World Library and Information Congress: 70th IFLA General Conference and Council

22-27 August 2004
Buenos Aires, Argentina

Programme: http://www.ifla.org/IV/ifla70/prog04.htm

Code Number: 009-S
Meeting: 89. Cataloguing
Simultaneous Interpretation: -

El Proyecto Paradigma y su búsqueda de soluciones para metadatos y servicios a usuarios

Carol Van Nuys, Ketil Albertsen, Linda Pedersen et Asborg Stenstad

El Proyecto Paradigma, Biblioteca Nacional de Noruega
carol.vannuys@nb.no, ketil.albertsen@nb.no, linda.pedersen@nb.no, asborg.stenstad@nb.no

Resumen:

El Proyecto Paradigma, de la Biblioteca Nacional de Noruega, trabaja para asegurar un Depósito Legal satisfactorio de todos los tipos de documentos digitales, incluidos los millones de documentos publicados en los dominios noruegos de Internet. Es de esperar que Noruega sea capaz de preservar su patrimonio digital cultural para el futuro, dando a los investigadores acceso al archivo de Internet, por ejemplo mediante metadatos y búsquedas a texto completo. Esta ponencia ofrece una breve descripción del Proyecto en sí, antes de iniciar la discusión de los problemas con que se encuentran en su búsqueda de estándares de metadatos de búsqueda, preservación a largo plazo, etc.

Se presentará el uso que el Proyecto hace de las entidades a nivel de obra, expresión, manifestación y ejemplar - establecidos por los Requisitos Funcionales de los Registros Bibliográficos de IFLA (en adelante FRBR) - en el diseño del archivo, así como las propuestas para futuros servicios: Servicio de autenticación y verificación y Servicio de asignación de identificadores – ambos disponibles a través de Internet.

1.- Introducción

1.1.- El Archivo Web en otros países

Los documentos digitales están desapareciendo continuamente; un estudio[1] muestra que sólo

[1] Mannerheim, Johan. The WWW and our digital heritage [online]. – URL: http://ifla.org/IV/ifla66/papers/158-157e.htm (se accedió el 15 de abril de 2004)

el 20% de los documentos encontrados en la red se mantienen - sin cambios - pasado un año. Por lo tanto, las posibilidades para las nuevas generaciones de lectores de estudiar en el futuro documentos digitales actuales también desaparecen. La preservación del patrimonio cultural digital supone un desafío de creciente importancia; y la Biblioteca Nacional de Noruega[2] es una de las muchas instituciones que trabaja para encontrar respuestas a los problemas legales, técnicos y bibliográficos que la acompañan.

Uno de los aspectos del trabajo de preservación digital es la recopilación y archivo de documentos publicados en los dominios nacionales de Internet. En este sentido, cada país establece sus propias estrategias de recopilación: Dinamarca [1] y Australia [2] han establecido una metodología selectiva, mientras Suecia [3], Islandia y Finlandia han decidido recopilar todo su espacio web nacional. Noruega es una de las Bibliotecas Nacionales europeas que recoge y archiva documentos digitales procedentes de sus dominios nacionales de Internet basándose en la legislación existente sobre depósito legal[3].

1.2.- La Ley de Depósito Legal

El propósito de la Ley de Depósito Legal [5] es:

[…] asegurar que documentos que contengan información de acceso general sean depositados en colecciones nacionales, de tal manera que esos registros de la vida cultural y social noruega sean preservados y puedan estar disponibles como fuente de información con propósitos de investigación y documentación." (§ 1)

La actual Ley de Depósito Legal, considerada sumamente avanzada cuando se aprobó en 1989, cubre todos los documentos noruegos *de acceso general* almacenados en *cualquier* soporte: por ejemplo, papel, microformas, fotografías, documentos combinados, archivos sonoros, películas, vídeos, documentos digitales y programas de televisión. Los documentos publicados en el extranjero por editores noruegos y aquellos especialmente adaptados para el público noruego, también están cubiertos por la ley.

Por supuesto, la WWW aún no había aparecido en Internet en 1989. Los documentos digitales - en su mayor parte en forma de bases de datos - constituían un número muy pequeño comparado con los millones de documentos actualmente publicados en Internet, pero ya presentaban problemas tecnológicos en su gestión. Actualmente, el Depósito de Preservación a Largo Plazo de la Biblioteca Nacional de Noruega, tiene capacidad para almacenar 100TBytes de información: el equivalente a un número realmente muy elevado de documentos digitales.

2.- El Proyecto Paradigma

La Biblioteca Nacional de Noruega inició el Proyecto Paradigma[4] en agosto de 2001. Su

[2] Para más información sobre la Biblioteca Nacional Noruega, consulte URL: http://www.kb.nl/gabriel/libraries/pages_generated/no_en.html (se accedió el 15 de abril de 2004)

[3] Halgrímsson, Torsteinn (2003, februar 28). Web Archiving in Europe [discussion]. – NWA [online].- E-mail address: nwa@nb.no

[4] Para más información sobre el Proyecto Paradigma consulte URL: http://www.nb.no/paradigma/eng_index.html (se accedió el 15 de abril de 2004)

objetivo es asegurar el Depósito Legal satisfactorio de los documentos digitales noruegos, lo cual incluye el desarrollo de tecnología, metodología y rutinas de trabajo para la selección, colección, descripción e identificación de *todos* los tipos de documentos digitales – incluyendo aquellos documentos de acceso general a través de Internet. Además, el Proyecto pretende proporcionar a los usuarios acceso al archivo de Internet en conformidad con la legislación en vigor.

Las actividades del Proyecto se basan en buena medida en anteriores trabajos de la Biblioteca Nacional en algunas áreas de importancia, con cuatro personas trabajando en ello a tiempo completo. Aproximadamente la tercera parte del personal de la Biblioteca está implicado en actividades de algún tipo relacionadas con el Proyecto, que está programado hasta el 31 diciembre de 2004.

Los siguientes epígrafes describirán brevemente el trabajo en curso del proyecto para seleccionar, recopilar y dar acceso al material digital recogido de Internet por depósito legal, así como la naturaleza y el tamaño del dominio noruego de Internet.

2.1. Estrategias de Colección y Selección

2.1.1. Colección

Basándose en la Ley de Depósito Legal y las recomendaciones del Proyecto Paradigma, la Biblioteca Nacional de Noruega ha decidido iniciar la recopilación de *todos los documentos digitales de acceso general* procedentes del espacio web noruego (".no"). En el futuro, los documentos recuperados desde dominios como ".com", ".org" y ".net" serán también recogidos.

