2009

República Checa

Proyectos Manuscriptorium, Kramerius y Web Archiv

Pablo Rodríguez Gordo

GOBIERNO DE ESPAÑA
MINISTERIO DE CULTURA

República Checa

Proyectos Manuscriptorium, Kramerius y Web Archiv

Biblioteca Nacional de la República Checa
Praga, 11-15 de mayo de 2009

1. Presentación

El plan de trabajo consistió en estudiar tres proyectos de la Biblioteca Nacional de la República Checa (en adelante BNRC), todos relacionados con las nuevas tecnologías. Estos tres proyectos son:

- Manuscriptorium (www.manuscriptorium.com): biblioteca digital de manuscritos e impresos antiguos;
- Kramerius (http://kramerius.nkp.cz): biblioteca digital de prensa histórica y monografías modernas;
- Web Archiv (www.webarchiv.cz): archivo de la web checa.

2. Objetivos e historia de Manuscriptorium y Kramerius

Actualmente, Manuscriptorium y Kramerius son las dos bibliotecas digitales que aglutinan los esfuerzos de digitalización del Patrimonio Bibliográfico de la República Checa.

2.1 Historia hasta el año 2000

Las primeras digitalizaciones de fondo antiguo de la República Checa se llevaron a cabo en el marco de pequeños proyectos que tenían el apoyo del Programa Memory of the World de la UNESCO[1], durante los años 1992-1993.

En los años posteriores se hizo necesario una estrategia permanente para la digitalización de manuscritos e impresos antiguos, así que:

- Se diseñó una estrategia de digitalización para la BNRC dividida en dos proyectos:
  - Memoria, dirigido a manuscritos e impresos antiguos (hasta el siglo XIX)
  - Kramerius, dirigido a publicaciones periódicas e impresos del siglo XIX y XX deteriorados por la acidificación.
- Para ambos planes se creó un modelo de metadatos para la descripción de las obras digitalizadas, una DTD (Document Type Definition) de SGML (Standard Generalized Markup Language) llamada DOBM (Description of Old Books, Manuscripts and other documents). DOBM será el origen de los esquemas de metadatos propietarios utilizados hoy en día en Manuscriptorium y en Kramerius. Ambos esquemas evolucionaron por separado posteriormente, debido a que tenían por objetivo documentos con características diferentes.

[1] http://portal.unesco.org/ci/en/ev.phpURL_ID=1538&URL_DO=DO_TOPIC&URL_SECTION=201.html

- En los años 1995-96, la BNRC construyó el Depósito Central de Hostivar como centro de microfilmación y digitalización (http://digit.nkp.cz).

En la segunda mitad de la década de los 90, varias bibliotecas, archivos y museos checos comenzaron a integrarse en Memoria y Kramerius y a utilizar DOBM para sus trabajos de descripción.

En 1999 esta DTD fue recomendada por la UNESCO como esquema de metadatos para los programas de digitalización que debían realizarse bajo los auspicios de "Memory of the World".

En el año 2000 el Ministerio de Cultura checo convirtió a Memoria y Kramerius en los programas de digitalización nacionales de la República Checa.

2.2 Evolución desde el año 2000

Durante los primeros años de la nueva década, la BNRC actualizó DOBM a XML (eXtended Markup Language), el nuevo estándar emergente en lenguajes de marcado para Internet, más fácil de manejar que SGML. El trabajo se llevó a cabo en el marco de MASTER (Manuscript Access through Standard for Electronic Records) un proyecto internacional liderado por las Universidades de Montfort y Oxford. El fruto de MASTER fue un esquema de descripción de manuscritos y libro antiguo basado en TEI P4. TEI son las iniciales de Text Encoding Initiative,

La migración de ambos sistemas a MASTER se completó en 2003, y al año siguiente se lanzaron sendas bibliotecas digitales: Manuscriptorium (heredera del programa Memoria) y Kramerius.

En 2006 ambas bibliotecas fueron incluidas en el proyecto TEL (The European Library).

