¿Para qué los utiliza? Aunque invitamos a los usuarios a opinar sobre nuestro servicio y a contarnos sus expectativas y experiencias, la mayoría de ellos eligieron no hacerlo. Una consecuencia del concepto de acceso anónimo del proyecto Linked Data es que sólo podemos mantener relación con aquellos usuarios que deciden contactar con nosotros. Esto también significa que no podemos prestar ayuda específica a otros usuarios: un desarrollo extraño que observamos fue que al parecer alguien realizó un programa para rastrear nuestros datos vinculados usando un enfoque (muy) naïve (intentando todas las combinaciones numéricas posibles que podrían constituir IDN para generar URIs); este programa funcionará durante varios meses antes de completarse. Hubiera sido mucho más fácil redirigir a este usuario a la versión descargable que proporcionamos de nuestros datos, pero como está utilizando una conexión dial-up, no podemos identificarle ni ponernos en contacto con él/ella.
- *Los datos modelados correctamente son muy útiles*. Una vez que se ha completado la modelización de los datos y estos son accesibles, el sistema puede ser utilizado por otros. Un colega de la Universidad Técnica de Braunschweig ha demostrado que los datos modelados correctamente pueden resultar muy útiles en varias aplicaciones: un día, él importó nuestros datos a una base de datos, les añadió un interface web y así creó un acceso mediante búsqueda a nuestra información.

En definitiva, podemos dibujar un resumen positivo de nuestra experiencia a pesar de los retos que acabamos de señalar. Una vez superados los obstáculos, es relativamente fácil utilizar los datos y ampliar el servicio.

6. Trabajo futuro

Al establecer un servicio de datos vinculados, hemos logrado un paso hacia la visión de un grafo cultural global –pero quedan todavía muchas cosas por hacer, y los objetivos a largo plazo solo son alcanzables una vez que otras instituciones se hayan unido al proyecto. Las organizaciones culturales patrimoniales en su conjunto deben adecuar su

[12] http://www.loc.gov/standards/sru/
[13] http://www.opensearch.org/Home
[14] http://www.openarchives.org/OAI/openarchivesprotocol.html

infraestructura técnica, sus procesos de negocio, sus normas, sus políticas y las licencias de sus metadatos para lograr adaptarse a los requerimientos de la web. Este cambio fundamental debe llevarse a cabo paso a paso y de manera manejable. Ante todo, hay que hablar de cuáles son los desafíos para las instituciones culturales patrimoniales y una vez que la nube semántica crezca, se podrán desarrollar estrategias globales.

En este apartado, mostramos cuáles son nuestros planes de futuro y cuáles son los objetivos que solo podremos conseguir en cooperación con otras instituciones.

6.1 Objetivos a corto plazo

- *Revisar la infraestructura*: nuestro objetivo inmediato para el futuro es mejorar el actual servicio tanto en términos de infraestructura como de visualización de la información y añadir nuevos grupos de datos. La siguiente versión del servicio incluirá ya mecanismos de actualización automática (ver más arriba) de nuestros datos y potencialmente también de los enlaces con otras fuentes externas. Así mismo se podrá encontrar una arquitectura con mayor escalabilidad y proporcionaremos un interface SRU para lograr un acceso más amplio a la información.
- *Nuevos grupos de datos*: añadiremos clasificaciones y mejoraremos la modelización de los datos. Los nuevos grupos de datos contendrán, entre otras cosas, un subgrupo seleccionado de la bibliografía nacional alemana y de la traducción al alemán de la Clasificación Decimal de Dewey. En lo que se refiere a los datos del título está pendiente una decisión relacionada con el esquema de URIs y el establecimiento de un trabajo de registro que es compatible con los planes actuales de optimización del intercambio y de reutilización de los datos bibliográficos en Alemania. Hoy por hoy, estamos discutiendo este asunto con los centros bibliotecarios alemanes y esperamos llegar pronto a un acuerdo. Otra cuestión importante será la modelización de la ontología para los datos del título. De nuevo, tenemos que encontrar un compromiso entre la reutilización de los vocabularios existentes, tales como Dublin Core[15] o la ontología bibliográfica[16], la inclusión de estándares como RDA y la utilización de nuestras propias estructuras de datos.
- *Servicios para el usuario final*: una razón para poner a disposición del usuario nuestros datos a través de la web es que sea posible utilizarlos para construir servicios mucho más elaborados. Las instituciones culturales patrimoniales no deben ser las únicas que ofrezcan servicios de información; el potencial creativo de la comunidad web puede lograr mucho más que cada una de estas instituciones por separado. Nuestro objetivo es proporcionar al fin un servicio que ilustre el potencial total de las bases de datos bibliotecarias vinculadas entre sí. Este servicio tiene dos objetivos: motivar a otras instituciones culturales patrimoniales (especialmente en Alemania) para que contribuyan al grafo cultural global e intentar atraer la atención de la comunidad web a estas bases de datos tan valiosas. El servicio decidirá una URI que identifique a cada tipo de entidad (por ejemplo, libro, obra, persona, entidad corporativa, materia, etc…) y devolverá una página de información con enlaces a todas las bases de datos registradas que contengan (directamente) recursos relacionados. Este servicio