Existen varias razones para adoptar esta postura de recopilación general: primero, no es posible predecir de antemano que documentos tendrán valor en futuras investigaciones y series documentales; en segundo lugar, el almacenamiento digital cada día resulta más barato; en tercer lugar, la recopilación sin filtros ahorra recursos – la recopilación manual consume tiempo en la selección; y finalmente, un usuario del archivo de Internet puede encontrar documentos mediante las consultas a texto libre, lo que permite la revisión de todos los documentos, incluyendo aquellos que no reúnen las condiciones para la catalogación manual. Además, los criterios de selección para cualquier uso, al igual que las descripciones bibliográficas exhaustivas, pueden ser abordados y modificados en cualquier momento. Esto, por supuesto, sería imposible si los documentos son excluidos en el momento de la recopilación.

La Sección de Depósito Legal ha recopilado una selección de documentos web de manera semimanual desde 2001 usando el software HTTrack[5], y estos documentos están catalogados en el catálogo de Biblioteca Nacional (BIBSYS[6]). Esta tarea continuará hasta que las actividades de recopilación general establecidas por el Proyecto Paradigma y los procedimientos relacionados estén completamente establecidos. La misma sección también se

[5] Para más información sobre el software HTTrack consulte URL: http://www.httrack.com/ se accedió el 15 de abril de 2004)

[6] Para más información sobre BIBSYS consulte URL: http://www.bibsys.no/english.html (se accedió el 15 de abril de 2004)

ocupa de la recopilación de documentos basados en acontecimientos, por ejemplo, los sitios web de los partidos políticos antes, durante y después de las elecciones. Otras secciones también se ocupan de actividades relacionadas con el depósito legal de documentos digitales, y el Archivo de Imagen y Sonido de la Biblioteca Nacional de Noruega trabaja para encontrar soluciones sobre el depósito legal de los programas de radio y televisión "de origen digital" , en cooperación con la Norwegian Broadcasting Corporation.

Un tema particularmente estimulante es el depósito de la "*Internet invisible*" (deep web), por ejemplo, los periódicos digitales, emisiones digitales, documentos de webcam, medios interactivos y materiales electrónicos de todo tipo almacenados en bases de datos. El Proyecto Paradigma ha iniciado la recopilación diaria de aproximadamente 65 periódicos digitales, y prevé llevar a cabo en un futuro próximo la descarga de diversas bases de datos completas de prensa, lo que permitirá complementar la "visión global" diaria. Estamos analizando los problemas derivados de la "Internet invisible" dentro del marco del International Internet Preservation Consortium[7], pero numerosos aspectos administrativos, legales y técnicos aún no han sido solventados.

En resumen, la Biblioteca Nacional de Noruega espera recibir documentos digitales a través de diversos canales: documentos recogidos automáticamente de Internet, actualizaciones de bases de datos entregadas por lotes, subscripciones de publicaciones periódicas y listas de distribución recibidas por correo electrónico, grupos de noticias y documentos entregados por medios físicos como CD-ROMs.

2.1.1.- Selección

Existen numerosos documentos valiosos que pueden ser encontrados en Internet, y actualmente estamos trabajando en definir *criterios de selección* para esos documentos que consideramos "merecen" algún tipo de descripción bibliográfica manual. Estos criterios de selección están basados en la legislación de depósito legal, así como en la política de obtención de documentos de la Biblioteca Nacional, tal y como aparece formulada en nuestros principios y planes estratégicos. Los criterios de selección para documentos digitales están siendo integrados con los que se refieren a documentos más tradicionales en el Manual de Selección de la Biblioteca Nacional.

El Proyecto Paradigma pretende implementar una estructura de sistemas que permita un proceso de *selección* en tres fases, de tal modo que los bibliotecarios reciban ayuda técnica para buscar aquellos documentos que deben ser catalogados en un cierto nivel. La primera fase busca y recoge documentos de Internet en noruego y sami. La segunda fase ofrece a los bibliotecarios la oportunidad de producir *automáticamente* listas sistematizadas basadas en consultas específicas. Estas listas están basadas en el uso de vectores que contienen metadatos, extraídos automáticamente a partir de los documentos recogidos. En la tercera fase, los bibliotecarios eligen documentos específicos de las listas sistematizadas para su registro manual en un cierto nivel, usando los criterios de selección anteriormente mencionados. En un futuro, podremos incluso ser capaces de supervisar los recursos integrables que hayan sido catalogados manualmente, lo que ayudará a los bibliotecarios a identificar y modificar estos registros bibliográficos, por ejemplo, en determinados intervalos de tiempo, cuando los cambios en el texto excedan un determinado porcentaje, etc.

[7] Para más información sobre las actividad relacionada con la "Internet invisible" consulte URL: http://www.nla.gov.au/ntwkpubs/gw/66/html/p15a01.html (se accedió el 15 de abril de 2004)

2.2. El dominio noruego de Internet

El tamaño exacto del dominio noruego de Internet aún no se conoce. La primera ronda de recopilación llevada a cabo por el Proyecto Paradigma entorno a diciembre de 2002 y enero de 2003 ofreció un resultado aproximado de 3.1. millones de URLs (es decir, de ficheros), de los cuales aproximadamente el 53% son imágenes (.jpg, .gif, .png). El software de recopilación NEDLIB[8] empezó con cerca de 1000 URLs iniciales, y la recopilación se limitó a: protocolo HTTP, dominios nacionales noruegos (".no"), URLs sin parámetros. La segunda ronda de recopilación se desarrolló en Agosto de 2003 y ofreció un resultado de aproximadamente 4.1 millones de URLs. La tercera fase de recopilación se está desarrollando actualmente y aún no están disponibles los resultados estadísticos.

Asumiendo que los resultados obtenidos en similares rondas de recopilación desarrolladas en Suecia y Finlandia sean aplicables en el caso de Noruega, esperamos encontrar entre un 45% y un 55% de sitios web fuera de los dominios ".no". Resulta obvio por tanto que el manejo y evaluación de cada documento individualmente no es posible; la mayor parte de ellos deberán ser procesados automáticamente.

2.3. La estrategia de acceso

2.3.1. ¿*Quién* consultará nuestro Archivo y *para qué*?