En 2007 nació el proyecto ENRICH, que pretende crear una biblioteca digital virtual internacional de manuscritos mediante la agregación de datos de diversos socios extranjeros a Manuscriptorium.

3. MANUSCRIPTORIUM

Manuscriptorium es una biblioteca digital que utiliza una interfaz de gestión online y MASTER como esquema de descripción de recursos. Gracias al proyecto ENRICH (http://enrich.manuscriptorium.com) está en proceso de convertirse en una biblioteca digital internacional, que utilice una interfaz de gestión mejorada y TEI P5 como esquema de metadatos.

En 2009 las siguientes instituciones participaban en Manuscriptorium, bajo la coordinación de la BNRC:

- 33 bibliotecas y museos de la República Checa. Se incluyen bibliotecas especializadas y universitarias entre ellas.
- 11 bibliotecas extranjeras, todas de países del este de Europa, entre las que destacan las Bibliotecas Nacionales de Polonia y Turquía, y las bibliotecas de las universidades de Budapest (Hungría) y Wroclaw (Polonia).

El socio técnico es AiP Beroun. Sus tareas dentro de Manuscriptorium son:

- Mantenimiento de los servidores
- Preparación de las mejoras del proyecto, lo que comprende:
    - Implementación de los formatos de metadatos utilizados: MASTER, TEI P5.

        - Diseño de software para la agregación de los datos de los diferentes socios.
    - Digitalización de fondos para el proyecto

La BNRC se encarga de la coordinación de los contenidos, es decir, las relaciones con los participantes y los contactos con nuevos socios potenciales. También asesoran a las bibliotecas que lo necesiten en materia de catalogación de fondo antiguo.

3.1 La agregación de datos

Normalmente es AiP Beroun quien realiza la digitalización y creación de metadatos de los nuevos documentos que se incluyen en Manuscriptorium procedentes de bibliotecas checas participantes. Para ello utilizan una herramienta ad hoc llamada M-Tool, que permite escribir los archivos XML con los metadatos directamente en el formato de Manuscriptorium.

Un caso diferente es la agregación de datos y metadatos procedentes de los socios extranjeros de Manuscriptorium. En este caso, la creación de metadatos se puede hacer de estas dos formas.

a. Cada biblioteca crea sus propios metadatos con M-Tool y los carga en la base de datos central a través de la web de Manuscriptorium.
b. Los metadatos de las instituciones extranjeras son recolectados mediante OAI-PMH y después convertidos automáticamente al formato MASTER gracias a unos programas llamados “conectores”, creados por AiP.

En ninguno de los dos casos los metadatos son sometidos a revisión por parte de la BNRC. Los participantes tienen a su disposición un entorno de pruebas en la web, una cadena editorial en la que comprobar que los datos han sido cargados y, en su caso, convertidos correctamente. Si todo esta bien, marcan sus datos como válidos, indicando al administrador del sistema en la BNCR que puede a su vez validarlo para que sea recuperable para los usuarios finales de Manuscriptorium.

El método de trabajo en el proyecto ENRICH es similar, salvo que a los dos tipos de participación mencionados en a) y b) hay que sumar el caso de las instituciones que ya utilicen TEI P5. En este caso no es necesario ningún “conector”, porque la biblioteca digital resultante de ENRICH y la biblioteca en cuestión comparten el mismo formato de metadatos.

Las imágenes de las obras aportadas por los socios participantes deben ser accesibles vía HTTP, en todo caso.

3.2 Los conectores

Estos programas ejecutan en batch (procesamiento por lotes, fuera de línea) un proceso de transformación XSL (eXtensible Stylesheet Language Transformation) que convierte los archivos del formato local a MASTER y próximamente a TEI P5, en el marco de ENRICH.