[15] http://dublincore.org/documents/dc-rdf/
[16] http://bibliontology.com/

seguirá los convenios de la vinculación de datos y ofrecerá una representación RDF de las entidades.

6.2 Objetivos a largo plazo

Es obvio que los objetivos a largo plazo solo serán posibles cuando las organizaciones culturales patrimoniales, la industria del software, las instituciones de investigación y las autoridades públicas actúen conjuntamente. La Biblioteca Nacional Alemana usará toda su influencia y su competencia técnica para intentar lograr este trabajo en colaboración. Los asuntos tratados a continuación son aquellos que nosotros consideramos como más urgentes:

- *Modelo de Licencia Compartido*: las bibliotecas están indizando cada vez más y más en colaboración. La creación de la Deutsche Digitale Bibliothek[17] hará que las bibliotecas alemanas, los museos y los archivos crezcan y se acerquen todavía más las unas de las otras. En un mundo en el que se tiende a generar metadatos de manera conjunta, los modelos de licencias más adecuados solo pueden ser aquellos que se establecen y se desarrollan cooperativamente. Si esto no fuera posible, la reutilización de la información en la nube semántica estaría tremendamente limitada, hasta el punto de que necesitaríamos manejar licencias de datos no a nivel de los registros sino a nivel de los metadatos aislados (por ejemplo, en los casos en los que una institución enriquece los datos de otra institución). En la práctica, este mundo resultaría demasiado complicado y costoso.
- *La adopción de un flujo de trabajo y unas políticas*: las organizaciones culturales patrimoniales deben generar sus propios flujos de trabajo para hacer sus bases de datos menos redundantes y más consistentes a largo plazo (especialmente los datos del título en los niveles de Obra y de Manifestación). Es necesaria la existencia de identificadores públicos que se puedan citar, que sean acordados por todos y que todos podamos utilizar en un futuro. La institución alemana Gemeinsame Normdatei (GND) podría servir como modelo para este tipo de iniciativas.
- *Más diseños ensayados de patrones para la modelización de la ontología*: las instituciones culturales patrimoniales no pueden solucionar por sí solas los desafíos más importantes ligados a la modelización de los datos (por ejemplo, las declaraciones sobre las declaraciones) y a la realización técnica, pero sí que pueden aportar casos prácticos de uso. Los proveedores de software y las instituciones de investigación son las adecuadas para encontrar soluciones a estos problemas.
- *Soluciones técnicas avanzadas*: el estado del arte actual en lo que se refiere a las soluciones técnicas no se adecua a los entornos productivos de las organizaciones culturales patrimoniales. Se necesitan más herramientas y estructuras que faciliten la introducción de los datos en la nube semántica y la utilización de las fuentes disponibles. La tecnología de los datos vinculados debe poder integrarse sin problemas en los entornos y en los flujos de trabajo ya existentes. Para ello, los nuevos desarrollos deben hacerse en una plataforma independiente, deben seguir estándares abiertos y deben utilizar una arquitectura en capas altamente modularizada con APIs abiertos. Al mismo tiempo, los sistemas bibliotecarios existentes tienen que evolucionar en paralelo ya que en el

[17] http://www.deutsche-digitale-bibliothek.de/ (En alemán)

futuro deberán ser capaces de manejar información granular referente a la procedencia y a las versiones.

7. Resumen

La vinculación abierta de datos ofrece un gran potencial a las instituciones culturales patrimoniales como las bibliotecas. No es solo que la información pueda difundirse mucho más gracias a la utilización de esta tecnología, sino que además, gracias a los conjuntos de datos enlazados entre sí, aumenta el valor de la nube de información resultante. Basándose en ella, las bibliotecas y otras organizaciones pueden crear servicios nuevos mucho más completos. Con la creación de este servicio de datos vinculados, hemos logrado dar un paso hacia el futuro, pero el objetivo a largo plazo del *grafo cultural global* sólo se conseguirá cuando este proyecto que aboga por la publicación de datos y su intercambio reciba un apoyo mucho más generalizado.