Cuando intentamos encontrar soluciones en cuanto a metadatos para describir la riqueza y variedad del material digital almacenado en nuestro archivo es importante cuestionarse: ¿Quién usará este material y con qué propósito?. Es muy difícil imaginar las consultas que realizarán los investigadores dentro de 10, 20 o 50 años pero sí podemos imaginar *grupos de usuarios* y *tipos* de consultas.

Un grupo de usuarios estará constituido por los usuarios interesados en estudiar Internet y los materiales digitales como *medios*, es decir, porque el material al que nos referimos ha sido recuperado de Internet y porque muestra las características propias de este medio. Algunos usuarios de este grupo pueden necesitar estudiar el uso del idioma en la red y las relaciones entre diferentes idiomas; los investigadores de medios pueden querer estudiar las relaciones entre los materiales impresos y digitales o entre las tendencias en las novedades tecnológicas y el contenido; los usuarios que analizan el diseño de páginas web pueden estar interesados en el uso de los anuncios, la composición, etc.; los investigadores del área de la ciencia informática querrán estudiar los diferentes protocolos de comunicaciones, el uso de formatos a lo largo del tiempo e incluso datos referentes a virus; los sociólogos pueden estar interesados en cómo la información disponible en Internet ha influido en la sociedad y viceversa. Por supuesto, somos conscientes de que muchos usuarios tendrán intereses que afecten a varias áreas al mismo tiempo.

Otro grupo de usuarios puede estar compuesto por quienes necesiten utilizar los documentos digitales como *fuentes documentales* – tal y como se emplean las fuentes tradicionales actualmente. Este grupo estaría compuesto por investigadores de todas las áreas temáticas, por lo tanto estará interesado en cubrir en detalle sus expectativas sobre los documentos digitales. ¿Están los documentos relevantes sólo en formato digital?. ¿Tiene importancia el contenido

[8] Para más información sobre el software de recopilación NEDLIB consulte URL: http://www.csc.fi/sovellus/nedlib/ver11/documentation11.doc (se accedió el 15 de abril de 2004)

dinámico, las animaciones, las pantallas de visualización interactivas, el sonido y vídeo integrado, etc.?. ¿Necesitan los investigadores acceder al material mediante consultas a texto libre o correlacionar gran cantidad de información de diferentes fuentes?.

2.3.2.- La legislación actual

Ofrecer acceso a los usuarios al depósito legal del archivo de Internet es un tema complejo, y la Biblioteca Nacional debe encontrar soluciones satisfactorias a pesar de las muchas, y en ocasiones contradictorias, normas recogidas en la Ley de Depósito Legal, la Ley de Copyright y la Ley de Datos Personales.

Estamos intentando encontrar respuesta a cuestiones como: ¿Que usuarios pueden tener acceso a diferentes tipos de materiales digitales?. ¿Pueden los usuarios acceder a estas colecciones desde ordenadores externos a la Biblioteca Nacional?.

2.3.3. Herramientas de Acceso

Los requerimientos de los usuarios, tal y como se señala anteriormente, son importantes para nosotros, ya que estamos desarrollando herramientas de acceso para la consulta de nuestro archivo de Internet. Por supuesto, tenemos que tener en cuenta que los bibliotecarios sólo catalogarán una minoría de los documentos disponibles.

En una planificación más técnica, el Proyecto Paradigma, espera ofrecer acceso a los usuarios al archivo de Internet a través del Nordic Web Archive´s[9] (NWA) Access Tool (ver figura 1). Actualmente, son opciones estándar la consulta a texto libre con operadores booleanos, la consulta de ciertas URLs, y la presentación del histórico del documento a través de una línea temporal. Esperamos que esta herramienta nos permita más posibilidades en el futuro: el uso de combinaciones de operadores booleanos para combinar diferentes listas de ocurrencias, búsquedas paralelas entre documentos catalogados en catálogos bibliográficos externos, búsquedas de metadatos automáticamente extraídos, navegación programada avanzada y disponibilidad de parámetros de búsqueda preprogramados, opciones que nos permitan almacenar las listas de ocurrencias en un "proyecto de biblioteca", accesos a búsquedas de grupos de documentos sisrtematizados de acuerdo a diferentes criterios (editor, etc.), máximo de una ocurrencia para los documentos duplicados, agrupación lógica en una sola ocurrencia de documentos consistentes en varias páginas web separadas, etc.

Planeamos adaptar el interfaz del NWA´s Access Tool para que se adecue a las funcionalidades de los diversos usuarios, y el uso del modelo FRBR de IFLA jugará un papel importante en cómo ofrecer acceso al material archivado en el futuro.

3. La búsqueda de soluciones respecto a los metadatos

El Proyecto Paradigma se halla a medio camino de encontrar formatos y soluciones respecto a los metadatos apropiados. La definición de metadatos para *búsqueda* ha sido una de nuestras principales actividades durante el pasado año, junto con nuestra investigación de soluciones satisfactorias para la extracción automática de metadatos técnicos. En el siguiente apartado intentaremos ofrecer una breve idea de *por qué* y *cómo* planeamos describir la mayor parte de

[9] Para más información sobre el Proyecto Nordic Web Archive consulte URL: http://nwa.nb.no (se accedió el 15 de abril de 2004)

los documentos digitales de nuestro archivo de Internet.

3.1. ¿Por qué deberíamos catalogar los recursos de Internet?

En la introducción de su libro Cataloging Internet Resources [3] Nancy Olsen ofrece tres razones básicas sobre por qué los recursos de Internet deben ser catalogados

1. Hay gran cantidad de información valiosa en Internet
2. Estos recursos deben ser organizados para resultar accesibles
3. El método más eficaz para acceder a estos recursos pasa por emplear las actuales técnicas y procedimientos bibliográficos y por crear registros para su recuperación a través de los catálogos en línea existentes.

Estamos de acuerdo con los tres puntos de Nancy Olsen, pero al mismo tiempo estimamos que *bastante menos del 1%* del material obtenido del dominio noruego de Internet será objeto de registro bibliográfico a algún nivel. Esto es debido, por supuesto, a la tremenda magnitud de documentos en el archivo. Tratamos de consolarnos con el siguiente pensamiento: aunque un porcentaje mucho más elevado de materiales más tradicionales es objeto de registro bibliográfico, diferentes materiales son de hecho tratados de diferentes maneras: al material efímero se le da un nivel de registro muy simple mientras que a los libros y publicaciones periódicas se les aplica niveles mucho más exhaustivos de catalogación.