El principal problema que deben resolver los conectores es la creación del “mapa estructural” de los datos, en caso de que no exista en los metadatos suministrados por el socio participante. Este mapa estructural es un conjunto de etiquetas XML que contienen la numeración de las páginas de la obra y el vínculo a la imagen correspondiente en la base de datos del socio. Recordamos que en los manuscritos y libros antiguos la numeración no es tan evidente como en el libro moderno, por ej. “folio 23r” equivaldría a la página 45, es decir, el recto del folio 23.

AiP estudia los metadatos de cada participante, expresándolos a un esquema XML hecho con Altova XML Spy. A partir de ahí diseña una transformación XML que establece las correspondencias (mapea) entre los elementos de ambos esquemas y que convierte los registros recolectados del formato local a MASTER. Actualmente están trabajando en el diseño de los conectores que permitirán la conversión a TEI P5 de los participantes en ENRICH.

Vamos a centrarnos en dos ejemplos, la Biblioteca de la Universidad de Heidelberg (Alemania) y la Biblioteca Nacional de España, por ser el socio español del proyecto. Ambos colaboran en Manuscriptorium desde antes del inicio del proyecto ENRICH, por tanto sus metadatos son convertidos a MASTER y no a TEI P5.

En el caso de Heidelberg lo que se recolecta a través de OAI-PMH son unos registros METS que contienen metadatos descriptivos en MODS y metadatos administrativos en un formato local llamado UBHD.

Estos ficheros METS tienen una sección de archivos (etiqueta <fileSec>) que agrupa los archivos con las imágenes por su calidad, en cinco grupos. Luego en el mapa estructural (etiqueta <structMap>) presenta una organización física y lógica. En el mapa físico, se crea un elemento por cada página del original, que enlaza con cinco ficheros con la imagen de la página en las distintas calidades. Además hay un mapa lógico también, que utiliza un para cada capítulo, vinculándolo a las imágenes. Una última sección <structLink> crea los vínculos entre el mapa lógico y el físico. Para los ejemplos véase el Folleto 5.2 del proyecto ENRICH (http://enrich.manuscriptorium.com/index.php?q=node/22).

Los responsables del proyecto, en colaboración con personal de la Universidad de Heidelberg, prepararon una transformación XML (XSLT) que mapeaban los elementos de estos ficheros METS a MASTER. Para comprobar las diferencias entre los ficheros METS importados y los registros de ENRICH, ofrecemos más adelante una visión del esquema de metadatos TEI P5.

Por su parte, la Biblioteca Nacional ofrece metadatos descriptivos en MARC 21 XML, pero sus registros no tienen mapa estructural. No hay recolección OAI-PMH, los datos se intercambian por FTP o correo electrónico. Para crear los metadatos estructurales la BNE ha desarrollado un proceso que lee los directorios de las imágenes y los lista en una hoja de cálculo Excel. El nombre de los ficheros incluye el número de página del documento, (por ej., 1096521_Vitr_000006-006_195r.jpg) por lo que la transformación XML del conector puede crear fácilmente las etiquetas MASTER con el mapa estructural:

```
<page>
<pgFoliation>195r</pgFoliation>
<pgImage id="ID1096521"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096521_Vitr_000006-006_195r.jpg"
quality="Normal"/>
</page>
<page>
<pgFoliation>195v</pgFoliation>
<pgImage id="ID1096522"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096522_Vitr_000006-006_195v.jpg"
quality="Normal"/>
</page>
<page>
<pgFoliation>196r</pgFoliation>
<pgImage id="ID1096523"
href="http://www2.bne.es:81/Enrich/ENRICH/1096519/1096524_Vitr_000006-006_196v.jpg"
quality="Normal"/>
</page>
```

Para conocer mejor el trabajo de los conectores, pasamos a describir someramente las etiquetas de TEI P5.

3.3 La codificación de las obras: TEI P5

MASTER (Manuscript Access through Standard for Electronic Records) fue un proyecto internacional liderado por las Universidades de Montfort (Leicester) y Oxford, ambas en el Reino Unido. Creó una DTD de XML con una serie de etiquetas para la descripción de manuscritos y libros antiguos. El proyecto ENRICH utilizará TEI P5 en lugar de MASTER.