Como contraste, el 100% de los documentos de Internet estará totalmente indizado mediante el software de indización FAST[10] tras su recopilación. Este software permitirá al personal de la Biblioteca y a los investigadores buscar en el archivo de Internet tanto a texto libre como por otros índices. Una mínima parte de documentos de Internet catalogados manualmente estará disponible por un lado a texto completo en el archivo y por otro como registros bibliográficos en el catálogo de la Biblioteca - esperemos que enlazados de manera lo más amigablemente posible para el usuario.

Además de la indización de todos los documentos y de la catalogación manual de algunos de ellos, recopilaremos los metadatos integrados existentes así como los documentos de Internet que describen. La Biblioteca Nacional planea un futuro servicio que permitirá a los editores generar y entregar metadatos con sus documentos en el momento del depósito.

3.2. ¿Qué son los metadatos?

La investigación de soluciones respecto a los metadatos nos ha llevado, por supuesto, a la búsqueda de definiciones adecuadas. El término “metadatos” ha sido definido y redefinido en la literatura. “Datos de datos” parece ser la definición más recurrente, y este término abarca un amplio espectro de tipos de información[11]. Hemos descubierto que los esquemas de metadatos son tan abundantes como diversos, pero todos ellos tienen una cosa en común: nos ayudan a *describir* y *encontrar* muchos documentos valiosos en nuestra colección, incluyendo aquellos que no fueron seleccionados para un nivel más exhaustivo de catalogación.

---

[10] Para más información sobre FAST Searc & Transfer (FAST) ASA, consulte URL: http://www.fast.no (se accedió el 15 de abril de 2004)

[11] Uno de los estudios sobre metadatos analizados es DESIRE: A review of metadata: a survey of current resource description formats. (1997). Consulte URL: http://www.ukoln.ac.uk/metadata/desire/overview/rev_toc.htm (se accedió el 15 de abril de 2004)

3.3. ¿Qué es un documento de Internet?

3.3.1. Definición de Documento de Internet desde un punto de vista técnico

Cuando un documento de Internet es seleccionado para ser captado y consecuentemente archivado, la semántica del término "documento" puede ser algo ambigua: ¿Qué componentes deben ser recogidos y archivados como partes integrantes del documento?, ¿Qué componentes deben ser objeto de una evaluación individual?. Consideramos que cualquier componente que afecte al "diseño" (incluyendo *sonido* y otros elementos "*no gráficos*") de una página web debe ser incluido incondicionalmente si esta web es seleccionada, y eso significa incluir imágenes de fondo, el contenido de los marcos, las imágenes de los botones, etc.

Los documentos referenciados mediante enlaces, aunque relacionados, son diferentes de los documentos que les hacen referencia. A un nivel semántico superior, a menudo trataremos un grupo completo de documentos enlazados entre sí como un único documento muy extenso. Si tratamos estos documentos relacionados entre sí de manera completamente independiente corremos el riesgo, por ejemplo, de recoger algunos capítulos de un informe y rechazar otros (debido a que en muchos casos los informes contienen prolongadas citas, sumarios, etc. en idiomas diferentes al noruego).

Así que, para responder a la pregunta de "qué compone un documento de Internet" debemos decir que un documento de Internet consiste en muchas partes o archivos relacionados, es decir, texto, imagen, sonido, animaciones, etc. y que estas partes están en la mayoría de los casos conectadas a través de enlaces y en algunos casos contenidas en conjuntos de marcos.

3.3.2. Definición de Documento de Internet desde un punto de vista bibliográfico

Es obvio que nunca podemos depender de un ordenador para que nos diga dónde empieza y dónde acaba un documento de Internet, incluso aunque lo hayamos programado para seguir ciertas instrucciones con ese objetivo in mente. Afortunadamente los bibliotecarios son muy acertados decidiendo qué partes de un documento de Internet conforman una unidad lógica. Por lo tanto, desde un punto de vista bibliográfico, podemos definir documento de Internet como una unidad de información que puede ser descrita bibliográficamente. Esta definición *no* especifica deliberadamente ningún tipo de componente de documento único o definitivo, pero en lugar de ello permite a los bibliotecarios identificar el objeto que deben describir: un sitio web completo puede ser descrito con un único registro y también se puede describir un recurso concreto de ese sitio web. El bibliotecario puede incluir u omitir en un documento los sonidos de fondo, las hojas de estilo, etc. y puede seleccionar otras páginas web relacionadas (por ejemplo capítulos de un informe). Nuestro futuro proceso automatizado sugerirá al bibliotecario definiciones de documentos basadas en el análisis del contenido, clases de enlaces, etc.: en el documento se incluyen por defecto las imágenes integradas, los sonidos y videos directamente referenciados y las hojas de estilo. También se incluyen los enlaces de ciertos tipos que identifican a una página web referenciada (por ejemplo, como una tabla de contenidos o como una sección).

Por lo tanto, una unidad de información que puede ser descrita bibliográficamente es el punto de partida para hacer una descripción de metadatos, tanto si el material digital está depositado o grabado en un soporte como CD-ROMs, DVDs o como si el documento es ofrecido como archivos separados a través de Internet. Esto significa que todos los documentos digitales –

desde los tipos más *tradicionales* como las monografías o las tesis etc., documentos de tipo "*transitorio*" como periódicos de Internet, hiper-poesía, hiper-drama, etc. y *nuevos* tipos de documentos como las páginas de inicio, los weblog (es decir, blog), etc.- son candidatos para descripciones de metadatos en el marco de nuestro archivo de Internet.

3.4. Una encuesta sobre metadatos y trabajos relacionados

3.4.1. ¿Qué tipos de metadatos necesitamos?

Consideramos interesante preguntarnos qué tipos de formatos de metadatos utiliza actualmente la Biblioteca Nacional para la descripción de diferentes tipos de materiales digitales. Esta información será muy útil dado que esperamos poder importar y exportar algún día datos de nuestro archivo. Los resultados de nuestra encuesta muestran que se utilizan diversos formatos de metadatos: BIBSYS-MARC (el formato MARC del sistema BIBSYS) para los textos digitales, Dublin Core Metadata Element Set[12] para los programas de radio, MAVIS[13] (un formato y sistema australiano) para el material de televisión, sonidos e imágenes, y diversos formatos más diseñados para sistemas locales.