Estas son las principales etiquetas de TEI P5 para la descripción de manuscritos:

<msDesc> (manuscript description)
<msIdentifier> identificador del manuscrito,
<country>
<región>
<settlement> localidad
<institution> institución (usualmente, una biblioteca)
<idno> número de identificación, normalmente la signatura del manuscrito
<msName> nombre alternativo, por ej. “Codex Gigas”
<head> (heading) encabezamiento, consiste una breve descripción como la realizada en los inventarios de manuscritos. Consta normalmente de autor, título y signatura o referencia del segundo folio. Por ej.: Arch. B.3.2.: Evangelium Matthei cum glossa.
<msContents> se describe en contenido intelectual del manuscrito, dentro de uno o varios elementos <msItem>
<msItem>
<author>: el autor del manuscrito
<respStmt> : otra mención de responsabilidad
<title>
<incipit>: el íncipit, es decir, las primeras palabras del texto
<explicit>: el explicit, esto es, las últimas palabras del texto
<rubric>: rúbricas del manuscrito, palabra o palabras destacadas que marcan una división en el seno del texto
<decoNote>: nota sobre la decoración
<filiation>: información sobre la filiación, o sea, las relaciones con otros manuscritos conservados del mismo texto
<textLang>: lenguas y escrituras del texto
<physDesc>: aglutina la descripción física del manuscrito
<objectDesc>: agrupa las etiquetas que caracterizan el soporte físico y la organización física del manuscrito dentro de aquél
<supportDesc>: descripción del soporte
<extent>: longitud del texto
<collation>: colación, o sea la ordenación física de los bifolios
<foliation>: foliación o sistema para que el lector localice los folios, páginas, columnas o líneas
<condition>: estado de conservación
<layoutDesc>: distribución del texto sobre las hojas
<handDesc>: información sobre las diferentes “manos” o clases de escritura utilizadas
<typeDesc>: detalles sobre los tipos de imprenta utilizados en un impreso antiguo
<decoDesc>: decoración del menuscrito
<bindingDesc>: información sobre la encuadernación
<sealDesc>: comentarios sobre los sellos
<history>: historia del manuscrito

<origin>: datos sobre su creación
<provenance>: otros sucesos de su historia
<acquisition>: información sobre cambios de propietario
<additional>
<adminInfo>: información administrativa
<recordHist>: histórico de sucesos relacionados con el propio registro TEI
<availability>: disponibilidad del texto, es decir, restricciones por aspectos de propiedad intelectual, estado de conservación…
<custodialHist>: diferentes posesiones por las que ha pasado el manuscrito

Es un esquema pensado para manuscritos pero que permite describir también incunables e impresos antiguos. Contiene también metadatos administrativos y estructurales, a la manera de METS.

Los registros de la base de datos de ENRICH tendrán la siguiente estructura general:

```
<TEI>
        <teiHeader>
                Aquí se introducirán los metadatos que describen el manuscrito
        </teiHeader>
        <facsimile>
```

Esta etiqueta se utiliza para hacer referencia a imágenes digitales que reproducen el original del manuscrito, es decir, la fuente primaria del documento.

Aquí se incluirán metadatos de la imagen. Más abajo se proporciona un ejemplo:

```
</facsimile>
<text>
```

Opcionalmente, dentro de esta etiqueta se incluirá una transcripción del manuscrito, obtenida por OCR cuando sea posible.

```
</text>
</TEI>
```

Ejemplo de etiqueta <facsimile>

```
<facsimile xml:base="http://www.handrit.org/AM/fol/">
<surface xml:id="LSB-1r" ulx="0" uly="0" lrx="200" lry="300">
<graphic mimeType="jpeg" xml:id="AM02-5000-1r" url="AM02-5000-1r.jpg"/>
<graphic mimeType="jpeg" url="AM02-5000-1rthumb.jpg" width="1in" decls="#thumb"/>
 </surface>
 <surface start="#LSB-1v" ulx="0" uly="0" lrx="200" lry="300">
<graphic mimeType="jpeg" xml:id="AM02-5000-1v" url="AM02-5000-1v.jgp"/>
<graphic mimeType="jpeg" url="AM02-5000-1vthumb.jpg"
decls="http://www.enrich.org/imageDescs#thumb"/>
</surface>
</facsimile>
```

En este ejemplo, la etiqueta <facsimile> agrupa todas las imágenes del manuscrito correspondiente. Mediante el atributo xml:base nos muestra la URL raíz de la que cuelgan los archivos con las imágenes.