Estos formatos de metadatos se adecuan correctamente al uso que se les da, pero no existen soluciones satisfactorias para todas nuestras necesidades de metadatos. El archivo de Internet necesita muchos tipos de metadatos: metadatos *administrativos* relacionados por ejemplo, con la creación y modificación de registros de metadatos; metadatos para la *gestión de derechos y accesos* para almacenar la información sobre copyright y definir qué grupos de usuarios pueden acceder al archivo y qué documentos se pueden leer; metadatos *estructurales* para mostrar las relaciones lógicas entre objetos, entre metadatos o entre objetos y metadatos; metadatos sobre la *preservación a largo plazo* para la especificación, por ejemplo, de los tipos de archivos, el software necesario y el histórico conversión/migración, y finalmente metadatos *técnicos* para especificar el tamaño del documento, scripts, detalles de comunicación, etc. Por último, pero no menos importante, necesitamos metadatos *descriptivos* y *analíticos* para la consulta y recuperación de información.

3.4.2. ¿Qué modelos de descripción deberíamos elegir?

Existen muchas opiniones sobre qué nivel de descripción le debemos dar a los documentos digitales. En nuestra tarea de definir metadatos para los metadatos descriptivos y analíticos, hemos tomado en consideración dos modelos alternativos. La primera opción es usar tres niveles de descripción:

1. Catalogación para su inclusión en la Bibliografía Nacional, en el catálogo de la Biblioteca Nacional BIBSYS o en otras bases de datos específicas.
2. Catalogación a un nivel más simple y en un formato común.
3. Extracción automática de metadatos del propio documento al igual que de los protocolos de comunicación, etc.

Otra alternativa es usar un modelo a dos niveles, es decir "catalogar o no catalogar"

[12] Para más información sobre Dublin Core Metadata Iniciative, consulte URL: http://www.dublincore.org (se accedió el 15 de abril de 2004)
[13] Para más información sobre el sistema Wizard´s MAVIS, consulte URL: http://www.wizardis.com.au/ie4/products/mavis/introducingmavis.html (se accedió el 15 de abril de 2004)

1. Catalogación para su inclusión en la Bibliografía Nacional, en el catálogo de la Biblioteca Nacional BIBSYS o en otras bases de datos específicas
2. Extracción automática de metadatos del propio documento al igual que de los protocolos de comunicación, etc.

Existen numerosos argumentos para optar por la segunda alternativa: 1) La recuperación de materiales digitales (mediante el texto libre, etc) no depende de su registro, como ocurre con materiales análogos sin identificar. 2) La Biblioteca no necesita registrar los materiales para dejar constancia de su logística (por ejemplo, de qué bibliotecas universitarias han recibido copias). 3) Siempre podemos reconsiderar nuestra decisión de no catalogar cierto tipo de materiales digitales.

A continuación se incluye una breve descripción de los tres niveles de catalogación:

- **Catalogación para su inclusión en la Bibliografía Nacional, etc.**

Por ahora, nuestras sugerencias sobre qué documentos deben ser catalogados a este exhaustivo nivel son incompletas, pero sí podemos decir con certeza que un pequeño número de documentos digitales claramente valiosos continuarán siendo catalogados en algún formato MARC para su inclusión en la Bibliografía Nacional (Podemos mencionar que la versión noruega del formato MARC se llama NORMAC, que algunos sistemas han adoptado versiones locales, por ejemplo BIBSYS MARC, y que el empleo de MARC21[14] está siendo objeto de discusión a nivel nacional. Las reglas de catalogación noruegas se basan en la segunda edición de las Anglo-American Cataloguing Rules (AACR2) cuyos capítulos 9 y 12 están actualmente disponibles en noruego).

- **Catalogación a un nivel más simple en un formato común**

Como se mencionó anteriormente, la Biblioteca Nacional planea un futuro servicio que permita a los editores generar y entregar metadatos en el momento del depósito de materiales. Actualmente, el Proyecto Paradigma trabaja para definir formato(s) de metadatos que posibiliten la creación de una futura herramienta "amigable" proporcionada por este servicio. Finalmente, los bibliotecarios pueden manejar los registros de metadatos ofrecidos por los editores empleándolos como base para un registro bibliográfico a más alto nivel.

Hemos analizado y comparado algunos formatos de metadatos para intentar descubrir qué solución se ajusta más a nuestro trabajo: MAchine Readable Cataloguing (MARC) y Dublin Core Metadata Element Set (DCMES) ya que ambos se usan en bibliotecas e instituciones análogas; Metadata Object Description Schema (MODS[15]) y Metadata Encoding & Transmission Standard (METS[16]) ya que ambos han sido desarrollados por bibliotecas para la

[14] Para más información sobre MARC21, consulte URL: http://www.loc.gov/marc/bibliographic/ecbdhome.html (se accedió el 15 de abril de 2004)

[15] Para más información sobre MODS, consulte URL: http://www.loc.gov/standards/mods/ (se accedió el 15 de abril de 2004)

[16] Para más información sobre METS, consulte URL: http://www.loc.gov/standards/mets/ (se accedió el 15 de abril de 2004)

comunidad bibliotecaria; y Online Information eXchange (ONIX[17]) como formato desarrollado por los editores y la industria del libro. También hemos tenido en cuenta que la comunidad del ISBN ha sugerido que en el futuro los registradores deberán entregar a las agencias del ISBN metadatos compatibles a través de ONIX en relación con la asignación de cada ISBN.

Hemos comparado los formatos antes mencionados preguntándonos: ¿Quién es el responsable del mantenimiento del formato?. ¿Es un estándar internacional?, ¿En qué área se utiliza?, ¿Incluye definiciones semánticas y/o sintácticas?, ¿Cómo describe las relaciones existentes entre documentos?, ¿Es un formato dependiente de reglas o códigos específicos?, ¿Es compatible o está relacionado con otros formatos?, ¿Cómo está de extendido su uso y por cuáles comunidades?

Esperamos que este estudio provoque en la Biblioteca una discusión más amplia sobre los metadatos en relación con la evaluación actualmente en curso. También planeamos analizar en profundidad cómo satisfacer los requisitos funcionales de nuestros usuarios mediante la aplicación de *Common core records* propuestas por el Grupo de Trabajo sobre el Uso de Esquemas de Metadatos de IFLA [6] y *el Informe Final de los Requisitos Funcionales de los Registros Bibliográficos (FRBR)* [5]. En cuanto a las soluciones respecto a los metadatos, también tenemos en agenda la colaboración con el proyecto bibliográfico en curso de la Biblioteca Nacional misma, así como con *La Biblioteca Digital Noruega*, otro proyecto a nivel nacional. Finalmente, esperemos que este trabajo tenga como resultado la recomendación de formatos de metadatos para la descripción a diferentes niveles.