Cada imagen está marcada mediante una etiqueta <surface>, que contiene 4 atributos indicando las coordenadas que delimitan la superficie rectangular de esa imagen. Esta delimitación tan precisa permite la adición de etiquetas <surface> que suministren metadatos

sobre determinadas áreas de una página ; por ejemplo, en el caso de una letra inicial decorada, contendrá una o más etiquetas <surface>, una por cada página, que definen una superficie del manuscrito.

Bajo cada etiqueta <surface> se sitúan los elementos <graphic>, que contienen el nombre del archivo con la imagen, en el atributo url. En este ejemplo vemos que para la página primera del manuscrito ("LSB" es una abreviatura de su nombre) hay dos imágenes JPG, cada una en su etiqueta <graphic>, la primera en tamaño real y la segunda en modo miniatura (thumbnail). El atributo decls de la segunda lleva a unos metadatos adicionales sobre las imágenes de miniatura, metadatos válidos para todas las imágenes de este tipo referenciadas bajo la etiqueta <facsimile>.

El siguiente elemento <surface> atañe a la segunda página, la primera versión, y contiene igualmente dos imágenes, en sendas etiquetas <graphic>. El atributo start y las coordenadas delimitan una zona que será transcrita bajo la etiqueta principal <text> del registro. La transcripción de estas dos páginas se codificaría así:

```
<text>
        <div facs="#LSB-1r">
                <pb n="1r"/>
<p>Maðr hét Ludovícus, sonr Bernharðs greifa, erkallaðr var loðinbjörn [... resto de la transcripción]
<pb n="1v" xml:id="LSB-1v"/>

[Transcripción de la segunda página]

 </p>
</div>
```

El elemento contiene la transcripción de la hoja completa, al recto y al verso. Cada cara de la hoja se codifica dentro de una etiqueta <pb> (pagebreak, salto de página). La etiqueta <pb> de la segunda hoja contiene un atributo xml:id que hace referencia al elemento <surface> de la misma hoja bajo <facsimile>.

Esta es, vista muy someramente, la codificación que utilizará el proyecto ENRICH y la nueva versión de Manuscriptorium. Gracias a los conectores sus transformaciones automatizadas, será posible convertir grandes cantidades de datos en poco tiempo al formato común de ENRICH.

4. KRAMERIUS

A partir de las inundaciones de la República Checa del año 2002, se decidió potenciar el programa nacional de digitalización para monografías y publicaciones seriadas modernas, que ya había comenzado su andadura en los años 90 con la creación de las DTD iniciales para estos documentos, todavía en el marco de DOBM.

Este programa de digitalización desembocó en la creación de una serie de bibliotecas digitales independientes, pertenecientes a importantes bibliotecas checas. Todas ellas tienen en común la utilización de un gestor de contenidos llamado Kramerius. Václav Matěj Kramerius (1753-1808) fue un escritor y periodista checo.

En adelante se hablará de la biblioteca digital Kramerius de la BNRC. Mientras que Manuscriptorium almacena manuscritos y libro antiguo anterior al año 1801, Kramerius se ocupa de las publicaciones seriadas publicadas desde su nacimiento (1719) hasta la actualidad, y de las monografías modernas.

4.1 Proceso de digitalización

Esta biblioteca digital tiene su sede en el Repositorio Digital Central de la BNRC, situado en Hostivar, a las afueras de Praga.