Mientras tanto, hemos estado trabajando para especificar los requerimientos técnicos de metadatos para el sistema de software de nuestro archivo. Hemos identificado diversos factores que pueden influir en nuestra elección de formatos de metadatos para la catalogación a nivel más básico. He aquí algunos factores técnicos deseables:

- Interoperabilidad semántica con MARC: es importante que los atributos del formato de metadatos estén semánticamente armonizados con el formato MARC dominante en la comunidad bibliotecaria. Si es posible, el formato debería ser un subconjunto funcional del MARC, lo que facilitaría el intercambio de datos.
- Simplicidad pero riqueza: es importante hallar un formato de metadatos que resulte fácil de utilizar pero que sea lo suficientemente rico para permitirnos representar una adecuada cantidad de detalles.
- Facilidad de conversión a otros formatos: las correspondencias de conversión entre los formatos elegidos y el formato MARC deben estar disponibles, o al menos ser relativamente sencillas de definir. Sabemos que existen ya algunos sistemas de correspondencias entre MARC21 y MODS y entre MARC21 y ONIX, al igual que entre otros formatos de Dublin Core no evaluados y MODS.
- Compatibilidad con XML: XML es considerado de hecho un estándar y un formato compatible con XML nos permitirá manejar este formato con el software disponible. De este modo, sería posible definir un marco estructural más amplio que permitiera a nuestro archivo aceptar metadatos de diferente fuentes, manejar modificaciones de metadatos, definir metadatos originales, dejar constancia del histórico de las versiones, etc. (por ejemplo, METS)

[17] Para más información sobre ONIX, consulte URL: http://www.loc.gov/standards/mets/ (se accedió el 15 de abril de 2004)

- Extensibilidad: un formato de metadatos debe permitirnos definir nuevos elementos cuando sea necesario.
- Elementos clave: es importante definir los elementos clave de los metadatos, es decir, un denominador común puede facilitar la consulta y recuperación de diferentes tipos de materiales.

Si comparamos estos factores con los formatos de metadatos descritos en nuestro estudio, nos damos cuenta de que los formatos compatibles con MARC y XML son los más deseables. Sin embargo, no hay una receta sencilla. Es necesario definir nuevos elementos sobre la técnica, estructura, derechos y gestión de acceso de los metadatos, quizás en combinación con el marco METS. Por supuesto, esto también es aplicable para los metadatos sobre preservación a largo plazo. En este sentido, el Depósito Digital a Largo Plazo de la Biblioteca debe usar metadatos que se adapten a OAIS[18].

➢ **Extracción automática de metadatos**

Desafortunadamente, los bibliotecarios nunca catalogarán el 99% de los documentos de Internet para nuestro archivo. Por esta razón continuamos investigando el uso de sistemas automáticos de análisis y extracción de metadatos de documentos de Internet como parte de nuestro trabajo con metadatos y diseño de sistemas. Los metadatos extraídos serán almacenados junto con los objetos digitales y otras descripciones de metadatos, y estarán disponibles para búsquedas estructuradas en el archivo de Internet.

La tecnología aún no está suficientemente desarrollada para decidir automáticamente un tipo de documento, pero puede ayudar a reducir el número de documentos que serán objeto de análisis humano en la segunda fase de la selección. Ejemplos de características de las clases de documentos son: 1) idioma, vocabulario y gramática, 2) tamaño del documentos y estructura, 3) fuente, editor y servidor web, 4) uso de "cookies", 5) antigüedad y esperanza de vida de un documento, 6) sonido, imágenes, animaciones y otros tipos avanzados de información 7) interacción con el usuario como formularios, botones, etc. 8) número, tipo y fuente de los enlaces, 9) valores de la URL, como por ejemplo el uso de caracteres y palabras especiales en la misma, 10) uso de scripts del lado del cliente, 11) detalles técnicos de la comunicación.

La tecnología para analizar el vocabulario y la gramática está mejorando y consideramos que este tipo de análisis puede ser un elemento importante para futuros procedimientos automatizados. Finalmente, la elección automática de características estará disponible para búsquedas estructuradas en el archivo de Internet. El valor de estas propiedades será limitado pero, en combinación con otros criterios de búsqueda, resultarán realmente útiles.

4 El papel de las normas FRBR en el archivo de Internet

El Proyecto Paradigma quiere presentar los metadatos y documentos digitales archivados de una manera organizada y estructurada que facilite la navegación de los usuarios. Hemos hallado que el modelo FRBR de IFLA es una herramienta esencial en esta tarea y lo usaremos como modelo para sentar las bases del diseño del archivo de Internet.

[18] Para más información sobre el Modelo de Referencia OAIS, consulte URL: http://ssdoo.gsfc.nasa.gov/nost/wwwclasic/documents/pdf/CCSDS-650.0-B-1.pd (se accedió el 15 de abril de 2004)

Creemos que añadir mecanismos adicionales de modelización al modelo FRBR beneficiará nuestro trabajo con medios dinámicos tales como los documentos de Internet, documentos multimedia y otros recursos continuados. Los mecanismos agregados pueden ser implementados como extensiones puras del modelo sin que requieran cambios significativos en los actuales conceptos FRBR. A lo largo de este año se publicará en la sección dedicada a las FRBR de *Cataloging & Classification Quaterly* un artículo sobre los mecanismos agregados que proponemos.

Para adaptar el modelo FRBR a los documentos dinámicos de Internet es necesaria una ligera reinterpretación de los conceptos de *manifestación* y *ejemplar*. Estos conceptos se describen en la siguiente sección.

4.1. Adaptación de las FRBR para su uso con documentos dinámicos de Internet

4.1.1. Documentos dinámicos

Los documentos de Internet son muchas veces *dinámico;,* por ejemplo, un periódico de Internet actualizado varias veces a lo largo del día. Un usuario puede referirse a este tipo de documento dinámico como a un foro o canal de información. "El *Daily News* informa que...". Quizá podríamos decir que un documento dinámico se corresponde, aproximadamente, con una URL. Conceptos como "entregas" y sucesivas "ediciones" deben igualmente ser revisados en el contexto de Internet: desde un punto de vista formal, una actualización de una página web puede ser similar a una nueva edición de un libro. Sin embargo, los lectores ven por ejemplo los continuos cambios en las portadas principales de los periódicos de Internet como una sola entidad cambiante, no como diferentes y separadas ediciones.