Una vez se ha decidido digitalizar una determinada obra, el personal del Repositorio lo revisa para descubrir páginas estropeadas o perdidas, errores en la paginación o ilustraciones interiores valiosas. El libro queda así marcado con esta información adjunta.

Las digitalizaciones más difíciles o los libros más valiosos se escanean en el Repositorio Digital Central. En los demás casos -la mayoría- la digitalización está externalizada.

4.2 Metadatos

Kramerius puede importar metadatos en el formato MARC checo, basado en UNIMARC.

Dentro de Kramerius se utilizan los siguientes formatos de metadatos:

- Formato interno, expresado en un esquema XML en las siguientes direcciones:
  - http://www.digit.nkp.cz/DigitizedPeriodicals/DTD/2.10/DocumentationPeriodical/Periodical.html [esquema para publicaciones periódicas]
  - http://www.digit.nkp.cz/Monographs/DTD/2.10/Monograph.xsd [esquema para monografías]
- MARC 21
- Dublin Core

La herramienta utilizada para crear los metadatos se llama Sirius. Este software, creado por la empresa checa Elsyst (www.ee.cz) sirve para la digitalización y adición de metadatos.

Además se puede generar un documento METS instantáneamente a partir de cada título, número, capítulo o página.

4.3 La interfaz de búsqueda

La búsqueda de una cadena de texto en la sección de publicaciones periódicas no recupera títulos, sino números de un título. Esto provoca listas inusualmente largas de resultados, aunque si se combina la búsqueda con la fecha de publicación del número que nos interesa el conjunto de resultados gana mucha precisión.

A su vez, la búsqueda de monografías recupera tanto títulos como capítulos o artículos de monografías, marcados como “Internal component part”.

Ambas búsquedas (monografías y publicaciones periódicas) permiten recuperar también todas las páginas en las que la cadena de texto está presente. Es decir, permite una búsqueda a texto completo, en la cual los resultados se ordenan por relevancia.

Además del mencionado motor de búsqueda “Apache Lucene” existía la posibilidad de realizar una búsqueda alternativa a través del buscador semántico llamado “Convera”. Esta última funcionalidad actualmente no está en funcionamiento, debido a que no hay presupuesto para su mantenimiento.

4.4 El gestor de contenidos Kramerius

Kramerius fue desarrollado por la empresa checa QBIZM. Es de código abierto.

Permite en esencia la importación y exportación de los datos y metadatos y su publicación en la página web. Interactúa con el sistema gestor de la base de datos para hacer todo esto.

La publicación web se puede configurar, estableciendo la fecha a partir de la cual las obras pasan a formar parte del dominio público y, por lo tanto, pueden consultarse en línea a texto completo.

Kramerius es un proveedor de datos OAI-PMH, pero no es un proveedor de servicios. Los registros recolectables mediante OAI-PMH se corresponden con páginas, no con números ni títulos.

En el futuro se quiere unificar todas las instancias de Kramerius existentes actualmente, para que sean accesibles a través de una única página.

4.5 La base de datos de Kramerius

Dentro de la base de datos se puede identificar sin ambigüedad y recuperar cada título (revista o monografía), volumen, número, página y en ocasiones capítulo mediante un URI (Uniform Resource Identifier) basado en la tecnología Handle.

Los ficheros resultantes son almacenados en una base de datos cuyo sistema de gestión es PostgreSQL, un sistema de tipo objeto-relacional.

En resumen, la base de datos almacena lo siguiente:

- Metadatos:
  - Descriptivos: formato interno Kramerius, formato MARC21, Dublin Core;
  - Administrativos: en un formato interno Kramerius. Véase el esquema en http://digit.nkp.cz/Kramerius/AdminMetaData/1.0/AdminMetaData.xsd
- Datos:
  - Imágenes JPEG, para conservación
  - Imágenes DjVu, para visualización
  - Ficheros de texto resultado del OCR