Usando el modelo FRBR con las extensiones para los componentes agregados, hemos definido el concepto de *documento dinámico* como "el ciclo vital completo de una página web o un documento de Internet similar en continuo cambio".

Si catalogáramos una actualización de un documento web de este tipo de acuerdo con las AACR2, utilizaríamos normalmente las reglas para los recursos integrables, es decir, un recurso bibliográfico que se añade a, o que se modifica por medio de actualizaciones, que no permanece independiente y que se integra en el conjunto. Sin embargo, los periódicos de Internet son más parecidos a una emisora de radio, un flujo de información pasajera en continuo cambio. No están "integrados en el conjunto". La captura de contenidos de un documento en continuo cambio en un determinado momento es como grabar un *"ejemplo"* de una emisión efímera. Consideramos cada ejemplo o fotografía como un *documento específico.*

Cuando se accede a un documento dinámico a través de Internet el *ejemplar* (item) recuperado por un usuario puede ser diferente de otros *ejemplares* del mismo documento: depende de la combinación de diversos factores: la identidad del usuario, la herramienta de acceso empleada (el navegador web), la información sobre anteriores accesos al mismo documento (las cookies almacenadas), los parámetros explícitamente especificados por el usuario (por ejemplo, en un formulario), y por último pero no menos importante, el estado actual de la base de datos. A menudo, el ejemplar es generado "al vuelo" cuando el usuario realiza una petición.. En otras palabras, una petición HTTP actúa como un servicio de "impresión bajo demanda": la copia entregada refleja el contenido de la base de datos del

documento en el momento de la impresión. La base de datos puede ser considerada como una representación física (semi)permanente de un documento dinámico, del cual se obtienen *ejemplares* específicos. Estos *ejemplares* por sí mismos no tienen una representación permanente, son información transitoria hasta que sean preservados, por ejemplo, en un archivo de Internet.

4.1.2. Documentos específicos

Hemos definido un *ejemplar* mostrando un documento dinámico como un *documento específico,* de manera diferente al *ejemplar* tradicional en un sentido primordial: es un miembro de un grupo de *ejemplares* que ilustran el mismo documento dinámico. Un documento almacenado en un archivo o visualizado por un usuario es, obviamente, un documento específico: una búsqueda a texto completo nos ofrecerá como máximo una entrada en la lista sistematizada de documentos dinámicos. Si el usuario solicita la visualización de un elemento de la lista, el documento dinámico es presentado como una unidad, y el usuario entonces puede seleccionar un *ejemplar* específico en la *línea temporal*, es decir, la línea de menú que representa la vida útil del documento. Cada versión preservada, es decir, cada documento específico, se indica en esta línea temporal con un marcador. El usuario puede acceder a cualquier documento específico pinchando en el marcador de una determinada fecha/hora recuperando de este modo el *ejemplar*. (Ver figura 1).

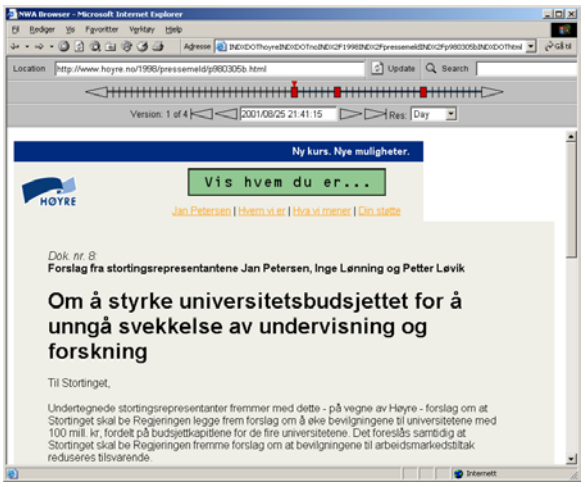

Figura 1. Presentación de un documento dinámico en la interfaz de usuario de la Herramienta de Acceso NWA.

4.2. Definiciones de editores y usuarios sobre documentos y metadatos

La presentación de documentos archivados con fines documentales y de investigación es sólo uno de los servicios que ofrece la Biblioteca Nacional Noruega. Además, y basándonos en las ideas mencionadas anteriormente, hemos sugerido revisar el servicio de asignación de identificadores existente en la Biblioteca. Actualmente, este servicio asigna URN:NBNs para las universidades e instituciones de la sección noruega del espacio de nombres URN:NBN. Sin embargo, vemos posibilidades para la asignación igualmente en este servicio de ISBNs autónomos.

4.2.1. Funcionalidades futuras - un escenario

Un escenario mostrando futuras funcionalidades es el siguiente: las series de identificadores primarios asignadas por este servicio requieren que el usuario/peticionario facilite por un lado un juego mínimo de *metadatos* y una definición exacta del documento identificado.

Se deben asignar los identificadores de *obra, expresión y manifestación* (incluyendo las definiciones de documento dinámico) y *ejemplar* (incluyendo las definiciones de documentos específicos). Los *ejemplares* (documentos específicos) deben estar especificados mediante una lista completa de elementos (por ejemplo, un archivo HTML, archivos de imágenes, archivos de sonido, etc); las *manifestaciones* (documentos dinámicos) también deben ser especificadas por normas del tipo "la portada de los periódicos de Internet en este URL y todas las páginas enlazadas directamente desde esta portada que residen en el mismo sitio Web".

Para los identificadores de *expresión* y *obra*, el usuario puede identificar opcionalmente *expresiones*/documentos dinámicos y *ejemplares*/documentos específicos, los cuales son ilustraciones de esta *obra/expresión.*

Las definiciones especificadas por editores y usuarios se consideran más definitivas que las propuestas automáticamente. Se archiva la identidad del editor o usuario que asigna el identificador; una definición de documento especificada por una editorial o universidad reconocida debe ser considerada más significativa que otra solicitada al arbitrio del usuario.