4.6 La preparación de las obras

El proceso técnico de las obras sigue las siguientes etapas:

a. Digitalización en formato JPG

b. Los ficheros resultantes se someten a un proceso de reconocimiento de caracteres (OCR). El resultado son ficheros de texto, que son indizados por el motor de búsqueda Apache Lucene. Este índice es el soporte de la utilidad de búsqueda a texto completo

c. Se preparan los metadatos utilizando Sirius

d. Validación de los archivos XML contra la DTD o esquema correspondiente utilizando XMetaL Author

e. Conversión de las imágenes al formato DjVu y modificación final de los metadatos en XML

5. WEB ARCHIV

Web Archive es un proyecto que pretende almacenar y preservar páginas web checas cuyo contenido sea interesante para futuros investigadores de la realidad checa. Las páginas web de las que se ocupa WebArchiv son:

- Las publicadas por autores o editores checos en cualquier parte del mundo;
- Las publicadas en la República Checa;

- Aquellas cuyo tema esté relacionado con la República Checa.

Asimismo, realizan campañas extraordinarias de recolección de sitios web cuando suceden acontecimientos de especial interés para Chequia, como por ejemplo: elecciones, presidencia de turno de la Unión Europea en 2009, etc.

5.1 Estrategias de rastreo de la web

El rastreo y recolección de páginas web realizado por Web Archiv sigue una estrategia doble:

5.1.1. Estrategia cuantitativa

Mediante un robot (crawler) rastrean y "recolectan" páginas web del domino .cz. El rastreo comienza utilizando la lista de webs registradas por la autoridad de Internet checa. El robot se llama Heritrix (http://crawler.archive.org) y es un software de código abierto creado por el proyecto Internet Archive.

El robot no sólo recolecta estos dominios, sino que recolecta también las webs que están enlazadas en estas 550.000 webs, es decir, profundiza un nivel en las webs iniciales. Después pasa a recolectar las webs enlazadas por las webs del segundo nivel, y así sucesivamente hasta 15 niveles.

5.1.2. Estrategia cualitativa

Actualmente sólo se pueden consultar la totalidad de las webs archivadas desde ordenadores situados dentro de la BNRC. Esto se debe a que la legislación de propiedad intelectual checa obliga a obtener el permiso del autor para la publicación en Internet de una sede web por un tercero. En la práctica, es muy difícil localizar a los autores de centenares de miles de webs y solicitarles autorización. Por eso se utiliza además del rastreo cuantitativo uno cualitativo.

Este rastreo consiste en la búsqueda y selección de sedes web de especial calidad e interés, para las que sí se realiza el esfuerzo de localizar a sus autores y solicitarles autorización. A lo largo de la vida del proyecto se han seleccionado unas 3000 sedes de este tipo. De ellas, se ha conseguido localizar al autor y obtener su plácet para unas 1000 sedes.

Estas 1000 webs son completamente accesibles a través de www.webarchiv.cz, y a través del catálogo de la BNRC. A finales de 2009 los responsables del proyecto prevén llegar a 1300.

Un pequeño equipo de bibliotecarios se ocupa de la primera selección de sedes, de entre las recolectadas por el robot de búsqueda. Dos bibliotecarios a tiempo completo y tres a tiempo parcial se ocupan de estas tareas.

Estas webs son clasificadas mediante el sistema Conspectus, que tiene 24 categorías. Los bibliotecarios se reparten estas categorías entre ellos formando grupos de 8, de forma que uno o dos se ocupan de un grupo de 8. Una lista en inglés de las 24 categorías de Conspectus está en: http://www.nkp.cz/pages/page.php3?page=fond_Mdt_tabulky1.htm

Por ejemplo: Educación, Informática, Agricultura...

5.2 Base de datos interna

El almacenamiento de las sedes web se realiza en una gran base de datos en formato MySQL. Los metadatos disponibles de cada sede web dependen de si ha sido recolectada con la estrategia cuantitativa o con la cualitativa.

a. Webs de la estrategia cuantitativa: únicamente la URL

b. Webs de la estrategia cualitativa: URL, título y autor, siempre que los dos últimos datos puedan averiguarse. Siempre que existen, se utilizan las etiquetas "meta" del sitio web para discernir estos datos.