4.2.2 Campos de metadatos.

La obligatoriedad y no obligatoriedad de los campos de metadatos podría estar disponible para la descripción del documento en cada nivel de las FRBR, y cada nivel estaría identificado con una URN:NBN. Los valores de metadatos serán almacenados con el identificador, y los usuarios de nuestro servicio a través de Internet podrán encontrar el documento mediante este número.

Tras rellenar toda la información en los campos de metadatos de la futura herramienta de asignación de metadatos/identificadores, los editores podrán ver estos metadatos en una ventana independiente en formato HTML simplemente pinchando un botón, por ejemplo <HTML Dublin Core>. A partir de aquí, el usuario puede copiar y pegar los metadatos en el elemento <HEAD> del documento web que se está describiendo antes de continuar con el proceso de asignación de identificadores. Una vez guardado el documento digital, que ya incluye los metadatos insertados, el usuario puede almacenar fácilmente una copia del documento enriquecido con los metadatos en el archivo de la biblioteca pinchando en el botón de actualización del navegador.

4.3. ¿Posibles servicios de verificación de documentos y autenticación?

Hemos oído historias de autoridades que revisan los comunicados oficiales de Internet, y después rechazan reconocer la existencia de versiones anteriores. Incluso hemos oído que firmas comerciales anuncian sus productos a un determinado precio, pero luego le cobran a sus clientes un precio muy superior.

Con estas y otras historias in mente, el Proyecto Paradigma propone un servicio de

autenticación y verificación que permita a los usuarios solicitar la descarga de un determinado documento de Internet, es decir, una fotografía de una página web que contenga un determinada oferta comercial, un comunicado de responsabilidad legal, calumnia, etc.

Si tiempo después aparecen dudas sobre el contenido de estos documentos, la Biblioteca puede confirmar (o rechazar) cualquier demanda en este sentido. Incluso cuando no hay aspectos legales implicados, la preservación de un determinado *ejemplar* de documento puede servir como imagen claramente definida de un documento dinámico en un determinado momento, por ejemplo, para propósitos de cita o referencia. Esto es importante especialmente cuando tenemos en cuenta que la mayor parte de los documentos de Internet no tienen número de página, nº de versión, etc.

En nuestro archivo de Internet, un documento específico se describe tal y como haya sido recibido desde el servidor web. Existe una cadena bien definida de bits para cada componente del documento (texto, imágenes, etc.). La representación gráfica del documento *no* es parte de su definición – este proceso se deja a la herramienta de acceso. El documento específico es identificado como el contenido de un documento dinámico proporcionado por determinados componentes y metadatos:

- la fuente de cada componente (por ejemplo, la URL)
- todos los parámetros especificados por el cliente cuando recupera los componentes
- el momento temporal en que cada componente fue recuperado
- el conjunto de componentes incluidos en el documento.

5. Conclusión

El Proyecto Paradigma de la Biblioteca Nacional de Noruega va a trabajar con la intención de establecer tecnología, metodologías y rutinas de trabajo satisfactorias para el depósito legal de todos los tipos de documentos digitales – incluyendo los millones de documentos hallados en el dominio noruego de Internet – durante el restante periodo del proyecto. Esperamos ser capaces, ya en 2005, de ofrecer a nuestros usuarios acceso al material archivado a través de registros bibliográficos, diversos tipos de metadatos y herramientas de consulta a texto completo.

Nuestro archivo de Internet estructurado a partir de los FRBR será ciertamente uno de los primeros de este tipo e incluso esperamos llevar a la práctica nuestras ideas sobre el uso de los niveles de *obra, expresión, manifestación* y *ejemplar* de los FRBR en un futuro servicio de asignación de identificadores. ¿Podrían nuestras ideas sobre un servicio de verificación/autenticación en Internet ser factibles en un futuro?. El tiempo lo dirá pero, mientras tanto, la Biblioteca Nacional continuará explorando nuevas formas de preservar el patrimonio cultural digital noruego y ofreciendo a los usuarios herramientas que puedan abrir las puertas de esta excitante biblioteca digital.

Referencias Seleccionadas

(Todas las URLs eran accesibles el 15 de abril de 2004)

[1] Final Report for the Pilot project "Netarkivet.dk" [online]. – URL: http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf

[2] Guidelines for the selection of online Australian publications intended for preservation by

the National Library of Australia [online]. – URL: http://pandora.nla.gov.au/selectionguidelines.html

[3] The Kulturarw3 Project – The Royal Swedish Web Archiw3e – An example of "complete" collection of web pages [online]. – URL: http://www.ifla.org/IV/ifla66/papers/154-157e.htm

[4] Olsen, Nancy (2002). Cataloging Internet Resources : A Manual and Practical Guide [online]. - OCLC. - URL: http://www.oclc.org/support/documentation/worldcat/cataloging/internetguide/1/1.htm

[5] Norway. [The Legal Deposit Act (1989)] Act relating to the legal deposit of generally available documents : no. 32 of 9 June 1989 : with regulations / [published by the Ministry of Church and Cultural Affairs ; unofficial English translation published by the National Library of Norway. - [Oslo] : National Library of Norway, 1997. - 21 s.

[6] IFLA Cataloguing Section Working Group on the Use of Metadata Schemas (2003). Guidance on the structure, content, and application of metadata records for digital resources and collections : draft for worldwide review 27 October, 2003 [online]. – URL: http://www.ifla.org/VII/s13/guide/metaguide03.pdf

[7] RFC 3188 Using National Bibliography Numbers as Uniform Resource Names [online] / J. Hakala, 2001. – URL: http://www.ietf.org/rfc/rfc3188.txt

[8] Van Nuys, Carol (2003). Identification of network accessible documents : problem areas and suggested solutions [online] / Carol van Nuys, Ketil Albertsen. – S. 13-25. – *I*: Proceedings : in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries, ECDL 2003 / Julien Masanès, Andreas Rauber, Gregory Cobena (eds). – URL: http://bibnum.bnf.fr/ecdl/2003/index.html

[9] Albertsen, Ketil (2003). The Paradigma web harvesting environment. – S. 49-62. – *I*: Proceedings : in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries, ECDL 2003 / Julien Masanès, Andreas Rauber, Gregory Cobena (eds). – URL: http://bibnum.bnf.fr/ecdl/2003/index.html

[10] Van Nuys, Carol (2003). The Paradigma project [online]. – *I*: RLG DigiNews. - Vol. 7, no. 2. – URL: http://www.rlg.org/preserv/diginews/v7_n2_feature2.html