En www.webarchiv.cz la interfaz de búsqueda sólo dispone de un campo: URL, para ambos tipos de webs.

5.3 Recolecciones

Las recolecciones se hacen anualmente. Cuando Heritrix encuentra una web que ya está en la base de datos, la recolecta tal como está y le añade la fecha de la recolección. Al recuperar esa sede web durante una búsqueda, el portal mostrará una lista de todas las versiones disponibles, correspondientes al estado de la web en una determinada fecha.

En el caso de las webs "cualitativas", se hacen 6 recolecciones al año.

5.4 Almacenamiento

Otro programa, WayBack Machine, guarda las webs recolectadas en archivos con formato ARC. Estos ficheros tienen un tamaño limitado a 100 MB.

WayBack Machine crea un índice de las páginas web de las que hay información en cada fichero. Una sede web que tenga varios años de antigüedad y que, por tanto, haya sido recolectada varias veces, tendrá su información dispersa entre varios ficheros ARC. Incluso es probable que los datos de una web en un momento dado estén también repartidos entre varios ficheros ARC.

Wayback Machine también permite visualizar las webs, tanto al usuario final como al administrador de Web Archiv. Es decir, detrás del sencillo formulario de búsqueda por URL de www.webarchiv.cz está Wayback Machine como intermediario entre el usuario y la base de datos de ficheros ARC.

5.5 Futuras mejoras previstas

El principal desarrollo que desean implementar los responsables es un buscador a texto completo en las webs, como Google o cualquier otro motor de búsqueda de Internet. Los principales proyectos de archivado de web del mundo (Government of Canada Web Archive, Pandora, etc.) ya lo han implementado, pero en Web Archiv hay un problema por solucionar. Cuando entramos en una página web a través de una página de resultados de búsqueda, se abre Wayback para mostrarnos la web. Hasta aquí todo bien. Pero el problema es que, si accedemos a través de esos resultados, no se reescriben los vínculos de la web archivada en el formato "Wayback"; por lo cual, si hacemos clic en cualquier vínculo, saldremos de Wayback y veremos la página destino del enlace tal como es en la actualidad, si es que continúa existiendo. Este problema no existe entrando a una web archivada desde la pantalla de resultado de una búsqueda por URL, sólo cuando se entra a partir de los resultados arrojados por una búsqueda a texto completo.

5.6 Relaciones con otros proyectos

El responsable de Web Archiv es miembro del Internet Archive Consortium, organización que agrupa a 39 miembros, la mayoría de ellos bibliotecas nacionales.

También cooperan en otros proyectos del I.P.C.: grupos de trabajo sobre recolección de webs, preservación, etc. [Más información en http://netpreserve.org].

6. CONCLUSIÓN

Estos tres proyectos constituyen el conjunto de estrategias del Estado checo para la conservación de su patrimonio bibliográfico y documental. Enrich es el proyecto más avanzado a día de hoy y en nuestra humilde opinión se puede considerar que su apuesta por TEI es audaz vista en comparación con proyectos similares españoles, que no han explotado esta opción aún. Por otro lado será interesante ver como Kramerius camina hacia la unificación de todas sus instancias bajo una única interfaz, utilizando Fedora como plataforma. Esta unión aumentará su potencial como proyecto que pueda algún día seguir los pasos de Manuscriptorium y trascender las fronteras checas. Web Archiv por su parte es un proyecto que no es puntero en relación con otros que tratan el problema de la preservación de páginas web, pero su actividad es meritoria teniendo en cuenta los medios de que disponen. Estudiar más a fondo este proyecto, el PADICAT (Patrimonio Digital de Cataluña) y otros punteros como Pandora o Canada Government Web Archive, podría desembocar en la creación de un archivo similar para toda España, con las previsibles favorables consecuencias para la preservación de la web española.

Madrid, 22 de agosto de 2009